Understanding and Mitigating Biases in Evaluation

Jingyan Wang

CMU-RI-TR-21-48

August 2021

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Nihar Shah (Chair)
Artur Dubrawski
Jeff Schneider
Ariel Procaccia, Harvard University
Avrim Blum, Toyota Technological Institute at Chicago

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

Many applications in real life involve collecting and aggregating evaluation from people, such as in hiring, peer grading and conference peer review. In this thesis, we focus on three sources of biases that arise in such problems: people, estimation and policies. Specifically, people provide evaluation data; estimation procedures perform inference and draw conclusions from the provided data; policies specify all the details that are needed in order to execute the entire process. We model and analyze these biases, and subsequently propose methods to mitigate them.

First, we study human bias, that is, the bias in the evaluation data introduced by human evaluators. We consider the miscalibration aspect, meaning that different people have different calibration scales. We propose randomized algorithms that provably extract useful information under a general model we propose for arbitrary miscalibration. Building upon these results, we also propose a heuristic that is applicable to a broader range of settings. In addition to miscalibration, we also consider the bias induced by the "outcome" experienced by people. As an example, when students rate their course instructors, the students' ratings are influenced by the grades that the students receive in these courses. We make mild assumptions to model such biases, and propose an adaptive algorithm that corrects this bias using knowledge about the "outcomes".

Second, we study estimation bias, that is, when algorithms exhibit different behaviors on different subgroups of the population. We consider the problem of estimating the quality of individual items from pairwise comparison data. We analyze the statistical bias (defined as the expectation of the estimated value minus the true value) when using the maximum-likelihood estimator, and then propose a simple modification to the estimator to reduce the bias.

Third, we study policy bias, that is, when the rules dictating the evaluation process induce undesirable outcomes. We examine large-scale multi-attribute evaluation tasks. As an example, in graduate admissions, the evaluation criteria often consist of multiple attributes, such as school GPAs, standardized test scores, recommendation letters, research experience, etc. The number of applications is large, and therefore the evaluation task needs to be divided and assigned to many reviewers in a distributed fashion. It is common practice to assign each reviewer a subset of the applications, and ask them to assess all relevant information for their assigned subset. In contrast, we propose an alternative approach where each reviewer evaluates more applicants but fewer attributes per applicant. We establish various tradeoffs between these two approaches, and identify conditions under which our proposed approach results in better evaluation.

Finally, we briefly describe our outreach efforts to improve the peer review process – reducing the bias caused by the alphabetical-ordering authorship in scientific publications, and analyzing the gender distribution of the recipients of conference paper awards.

Contents

Chapter 1

Introduction

1.1 Background and motivation

Many applications in real life involve some form of evaluation and selection – given a pool of items (or applicants), the goal is to evaluate the quality of items, and potentially make downstream decisions about the items based on the evaluation. As a qualifying note, we focus on applications where the evaluation involves both an objective and a subjective component. To understand the spectrum, a math exam of all multiple-choice questions is an objective task, where there is a clear correct answer to each question. On the other hand, voting for elections is a subjective task, where different political views cannot be simplified classified as correct or incorrect. Many evaluation tasks include both an objective and subjective components. For example, in graduate admissions, consensus may be reached relatively easily on a small fraction of very outstanding applicants. However, for a large number of applicants who do not clearly make the cut but still are above-average, whom to admit depends on subjective interpretation of their track records and what specific caliber the committee puts more weight on. In addition to education (admissions), many applications in this regime are also high-stakes decision-making problems, such as healthcare (what treatments to give to the patients, and which patients to prioritize), banking (which applications are approved for a loan, and what the interest rates are) and law (which defendants are released on bail, and what the associated bail amounts are). It is self-evident that these decisions have long-lasting influence on the individuals involved, and for these applications we as a society desire the decisions to be equitable.

In this thesis, we consider three major components that play a role in such evaluation processes: *people* (expressing opinions and providing data), *estimation* (performing inference and making conclusions based on the provided data) and *policies* (specifying the evaluation rules and the guidelines). To understand the definitions and connections between these components, we consider the following two more examples:

- In *conference peer review*, the goal is to identify and accept the top papers submitted to the conference. Reviewers (*people*) read the papers and provide their scores and comments about the papers. Given these reviews, the program chairs or area chairs form summarized views of the papers and make the acceptance decisions accordingly (*estimation*). The program chairs also need to decide on different aspects of the review process (*policy*), such as how to assign papers to reviewers (e.g., whether to use automated matching algorithms such as Toronto Paper Matching System [35]; whether to use bidding), what types of data to elicit in the review forms (e.g., on what scale the reviewers provide numerical scores; how to instruct the reviewers to write free-form text reviews), how the chairs make paper decisions based on the reviews (e.g., how many chairs are involved in making the final decision of each paper; what the conference puts emphasis on, if a paper has both clear strengths and weaknesses), whether the conference is single- or double-blind, etc.

- In *grading*, the goal is to assign accurate scores to students that reflect their performance in a homework assignment or an exam (e.g., writing an essay). The graders (*people*) can be the instructors, the TAs or the students themselves (termed "peer grading"). The graders read and give scores to the assignments. Let us consider the case where each assignment is graded by two graders to ensure accuracy. Then one natural approach to deriving a final score to each student is to take the mean (*estimation*) of the two scores given by the two graders. In the case where each assignment is graded by only one grader for efficiency, one approach to estimating a final score for each student is to trivially just take this score given by the reviewer. The instructors specify the rules (*policy*) for the grading process, such as how to define the grading rubrics for each question in the homework, how to distribute the assignments to the graders (e.g., whether to combine instructor/TA/student graders; whether to alternate graders so that each student is assigned different graders for different homeworks they submit), and whether to allow regrading requests or not.

The three components of people, estimation and policies also naturally apply to other applications, such as admissions and hiring. The high-level ideas are very similar, so we omit further details.

Many sources of biases are involved in such evaluation process. In this thesis, we use the term "bias" to refer to, broadly speaking, systematic errors that are based on external factors independent of the true quality that we want to evaluate. We analyze the biases associated with each of the three components. Specifically, we consider

- **Human bias:** People are biased, and the bias can be conscious or subconscious. For example, people are miscalibrated; they do not always map their scores to the pre-defined scale that they are instructed to calibrate to. People are noisy; they lose attention to the evaluation task and give scores that are different from what they would have given if they paid attention. People are subjective; their scores reflect their personal opinions and value systems. People are strategic; they may intentionally manipulate their evaluation due to self interests. There are also well-known types of biases that are of a discriminative nature, including gender bias [125, 175] and racial bias [20, 55]. Moreover, there are also biases that apply to more specific contexts. For example, in peer review, the reviewers tend to be in favor of the papers whose perspectives align with the reviewers' own perspectives, termed the "confirmation bias" [112]; people also tend to put more emphasis on the quantity than the quality of the papers when perceiving the success of universities or individual researchers [148], etc.

  So far we have described many types of biases inherited by a single reviewer. Notably, the issue with human bias is further exacerbated by the fact that in many applications, the evaluation is done in a distributed fashion. That is, the number of items under evaluation is large, and therefore the items are distributed to multiple evaluators in parallel. The discrepancy between reviewers' evaluation due to bias translates to the discrepancy of how the items are reviewed and treated differently. This potentially leads to unfairness in the evaluation process, and therefore calls for a need to carefully address human bias.

- **Estimation bias:** Estimators and algorithms may be biased, that is, they may yield different performance on different subgroups of the population. For example, in machine learning, the objective is often to minimize the average loss. As a result, algorithms trained to minimize the average loss may optimize their performance on frequently-seen samples, but do not work as well on the less common samples, termed "sample size disparity" (e.g., [13, Chapter 2]). Another example is the definition of bias in statistics – the (statistical) bias of an estimator is defined as the expectation of the estimated value minus the true value. A non-zero bias means that the estimator systematically overestimates or underestimates the parameter of interest.
- **Policy bias:** We broadly interpret "policies" as the design of rules associated with the evaluation process. Inappropriate policies, rules and practices may lead to undesirable outcomes and inequity against certain subgroups of the population. Reasons for the policies to induce undesirable outcomes include: introducing misaligned incentives of the participating agents, suboptimally using evaluators' ability and expertise, and triggering specific forms of human bias. On the other hand, appropriate policies can also be established to reduce bias. Some well-known policies include the affirmative action in college admissions and Rooney rule in hiring.

On these three sources of biases, we combine tools from statistics (for establishing theoretical guarantees), computer science (for designing algorithms) and policies (for outreach efforts). We provide theoretical and experimental results that aim to answer the following questions:

- What is the bias? To what extent is the bias? Why does the bias exist? (*understanding*)
- What approach can we take to reduce or correct the bias? What is the outcome of the approach? (*mitigation*)

In summary, this thesis aims to

understand and mitigate biases arising from human, estimation and policies in evaluation and decision-making problems.

1.2 Organization of thesis

The thesis is organized into three parts. Each part discusses one of the three sources of biases. Each problem is presented with a motivating application, but the models are intended to be general and not restrictive to any particular application.

Human Bias

Chapter 2 studies the miscalibration of people. Specifically, we assume that the miscalibration of people can be arbitrary, and make minimal assumptions on the nature of miscalibration. We consider a general assumption of monotonicity, where people can essentially give arbitrary scores, as long as they are "somewhat reasonable" – being more likely to give higher scores to items of higher quality. The proposed calibration methods provably extract useful information from people's scores that are arbitrarily miscalibrated, hence making the evaluation more fair. This chapter is based on joint work with Nihar Shah [186].

Chapter 3 builds upon the results in Chapter 2, and proposes a heuristic to address miscalibration in more general settings compared to the algorithms in Chapter 2. In this problem, we continue to characterize miscalibration as a monotonic constraint, and consider its least-squares estimator which is known to have various desirable statistical guarantees. No computationally-efficient methods for computing the least-squares estimator is known up to date, and the prior work primarily focuses on alternative estimators that are computationally more efficient but have weaker statistical guarantees. Instead, we provide a new perspective to the problem by computing an approximate solution to the least-squares estimator. We capture the monotonic constraint by incorporating a new regularizer term in the optimization objective. We present desirable properties focused on the stationary points of the optimization problem, and conduct simulation to demonstrate the effectiveness of our method. This chapter is based on joint work with Komal Dhull, Nihar Shah, Yuanzhi Li, and R. Ravi [53].

Chapter 4 studies the bias induced by people's experience. One prevalent example is teaching evaluation, where universities survey students at the end of each semester to evaluate the teaching quality of their instructors. However, prior studies have shown that instructors' grading practices have a significant influence on the end-of-course teaching evaluations: students who receive higher grades in a course often give higher ratings, and the students who receive lower grades often give lower ratings. We again make mild monotonic assumptions on the correlation between the student ratings and the grades they receive. We propose a cross-validation debiasing algorithm that provably adapts to different extents of the bias in the data without prior knowledge. The algorithm is also shown to perform favorably compared to standard baselines, in a semi-synthetic experiment using real grading statistics from the Indiana University Bloomington [90]. This chapter is based on joint with Ivan Stelmakh, Yuting Wei and Nihar Shah [188].

Estimation Bias

Chapter 5 considers the problem of estimating the quality of individual items from pairwise comparison data, and studies the bias introduced by the maximum-likelihood (ML) estimator on this problem. Here the term "bias" refers to the standard definition in statistics, which is defined as the expectation of the estimated value minus the true value. While prior work has shown that the ML estimator is minimax-optimal in terms of the squared Euclidean error, we show that the ML estimator incurs a suboptimal rate in terms of its bias. Moreover, by a simple modification to the ML estimator, we derive a class of estimators that achieve a significantly better rate on the bias and at the same time maintain minimax-optimality in the squared Euclidean error. Hence, our modified estimator provably improves fairness while maintaining estimation accuracy. This chapter is based on joint work with Nihar Shah and R. Ravi [187].

Policy Bias

Chapter 6 studies large-scale multi-attribute evaluation problems. In applications such as hiring and educational admissions, the number of applicants is often large, thereby making it infeasible for a single reviewer to evaluate all applications. The common practice is to assign the evaluation task to multiple reviewers in a distributed fashion. Specifically, each reviewer is assigned a subset of the applications, and asked to assess all relevant information for their assigned subset. We propose an alternative approach to assigning applicants to reviewers. Our approach is based on the observation that the evaluation criteria often consist of multiple attributes, such as – in admissions – student GPAs, standardized test scores, recommendation letters and essays. Our approach requires each reviewer to evaluate more applicants but fewer attributes per applicant. We compare our proposed approach to the traditional aforementioned approach on several dimensions via theoretical and experimental methods. We establish various tradeoffs between these two approaches, and identify conditions under which our proposed approach has an advantage. This chapter is based on joint work with Carmel Baharav, Nihar Shah, Anita Woolley, and R. Ravi.

Chapter 7 discusses the bias arising from the alphabetical-ordering of the authorship in scientific publications, and describes an outreach work to mitigate the bias by using appropriate citation styles and ordering of individuals. This chapter is based on joint work [1] with Nihar Shah.

Chapter 8 provides an analysis on the gender statistics of the authors of award-winning conference papers. Our results suggest a notable discrepancy between men and women, providing a complementary evidence point in addition to prior work in understanding gender bias in peer review, and more generally in academia. This chapter is based on joint work [2] with Nihar Shah.

Finally, we conclude with a discussion on directions for future work.

[1] https://researchonresearch.blog/2018/11/28/theres-lots-in-a-name/

[2] https://researchonresearch.blog/2019/06/18/gender-distributions-of-paper-awards/

Part I

Human Bias

Chapter 2

Handling Arbitrary Miscalibrations in Ratings

Cardinal scores (numeric ratings) collected from people are well known to suffer from miscalibrations. A popular approach to address this issue is to assume simplistic models of miscalibration (such as linear biases) to de-bias the scores. This approach, however, often fares poorly because people's miscalibrations are typically far more complex and not well understood. In the absence of simplifying assumptions on the miscalibration, it is widely believed by the crowdsourcing community that the only useful information in the cardinal scores is the induced ranking. In this chapter, inspired by the framework of Stein's shrinkage, empirical Bayes, and the classic two-envelope problem, we contest this widespread belief. Specifically, we consider cardinal scores with arbitrary (or even adversarially chosen) miscalibrations which are only required to be consistent with the induced ranking. We design estimators which despite making no assumptions on the miscalibration, strictly and uniformly outperform all possible estimators that rely on only the ranking. Our estimators are flexible in that they can be used as a plug-in for a variety of applications, and we provide a proof-of-concept for A/B testing and ranking. Our results thus provide novel insights in the eternal debate between cardinal and ordinal data.

2.1 Introduction

"A raw rating of 7 out of 10 in the absence of any other information is potentially useless." [123]

"The rating scale as well as the individual ratings are often arbitrary and may not be consistent from one user to another." [7]

Consider two items that need to be evaluated (for example, papers submitted to a conference) and two reviewers. Suppose each reviewer is assigned one distinct item for evaluation, and this assignment is done uniformly at random. The two reviewers provide their evaluations (say, in the range $[0, 1]$) for the respective item they evaluate, from which the better item must be chosen. However, the reviewers' rating scales may be miscalibrated. It might be the case that the first reviewer is lenient and always provides scores in $[0.6, 1]$ whereas the second reviewer is more stringent and provides scores in the range $[0, 0.4]$. Or it might be the case that one reviewer is moderate whereas the other is extreme – the first reviewer's 0.2 is equivalent to the second reviewer's 0.1 whereas the first reviewer's 0.3 is equivalent to the second reviewer's 0.9. More generally, the miscalibration of the reviewers may be arbitrary and unknown. Then is there any hope of identifying the better of the two items with any non-trivial degree of certainty?

A variety of applications involve collection of human preferences or judgments in terms of cardinal scores (numeric ratings). A perennial problem with eliciting cardinal scores is that of miscalibration – the systematic errors introduced due to incomparability of cardinal scores provided by different people (see [77] and references therein).

This issue of miscalibration is sometimes addressed by making simplifying assumptions about the form of miscalibration, and post-hoc corrections under these assumptions. Such models include one-parameter-per-reviewer additive biases [10, 65, 111, 135], two-parameters-per-reviewer scale-and-shift biases [135, 147] and others [61]. The calibration issues with human-provided scores are often significantly more complex causing significant violations to these simplified assumptions (see [77] and references therein). Moreover, the algorithms for post-hoc correction often try to estimate the individual parameters which may not be feasible due to low sample sizes. For instance, John Langford notes from his experience as the program chair of the ICML 2012 conference [105]:

> *"We experimented with reviewer normalization and generally found it significantly harmful."*

This problem of low sample size is exacerbated in a number of applications such as A/B testing where every reviewer evaluates only one item, thereby making the problem underdetermined even under highly restrictive models.

It is commonly believed that when unable or unwilling to make any simplifying assumptions on the bias in cardinal scores, the only useful information is the ranking of the scores [7, 64, 80, 123, 126, 146]. This perception gives rise to a second approach towards handling miscalibrations – that of using only the induced ranking or otherwise directly eliciting a ranking and not scores from the use. As noted by Freund et al. [64]:

> *"[Using rankings instead of ratings] becomes very important when we combine the rankings of many viewers who often use completely different ranges of scores to express identical preferences."*

These motivations have spurred a long line of literature on analyzing data that takes the form of partial or total rankings of items [7, 15, 44, 126, 141, 154, 156].

In this chapter, we contest this widely held belief with the following two fundamental questions:

- In the absence of simplifying modeling assumptions on the miscalibration, is there any estimator (based on the scores) that can outperform estimators based on the induced rankings?
- If only one evaluation per reviewer is available, and if each reviewer may have an arbitrary (possibly adversarially chosen) miscalibration, is there hope of estimation better than random guessing?

We show that the answer to both questions is "Yes". One need not make simplifying assumptions about the miscalibration and yet guarantee a performance superior to that of any estimator that uses only the induced rankings.

In more detail, we consider settings where a number of people provide cardinal scores for one or more from a collection of items. The calibration of each reviewer is represented by an unknown monotonic function that maps the space of true values to the scores given by this reviewer. These functions are arbitrary and may even be chosen adversarially. We present a class of estimators based on cardinal scores given by the reviewers which *uniformly* outperforms any estimator that uses only the induced rankings. A compelling feature of our estimators is that they can be used as a plug-in to improve ranking-based algorithms in a variety of applications, and we provide a proof-of-concept for two applications: A/B testing and ranking.

The techniques used in our analyses draw inspiration from the framework of Stein's shrinkage [91, 166] and empirical Bayes [145]. Moreover, our setting with 2 reviewers and 2 papers presented subsequently in the chapter carries a close connection to the classic two-envelope problem (for a survey of the two-envelope problem, see [71]), and our estimator in this setting is similar in spirit to the randomized strategy [45] proposed by Thomas Cover. We discuss connections with the literature in more detail in Section 2.3.1.

Our work provides a new perspective on the eternal debate between cardinal scores and ordinal rankings. It is often believed that ordinal rankings are a panacea for the miscalibration issues with cardinal scores. Here we show that ordinal estimators are not only inadmissible, they are also strictly and uniformly beaten by our cardinal estimators. Our results thus uncover a new point on the bias-variance tradeoff for this class of problems: Estimators that rely on simplified assumptions about the miscalibration incur biases due to model mismatch, whereas the absence of such assumptions in our work eliminates the modeling bias. Moreover, in this minimal-bias regime, our cardinal estimators incur a strictly smaller variance as compared to estimators based on ordinal data alone.

Finally, a note qualifying the scope of the problem setting considered here. In applications such as crowdsourced microtasks where workers often spend very little time answering every question, the cardinal scores elicited may not necessarily be consistent with the ordinal rankings, and moreover, ordinal rankings are often easier and faster to provide. These differences cease to exist in a variety of applications such as peer-review or in-person laboratory A/B tests which require the reviewers to spend a non-trivial amount of time and effort in the review process, and these applications form the motivation of this work.

2.2 Preliminaries

Consider a set of n items denoted as $\{1, \ldots, n\}$ or $[n]$ in short.[1] Each item $i \in [n]$ has an unknown value $x_i \in \mathbb{R}$. For ease of exposition, we assume that all items have distinct values. There are m reviewers $\{1, \ldots, m\}$ and each reviewer evaluates a subset of the items. The calibration of any reviewer $j \in [m]$ is given by an unknown, strictly-increasing function $f_j : \mathbb{R} \to \mathbb{R}$. (More generally, our results hold for any non-singleton intervals on the real line as the domain and range of the calibration functions). When reviewer j evaluates item i, the reported score is $f_j(x_i)$. We make no other assumptions on the calibration functions $f_1, \ldots, f_m$. We use the notation $\succ$ to represent a relative order of any items, for instance, we use "$1 \succ 2$" to say that item 1 has a larger value (ranked higher) than item 2. We assume that m and n are finite.

[1] We use the standard notation of $[\kappa]$ to denote the set $\{1, \ldots, \kappa\}$ for any positive integer κ.

Every reviewer is assigned one or more items to evaluate. We denote the assignment of items to reviewers as $A = (S_1, \ldots, S_m)$, where $S_j \subseteq [n]$ is the set of items assigned to reviewer $j \in [m]$. We use the notation Π to represent the set of all permutations of n items. We let $\pi^* \in \Pi$ denote the ranking of the n items induced by their respective values $(x_1, \ldots, x_n)$, such that $x_{\pi^*(1)} > x_{\pi^*(2)} > \cdots > x_{\pi^*(n)}$. The goal is to estimate this ranking π^* from the evaluations of the reviewers. We consider two types of settings: an ordinal setting where estimation is performed using the rankings induced by each reviewer's reported scores, and a cardinal setting where the estimation is performed using the reviewers' scores (which can have an arbitrary miscalibration and only need to be consistent with the rankings). Formally:

- **Ordinal:** Each reviewer j reports a total ranking among the items in S_j, that is, the ranking of the items induced by the values $\{f_j(x_i)\}_{i \in S_j}$. An ordinal estimator observes the assignment A and the rankings reported by all reviewers.
- **Cardinal:** Each reviewer j reports the scores for the items in S_j, that is, the values of $\{f_j(x_i)\}_{i \in S_j}$. A cardinal estimator observes the assignment A and the scores reported by all reviewers.

Observe that the setting described above considers "noiseless" data, where each reviewer reports either the scores $\{f_j(x_i)\}$ or the induced rankings. We provide an extension to the noisy setting in Section 2.5.1.

In order to compare the performance of different estimators, we use the notion of *strict uniform dominance*. Informally, we say that one estimator strictly uniformly dominates another if it incurs a strictly lower risk for all possible choices of the miscalibration functions and the item values.

In more detail, suppose that you wish to show that an estimator $\widehat{\pi}_1$ is superior to estimator $\widehat{\pi}_2$ with respect to some metric for estimating π^*. However, there is a clever adversary who intends to thwart your attempts. The adversary can choose the miscalibration functions of all reviewers and the values of all items, and moreover, can tailor these choices for different realizations of π^*. Formally, the adversary specifies a set of values $\{f_1^\pi, \ldots, f_m^\pi, x_1^\pi, \ldots, x_n^\pi\}_{\pi \in \Pi}$. The only constraints in this choice are that the miscalibration functions $f_1^\pi, \ldots, f_m^\pi$ must be strictly monotonic and that the item values $x_1^\pi, \ldots, x_n^\pi$ should induce the ranking π. In the sequel, we consider two ways of choosing the true ranking π^*: In one setting, π^* can be chosen by the adversary, and in the second setting π^* is drawn uniformly at random from Π. Once this ranking π^* is chosen, the actual values of the miscalibration functions and the item values are set as $f_1^{\pi^*}, \ldots, f_m^{\pi^*}$ and $x_1^{\pi^*}, \ldots, x_n^{\pi^*}$. The items are then assigned to reviewers according to the (possibly random) assignment A. The reviewers now provide their ordinal or cardinal evaluations as described earlier, and these evaluations are used to compute and evaluate the two estimators $\widehat{\pi}_1$ and $\widehat{\pi}_2$. We say that estimator $\widehat{\pi}_1$ strictly uniformly dominates $\widehat{\pi}_2$, if $\widehat{\pi}_1$ is always guaranteed to incur a strictly smaller (expected) error than $\widehat{\pi}_2$. Formally:

Definition 2.1 (Strict uniform dominance). *Let $\widehat{\pi}_1$ and $\widehat{\pi}_2$ be two estimators for the true ranking π^*. Estimator $\widehat{\pi}_1$ is said to strictly uniformly dominate estimator $\widehat{\pi}_2$ with respect to a given loss $L : \Pi \times \Pi \to \mathbb{R}$ if*

$$\mathbb{E}[L(\pi^*, \widehat{\pi}_1)] < \mathbb{E}[L(\pi^*, \widehat{\pi}_2)] \qquad \textit{for all permissible } \{f_1^\pi, \ldots, f_m^\pi, x_1^\pi, \ldots, x_n^\pi\}_{\pi \in \Pi}. \tag{2.1}$$

The expectation is taken over any randomness in the assignment A and the estimators. If the true *ranking π^* is drawn at random from a fixed distribution, then the expectation is also taken over this distribution; otherwise, inequality* (2.1) *must hold for all values of* π^*.

Note that strict uniform dominance is a stronger notion than comparing estimators in terms of their minimax (worst-case) or average-case risks. Moreover, if an estimator $\widehat{\pi}_2$ is strictly uniformly dominated by some estimator $\widehat{\pi}_1$, then the estimator $\widehat{\pi}_2$ is inadmissible.

Finally, for ease of exposition, we focus on the 0-1 loss in the main text:

$$L(\pi^*, \pi) = \mathbb{1}\{\pi^* \neq \pi\},$$

where we use the standard notation $\mathbb{1}\{A\}$ to denote the indicator function of an event A, where $\mathbb{1}\{A\} = 1$ if the event A is true, and 0 otherwise. Extensions to other metrics of Kendall-tau distance and Spearman's footrule distance are provided in Section 2.5.2.

2.3 Main results

In this section we present our main theoretical results. All proofs are provided in Chapter 9.

2.3.1 A canonical setting

We begin with a canonical setting that involves two items and two reviewers (that is, $n = 2$, $m = 2$), where each reviewer evaluates one of the two items. Our analysis for this setting conveys the key ideas underlying our general results. These ideas are directly applicable towards designing uniformly superior estimators for a variety of applications, and we subsequently demonstrate this general utility with two applications.

In this canonical setting, each of the two reviewers evaluates one of the two items chosen uniformly at random without replacement, that is, the assignment A is chosen uniformly at random from the two possibilities $(S_1 = 1, S_2 = 2)$ and $(S_1 = 2, S_2 = 1)$. Since each reviewer is assigned only one item, the ordinal data is vacuous. Then the natural ordinal baseline is an estimator which makes a guess uniformly at random:

$$\widehat{\pi}_{\text{can}}(A, \{\}) = \begin{cases} 1 \succ 2 & \text{with probability } 0.5 \\ 2 \succ 1 & \text{with probability } 0.5. \end{cases}$$

In the cardinal setting, let y_1 denote the score reported for item 1 by its respective reviewer, and let y_2 denote the score for item 2 reported by its respective reviewer. Since the calibration functions are arbitrary (and may be adversarial), it appears hopeless to obtain information about the relative values of x_1 and x_2 from just this data. Indeed, as we show below, standard estimators such as the sign test — ranking the items in terms of their reviewer-provided scores — provably fail to achieve this goal. More generally, the following theorem holds for the class of all deterministic estimators, that is, estimators given by deterministic mappings from $\{A, y_1, y_2\}$ to the set $\{1 \succ 2, 2 \succ 1\}$.

Theorem 2.2. *No deterministic (cardinal or ordinal) estimator can strictly uniformly dominate the random-guessing estimator* $\widehat{\pi}_{\textit{can}}$.

This theorem demonstrates the difficulty of this problem by ruling out all deterministic estimators. Our original question then still remains: is there any estimator that can strictly uniformly outperform the random-guessing ordinal baseline?

We show that the answer is yes, with the construction of a randomized estimator for this canonical setting, denoted as $\widetilde{\pi}^{\text{our}}_{\text{can}}$. This estimator is based on a function $w : [0, \infty) \to [0, 1)$ which may be chosen as any arbitrary strictly-increasing function. For instance, one could choose $w(x) = \frac{x}{1+x}$ or w as the sigmoid function. Given the scores y_1, y_2 reported for the two items, let $\widehat{i}^{(1)} \in \operatorname{argmax}_{i \in \{1,2\}} y_i$ denote the item which receives the higher score, and let $\widehat{i}^{(2)}$ denote the remaining item (with ties broken uniformly). Then our randomized estimator outputs:

$$\widetilde{\pi}^{\text{our}}_{\text{can}}(A, y_1, y_2) = \begin{cases} \widehat{i}^{(1)} \succ \widehat{i}^{(2)} & \text{with probability } \frac{1+w(|y_1-y_2|)}{2} \\ \widehat{i}^{(2)} \succ \widehat{i}^{(1)} & \text{otherwise.} \end{cases} \tag{2.2}$$

Note that the the output of this estimator is independent of the assignment A, so in the remainder of this chapter we also denote this estimator as $\widetilde{\pi}^{\text{our}}_{\text{can}}(y_1, y_2)$.

The following theorem now proves that our proposed estimator indeed achieves the stated goal.

Theorem 2.3. *The randomized estimator $\widetilde{\pi}^{our}_{can}$ strictly uniformly dominates the random-guessing baseline $\widehat{\pi}_{can}$.*

While this result considers a setting with "noiseless" observations (that is, where $y = f(x)$), in Section 2.5.1 we show that the guarantee for $\widetilde{\pi}^{\text{our}}_{\text{can}}$ continues to hold when the observations are noisy.

Having established the positive result for this canonical setting, we now discuss some connections and inspirations in the literature.

Connections to the literature

The canonical setting has a close connection to the randomized version of the two-envelope problem [45]. In the two-envelope problem, there are two arbitrary numbers. One of the two numbers is observed uniformly at random, and the other remains unknown. The goal is to estimate which number is larger. This problem can also be viewed from a game-theoretic perspective [71] as ours, where one player picks an estimator and the other player picks the two values. Cover [45] proposed a randomized estimator whose probability of success is strictly larger than 0.5 uniformly across all arbitrary pairs of numbers. The proposed estimator samples a new random variable Z whose distribution has a probability density function p with $p(z) > 0$ for all $z \in \mathbb{R}$. Then if the observed number is smaller than Z, the estimator decides that the observed number is the smaller number; if the observed number is larger than Z, the estimator decides that the observed number is the larger number.

Our canonical setting can be reduced to the two-envelope problem as follows. Consider the two values $f_1(x_1) - f_2(x_2)$ and $f_1(x_2) - f_2(x_1)$. Since the two items are assigned to the two reviewers uniformly at random, we observe one of these two values uniformly at random. By the assumption that f_1 and f_2 are monotonically increasing, we know that these two values are distinct, and furthermore, $f_1(x_1) - f_2(x_2) > f_1(x_2) - f_2(x_1)$ if and only if $x_1 > x_2$. Hence, the relative ordering of these two values is identical to the relative ordering of x_1 and x_2, reducing our canonical setting to the two-envelope problem. Our estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ also carries a close connection to Cover's estimator to the two-envelope problem. Specifically, Cover's estimator can be equivalently viewed as being designated by a "switching function" [121]. This switching function specifies the probability to "switch" (that is, to guess that the unobserved value is larger), and is a monotonically-decreasing function in the observed value. The use of the monotonic function w in our estimator in (2.2) is similar in spirit.

The two-envelope problem can also be alternatively viewed as a secretary problem with two candidates. Negative results have been shown regarding the effect of cardinal vs. ordinal data when there are more than two candidates [72, 165], and positive result has been shown on extensions of the secretary problem to different losses [73].

Our original inspiration for our proposed estimator arose from Stein's phenomenon [166] and empirical Bayes [145]. This inspiration stems for the fact that the two items are not to be estimated in isolation, but in a joint manner. That said, a significant fraction of the work (e.g., [11, 22, 91, 145, 166, 178]) in these areas is based on deterministic estimators. In comparison, our negative result for all deterministic estimators (Theorem 2.2) and the positive result for our randomized estimator (Theorem 2.3) provide interesting insights in this space.

2.3.2 A/B testing

We now demonstrate how to use the result in the canonical setting as a plug-in for more general scenarios. Specifically, we construct simple extensions to our canonical estimator, as a proof-of-concept for the superiority of cardinal data over ordinal data in A/B testing (this section) and ranking (Section 2.3.3). A/B testing is concerned with the problem of choosing the better of two given items, based on multiple evaluations of each item, and is used widely for the web and e-commerce (e.g. [103]). In many applications of A/B testing, the two items are rated by disjoint sets of individuals (for example, when comparing two web designs, each user sees one and only one design). It is therefore important to take into account the different calibrations of different individuals, and this problem fits in our setting with $n = 2$ items and m reviewers. For simplicity, we assume that m is even. We consider the assignment obtained by assigning item 1 to some $m/2$ reviewers chosen uniformly at random (without replacement) from the set of m reviewers, and assigning item 2 to the remaining $m/2$ reviewers.[2]

As in the canonical setting we studied earlier, in the absence of any direct comparison between the two items, a natural ordinal estimator in the A/B testing setting is a random guess:

$$\widehat{\pi}_{\text{ab}}(A, \{\}) = \begin{cases} 1 \succ 2 & \text{with probability } 0.5 \\ 2 \succ 1 & \text{with probability } 0.5. \end{cases}$$

For concreteness, we consider the following method of performing the random assignment of the two items to the m reviewers. We first perform a uniformly random permutation of the m reviewers, and then assign the first $m/2$ reviewers in this permutation to item 1; we assign the last $m/2$ reviewers in this permutation to item 2. We let $y_1^{(1)}, \ldots, y_1^{(m/2)}$ denote the scores

[2] Our results also hold in the following settings: (a) Each reviewer is assigned one of the two items independently and uniformly at random. (b) Reviewers are grouped (in any arbitrary manner) into $m/2$ pairs, and within each pair, the two reviewers are assigned one distinct item each uniformly at random.

given by the $m/2$ reviewers to item 1, and let $y_2^{(1)}, \ldots, y_2^{(m/2)}$ denote the scores given by the $m/2$ reviewers assigned to item 2. Namely, the reviewers (in the permuted order) provide the scores $[y_1^{(1)}, \ldots, y_1^{(m/2)}, y_2^{(1)}, \ldots, y_2^{(m/2)}]$. Now consider the following standard (deterministic) estimators:

- *Sign estimator:* The sign estimator outputs the item which has more pairwise wins: $\sum_{j=1}^{m/2} \mathbb{1}\{y_1^{(j)} > y_2^{(j)}\} \underset{2 \succ 1}{\overset{1 \succ 2}{\gtrless}} \sum_{j=1}^{m/2} \mathbb{1}\{y_2^{(j)} > y_1^{(j)}\}$.
- *Mean estimator:* The mean estimator outputs the item with the higher mean score: $\mathrm{mean}(y_1^{(1)}, \ldots, y_1^{(m/2)}) \underset{2 \succ 1}{\overset{1 \succ 2}{\gtrless}} \mathrm{mean}(y_2^{(1)}, \ldots, y_2^{(m/2)})$.
- *Median estimator:* The median estimator outputs the item with the higher median score (upper median if there are multiple medians)[3]: $\mathrm{median}(y_1^{(1)}, \ldots, y_1^{(m/2)}) \underset{2 \succ 1}{\overset{1 \succ 2}{\gtrless}} \mathrm{median}(y_2^{(1)}, \ldots, y_2^{(m/2)})$.

In each estimator, ties are assumed to be broken uniformly at random.

We now show that despite using the scores given by all m reviewers, where m can be arbitrarily large, these natural estimators fail to uniformly dominate the naïve random-guessing ordinal estimator.

Theorem 2.4. *For any (even) number of reviewers, none of the sign, mean, and median estimators can strictly uniformly dominate the random-guessing estimator $\widehat{\pi}_{ab}$.*

The negative result of Theorem 2.4 demonstrates the challenges even when one is allowed to collect an arbitrarily large number of scores for each item. Intuitively, the more reviewers there are, the more miscalibration functions they introduce. Even if the statistics used by these estimators converge as the number of the reviewers m grows large, these values are not guaranteed to be informative towards comparing the values of the items due to the miscalibrations.

The failure of these standard estimators suggests the need of a novel approach towards this problem of A/B testing under arbitrary miscalibrations. To this end, we build on top of our canonical estimator $\widetilde{\pi}_{\text{can}}^{\text{our}}$ from Section 2.3.1, and present a simple randomized estimator $\widetilde{\pi}_{\text{ab}}^{\text{our}}$ as follows:

(1) For every $j \in [m/2]$, use the canonical estimator $\widetilde{\pi}_{\text{can}}^{\text{our}}$ on the j^{th} pair of scores $(y_1^{(j)}, y_2^{(j)})$ and obtain the estimate $r_j := \widetilde{\pi}_{\text{can}}^{\text{our}}(y_1^{(j)}, y_2^{(j)}) \in \{1 \succ 2, 2 \succ 1\}$.

(2) Set the output $\widetilde{\pi}_{\text{ab}}^{\text{our}}$ as the outcome of the majority vote among the estimates $\{r_j\}_{j \in [m/2]}$ with ties broken uniformly at random.

The following theorem now shows that the results for the canonical setting from Section 2.3.1 translate to this A/B testing application.

Theorem 2.5. *For any (even) number of reviewers, the estimator $\widetilde{\pi}_{ab}^{our}$ strictly uniformly dominates the random guessing estimator $\widehat{\pi}_{ab}$.*

This result thus illustrates the use of our canonical estimator $\widetilde{\pi}_{\text{can}}^{\text{our}}$ as a plug-in for A/B testing. So far we have considered settings where there are only two items and where each reviewer is

[3] For values $a_1 \geq \cdots \geq a_n$, we define the median function as the upper median, $\mathrm{median}(a_1, \ldots, a_n) = a_{\lfloor (n+1)/2 \rfloor}$. Theorem 2.4 also holds instead for the lower median $a_{\lfloor (n+2)/2 \rfloor}$, and the median defined as the mean of the two middle values, $(a_{\lfloor (n+1)/2 \rfloor} + a_{\lfloor (n+2)/2 \rfloor})/2$.

assigned only one item, thereby making the ordinal information vacuous. We now turn to an application that is free of these restrictions.

2.3.3 Ranking

It is common in practice to estimate the partial or total ranking for a list of items by soliciting ordinal or cardinal responses from individuals. In conference reviews or peer-grading, each reviewer is asked to rank [54, 155, 159] or rate [65, 136, 159] a small subset of the papers, and this information is subsequently used to estimate a partial or total ranking of the papers (or student homework). Other applications for aggregating rankings include voting [139, 193], crowdsourcing [154, 156], recommendation systems [64] and meta-search [56].

Formally, we let $n > 2$ denote the number of items and m denote the number of reviewers. For simplicity, we focus on a setting where each reviewer reports noiseless evaluations of some pair of items, and the goal is to estimate the total ranking of all items. We consider a random design setup where the pairs compared are randomly chosen and randomly assigned to reviewers. We assume $1 < m < \binom{n}{2}$ so that the problem does not degenerate. Each reviewer evaluates a pair of items, and these pairs are drawn uniformly without replacement from the $\binom{n}{2}$ possible pairs of items. We let $A = (S_1, \ldots, S_m)$ denote these m pairs of items to be evaluated by the m respective reviewers, where $S_j \in [n] \times [n]$ denotes the pair of items evaluated by reviewer $j \in [m]$. For each pair $S_j = (i, i')$, denote the cardinal evaluation as $y(S_j) = (f_j(x_i), f_j(x_{i'}))$, and the ordinal evaluation as the induced ranking $b(S_j) \in \{i \succ i', i' \succ i\}$. Denote the set of ordinal observations as $\mathcal{B} = \{b(S_j)\}_{j=1}^m$, and the set of cardinal observations as $\mathcal{Y} = \{y(S_j)\}_{j=1}^m$. The input to an ordinal estimator is the ordinal information $\mathcal{B}$. The input to a cardinal estimator is the reviewer assignment A and the set of cardinal observations $\mathcal{Y}$. Finally, let $\mathcal{G}(\mathcal{B})$ denote a directed acyclic graph (DAG) with nodes comprising the n items and with an edge from any node i to any other node i' if and only if $\{i \succ i'\} \in \mathcal{B}$. A topological ordering on $\mathcal{G}$ is any total ranking of its vertices which does not violate any pairwise comparisons indicated by $\mathcal{B}$.

We now present our (randomized) cardinal estimator $\widetilde{\pi}_{\text{rank}}^{\text{our}}(A, \mathcal{Y})$ in Algorithm 1. In words, this algorithm start from any topological ordering of the items as the initial estimate of the true ranking. Then the algorithm scans one-by-one over the pairs with adjacent items in the initial estimated ranking. If a pair can be flipped (that is, if the ranking after flipping this pair is also a topological ordering), we uniformly sample a pair of scores for these two items from the cardinal observations $\mathcal{Y}$, and use the randomized estimator $\widetilde{\pi}_{\text{can}}^{\text{our}}$ to determine the relative order of the pair. After $\widetilde{\pi}_{\text{can}}^{\text{our}}$ is called, the positions of this pair are finalized. We remove all scores of these two reviewers from future use, and jump to the next pair that does not contain these two items.

The following theorem now presents the main result of this section.

Theorem 2.6. *Suppose that the true ranking π^* is drawn uniformly at random from the collection of all possible rankings, and consider any ordinal estimator $\widehat{\pi}_{rank}$ for π^*. Then the cardinal estimator $\widetilde{\pi}_{rank}^{our}$ strictly uniformly dominates the ordinal estimator $\widehat{\pi}_{rank}$.*

We note that Algorithm 1 runs in polynomial time (in the number of items n) because the two major operations of this estimator – finding a topological ordering, and checking if a ranking is a topological ordering on the DAG – can be implemented in polynomial time [50]. Theorem 2.6 thus demonstrates again the power of the canonical estimator $\widetilde{\pi}_{\text{can}}^{\text{our}}$ as a plug-in component to be

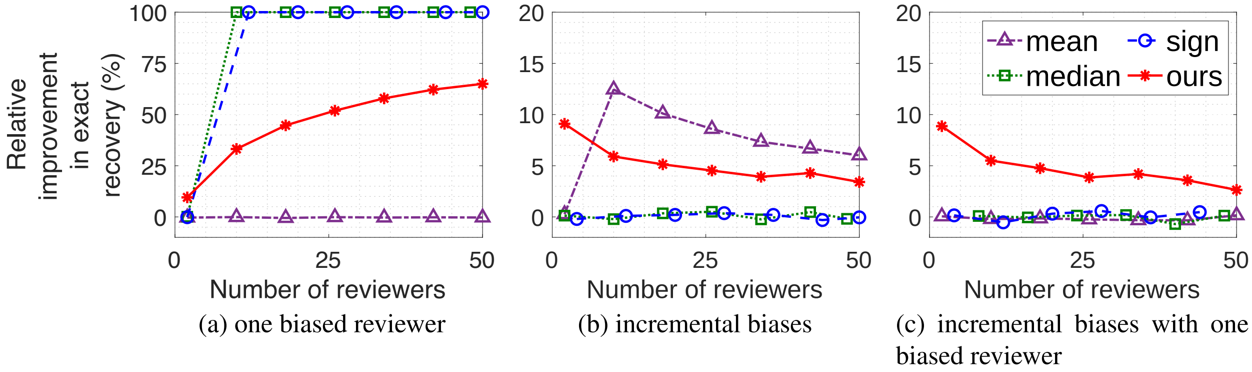
Algorithm 1: Our cardinal ranking estimator $\widetilde{\pi}^{\text{our}}_{\text{rank}}(A, \mathcal{Y})$.

```
Deduce the ordinal observations $\mathcal{B}$ from the cardinal observations $\mathcal{Y}$.
Compute a topological ordering $\widehat{\pi}$ on the graph $\mathcal{G}(\mathcal{B})$, with ties broken in order of the
    indices of the items.
$t \leftarrow 1$.
while $t < n$ do
    Let $\widehat{\pi}_{\text{flip}}$ be the ranking obtained by flipping the positions of the $t^{th}$ and the $(t+1)^{th}$
        items in $\widehat{\pi}$.
    if $\widehat{\pi}_{flip}$ is a topological ordering on $\mathcal{G}(\mathcal{B})$, and both the $t^{th}$ and $(t+1)^{th}$ items are
        evaluated by at least one reviewer each in $\mathcal{Y}$ then
        From all of the scores of the $t^{th}$ item in $\mathcal{Y}$, sample one uniformly at random and
            denote it as $y_{\widehat{\pi}(t)}$. Likewise denote $y_{\widehat{\pi}(t+1)}$ as a randomly chosen score of the
            $(t+1)^{th}$ item from $\mathcal{Y}$.
        Consider the two reviewers reporting the scores $y_{\widehat{\pi}(t)}$ and $y_{\widehat{\pi}(t+1)}$. Remove from
            $\mathcal{Y}$ all scores provided by these two reviewers.
        if $\widetilde{\pi}^{our}_{can}(y_{\widehat{\pi}(t)}, y_{\widehat{\pi}(t+1)})$ outputs $\widehat{\pi}(t+1) \succ \widehat{\pi}(t)$ then
            $\widehat{\pi} \leftarrow \widehat{\pi}_{\text{flip}}$.
        end
        $t \leftarrow t + 2$.
    else
        $t \leftarrow t + 1$.
    end
end
Output $\widetilde{\pi}^{\text{our}}_{\text{rank}}(A, \mathcal{Y}) = \widehat{\pi}$.
```

used in a variety of applications. An extension of our results to the setting where π^* can be arbitrary (adversarially chosen) is presented in Section 2.5.3.

2.4 Simulations

We now experimentally evaluate our proposed estimators for A/B testing and ranking. Since the performance of the ordinal estimators vary significantly in different problem instances, we use the notion of "relative improvement". The relative improvement $\rho_{\widehat{\pi}}(\widetilde{\pi})$ of an estimator $\widetilde{\pi}$ as compared to a baseline estimator $\widehat{\pi}$ is defined as: $\rho_{\widehat{\pi}}(\widetilde{\pi}) = \frac{\mathbb{E}[L(\pi^*,\widehat{\pi})]-\mathbb{E}[L(\pi^*,\widetilde{\pi})]}{\mathbb{E}[L(\pi^*,\widehat{\pi})]} \times 100\%$. A positive value of the relative improvement $\rho_{\widehat{\pi}}(\widetilde{\pi})$ indicates the superiority of estimator $\widetilde{\pi}$ over the estimator $\widehat{\pi}$. A relative improvement of zero indicates an identical performance of the two estimators. In our proposed estimators, the function w is set as $w(x) = \frac{x}{1+x}$.



(a) one biased reviewer

(b) incremental biases

(c) incremental biases with one biased reviewer

Figure 2.1: Relative improvement in exact recovery of various estimators as compared to the random-guessing ordinal estimator $\widehat{\pi}_{\text{ab}}$ for A/B testing. Each point is an average over $10,000$ trials. The error bars are too small to display.

2.4.1 A/B testing

We now present simulations to evaluate various points on the bias-variance tradeoff. For A/B testing, we compare our estimator $\widetilde{\pi}^{\text{our}}_{\text{ab}}$ with other standard estimators — the sign, mean and median estimators introduced in Section 2.3.2. The item values x_1 and x_2 are chosen independently and uniformly at random from the interval $[0, 1]$. The calibration functions are linear and given by:

(a) *One biased reviewer:* One reviewer gives an abnormally (high or low) score. Formally, $f_j(x) = x$ for $j \in [m-1]$, and $f_m(x) = x + m$.

(b) *Incremental biases:* Calibration functions of reviewers are shifted from each other. Formally, $f_j(x) = x + j$ for $j \in [m]$.

(c) *Incremental biases with one biased reviewer:* A combination of setting (a) and setting (b). Formally, $f_j(x) = x + (j-1)$ for $j \in [m-1]$, and $f_m(x) = x + \frac{m(m-1)}{2}$.

We simulate and compute the relative improvement of the different estimators as compared to the random-guessing estimator $\widehat{\pi}_{\text{ab}}$. The results are shown in Figure 2.1. While the performance of the estimators vary with respect to each other, our estimator consistently beats the baseline whereas every other estimator fails. Our estimator thus indeed operates at a unique point on the bias-variance tradeoff with a low (zero) bias and a variance strictly smaller than the ordinal estimators, whereas all other estimators incur a non-zero error due to bias.

2.4.2 Ranking

Next, we evaluate the performance of our ranking estimator $\widetilde{\pi}^{\text{our}}_{\text{rank}}$ when the true ranking π^* is drawn from a uniform prior. We compare this estimator with an optimal ordinal estimator $\widehat{\pi}_{\text{rank}}$ which outputs a topological ordering with ties broken in order of the indices of the items (this ordinal estimator is optimal regardless of the tie-breaking strategy).

For any number of items n, we generate the values $x_1, \ldots, x_n$ of the items i.i.d. uniformly from the interval $[0, n]$. We set $m = \lfloor \frac{1}{2}\binom{n}{2} \rfloor$. We assume that the j^{th} reviewer has a linear

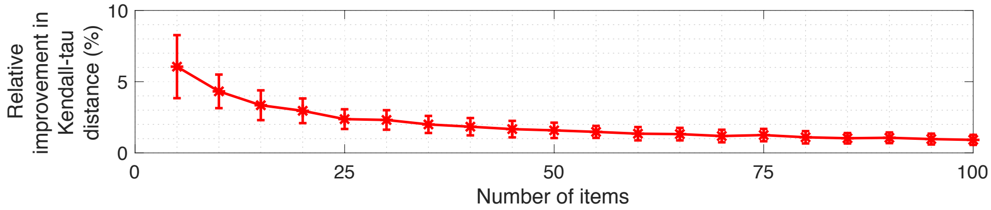


Figure 2.2: Relative improvement in Kendall-tau distance of our ranking estimator $\widetilde{\pi}_{\text{rank}}^{\text{our}}$ as compared to an optimal ordinal estimator $\widehat{\pi}_{\text{rank}}$ for ranking. Each point is an average over 100 trials, where in each trial the quantities $\mathbb{E}[L(\pi^*, \widetilde{\pi}_{\text{rank}}^{\text{our}})]$ and $\mathbb{E}[L(\pi^*, \widehat{\pi}_{\text{rank}})]$ are approximated by an empirical average over 1000 samples.

calibration function $f_j(x) = k_j x + b_j$, where we sample k_j and b_j i.i.d. uniformly from the interval $[0, 1]$.

We have previously proved that our estimator $\widetilde{\pi}_{\text{rank}}^{\text{our}}$ based on cardinal data can strictly uniformly outperform the optimal ordinal estimator for the 0-1 loss. We use these simulations to evaluate the efficacy of our approach for a different loss function – Kendall-tau distance. Specifically, Figure 2.2 compares these two estimators in terms of Kendall-tau distance (Section 2.5.2 provides a formal definition of this distance and associated theoretical results). We observe that our estimator $\widetilde{\pi}_{\text{rank}}^{\text{our}}$ is able to consistently yield improvements even for this loss. The reason that the improvement becomes smaller when the number of items is large is that by flipping pairs, our estimator only modifies the ranking in the neighborhood of the initial estimate. We strongly believe that it should be possible to design better estimators for the large n regime using the tools developed in this chapter. Having met our stated goal of outperforming ordinal estimators to handle arbitrary miscalibrations, we leave this interesting problem for future work.

2.4.3 Tradeoff between estimation under perfect calibration vs. miscalibration

In this section, we present a preliminary experiment showing the tradeoff between estimation under perfect calibration (all reviewers reporting the true values of the papers) and estimation under miscalibration. For simplicity, we consider the canonical setting from Section 2.3.1. We evaluate the performance of our estimator under two scenarios: (1) perfect calibration, where $f_j(x) = x$ for each $j \in \{1, 2\}$; (2) miscalibration with one biased reviewer, where $f_1(x) = x$ and $f_2(x) = x + 1$. We consider the function w in our estimator as $w(x) = \frac{\gamma x}{1+\gamma x}$, where $\gamma \in \{2^k \mid -10 \leq k \leq 10, k \in \mathbb{Z}\}$. We sample x_1 and x_2 uniformly at random from the interval $[0, 1]$.

Figure 2.3 shows the relative improvement of our estimator over the random-guessing baseline under perfect calibration and under miscalibration, where γ increases from left to right. Let us focus on a few regimes in this plot. First, on the left end of the curve, when γ is close to 0, we have $w(x)$ close to 0. The estimator is close to random-guessing. At the other extreme, on the right end of the curve, when γ goes to infinity, we have $w(x)$ close to 1. The estimator always outputs the item with the higher score, and hence gives perfect estimation under perfect calibration. Under miscalibration, this estimator always chooses the biased reviewer giving the higher score and hence performs the same as random guess. Past the maximum point of the function at approximately $(25\%, 9\%)$ when $\gamma = 1$, the value of the curve starts decreasing, suggesting a tradeoff of estimation accuracy under perfect calibration and under miscalibration. It is clear that points to the left of the maximum point are not Pareto-efficient, since there exist other points with the same accuracy under miscalibration but improved accuracy under perfect calibration.

We thus see that robustness under arbitrary miscalibration comes at a cost of lower accuracy under perfect calibration. Establishing a formal understanding of this tradeoff and designing estimators that are provably Pareto-efficient are important open problems.

2.5 Extensions

We now present three extensions of our problem setting and results from the main text.

2.5.1 Noisy data

In this section, we show that even when the scores given by the reviewers are noisy, our estimator in (2.2) continues to strictly uniformly dominate random guessing in the canonical setting (Section 2.3.1). We focus on the canonical estimator.

In the noisy setting, when reviewer $j \in [m]$ evaluates item $i \in [n]$, the reported score is

$$f_j(x_i) + \epsilon_{ij},$$

where ϵ_{ij} is a noise term. We assume that the noise terms $\{\epsilon_{ij}\}_{i \in [n], j \in [m]}$ are drawn i.i.d. from an unknown distribution. In this setting of noisy reported scores, we modify Definition 2.1 of strict uniform dominance, and let the expectation include the randomness in the noise.

The following theorem establishes the strict uniform dominance in the noisy setting for the cardinal estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ in (2.2) (cf. Theorem 2.3 for the noiseless setting).

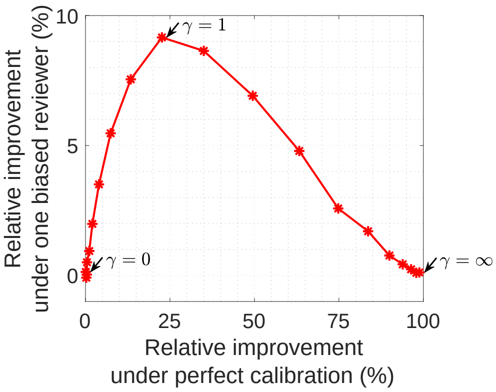


Figure 2.3: Relative improvement of our canonical estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ under perfect calibration and under miscalibration of one biased reviewer, with $w(x) = \frac{\gamma x}{1+\gamma x}$ and $\gamma \in \{2^k \mid -10 \leq k \leq 10, k \in \mathbb{Z}\}$, where γ increases from left to right in the plot. Each point is an average over 5×10^5 trials. The error bars are too small to display.

Theorem 2.7. *The canonical estimator $\widetilde{\pi}^{our}_{can}$ strictly uniformly dominates the random-guessing estimator $\widehat{\pi}_{can}$ in the presence of noise.*

The proof of this theorem is in Section 9.6. Observe that this result is quite general, since the noise distribution can be arbitrary and unknown.

2.5.2 Ranking under Kendall-tau and Spearman's footrule distance

In addition to the 0-1 exact recovery loss considered in Theorem 2.6, Kendall-tau distance and Spearman's footrule distance are also common metrics for ranking. Recall that a ranking of n items is defined by a function $\pi : [n] \to [n]$, such that $\pi(t)$ is the index of the t^{th} ranked item for each $t \in [n]$. Equivalently, we can define a ranking by the function $\sigma : [n] \to [n]$, such that $\sigma(i)$ is the rank of each item $i \in [n]$. With this notation, we have the relation $\sigma = \pi^{-1}$.

The Kendall-tau distance and the Spearman's footrule distance are usually defined in terms of the ranking σ. Hence for consistency with these definitions, throughout this section we focus on the rankings as defined by σ (instead of π as done throughout the remainder of the chapter). Kendall-tau distance and Spearman's footrule distance between any two rankings σ_1 and σ_2 of n items are defined as:

$$\text{Kendall-tau distance:} \qquad L_{\text{KT}}(\sigma_1, \sigma_2) = \sum_{\substack{i \in [n], i' \in [n]: \\ \sigma_1(i) < \sigma_1(i')}} \mathbb{1}\{\sigma_2(i) > \sigma_2(i')\}$$

$$\text{Spearman's footrule distance:} \qquad L_{\text{SF}}(\sigma_1, \sigma_2) = \sum_{i \in [n]} |\sigma_1(i) - \sigma_2(i)|.$$

The following theorem states that given any arbitrary ordinal estimator, there exists a cardinal estimator that performs strictly uniformly better than this ordinal estimator, simultaneously on Kendall-tau distance and Spearman's footrule distance (cf. Theorem 2.6 for 0-1 loss).

Theorem 2.8. *Suppose that the true ranking σ^* is drawn uniformly at random from the collection of all possible rankings. For any arbitrary ordinal estimator $\widehat{\sigma}_{rank}$, there exists a cardinal estimator with access to one call to the ordinal estimator $\widehat{\sigma}_{rank}$ that strictly uniformly dominates the ordinal estimator $\widehat{\sigma}_{rank}$ with respect to Kendall-tau distance and Spearman's footrule distance. The computatinal complexity of this cardinal estimator is polynomial in the number of items n, in addition to the time taken by one call to the ordinal estimator $\widehat{\sigma}_{rank}$.*

The proof of this result is in Section 9.7. This result demonstrates the generality of our results in the main text with respect to various (not only 0-1) loss functions.

2.5.3 Ranking under arbitrary true ranking

Theorem 2.6 in Section 2.3.3 compared our cardinal estimator with arbitrary ordinal estimators under a uniform prior over the true ranking. In this section, we present a result for ranking under any arbitrary true ranking. This setting is more similar to our results on the canonical setting (Theorem 2.3) and A/B testing (Theorem 2.5) in the main text. When the true ranking is arbitrary, a minimax-optimal ordinal estimator outputs uniformly at random a topoglocial ordering consistent with the pairwise comparisons. We denote this optimal ordinal estimator as $\widehat{\pi}_{\text{rank-unif}}$.

Given this ordinal estimator, we then construct a cardinal estimator $\widetilde{\pi}^{\text{our}}_{\text{rank-unif}}$ by simply setting the initial estimate $\widehat{\pi} = \widehat{\pi}_{\text{rank-unif}}(\mathcal{B})$ in Line 2 of Algorithm 1 (instead of executing the current Line 2). The following theorem states the desired result for strict uniform dominance of this cardinal estimator over the optimal ordinal estimator $\widehat{\pi}_{\text{rank-unif}}$.

Theorem 2.9. *When the true ranking is arbitrary, the cardinal estimator $\widetilde{\pi}^{our}_{rank\text{-}unif}$ strictly uniformly dominates the minimax-optimal ordinal estimator $\widehat{\pi}_{rank\text{-}unif}$.*

The proof of this Theorem is in Section 9.9. Importantly, we can think of this cardinal estimator as a post-processing step which builds on the output of the optimal ordinal estimator. This cardinal estimator takes polynomial time in the number of items n, in addition to the time taken by one call to the ordinal estimator $\widehat{\pi}_{\text{rank-unif}}$.

2.6 Discussion

Breaking the barrier of using only ranking data in the presence of arbitrary (and potentially adversarial) miscalibrations, we show that cardinal scores can yield strict and uniform improvements over rankings. This result uncovers a novel, strictly-superior point on the tradeoff between cardinal scores and ordinal rankings, and provides a new perspective on this eternally-debated tradeoff. Our estimator allows for easily plugging into a variety of algorithms, thereby yielding it a wide applicability.

The results of this chapter lead to several useful open problems. First, while our estimators indeed uniformly outperform ordinal estimators, in the future, a more careful design in our estimators (e.g. how to choose the function w in the canonical estimator, and how to design better estimators for A/B testing and ranking) may yield even better results. Second, it is of interest to obtain statistical bounds on the relative errors of the cardinal and ordinal estimators in terms of the unknown miscalibration functions. Third, a promising direction of future research is to design estimators that achieve the guarantees of our proposed estimator under arbitrary/adversarial miscalibrations while simultaneously being able to adapt and yield stronger guarantees when the calibration functions follow one of the popular simpler models of miscalibration (à la “win-win” models and estimators in prior work [162, Part I] [84, 154, 157, 158, 160]). Fourth, although we consider the rating scales as continuous intervals, it is not hard to see that our results extend to discrete scales (but with the strict inequality in Equation (2.1) sometimes replaced by a non-strict inequality to account for ties). Using our results to guide the choice of the scale used for elicitation is an open problem of interest. And finally, practical applications such as peer-review do not suffer from the problem of miscalibration in isolation. It is a useful and challenging open problem to address miscalibration simultaneously with other issues such as noise [167], subjectivity [130], strategic behavior [192] and others.

Acknowledgments

This work was supported in parts by NSF grants CRII: CIF: 1755656 and CCF: 1763734. We thank Bryan Parno for very useful discussions on biases in conference peer review. We thank Pieter Abbeel for pointing out the related work on the two-envelope problem.

Chapter 3

Calibration in General Settings: A Heuristic for Statistical Seriation

The estimators presented in Chapter 2 provide fundamental insights on the benefits of cardinal scores even in the presence of arbitrary miscalibration. However, these estimators do not straightforwardly generalize to settings. For example, the ranking estimator presented in Algorithm 1 of Section 2.3.3 starts with the assumption that there exists at least one topological ordering that is consistent with the scores given by all the reviewers. However, this assumption may not hold in real-life scenarios: it is natural to expect that some reviewers may not agree on the comparison of some pair of papers. In this chapter, we study calibration in general settings under the statistical seriation framework, where the goal is to estimate a matrix whose columns are assumed to satisfy an unknown permutation. This is a important classical problem, with close connections to statistical literature in permutation-based models. In addition to calibration, it also has wide applications ranging from archaeology to biology. Past work has shown that the least-squares estimator is optimal up to logarithmic factors, but efficient algorithms for computing the least-squares estimator remain unknown to date. We approach this important problem from a heuristic perspective. Specifically, we replace the combinatorial permutation constraint by a continuous regularization term, and then use projected gradient descent to obtain a local minimum of the non-convex objective. We show that the attained local minimum is the global minimum in certain special cases under the noiseless setting and preserves desirable properties under the noisy setting. Simulation results reveal that our proposed algorithm outperforms prior algorithms when (1) the underlying model is more complex than simplistic parametric assumptions such as low-rankedness, or (2) the signal-to-noise ratio is high. Under partial observations, the proposed algorithm requires an initialization, and different initializations may lead to different local minima. We empirically observe that the proposed algorithm yields consistent results regardless of intialization, even though different initializations start with different levels of quality.

3.1 Introduction

Seriation refers to the problem of identifying a sequential ordering of the data such that "the position of each unit reflects its similarity to other units" [118]. For example, in archaeology seriation is used to identify the chronological ordering of historical artifacts (see [118] and references therein). Other applications include ecology (identifying ages of fossil sites [114]), biology (discovering gene expression patterns [31]), and operations research (understanding the interactions between organizations [120]), just to name a few. From the statistical perspective, termed "statistical seriation", seriation is formulated as a matrix estimation problem, where the rows of the matrix are assumed to satisfy the same shape constraint after an unknown permutation of the columns [62]. One common shape constraint is that the rows are monotonically increasing after the permutation of the columns, and in this chapter we focus on this monotonic case. We refer the reader to the papers [62, 107] for surveys of (statistical) seriation in various applications.

Statistical seriation also forms a fundamental building block for many other problems, and ideas on solving statistical seriation may be applicable to estimation under closely-related "permutation-based" models, which involve matrices that are monotonic up to unknown permutations of rows and/or columns. Permutation-based models arise in a variety of applications including estimating pairwise comparison probabilities [108, 117, 158], crowdsourced labeling [161], matrix completion [160], passive [84] and active ranking [154]. A key challenge in these applications, as well as in the statistical seriation problem, is the presence of unknown permutations.

An additional application of statistical seriation is miscalibration in peer review [186]. This application involves a collection of reviewers and papers, where each reviewer provides ratings to their assigned subset of papers. In this context, the ratings of each reviewer is represented by a row in a matrix, and the papers represented by the columns inherit an ordering. The goal is to estimate an underlying ordering of the papers. A key challenge is that reviewers may be miscalibrated, that is, different reviewers may have different rating scales. One model for miscalibration is to assume that there exists an underlying true value for each paper, and each row of the matrix (respresenting a reviewer) is some monotonic transformation of these true values combined with noise. In such applications, one prominent benefit of the statistical seriation model is that the permutation-based assumption is general, and does not impose overly-simplistic assumptions such as the matrix being low rank or having a specific parameter-based form. Hence, the seriation model is robust in modeling a broad class of true matrices and has low bias in estimation compared to specialized models that make parameter-based assumptions.

3.1.1 Problem formulation

We now introduce the formulation of statistical seriation. Let n and d be positive integers, and let $Y \in \mathbb{R}^{n\times d}$ be a real-valued matrix. Let Π_d be the set of all permutations of size d. For any permutation $\pi \in \Pi_d$, let $\mathcal{M}_\pi \subseteq \mathbb{R}^{n\times d}$ be the set of all matrices whose columns satisfy the ordering given by π. That is, for every matrix $A \in \mathcal{M}_\pi$, we have $A_{i,\pi(1)} \leq A_{i,\pi(2)} \leq \ldots \leq A_{i,\pi(d)}$ for every $i \in [n]$. Let $\mathcal{M} := \cup_{\pi\in\Pi_d}\mathcal{M}_\pi$ denote the set of all $(n \times d)$ matrices whose columns can be permuted such that every row is non-decreasing from left-to-right after some permutation of the columns. Statistical seriation assumes that observations are made in the form of

$$Y = A^* + Z, \tag{3.1}$$

where we have an unknown true matrix $A^* \in \mathcal{M}$, and the unknown matrix Z is a zero-mean sub-Gaussian random matrix that represents the noise. The goal of statistical seriation is to estimate the matrix A^* (and/or the ordering $\pi^* \in \Pi_d$ associated with it). A natural estimator for this problem is the least-squares estimator [62]

$$\widehat{A}_{\text{LS}} \in \operatorname*{argmin}_{A \in \mathcal{M}} \|A - Y\|_F^2. \tag{3.2}$$

The aforementioned description assumed that the matrix Y was fully observed, but this is rarely the case especially in applications such as peer grading or peer review, where each reviewer only evaluates a small subset of the items. Therefore, we also consider the setting of partial observations, where only a subset of entries $\Omega \subseteq [n] \times [d]$ in Y is observed. To this end, for any matrix $X \in \mathbb{R}^{n \times d}$, let $\|X\|_\Omega$ denote the Frobenius norm restricted to the set Ω, defined as $\|X\|_\Omega^2 = \sum_{(i,j) \in \Omega} X_{ij}^2$. Then the least-squares estimator under the case of partial observations finds the matrix within the domain $\mathcal{M}$ that best fits the observed entries:

$$\widehat{A}_{\text{LS}} \in \operatorname*{argmin}_{A \in \mathcal{M}} \|A - Y\|_\Omega^2. \tag{3.3}$$

The least-squares estimators (3.2) and (3.3) have desirable statistical properties. When the noise is i.i.d. normal, then they correspond to the maximum likelihood estimator (MLE). Furthermore, Flammarion et al. [62] shows that the least-squares estimator (3.2) is optimal up to logarithmic factors and adapts to matrices with a certain natural structure. However, despite the generality of the seriation model and the strong theoretical guarantees of the least-squares estimator, the unknown permutation π in (3.2) imposes computational challenges in solving (3.2) efficiently. If the permutation π were known, then A can be solved by isotonic regression taking $O(nd)$ time [12]. However, in (3.2) the permutation π is unknown, and naively brute-forcing all possible choices of π takes exponential time in d. Computationally efficient algorithms for computing (3.2) are not known to date [62]. Moreover, no algorithms have been found that are both efficient and statistically optimal (whether using the least-squares formulation (3.2) or not), showing an unclosed statistical-computation gap for the statistical seriation problem.

3.1.2 Our contributions

In this section, we outline the main contributions of this chapter and summarize our results.

Approach: A Heuristic Approximation The goal of our work is to provide a practical algorithm that heuristically approximates the solution to (3.3). Specifically, we approach the problem by replacing the combinatorial permutation constraint in (3.3) by a continuous regularization term while still capturing the permutation constraint. Formally, we define the following objective function $L : \mathbb{R}^{n \times d} \to \mathbb{R}$, parameterized by a tuning parameter $\lambda \geq 0$:

$$L(A) = L_{Y,\Omega,\lambda}(A) := \|A - Y\|_\Omega^2 + \lambda R(A). \tag{3.4}$$

where $R : \mathbb{R}^{n \times d} \to \mathbb{R}^{\geq 0}$ is a carefully-designed regularizer term to be explained in Section 3.2. Then our solution is computed by minimizing the objective as

$$\operatorname*{argmin}_{A \in [0,1]^{n \times d}} L(A). \tag{3.5}$$

Following Shah et al. [158], we assume Bernoulli noise Z in (3.1), and therefore restrict the domain of optimization (3.5) to $[0, 1]^{n \times d}$. Now that the objective is continuous and the domain is a closed bounded set, we use projected gradient descent to obtain a local minimum of this non-convex objective. Our approach is quite different from past work – past work has primarily focused on designing efficient algorithms that reduce the gap from the optimal estimator in terms of the statistical rates. On the other hand, we directly provide a heuristic for approximating the optimal estimator. We thus provide a new point of comparison in terms of the statistical and computational trade-off. Our approach thus provides new insights in terms of possible research directions to understand and address this statistical-computational gap.

Theoretical results We first theoretically analyze the stationary points of (3.5), and show that projected gradient descent converges to a stationary point (Section 3.4). Specifically, the attained stationary point recovers the exact input data in the noiseless case (Theorem 3.2) and has other desirable theoretical properties in the noisy case (Proposition 3.3 and Theorem 3.4). These theoretical results hold generally for any $\lambda \geq 0$. The theoretical results thus provide insights into our approach (3.5) to approximating statistical seriation, and provide justification for its validity.

Simulation results We then empirically evaluate our algorithm by simulation. Specifically, we examine the following aspects:

- **Accuracy-computational tradeoff of λ** We first observe that the tuning parameter λ induces an accuracy-computational tradeoff (Section 3.5.2). Specifically, when the value of λ increases, estimation achieves higher accuracy but gradient descent takes more iterations to converge.
- **Advantage under non-parametric models and high SNR** We then compare our estimator with various baselines under various models (Section 3.5.2). We observe that our estimator performs well when the true data violates simplistic parametric assumptions. This is because our estimator inherits the general formulation of statistical seriation, giving low bias in estimation. On the other hand, although the parametric baselines perform well when the true data is generated from such parametric models, they incur a large bias when the true data is not. In addition, our estimator especially performs well when the SNR is high. This is also expected, as noise is of low-rank in nature. Therefore, when the signal level relative to the noise is low, the noise overshadows the non-parametric structure of the true matrix.
- **Partial observations and initialization of gradient descent** Finally, we consider the case when the data is only partially observed (Section 3.5.3). In this case, the gradient descent algorithm requires an initialization on the unobserved entries of the matrix. Since the objective (3.4) is non-convex, gradient descent may converge to different local optima based on the initialization. We empirically observe that our algorithm consistently improves the estimation accuracy for different choices of initialization, although the amounts of error at the beginning of gradient descent are different for different initializations.

Putting the theoretical and empirical results together, our work demonstrates the effectiveness of our approach to approximating the solution of the least-squares estimator, and the generality of the approach inherited by the generality of the seriation model.

3.2 Our proposed algorithm

We propose the following regularizer R for the objective (3.4):

$$R(A) = \sum_{i,i' \in [n] j,j' \in [d]} R_{i,i',j,j'}(A), \tag{3.6}$$

where $R_{ii'jj'}(A)$ is defined as

$$R_{i,i',j,j'}(A) := \begin{cases} 0 & \text{if } (A_{ij} - A_{ij'})(A_{i'j} - A_{i'j'}) \geq 0 \\ (A_{ij} - A_{ij'})^2 (A_{i'j} - A_{i'j'})^2 & \text{otherwise.} \end{cases} \tag{3.7}$$

The goal of the regularizer R is to capture the permutation constraint of the matrix. The main challenge with the constraint is that the permutation is unknown. In (3.7), we consider the four matrix entries in rows $\{i, i'\} \subseteq [n]$ and columns $\{j, j'\} \subseteq [d]$ of the matrix. We call these four entries as the "quadruple" (i, i', j, j'). We observe that $A \in \mathcal{M}$ if and only if the terms $(A_{ij} - A_{ij'})$ and $(A_{ij} - A_{ij'})$ have the same sign (or one or both of the terms equal 0) for all the quadruples in the matrix (including quadruples where some or all of the four entries are unobserved). Hence, the regularizer $R_{ii'jj'}$ is designed to penalize the difference in the sign between the pairs of terms $(A_{ij} - A_{ij'})$ and $(A_{i'j} - A_{i'j'})$. The quadratic form (3.7) of $R_{ii'jj'}$ can be viewed as a differentiable approximation to the step function $\mathbb{1}\{(A_{ij} - A_{ij'})(A_{i'j} - A_{i'j'}) < 0\}$. Finally, the regularizer R takes a summation over all the quadruples (i, i', j, j'). It can be verified that we have $A \in \mathcal{M}$ if and only if $R(A) = 0$.

Putting (3.4), (3.5) and (3.6) together, our estimator is defined as

$$\operatorname*{argmin}_{A \in [0,1]^{n \times d}} \|A - Y\|_\Omega^2 + \lambda \sum_{i,i' \in [n], j,j' \in [d]} R_{ii'jj'}(A), \tag{3.8}$$

where ties are broken arbitrarily. Equivalently, our estimator can be viewed as first reformulating the original problem (3.3) to an equivalent problem:

$$\operatorname*{argmin}_{\substack{A \in [0,1]^{n \times d} \\ R_{ii'jj'}(A) = 0 \quad \forall i,i' \in [n], j,j' \in [d]}} \|A - Y\|_\Omega^2. \tag{3.9}$$

Then optimization (3.8) can be considered as the Lagrangian of the optimization problem (3.9). Intuitively, a large value of λ corresponds to stricter enforcement of the permutation structure on the matrix A.

To solve (3.8) we use projected gradient descent. The projected gradient descent algorithm consists of two steps in each iteration. In the gradient step, the algorithm updates its current estimate by computing gradient of the objective and moving the current estimate in its objective-improving direction for a stepsize. In the projection step, the algorithm projects the current estimate back to the domain $[0, 1]^{n \times d}$. Formally, we denote $\gamma_t \in \mathbb{R}$ as the stepsize in each iteration $t \geq 1$. We have

$$\text{Gradient step: } A_t = A_{t-1} - \gamma_t \nabla L_A(A). \tag{3.10a}$$

$$\text{Projection step: } \begin{aligned} A_t &\leftarrow \max\{0, A_t\}, \\ A_t &\leftarrow \min\{1, A_t\}. \end{aligned} \tag{3.10b}$$

Note that we choose a quadratic form in (3.7) instead of a linear form such as the hinge loss, because the quadratic form is differentiable, and hence its gradient can be computed straightforwardly.

3.3 Related work

Seriation and estimation under monotonicity Flammarion et al. [62] proposes the statistical model for seriation, and then shows that the least-squares estimator (3.2) is optimal up to logarithmic factors when the underlying constraint is either monotonic or unimodal. More generally, there is a rich line of literature on estimation under permutation constraints, where the data obeys certain underlying orderings, but the orderings are unknown. For example, Mao et al. [117] consider the class of bivariate isotonic matrices, where the matrix follows an unknown row permutation and an unknown column permutation, and a subclass where one of the two permutations is known. Shah et al. [158] analyze the class of stochastic transitivity (SST) matrices, which are bivariate isotonic matrices that are (shifted) skew-symmetric. A multivariate generalization is considered in Pananjady and Samworth [132]. For such problems, the least-squares estimators are considered (e.g., [62, 158, 161]more citations here). However, efficient algorithms for computing such least-squares estimators are not known [62, 108, 117]. Due to the computational inefficency of the least-squares estimator, other computational efficient estimators are proposed [62, 108, 117]. Many of these efficient estimators are statistically suboptimal, with the exception of Liu and Moitra [108] and Pananjady and Samworth [132]. Specifically, Liu and Moitra [108] considers bivariate isotonic matrix estimation where one of the two permutations is known, and proposes an estimator that runs in linear time achieving the optimal rate up to an $n^{o(1)}$ factor. Pananjady and Samworth [132] proposes an estimator that is optimal when the dimension of the problem is $d \geq 3$ (but not $d = 2$). For statistical seriation, positive or negative results on efficient estimators achieving the optimal rate remains unknown [62].

Landscape design and properties of local optima Optimization-based approaches are widely used for many problems, where the solution is posed as the minimizer to an objective function and computed by standard techniques such as gradient descent. The objective often includes regularization terms. Designing proper regularization (also termed "landscape design") that has desirable properties has been considered problems such as low-rank approximation [66] and neural networks [67]. In particular, Ge et al. [66] considers low-rank approximation under a random design setting and proves that all local minima are global minima. Ma et al. [110] considers a specific crowdsourced labeling setting with a rank-1 (Dawid-Skene) model, and shows that under arbitrary fixed design, all local minima are global minima for rank-1 matrix completion [110]. These theoretical results suggest that gradient descent converges to the global optimum for their problems. [110] further proposes an exponentiated gradient descent algorithm to achieve polynomial-rate convergence. Since a rank-1 matrix is monotonic by definition (where the permutation is unknown), our theoretical results (Section 3.4) can be considered as a generalized setting of Ma et al. [110]. Our idea of using projected gradient descent is also inspired by Ma et al. [110].

On using regularization for permutation constraints, Tibshirani et al. [179] proposes a regularizer to captures the permutation constraint in isotonic regression, where the permutation is known. On the other hand, we consider the case where the permutation is unknown.

Data imputation In the partial observation setting, our algorithm starts with an initialization. This initialization is related to data imputation, which is used in domains such as clustering. Methods such as naively taking the mean, nearest-neighbor (NN) [19] and MICE [9] are proposed. In the simulation results, we consider initializing the missing data by the mean and the nearest-neighbor methods.

3.4 Theoretical properties

In this section, we present theoretical properties of our algorithm. Specifically, we analyze the stationary points of the non-convex objective (3.8). We show desirable properties of any stationary point under the noiseless and the noisy settings. These results provide theoretical backing that the regularized objective proposed in (3.8) provides a natural approach to approximating the solution of (3.2).

The following result connects stationary points and gradient descent, stating that the gradient of the iterates obtained by projected gradient descent converges to 0.

Theorem 3.1. *Consider any matrix $Y \in [0,1]^{n\times d}$, any non-empty observation set $\Omega \subseteq [n] \times [d]$, and any value of the parameter $\lambda \geq 0$. With any initialization, the gradient of the iterates given by projected gradient descent on objective* (3.8) *converges to 0. Specifically, with a proper choice of a constant stepsize (dependent on n, d and λ), for any $\epsilon > 0$, the solution of projected gradient descent satisfies* $\lim_{t\to\infty} \|\nabla L(\widehat{A}_t)\|_F^2 < \epsilon$.

The proof of this theorem is provided in Section 10.2. In what follows, we present properties of the stationary points of the objective (3.8). Note that the objective (3.8) is continuous and over a closed bounded set (that is, $[0,1]^{n\times d}$). Therefore, there always exists at least one global minimum [149, Theorem 4.16], and hence at least one local minimum. In Lemma 10.2 of Section 10.1.2 , we show that all local minima on the boundary of the domain $[0,1]^{n\times d}$ are stationary points, so there exists at least one stationary point.

3.4.1 The noiseless setting

We first consider the noiseless setting where we have $Y \in \mathcal{M}$. Our approach is inspired by the work of Ma et al. [110]. Specifically, Ma et al. [110] considers rank-1 matrix completion under any fixed-design, and shows that their proposed algorithm can perfectly recover the rank-1 matrix in the noiseless case. Without a second thought, one may be tempted to write off this result – there is a straightforward algorithm to perfectly recover noiseless rank-1 matrices, that is, picking any non-zero row of the matrix, and writing each remaining row as the product of a multiplicative factor and this row. However, the theoretical results in Ma et al. [110] still provide non-trivial theoretical contributions and useful insights – the straightforward algorithm is heavily tailored to the noiseless case, and quickly becomes inapplicable when the data deviates from being rank-1.

On the contrary, the theoretical guarantees by Ma et al. [110] are shown on a much more general algorithm with any initialization, applicable to any arbitrary matrix Y.

In our problem, under the noiseless setting, the set of global minima to (3.8) is the set of monotonic matrices whose entries equal to Y on the observed set Ω. The following result shows that all stationary points are global minima. Since rank-1 matrices are monotonic by definition, our result supplements the result of Theorem 2 in Ma et al. [110] by considering general monotonic matrices in small matrix sizes.

Theorem 3.2. *Consider any $Y \in \mathcal{M}$, any non-empty observation set $\Omega \subseteq [n] \times [d]$ and any value of the parameter $\lambda \geq 0$. Consider $n = 2$ or $d \leq 3$. Then any stationary point to the objective* (3.8) *is a global minimum.*

The proof of this theorem is provided in Section 10.3. The proof relies on the first-order optimality condition, and uses combinatorial arguments to derive contradictions if any stationary point were not a global minimum.

Similar to the setting in Ma et al. [110], under the noiseless setting, there also exists a straightforward algorithm to obtain all the global minima of (3.8) – by first finding the total ordering of the columns (or the set of all such total orderings) induced by the entries within each row, and filling each unobserved entry to be any value subject to this total ordering. On the contrary, our algorithm is applicable to any arbitrary matrix Y. With its generality, it is even unclear if the original noiseless matrix can be recovered under any arbitrary initialization without Theorem 3.2. Furthermore, the property of perfectly recovering noiseless data is not only natural but also important – given the generality of the seriation model, Theorem 3.2 contrasts our algorithm with prior approaches in matrix estimation and completion such as using parameter-based models or low-rank matrix decomposition, where a non-zero bias is incurred in this noiseless case.

3.4.2 The noisy setting

Now we move to consider the noisy setting where the matrix Y is not guaranteed to be monotonic. A quadruple (i, i', j, j') is called a "disagreement quadruple" if the signs of $(A_{ij} - A_{ij'})$ and $(A_{i'j} - A_{i'j'})$ are different. The following result shows that the set of disagreement quadruples at any stationary point to (3.8) is a subset of the disagreement quadruples in the original matrix Y.

Proposition 3.3. *Consider any matrix $Y \in [0,1]^{n\times d}$, any non-empty observation set $\Omega \subseteq [n]\times[d]$ and any value of the parameter $\lambda \geq 0$. Consider $n = 2$. Let $\widehat{A}$ be any stationary point of the objective* (3.8)*. For every $\{i, i'\} = \{1, 2\}$ and any $j, j' \in [d]$ such that*

$$\widehat{A}_{i,j} < \widehat{A}_{i,j'} \qquad \textit{and} \qquad \widehat{A}_{i',j} > \widehat{A}_{i',j'},$$

we have the same relation holds at the corresponding entries of the matrix Y:

$$\begin{aligned} &Y_{i,j} < Y_{i,j'}, \qquad \textit{if } (i,j),(i,j') \in \Omega \\ \textit{and} \quad &Y_{i',j} > Y_{i',j'}, \qquad \textit{if } (i',j),(i',j') \in \Omega. \end{aligned}$$

The proof of this result is provided in Section 10.4. In words, this result shows that our estimator only reduces the disagreement quadruples in the observations Y and never introduces new ones that do not exist in Y, thus revealing another natural desirable property of our estimator (3.8).

Using Proposition 3.3 as a building block, the following result considers the case where there is a partition of the columns, and there is a total ordering describing the dominance relation of these columns in the matrix Y. Specifically, a set of columns $S \subseteq [d]$ is said to "dominate" another set of columns $S' \subseteq [d]$, if we have $Y_{ij} > Y_{ij'}$, for every $i \in [n], j \in S$ and $j' \in S'$ such that $(i, j), (i, j') \in \Omega$. The following theorem shows that any stationary point to (3.8) retains this dominance relation.

Theorem 3.4. *Consider any matrix $Y \in [0, 1]^{n \times d}$, any non-empty observation set $\Omega \subseteq [n] \times [d]$ and any value of the parameter $\lambda \geq 0$. Consider $n = 2$. Assume there exists a partition of columns $[d] = S_1 \cup \ldots \cup S_m$, such that S_{k+1} dominates S_k for each $k \in [m-1]$. Assume that for each $k \in [m-1]$, and each $j \in S_k, j' \in S_{k+1}$, we have*

$$\exists i \in \{1, 2\} \text{ such that } (i, j), (i, j') \in \Omega. \tag{3.11}$$

Then at any stationary point $\widehat{A}$ to the objective (3.8)*, we have $\widehat{A}_{ij} < \widehat{A}_{ij'}$ for any $i \in \{1, 2\}$ and any $j \in S_k, j' \in S_{k+1}$ with any $k \in [m-1]$.*

The proof of this result is provided in Section 10.5. In words, the condition (3.11) in Theorem 3.4 requires that the ordering of two columns are directly comparable. Note that in the noiseless case, we can write the partition as $[d] = \{1\} \cup \ldots \cup \{d\}$. Hence, this result is a generalization of our result from the noiseless case (Theorem 3.2). Proposition 3.3 and Theorem 3.4 thus together show that in the noisy setting, any stationary point to the objective (3.8) has desirable properties under certain special cases. These theoretical properties are natural but at the same time non-trivial, providing theoretical insights and validation to our proposed estimator (3.8).

3.5 Simulations

In this section, we evaluate the performance of gradient descent on the objective (3.8) in different settings[1]. We first discuss the simulation set-up for a full-observation setting ($\Omega = [n] \times [d]$)) in Section 3.5.1. We provide the associated results in Section 3.5.2. In a nutshell, our algorithm performs better than the baselines when the underlying models do not satisfy specialized parametric assumptions, and also when the signal-to-noise (SNR) is high so that the noise does not overshadow the non-parametric structure of the data. We then simulate settings with only partial observations in Section 3.5.3. We consider several natural methods to initialize the matrix Y and we find that our algorithm consistently improves the performance as compared to the various common initialization methods. We also find that our algorithm is quite robust to the choice of the initialization method, although the choice of initialization could in theory lead to very different local minima.

3.5.1 Simulation etup

We now describe the design choices made for our estimator (3.8) and the simulation settings.

[1] The code for the implementation of our estimator and for evaluation is provided at `https://github.com/jingyanw/heuristic-seriation`.

Reparameterizing the hyperparameter λ Instead of the objective (3.8) that weighs the two terms by 1 and λ, we reparametrize the hyperparameter λ and now weigh the two terms by $(1-\widetilde{\lambda})$ and $\widetilde{\lambda}$ with $\widetilde{\lambda} \in [0, 1)$. That is, we consider the objective

$$\underset{A \in [0,1]^{n \times d}}{\text{argmin}} \quad (1 - \widetilde{\lambda}) \cdot \|Y - A\|_\Omega^2 + \widetilde{\lambda} R(A). \tag{3.12}$$

Note that this objective (3.12) is equivalent to the previous objective (3.8), with a one-to-one correspondence between the values of λ and $\widetilde{\lambda}$. The reparameterized objective (3.12) reduces the variation on the magnitude of the objective through the range $\widetilde{\lambda} \in [0, 1)$, making it easier to choose a simple constant stepsize for gradient descent independent of the specific choice of $\widetilde{\lambda}$. For all the subsequent simulation results, we consider this reparameterized objective (3.12).

Gradient descent For simplicity, we choose a constant stepsize of 0.1 with a momentum of 0.9. We use the initialization $A_0 = Y$ under full observations. The choice of the initialization under partial observations is further discussed in Section 3.5.3. We terminate the algorithm when the normalized squared Frobenius norm of the gradient is smaller than 10^{-8}, that is, when $\frac{1}{nd}\|\nabla_A \widetilde{L}(A)\|_F^2 < 10^{-8}$, where $\widetilde{L}$ denotes the reparameterized objective (3.12). We implement our objective (3.12) and run gradient descent in PyTorch [134].

Models We follow the observation models studied in Shah et al. [158], but with an additional parameter that controls the relative levels of signal and noise. We consider square matrices with $n = d$. Let $A^* \in [0, 1]^{n \times n}$ represent the true matrix whose value is specified later for different models. Bernoulli observations Y are generated[2] from A^*, that is, we have $\mathbb{P}(Y_{ij} = 1) = A^*_{ij}$ for each $i, j \in [n]$. We use the five SST models of A^* described in Shah et al. [158, Section 4]; we also include the descriptions below for completeness.

(a) **Uniform:** The diagonal entries are 0.5. Then $\binom{n}{2}$ values are drawn independently and uniformly at random from $[\beta, 1]$, for a fixed choice of $\beta \in [0.5, 1]$, and sorted in the increasing order. The entries immediately above the diagonal are filled with the smallest $(n - 1)$ values uniformly at random. Then the entries in the next diagonal above are filled uniformly at random with the smallest $(n - 2)$ of the remaining values, and so on. The entries below the diagonal are filled in to make A^* skew symmetric.

(b) **Thurstone:** A vector $w^* \in \mathbb{R}^n$ is chosen uniformly at randomly from the set of w^* such that $\langle w^*, 1\rangle = 0$ and all entries of w^* are between $-0.5 - \beta$ and $0.5 + \beta$, for a fixed choice of β. Then the matrix A^* is filled in via $A^*_{ij} = F(w^*_i - w^*_j)$ for each $i, j \in [n]$, where F is the CDF of the standard normal distribution.

(c) **BTL:** Identical to the Thurstone model, except that F is given by the sigmoid function.

(d) **Noisy sorting:** The diagonal entries are 0.5. All entries above the diagonal are β, and all entries below the diagonal are $1 - \beta$, for a fixed choice of $\beta \in [\frac{1}{2}, 1]$. This is a classic model proposed by Braverman and Mossel [27] and studied subsequently in the literature (e.g., [116]).

[2] Note that Shah et al. [158] only generates i.i.d. Bernoulli observations in the upper diagonal with $i \leq j$, and set the entries in the lower diagonal as $Y_{ji} = 1 - Y_{ij}$. This is because Shah et al. [158] requires the matrix to be skew-symmetric whereas with the seriation model we do not have this restriction.

(e) **Independent bands:** The diagonal entries are 0.5. The entries immediately above the diagonal are chosen i.i.d. uniformly at random from $[\beta, 1]$, for a fixed choice of $\beta \in [0.5, 1]$. The entries in the next diagonal is chosen uniformly randomly from the range lower bounded by the entries to its left and below. The entries below the diagonal are filled in a manner that makes A^* skew symmetric.

Metrics For any estimator $\widehat{A}$, we consider its risk in terms of the normalized squared Frobenius norm, $\frac{1}{nd}\|\widehat{A} - A^*\|_F^2$.

Baselines We compare our algorithm to the following baselines:

1. **Rank-1**: The estimate $\widehat{A}$ is computed as the rank-1 approximation of Y.
2. **Singular-value thresholding (SVT):** This estimator is studied in Shah et al. [158, Section 3.2] (and also in various other works such as Chatterjee [36]), with a parameter α denoting the threshold level applied on the singular values of Y. The value of α is required to be strictly greater than $2\sqrt{n}$, and Shah et al. [158] uses $\alpha = 2.01\sqrt{n}$. For our settings, we consistently observe that a smaller value of α gives better performance, so we set $\alpha = 2.0000001\sqrt{n}$.We consistently observe that the hard-thresholding performs better than the soft-thresholding, so we use the hard-thresholding in our simulation.

3.5.2 Results for full observations

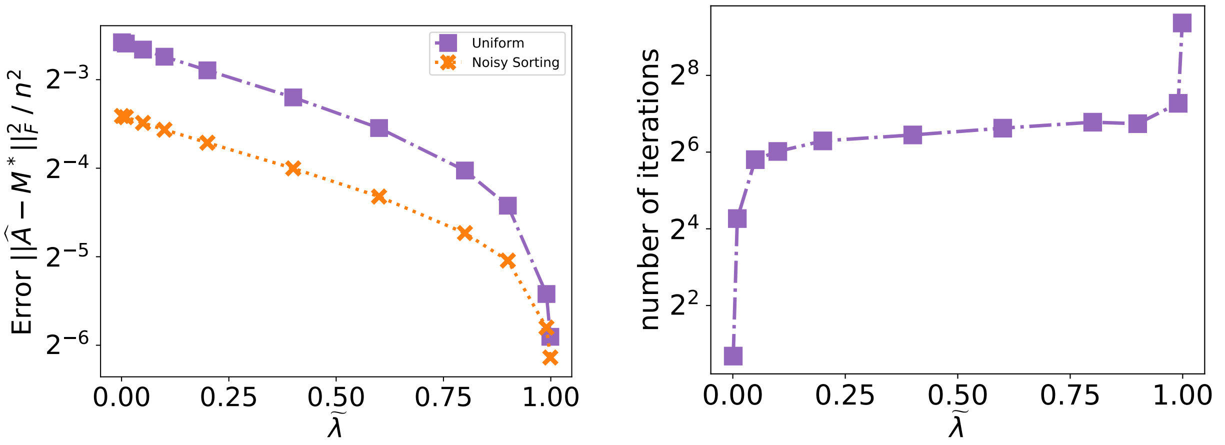
We now present the results from our simulations pertaining to the full-observation setting.

Accuracy-computation tradeoff induced by $\widetilde{\lambda}$

We first inspect the performance of our algorithm for different choices of $\widetilde{\lambda} \in [0, 1)$, in terms of the accuracy (measured by the Frobenius error of estimation) and the computational time (measured by the number of iterations taken till convergence of gradient descent), shown in Figure 3.1. We use $n = 64$ and $\beta = 0.5$ (which matches the setting in [158]).The error bars in Figure 3.1 and all subsequent results represent the standard error of the mean, computed over 10 trials. In Figure 3.1 and subsequent plots, the error bars are small and therefore not visible.

We observe from Figure 3.1 that there is a tradeoff between accuracy and the computational time. As the value of $\widetilde{\lambda}$ increases, our algorithm attains a lower error (Figure 3.1(a)), but takes more time (Figure 3.1(b)). This tradeoff is expected, because the original least-square estimator intuitively corresponds to setting $\widetilde{\lambda} = 1$, which is known to be optimal in estimation and conjectured computationally inefficient. On the other hand, setting $\widetilde{\lambda} = 0$ is equivalent to outputting the observation matrix Y without any computation. For clarity, only a few models are shown in Figure 3.1, but we consistently observe these trends for numerous settings not shown.larger legend

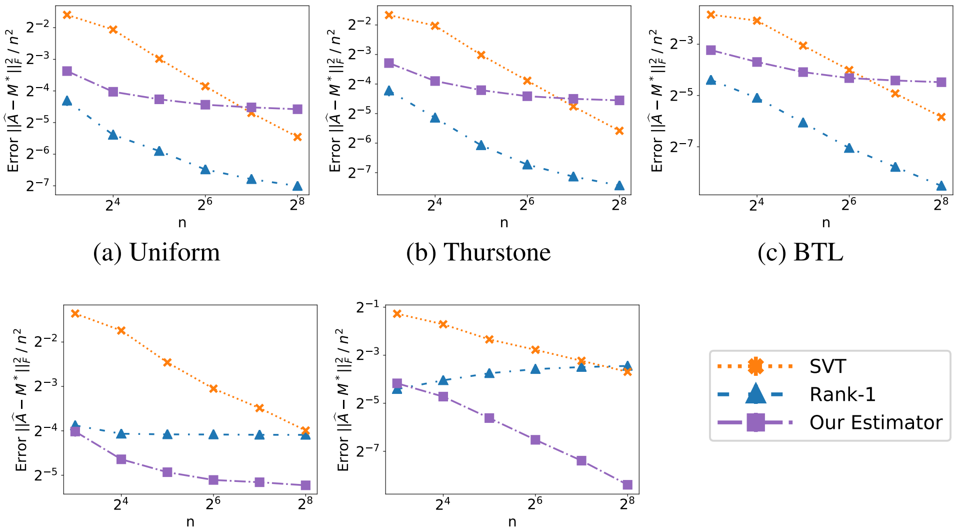
Consequently, for all subsequent simulations we set $\widetilde{\lambda} = 0.9$, which is a reasonably large value that attains low error without excessively slowing down the convergence. We now provide simulation results for the 5 models under the set-up described in Section 3.5.1.



(a) Estimation error as a function of $\widetilde{\lambda}$

(b) Number of iterations as a function of $\widetilde{\lambda}$ for the uniform model

Figure 3.1: Tradeoff between accuracy (estimation error) and time (number of iterations) for different values of $\widetilde{\lambda} \in [0, 1)$.



(a) Uniform

(b) Thurstone

(c) BTL

(d) Noisy Sorting

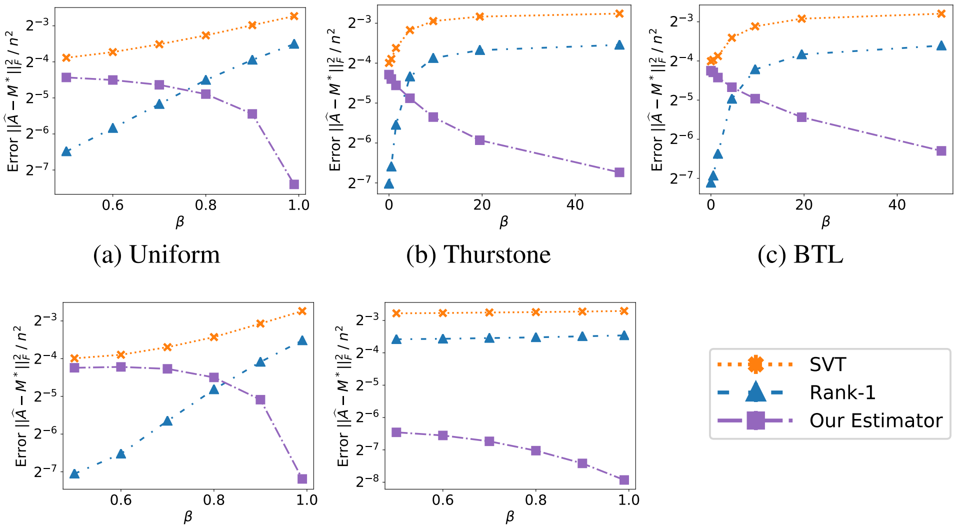
(e) Independent Bands

Figure 3.2: Estimation error of different algorithms for different models of A^*.

Comparison to baselines

We run simulations comparing the performance of our algorithm with the baselines on the aforementioned models in two ways: varying the matrix size n (for fixed $\beta = 0.5$) and varying the signal relative to noise, β (for fixed $n = 64$). The results are shown in Figure 3.2 and Figure 3.3, respectively (overlapping curves are slightly shifted horizontally for better readability). The key findings from these simulations are as follows:

- The baselines work well when the underlying model is parametric or similar (Figure 3.2(b)(c)),



(a) Uniform (b) Thurstone (c) BTL

(d) Noisy Sorting (e) Independent Bands

Figure 3.3: Estimation error of different algorithms under different levels of signal relative to noise.

but are inconsistent when such parametric assumptions do not hold (Figure 3.2(d)(e)). A similar observation about the Thurstone MLE is made in [158]. The rank-1 estimator and the (hard)-SVT estimator yield similar performance.

- Our estimator outperforms the baselines when the underlying model is more complex.
- When the noise level is high relative to the signal (smaller values of β in Figure 3.3), the baselines perform well. This is because the estimation error dominates, and the baselines trim off a lot of noise.
- When the noise level is low relative to the signal (larger values of β in Figure 3.3), our estimator offers substantial improvements. In this regime, the approximation error is the dominating source of error, and the baselines incur a large approximation error since they

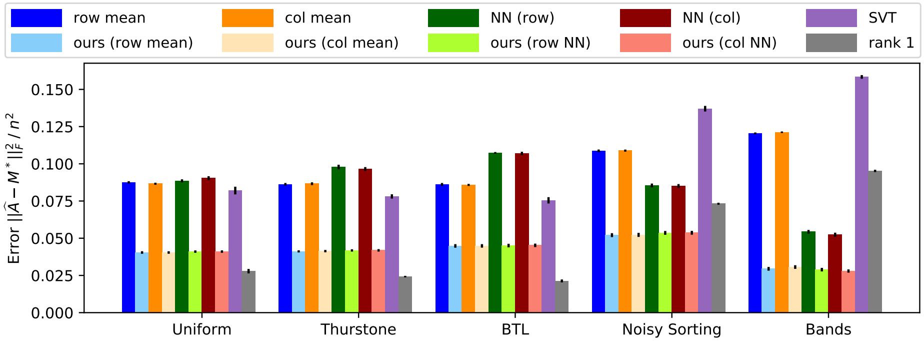


Figure 3.4: Performance with partial observations under different initialization methods. need to re-run for hard-SVT

also trim off a large part of the signal.

3.5.3 Partial observations

In what follows, we simulate settings where Y has missing entries, which is important in practice but has received much less attention in the literature. We consider our algorithm (3.8) and evaluate various initializations for gradient descent, as well as compare it to the baselines. The initialization potentially affects the performance of gradient descent, because gradient descent may converge to different local optima depending on the initialization.

Simulation setup

As before, we choose $n = 64$, and $\beta = 0.5$, matching the setting in Shah et al. [158].

Random-design observations We consider a random design to construct Ω so that each matrix entry is observed with probability 0.3 independent of all else.

Initialization methods We consider the following initialization methods:

- **Row mean:** Each unobserved entry is initialized to the mean of the observed entries in its row.
- **Column mean:** Each unobserved entry is initialized to the mean of the observed entries in its column.
- **Row kNN:** Each unobserved entry is imputed as the mean of the 5 nearest rows among the rows. The distance between rows is measured in terms of the normalized Euclidean distance.
- **Column kNN:** Each unobserved entry is imputed as the mean of the 5 nearest columns among the columns. The distance between columns is measured in terms of the normalized Euclidean distance.

Results for partial observations

The simulation results for partial observations are shown in Figure 3.4, where the bars for the same initialization before and after running our algorithm are coded in a pair of similar colors. We also compare the performance of our algorithm with the baselines described earlier in Section 3.5.1. The figure shows the performance of each baseline with the initialization method for which it performs the best (which happens to be both row and column kNNs for both baselines).what does prev sentence mean The salient findings from the simulations are as follows:

- The choice of the initialization method does not have strong influence on the performance of our algorithm.
- Our algorithm consistently improves upon different initialization methods.

- Similar to the full-observation setting, our method outperforms the baselines when the underlying model is more complex, whereas the baselines perform well when the underlying model is simpler.

3.6 Conclusion and discussion

In this work, we contribute a heuristic-based perspective with respect to the spectrum of the statistical-computational gap in the statistical seriation problem. In terms of open problems, on the theory front, it is still certainly of interest to accurately characterize the statistical-computational gap. On the applied side, a wide range of applications have application-specific characteristics. For example, in peer review, reviewers' behaviors may not be entirely monotonic due to subjectivity, so that the true matrix may have only a partially monotonic structure. Our heuristic-based approach can provide a useful tool to tackle such challenges that are even more complex than the open problem of statistical seriation.

Acknowledgments

This work was supported in part by an NSF CAREER award CIF: 1942124, NSF grant CIF: 1763734, the U.S. Office of Naval Research N00014-21-1-2243, the Air Force Office of Scientific Research FA9550-20-1-0080, and a grant from the CMU Block Center for Technology and Society.

Chapter 4

Debiasing Evaluations That Are Biased by Evaluations

It is common to evaluate a set of items by soliciting people to rate them. For example, universities ask students to rate the teaching quality of their instructors, and conference organizers ask authors of submissions to evaluate the quality of the reviews. However, in these applications, students often give a higher rating to a course if they receive higher grades in a course, and authors often give a higher rating to the reviews if their papers are accepted to the conference. In this work, we call these external factors the "outcome" experienced by people, and consider the problem of mitigating these outcome-induced biases in the given ratings when some information about the outcome is available. We formulate the information about the outcome as a known partial ordering on the bias. We propose a debiasing method by solving a regularized optimization problem under this ordering constraint, and also provide a carefully designed cross-validation method that adaptively chooses the appropriate amount of regularization. We provide theoretical guarantees on the performance of our algorithm, as well as experimental evaluations.

4.1 Introduction

It is common to aggregate information and evaluate items by collecting ratings on these items from people. In this work, we focus on the bias introduced by people's observable outcome or experience from the entity under evaluation, and we call it the "outcome-induced bias". Let describe this notion of bias with the help of two common applications – teaching evaluation and peer review.

Many universities use student ratings for teaching evaluation. However, numerous studies have shown that student ratings are affected by the grading policy of the instructor [23, 75, 95]. For instance, as noted in Johnson [95, Chapter 4]:

> *"...the effects of grades on teacher-course evaluations are both substantively and statistically important, and suggest that instructors can often double their odds of receiving high evaluations from students simply by awarding A's rather than B's or C's."*

As a consequence, the association between student ratings and teaching effectiveness can become negative [23], and student ratings serve as a poor predictor on the follow-on course achievement of the students [25, 33]:

> *"...teachers who are associated with better subsequent performance receive worst evaluations from their students."* [25]

The outcome we consider in teaching evaluation is the grades that the students receive in the course under evaluation[1] and the goal is to correct for the bias in student evaluations induced by the grades given by the instructor.

An analogous issue arises in conference peer review, where conference organizers survey authors to rate their received reviews in order to understand the quality of the review process. It is well understood that authors are more likely to give higher ratings to a positive review than a to negative review [97, 133, 189]:

> *"Satisfaction had a strong, positive association with acceptance of the manuscript for publication... Quality of the review of the manuscript was not associated with author satisfaction."* [189]

Due to this problem, an author feedback experiment [133] conducted at the PAM 2007 conference concluded that:

> *"...some of the TPC members from academia paralleled the collected feedback to faculty evaluations within universities... while author feedback may be useful in pinpointing extreme cases, such as exceptional or problematic reviewers, it is not quite clear how such feedback could become an integral part of the process behind the organization of a conference."*

With this motivation, for the application of peer review, the outcome we consider is the review rating or paper decision received by the author, and the goal is to correct for the bias induced by it in the feedback provided by the author.

Although the existence of such bias is widely acknowledged, student and author ratings are still widely used [17], and such usage poses a number of issues. First, these biased ratings can be uninformative and unfair for instructors and reviewers who are not lenient. Second, instructors, under the possible consideration of improving their student-provided evaluation, may be incentivized to "teach to the test", raising concerns such as inflating grades and reducing content [33]. Furthermore, author-provided ratings can be a factor for selecting reviewer awards [97], and student-provided ratings can be a heavily-weighted component for salary or promotion and tenure decision of the faculty members [17, 23, 33]. If the ratings are highly unreliable and sometimes even follow a trend that reverses the true underlying ordering, then naïvely using these ratings or simply taking their mean or median will not be sufficient. Therefore, interpreting and correcting these ratings properly is an important and practical problem.

The goal of this work is to mitigate such outcome-induced bias in ratings. Incidentally, in teaching evaluation and peer review, the "outcome" that people (students or authors) encounter in the process is the evaluation they receive (grades from instructors or reviews from reviewers), and hence we call this bias "evaluations that are biased by evaluations". That said, we note that the general problem we consider here is applicable to other settings with outcomes that are not necessarily evaluations. For example, in evaluating whether a two-player card game is fair or not, the outcome can be whether the player wins or loses the game [124].

[1] We use the term "grades" broadly to include letter grades, numerical scores, and rankings. We do not distinguish the difference between evaluation of a course and evaluation of the instructor teaching the course, and use them interchangeably.

The key insight we use in this work is that the outcome (e.g., grades and paper decisions) is naturally available to those conduct the evaluation (e.g., universities and conference organizers). These observed outcomes provide directional information about the manner that evaluators are likely to be biased. For example, it is known [23, 75, 95] that students receiving higher grades are biased towards being more likely to give higher ratings to the course instructor than students receiving lower grades. To use this structural information, we model it as a known partial ordering constraint on the biases given people's different outcomes. This partial ordering, for instance, is simply a relation on the students based on their grades or ranking, or on the authors in terms of acceptance decisions of their papers.

4.1.1 Our contributions

We identify and formulate a problem of mitigating biases in evaluations that are biased by evaluations (Section 4.2). Specifically, this bias is induced by observable outcomes, and the outcomes are formulated as a known partial ordering constraint. We then propose an estimator that solves an optimization jointly in the true qualities and the bias, under the given ordering constraint (Section 4.3). The estimator includes a regularization term that balances the emphasis placed on bias versus noise. To determine the appropriate amount of regularization, we further propose a cross-validation algorithm that chooses the amount of regularization in a data-dependent manner by minimizing a carefully-designed validation error (Section 4.3.2).

We then provide a theoretical analysis of the performance of our proposed algorithm (Section 4.4). First, we show that our estimator, under the two extremal choices of the regularization hyperparameter (0 and ∞), converges to the true value in probability under only-bias (Section 4.4.2) and only-noise (Section 4.4.3) settings respectively. Moreover, our estimator reduces to the popular sample-mean estimator when the regularization hyperparameter is set to ∞, which is known to be minimax-optimal in the only-noise case. We then show (Section 4.4.4) that the cross-validation algorithm correctly converges to the solutions corresponding to hyperparameter values of 0 and ∞ in probability in the two aforementioned settings, under various conditions captured by our general formulation. We finally conduct synthetic and semi-synthetic experiments that establish the effectiveness of our proposed approach via numerical experiments in more general settings not covered by the theoretical results (Section 4.5).

4.1.2 Related work

In terms of correcting rating biases, past work has studied the problem of adjusting student GPAs due to different grading policies across courses and disciplines. Proposed models include introducing a single parameter for each course and each student solved by linear regression [34], and more complicated parametric generative models [94]. Though grade adjustment seems to be a perfect counterpart of teaching evaluation adjustment, the non-parametric ordering constraint we consider is unique to teaching evaluation, and do not have obvious counterpart in grade adjustment. For the application of peer review, there are many works [59, 65, 93, 106, 115, 130, 167, 168, 180, 186] addressing various biases and other issues in the review process, but to the best of our knowledge none of them addresses biases in author-provided feedback. It is of interest in the future to design schemes that combine our present work with these past works in order to jointly address multiple problems such as simultaneous existence of outcome-dependent bias and miscalibration.

In terms of the models considered, one statistical problem related to our work is the isotonic regression, where the goal is to estimate a set of parameters under a total ordering constraint (see, e.g. 12, 78, 113, 196). Specifically, our problem becomes isotonic regression, if in our exact formulation (4.2) to be presented, we set $\lambda = 0, x = 0$ and the partial ordering to a total ordering.

Another type of related models in statistics literature concerns the semiparametric additive models (e.g. 49, 82, 191, 194) with shape constraints [39]. In particular, one class of semiparametric additive models involves linear components and components with ordering (isotonic) constraints [42, 87, 122, 150]. Our optimization (4.2) falls within this class of semiparametric models, if we set the second term of ℓ_2-regularization to 0. To see the connection, we write the first term of (4.2) in a linearized form as $\|y - Ax - b\|_2^2$, where $y, b \in \mathbb{R}^{dn}, x \in \mathbb{R}^d$ and $A \in \mathbb{R}^{dn \times d}$ is a $0/1$ matrix that specifies the course membership of each rating: if a rating is from course i, then in corresponding of row of A, the i^{th} entry is 1 and all other entries are 0. Past work has studied the least-squares estimator for this problem, but the results such as consistency and asymptotic normality rely on assumptions such as A being random design or each coordinate of x being i.i.d., which are not applicable to our setting. The special $0/1$ structure of A makes our problem unique and differ from past work in terms of the theoretical analysis.

In terms of the technical approach, our estimator (Equation 4.2) is partly inspired by permutation-based models [158, 162] which focuses only on shape constraints rather than parameters, but with the key difference that here we can exploit the crucial information pertaining to the ordering of the bias.

The idea of adopting cross-validation to select the right amount of penalization is classical in statistics literature (see, e.g. [81, 102, 171]). Yet, this generic scheme cannot be directly applied to models where training samples are not exchangeable—in which case, both the subsampling step and the test-error estimation are highly non-trivial. Therefore caution needs to be exercised when order restrictions, therefore non-exchangeability, are involved. The cross-validation algorithm proposed in this work is partly inspired by the cross-validation used in nearly-isotonic regression [179]. In nearly-isotonic regression, the hard ordering constraint is replaced by a soft regularization term, and the extent of regularization is determined by cross-validation. However, introducing the linear term of x as the quantity of interest significantly changes the problem. Thus, our cross-validation algorithm and its analysis are quite different.

4.2 Problem formulation

For ease of exposition, throughout the chapter we describe our problem formulation using the running example of course evaluation, but we note that our problem formulation is general and applies to other problems under outcome-induced bias as well. Consider a set of d courses. Each course $i \in [d]$ has an unknown true quality value $x_i^* \in \mathbb{R}$ to be estimated. Each course is evaluated by n students.[2] Denote $y_{ij} \in \mathbb{R}$ as the rating given by the j^{th} student in course i, for

[2] For ease of exposition, we assume that each course is evaluated by n students, but the algorithms and the results extend to regimes where the number of students is different across courses.

each $i \in [d]$ and $j \in [n]$. Note that we do not require the same set of n students to take all d courses; students in different courses are considered different individuals. We assume that each rating y_{ij} is given by:

$$y_{ij} = x^*_i + b_{ij} + z_{ij}, \tag{4.1}$$

where b_{ij} represents a bias term, and z_{ij} represents a noise term. We now describe these terms in more detail.

The term z_{ij} captures the noise involved in the ratings, assumed to be i.i.d. across $i \in [d]$ and $j \in [n]$. The term b_{ij} captures the bias that is induced by the observed "outcome" of student j experienced in course i. In the example of teaching evaluation, the outcome can be the grades of the students that are known to the university, and the bias captures the extent that student ratings are affected by their received grades. Given these observed outcomes (grades), we characterize the information provided by these outcomes as a known partial ordering, represented by a collection of ordering constraints $\mathcal{O} \subseteq ([d] \times [n])^2$. Each ordering constraint is represented by two pairs of (i, j) indices. An ordering constraint $((i, j), (i', j')) \in \mathcal{O}$ indicates that the bias terms obey the relation $b_{ij} \leq b_{i'j'}$. We say that this ordering constraint is on the elements $\{(i, j)\}_{i\in[d],j\in[n]}$ and on the bias $\{b_{ij}\}_{i\in[d],j\in[n]}$ interchangeably. We assume the terms $\{b_{ij}\}_{i\in[d],j\in[n]}$ satisfy the partial ordering $\mathcal{O}$. In teaching evaluations, the partial ordering $\mathcal{O}$ can be constructed by, for example, taking $((i, j), (i', j')) \in \mathcal{O}$ if and only if student j' in course i' receives a strictly higher grade than student j in course i.

For ease of notation, we denote $Y \in \mathbb{R}^{d\times n}$ as the matrix of observations whose $(i, j)^{\text{th}}$ entry equals y_{ij} for every $i \in [d]$ and $j \in [n]$. We define matrices $B \in \mathbb{R}^{d\times n}$ and $Z \in \mathbb{R}^{d\times n}$ likewise. We denote $x^* \in \mathbb{R}^d$ as the vector of $\{x^*_i\}_{i\in[d]}$.

Goal. Our goal is to estimate the true quality values $x^* \in \mathbb{R}^d$. For model identifiability, we assume $\mathbb{E}[z_{ij}] = 0$ and $\sum_{i\in[d],j\in[n]} \mathbb{E}[b_{ij}] = 0$. An estimator takes as input the observations Y and the partial ordering $\mathcal{O}$, and outputs an estimate $\widehat{x} \in \mathbb{R}^d$. We measure the performance of any estimator in terms of its (normalized) squared ℓ_2 error $\frac{1}{d}\|\widehat{x} - x^*\|_2^2$.

4.3 Proposed estimator

Our estimator takes as input the observations Y and the given partial ordering $\mathcal{O}$. The estimator is associated with a tuning parameter $\lambda \geq 0$, and is given by:

$$\widehat{x}^{(\lambda)} \in \operatorname*{argmin}_{x\in\mathbb{R}^d} \min_{\substack{B\in\mathbb{R}^{d\times n}\\ B \text{ satisfies } \mathcal{O}}} \|Y - x\mathbf{1}^T - B\|_F^2 + \lambda\|B\|_F^2, \tag{4.2}$$

where $\mathbf{1}$ denotes the all-one vector of dimension n. We let $\widehat{B}^{(\lambda)}$ denote the value of B that attains the minimum of the objective (4.2), so that the objective (4.2) is minimized at $(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$. Ties are broken by choosing the solution (x, B) such that B has the minimal Frobenius norm $\|B\|_F^2$. We show that the estimator under this tie-breaking rule defines a unique solution in Proposition 11.1 in Section 11.2.1. Furthermore, as explained in Section 4.7.1, the optimization (4.2) is a convex quadratic programming (QP) in (x, B), and therefore can be solved in polynomial time in terms of (d, n).

While the first term $\|Y - x\mathbf{1}^T - B\|_F^2$ of (4.2) captures the squared difference between the bias-corrected observations $(Y - B)$ and the true qualities $x\mathbf{1}^T$, the second term $\|B\|_F^2$ captures the magnitude of the bias. Since the observations in (4.1) include both the bias B and the noise Z, there is fundamental ambiguity pertaining to the relative contributions of the bias and noise to the observations. The penalization parameter λ is introduced to balance the bias and the variance, and at the same time preventing overfitting to the noise. More specifically, consider the case when the noise level is relatively large and the partial ordering $\mathcal{O}$ is not sufficiently restrictive — in which case, it is sensible to select a larger λ to prevent B overly fitting the observations Y.

For the rest of this section, we first describe intuition about the tuning parameter λ by considering two extreme choices of λ which are by themselves of independent interest. We then propose a carefully-designed cross-validation algorithm to choose the value of λ in a data-dependent manner.

4.3.1 Behavior of our estimator under some fixed choices of λ

To facilitate understandings of the estimator (4.2), we discuss its behavior for two important choices of λ — 0 and ∞ — that may be of independent interest.

$\boldsymbol{\lambda = 0}$: When $\lambda = 0$, intuitively the estimator (4.2) allows the bias term B to be arbitrary in order to best fit the data, as long as it satisfies the ordering constraint $\mathcal{O}$. Consequently with this choice, the estimator attempts to explain the observations Y as much as possible in terms of the bias. One may use this choice if domain knowledge suggests that bias considerably dominates the noise. Indeed, as we show subsequently in Section 4.4.2, our estimator with $\lambda = 0$ is consistent in a noiseless setting (when only bias is present), whereas common baselines are not.

$\boldsymbol{\lambda = \infty}$: We now discuss the other extremity, namely when λ approaches infinity. Intuitively, this case sets the bias term to zero in (4.2) (note that $\widehat{B} = 0$ trivially satisfies any partial ordering $\mathcal{O}$). Therefore, it aims to explain the observations in terms of the noise. Formally we define $(\widehat{x}^{(\infty)}, \widehat{B}^{(\infty)}) = \lim_{\lambda\to\infty}(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$. In the subsequent result of Proposition 4.7, we show that this limit exists, where we indeed have $\widehat{B}^{(\infty)} = 0$ and our estimator simply reduces to the sample mean as $[\widehat{x}^{(\infty)}]_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$ for every $i \in [d]$. We thus see that perhaps the most commonly used estimator for such applications — the sample mean — also lies in our family of estimators specified in (4.2). Given the well-known guarantees of the sample mean in the absence of bias (under reasonable conditions of the noise), one may use this choice if domain knowledge suggests that noise is highly dominant as compared to the bias.

$\boldsymbol{\lambda \in (0, \infty)}$: More generally, the estimator interpolates between the behaviors at the two extremal values $\lambda = 0$ and ∞ when both bias and noise is present. As we increase λ from 0, the magnitude of the estimated bias $\widehat{B}^{(\lambda)}$ gradually decreases and eventually goes to 0 at $\lambda = \infty$. The estimator hence gradually explains the observations less in terms bias, and more in terms of noise. Our goal is to choose an appropriate value for λ, such that the contribution of bias versus noise determined by the estimator approximately matches the true relative contribution that generates the observations. The next subsection presents a principled method to choose the value for λ.

4.3.2 A cross-validation algorithm for selecting λ

We now present a carefully designed cross-validation algorithm to select the tuning parameter λ in a data-driven manner. Our cross-validation algorithm determines an appropriate value of λ from a finite-sized set of candidate values $\Lambda \subseteq [0, \infty]$ that is provided to the algorithm. For any matrix $A \in \mathbb{R}^{d\times n}$, we define its squared norm restricted to a subset of elements $\Omega \subseteq [d] \times [n]$ as $\|A\|_\Omega^2 = \sum_{(i,j)\in\Omega} A_{ij}^2$. Let $\mathcal{T}$ denote the set of all total orderings (of the dn elements) that are consistent with the partial ordering $\mathcal{O}$. The cross-validation algorithm is presented in Algorithm 2. It consists of two steps: a data-splitting step (Lines 1-8) and a validation step (Lines 9-19).

Data-splitting step In the data-splitting step, our algorithm splits the observations $\{y_{ij}\}_{i\in[d],j\in[n]}$ into a training set $\Omega^{\text{t}} \subseteq [d] \times [n]$ and a validation set $\Omega^{\text{v}} \subseteq [d] \times [n]$. To obtain the split, our algorithm first samples uniformly at random a total ordering π_0 from $\mathcal{T}$ (Line 2). For every course $i \in [d]$, we find the sub-ordering of the n elements within this course (that is, the ordering of the elements $\{(i,j)\}_{j\in[n]}$) according to π_0 (Line 4). For each consecutive pair of elements in this sub-ordering, we assign one element in this pair to the training set and the other element to the validation set uniformly at random (Lines 5-7). We note that in comparison to classical cross-validation methods, our algorithm uses the total ordering π_0 to guide the split, instead of independently assigning each individual element to either the training set or the validation set uniformly at random. This splitting procedure ensures that for each element in the validation set there is an element that is “close” in the training set with respect to the partial ordering $\mathcal{O}$. This property is useful for interpolation in the subsequent validation step.

Validation step Given the training set and the validation set, our algorithm iterates over the choices of $\lambda \in \Lambda$ as follows. For each value of λ, the algorithm first computes our estimator with penalization parameter λ on the training set Ω^{t} to obtain $(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$. The optimization (Line 10) is done by replacing the Frobenius norm on the two terms in the original objective (4.2) by the Frobenius norm restricted to Ω^{t}. Note that this modified objective is independent from the parameters $\{b_{ij}\}_{(i,j)\in\Omega^{\text{v}}}$. Therefore, by the tie-breaking rule of minimizing $\|\widehat{B}^{(\lambda)}\|_F$, we have $[\widehat{B}^{(\lambda)}]_{ij} = 0$ for each $(i,j) \in \Omega^{\text{v}}$.

Next, our algorithm evaluates these choices of λ by their corresponding cross-validation (CV) errors. The high-level idea is to evaluate the fitness of $(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$ to the validation set Ω^{v}, by computing $\frac{1}{|\Omega^{\text{v}}|}\|Y - \widehat{x}^{(\lambda)}\mathbf{1}^T - \widehat{B}^{(\lambda)}\|_{\Omega^{\text{v}}}^2$. However, recall that the estimate $\widehat{B}^{(\lambda)}$ only estimates the bias on the training set meaningfully, and we have $\widehat{B}_{ij}^{(\lambda)} = 0$ for each element (i,j) in the validation set Ω^{v}. Therefore, we “synthesize” the estimated bias $\widetilde{B}^{(\lambda)}$ on the validation from the estimated bias $\widehat{B}^{(\lambda)}$ on the training set via an interpolation procedure (Lines 11-16), as explained below.

Algorithm 2: Cross-validation. Inputs: observations Y, partial ordering $\mathcal{O}$, and set Λ.

```
/* Step 1:  Split the data */
Initialize the training and validation sets as $\Omega^{\mathrm{t}} \leftarrow \{\}$, $\Omega^{\mathrm{v}} \leftarrow \{\}$.
Sample a total ordering of $\pi_0$ uniformly at random from the set $\mathcal{T}$ of all total orderings
  (of the $dn$ elements) consistent with the partial ordering $\mathcal{O}$.
foreach $i \in [d]$ do
    Find the sub-ordering of the $n$ elements in course $i$ according to $\pi_0$, denoted in
      increasing order as $(i, j^{(1)}), \ldots, (i, j^{(n)})$.
    for $t = 1, \ldots, \frac{n}{2}$ do
        Assign $(i, j^{(2t-1)}), (i, j^{(2t)})$ to $\Omega^{\mathrm{t}}$ and $\Omega^{\mathrm{v}}$, one each uniformly at random. If $n$ is
          odd, assign the last element $(i, j^{(n)})$ to the validation set.
    end
end
/* Step 2:  Compute validation error */
foreach $\lambda \in \Lambda$ do
    Obtain $(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$ as a solution to the following optimization problem:

        $$(\widehat{x}_\lambda, \widehat{B}^{(\lambda)}) \in \mathop{\mathrm{argmin}}_{\substack{x \in \mathbb{R}^d,\ B \in \mathbb{R}^{d \times n}, \\ B \text{ satisfies } \mathcal{O}}} \|Y - x\mathbf{1}^T - B\|_{\Omega^{\mathrm{t}}}^2 + \lambda \|B\|_{\Omega^{\mathrm{t}}}^2,$$

      where ties are broken by minimizing $\|\widehat{B}^{(\lambda)}\|_F$.
    foreach $(i, j) \in \Omega^{\mathrm{v}}$ do
        foreach $\pi \in \mathcal{T}$ do
            Find the element $(i^\pi, j^\pi) \in \Omega^{\mathrm{t}}$ that is closest to $(i, j)$ with respect to $\pi$, and
              set $[\widetilde{b}_\pi^{(\lambda)}]_{ij} = \widehat{b}_{i^\pi j^\pi}^{(\lambda)}$. There may be two closest elements at equal distance to
              $(i, j)$, in which case call them $(i_1^\pi, j_1^\pi)$ and $(i_2^\pi, j_2^\pi)$ and set
              $[\widetilde{b}_\pi^{(\lambda)}]_{ij} = \frac{\widehat{b}_{i_1^\pi j_1^\pi}^{(\lambda)} + \widehat{b}_{i_2^\pi j_2^\pi}^{(\lambda)}}{2}$.
        end
        Interpolate the bias as $\widetilde{B}^{(\lambda)} = \frac{1}{|\mathcal{T}|} \sum_{\pi \in \mathcal{T}} \widetilde{B}_\pi^{(\lambda)}$.
    end
    Compute the CV error $e^{(\lambda)} := \frac{1}{|\Omega^{\mathrm{v}}|} \|Y - \widehat{x}_\lambda \mathbf{1}^T - \widetilde{B}^{(\lambda)}\|_{\Omega^{\mathrm{v}}}^2$.
end
Output $\lambda_{\mathrm{cv}} \in \mathrm{argmin}_{\lambda \in \Lambda}\, e^{(\lambda)}$.        (Ties are broken arbitrarily)
```

Interpolation We now discuss how the algorithm interpolates the bias $\widetilde{b}_{ij}^{(\lambda)}$ at each element $(i,j) \in \Omega^{\text{v}}$ from $\widehat{B}^{(\lambda)}$. We first explain how to perform interpolation with respect to some given total ordering π (Line 13), and then compute a mean of these interpolations by iterating over $\pi \in \mathcal{T}$ (Line 15).

- **Interpolating with respect to a total ordering (Line 13):** Given some total ordering π, we find the element in the training set that is the closest to (i,j) in the total ordering π. We denote this closest element from the training set as (i^π, j^π), and simply interpolate the bias at (i,j) with respect to π (denoted $[\widetilde{b}_\pi^{(\lambda)}]_{ij}$) using the value of $\widehat{b}_{i^\pi j^\pi}$. That is, we set $[\widetilde{b}_\pi^{(\lambda)}]_{ij} = \widehat{b}_{i^\pi j^\pi}^{(\lambda)}$. If there are two closest elements of equal distance to (i,j) (one ranked higher than (i,j) and one lower than (i,j) in π), we use the mean of the estimated bias $\widehat{B}^{(\lambda)}$ of these two elements. This step is similar to the CV error computation in [179].
- **Taking the mean over all total orderings in $\mathcal{T}$ (Line 15):** After we find the interpolated bias $\widetilde{B}_\pi^{(\lambda)}$ on the validation set with respect to each π, the final interpolated bias $\widetilde{b}^{(\lambda)}$ is computed as the mean of the interpolated bias over all total orderings $\pi \in \mathcal{T}$. The reason for taking the mean over $\pi \in \mathcal{T}$ is as follows. When we interpolate by sampling a single ordering $\pi \in \mathcal{T}$, this sampling of the ordering introduces randomness in terms of which training elements are chosen for which validation elements, and hence increasing the variance of the CV error.[3] Taking the mean over all total orderings eliminates this source of the variance of the CV error due to sampling, and therefore leads to a better choice of λ.

After interpolating the bias $\widetilde{B}^{(\lambda)}$ on the validation set, the CV error is computed as $\frac{1}{|\Omega^{\text{v}}|}\|Y - \widehat{x}^{(\lambda)}\mathbf{1}^T - \widetilde{B}^{(\lambda)})\|_{\Omega^{\text{v}}}$ (Line 17). Finally, the value of $\lambda_{\text{cv}} \in \Lambda$ is chosen by minimizing the CV error (with ties broken arbitrarily). This completes the description of the cross-validation algorithm.

Implementation Now we comment on two important operations in Algorithm 2: sampling a total ordering from the set $\mathcal{T}$ of total orderings consistent with the partial ordering $\mathcal{O}$ (Line 2), and iterating over the set $\mathcal{T}$ (Line 12). For sampling a total ordering from $\mathcal{T}$ uniformly at random, many algorithms have been proposed that are approximate [29, 119] or exact [88]. For iterating over $\mathcal{T}$ which can be computationally intractable, we approximate the true mean over $\mathcal{T}$ by sampling from $\mathcal{T}$ multiple times, and take their empirical mean. In many practical settings, the partial ordering contains a structure on which these two operations are simple to implement and run in polynomial time – we discuss a subclass of such partial orderings termed "group orderings" in the theoretical results (Section 4.4.1); this subclass of partial orderings is also evaluated in the experiments (Section 4.5).

[3] In more detail, this variance on the CV error due to sampling causes the algorithm to choose an excessively large λ to underestimate the bias. A large λ shrinks the the magnitude of the estimated bias towards 0, and therefore the estimated bias becomes closer to each other, reducing this variance – in the extreme case, if the estimated bias is 0 on all elements from the training set, then the interpolated bias is 0 in the validation set regardless of the ordering π, giving no variance due to sampling π.

4.4 Theoretical guarantees

We now present theoretical guarantees for our proposed estimator (cf. (4.2)) along with our cross-validation algorithm (Algorithm 2). In Section 4.4.2 and 4.4.3, we establish properties of our estimator at the two extremal choices of λ ($\lambda = 0$ and $\lambda = \infty$) for no noise and no bias settings respectively. Then in Section 4.4.4, we analyze the cross-validation algorithm. The proofs of all results are in Chapter 11.

4.4.1 Preliminaries

Model assumptions: To introduce our theoretical guarantees, we start with several model assumptions that are used throughout the theoretical result of this chapter. Specifically, we make the following assumptions on the model (4.1):

(A1) **Noise:** The noise terms $\{z_{ij}\}_{i \in [d], j \in [n]}$ are i.i.d. $\mathcal{N}(0, \eta^2)$ for some constant $\eta \geq 0$.

(A2) **Bias:** The bias terms $\{b_{ij}\}_{i \in [d], j \in [n]}$ are marginally distributed as $\mathcal{N}(0, \sigma^2)$ for some constant $\sigma \geq 0$ unless specified otherwise, and obey one of the total orderings (selected uniformly at random from the set of total orderings) consistent with the partial ordering $\mathcal{O}$. That is, we first sample dn values i.i.d. from $\mathcal{N}(0, \sigma^2)$, and then sample one total ordering uniformly at random from all total orderings consistent with the partial ordering $\mathcal{O}$. Then we assign these dn values to $\{b_{ij}\}$ according to the sampled total ordering.

(A3) **Number of courses:** The number of courses d is assumed to be a fixed constant.

All theoretical results hold for any arbitrary $x^* \in \mathbb{R}^d$. It is important to note that the estimator (4.2) and the cross-validation algorithm (Algorithm 2) requires no knowledge of these distributions or standard deviation parameters σ and η.

Throughout the theoretical results, we consider the solution $\widehat{x}^{(\lambda_{\text{cv}})}$ as solution at $\lambda = \lambda_{\text{cv}}$ on the training set.

Our theoretical analysis focuses on a general subclass of partial orderings, termed "group orderings", where each rating belongs to a group, and the groups are totally ordered.

Definition 4.1 (Group ordering). *A partial ordering $\mathcal{O}$ is called a group ordering with r groups if there is a partition $G_1, \ldots, G_r \subseteq [d] \times [n]$ of the dn ratings such that $((i,j),(i',j')) \in \mathcal{O}$ if and only if $(i,j) \in G_k$ and $(i',j') \in G_{k'}$ for some $1 \leq k < k' \leq r$.*

Note that in Definition 4.1, if two samples are in the same group, we do not impose any relation restriction between these two samples.

Group orderings arise in many practical settings. For example, in course evaluation, the groups can be letter grades (e.g., $\{\text{A}, \text{B}, \text{C}, \text{D}, \text{F}\}$ or $\{\text{Pass}, \text{Fail}\}$), or numeric scores (e.g., in the range of $[0, 100]$) of the students. The group ordering intuitively says that a student receiving a strictly higher grade is more positively biased in rating than a student receiving a lower grade. A total ordering is also group ordering, with the number of groups equal to the number of samples. We assume that the number of groups is $r \geq 2$ since otherwise groups are vacuous.

Denote ℓ_{ik} as the number of students of group $k \in [r]$ in course $i \in [d]$. We further introduce some regularity conditions used in the theoretical results. The first set of regularity conditions is motivated from the case where students receive a discrete set of letter grades.

Definition 4.2 (Group orderings with the single constant-fraction assumption)**.** *A group ordering is said to satisfy the single c-fraction assumption for some constants $c \in (0, 1)$ if there exists some group $k \in [r]$ such that $\ell_{ik} > cn \; \forall \; i \in [r]$.*

Definition 4.3 (Group orderings with the all constant-fraction assumption)**.** *A group ordering of r groups is said to satisfy the all c-fraction assumption for some constant $c \in (0, \frac{1}{r})$, if $\ell_{ik} \geq cn \; \forall \; i \in [d], \; k \in [r]$.*

Note that group orderings with all c-fractions is a subset of group orderings with single c-fraction. The final regularity condition below is motivated from the scenario where student performances are totally ranked in the course.

Definition 4.4 (Total orderings with the constant-fraction interleaving assumption)**.** *Let $\mathcal{O}$ be a total ordering (of the dn elements $\{(i, j)\}_{i \in [d], j \in [n]}$). We define an interleaving point as any number $t \in [dn - 1]$, such that the t^{th} and the $(t + 1)^{th}$ highest-ranked elements according to the total ordering $\mathcal{O}$ belong to different courses. A total ordering $\mathcal{O}$ is said to satisfy the c-fraction interleaving assumption for some constant $c \in (0, 1)$, if there are at least cn interleaving points in $\mathcal{O}$.*

With these preliminaries in place, we now present our main theoretical results.

4.4.2 $\lambda = 0$ is consistent when there is no noise

We first consider the extremal case where there is only bias but no noise involved. The following theorem states that our estimator with $\lambda = 0$ is consistent in estimating the underlying quantity x^*, that is $\widehat{x}^{(0)} \to x^*$ in probability.

Theorem 4.5. *[Consistency in estimating x^*] Suppose the assumptions (A1), (A2) and (A3) hold. Suppose there is no noise, or equivalently suppose $\eta = 0$ in (A1). Consider any $x^* \in \mathbb{R}^d$. Suppose the partial ordering is one of:*

(a) any group ordering of r groups satisfying the all c-fraction assumption, where $c \in (0, \frac{1}{r}]$ is a constant, or

(b) any group ordering with $d = 2$ courses and 2 groups, or

(c) any total ordering.

Then for any $\epsilon > 0$ and $\delta > 0$, there exists an integer n_0 (dependent on $\epsilon, \delta, c, d, \eta$), such that for every $n \geq n_0$ and every partial ordering satisfying at least one of the conditions (a), (b) or (c):

$$\mathbb{P}\Big(\|\widehat{x}^{(0)} - x^*\|_2 < \epsilon \Big) \geq 1 - \delta.$$

The proof of this result is provided in Section 11.3. The convergence of the estimator to the true qualities x^* implies the following corollary on ranking the true qualities x^*. In words, our estimator $\widehat{x}^{(0)}$ is consistent in comparing the true qualities x^*_i and $x^*_{i'}$ of any pair of courses $i, i' \in [d]$ with $i \neq i'$, as long as their values are distinct.

Corollary 4.6 (Consistency on the ranking of x^*)**.** *Suppose the assumptions (A1), (A2) and (A3) hold. Consider any $x^* \in \mathbb{R}^d$. Assume there is no noise, or equivalently assume $\eta = 0$ in (A1). Then for any $\delta > 0$, there exists an integer n_0 (dependent on x^*, δ, c, d, η), such that for all $n \geq n_0$ and every partial ordering satisfying at least one of the conditions (a), (b) or (c) in Theorem 4.5:*

$$\mathbb{P}\Big(\operatorname{sign}(\widehat{x}_i - \widehat{x}_{i'}) = \operatorname{sign}(x^*_i - x^*_{i'}) \Big) \geq 1 - \delta \qquad \textit{for all } i, i' \in [d] \textit{ such that } i \neq i' \textit{ and } x^*_i \neq x^*_{i'}.$$

In Section 4.6.1, we also evaluate the mean estimator. We show that under the conditions of Theorem 4.5, the mean estimator is provably not consistent. This is because the mean estimator does not account for the biases and only tries to correct for the noise. In order to obtain a baseline that accommodates the outcome-dependent bias (since to the best of our knowledge there is no prior literature on it), in Section 4.6.2 we then propose a reweighted mean estimator. It turns out that our estimator at $\lambda = 0$ also theoretically outperforms this reweighted mean estimator (see Proposition 4.13 in Section 4.6.2).

4.4.3 $\lambda = \infty$ is minimax-optimal when there is no bias

We now move to the other extremity of $\lambda = \infty$, and consider the other extremal case when there is only noise but no bias. Recall that we define the estimator at $\lambda = \infty$ as $\widehat{x}^{(\infty)} = \lim_{\lambda\to\infty} \widehat{x}^{(\lambda)}$. The following proposition states that this limit is well-defined, and our estimator reduces to taking the sample mean at this limit.

Proposition 4.7 (Estimator at $\lambda = \infty$)**.** *The limit of* $(\widehat{x}^{(\infty)}, \widehat{B}^{(\infty)}) := \lim_{\lambda\to\infty}(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$ *exists and is given by*

$$\begin{aligned} [\widehat{x}^{(\infty)}]_i &= \frac{1}{n}\sum_{j=1}^{n} y_{ij}, \qquad \textit{for each } i \in [d], \textit{ and} \\ \widehat{B}^{(\infty)} &= 0. \end{aligned} \tag{4.3}$$

The proof of this result is provided in Section 11.4. With no bias, estimating the true quality x^* reduces to estimating the mean of a multivariate normal distribution with the covariance matrix $\eta^2 I_d$, where I_d denotes the identity matrix of size $d \times d$. Standard results in the statistics literature imply that taking the sample mean is minimax-optimal in this setting if d is a fixed dimension, formalized in the following proposition for completeness.

Proposition 4.8 (Implication of Example 15.8 in 185)**.** *Let $d \geq 1$ be a fixed constant. Let $Y = x^*\mathbf{1}^T + Z$, where $x^* \in \mathbb{R}^d$ is an unknown vector and each entry of Z is i.i.d. $\mathcal{N}(0, \eta^2)$ with unknown η. Then the sample mean estimator $\widehat{x} = \frac{1}{n}Y\mathbf{1}$ is minimax-optimal for the squared ℓ_2-risk $\frac{1}{d}\mathbb{E}\|\widehat{x} - x^*\|_2^2$, up to a constant factor that is independent of d.*

This concludes the properties of our estimator at the two extremal cases.

4.4.4 Cross-validation effectively selects λ

This section provides the theoretical guarantees for our proposed cross-validation algorithm. Specifically, we show that in the two extremal cases, cross-validation outputs a solution that converges in probability to the solutions at $\lambda = 0$ and $\lambda = \infty$, respectively. Note that the cross-validation algorithm is agnostic to the values of σ and η, or any specific shape of the bias or the noise.

The first result considers the case when there is only bias and no noise, and we show that cross-validation obtains a solution that is close to the solution using a fixed choice of $\lambda = 0$. The intuition for this result is as follows. The CV error $\|Y - \widehat{x}^{(\lambda)}\mathbf{1}^T - \widetilde{B}^{(\lambda)}\|_{\Omega^{\mathrm{v}}}^2$ measures the difference between the bias-corrected observations $Y - \widetilde{B}^{(\lambda)}$ and the estimated qualities $\widehat{x}^{(\lambda)}\mathbf{1}^T$.

By construction, the values in $\widehat{x}^{(\lambda)}\mathbf{1}^T$ are identical within each row. Hence, to minimize the CV error we want $\widetilde{B}^{(\lambda)}$ to capture as much variance as possible within each row of Y. Now consider $\lambda = 0$. In this case $\widehat{B}^{(\lambda)}$ correctly captures the intra-course variance of the bias on the training set due to the noiseless assumption. Due to the nearest-neighbor interpolation, we expect that the interpolated $\widetilde{B}^{(\lambda)}$ captures most of the intra-course variance of the bias on the validation set, giving a small CV error. However, for larger $\lambda > 0$, the bias estimated from the training set shrinks in magnitude due to the regularization term. The bias $\widehat{B}^{(\lambda)}$ and hence $\widetilde{B}^{(\lambda)}$ only capture a partial extent of the actual bias in the observations. The rest of the uncaptured bias within each course contributes to the residue $\|Y - \widehat{x}^{(\lambda)}\mathbf{1}^T - \widetilde{B}^{(\lambda)}\|_{\Omega^{\mathrm{v}}}^2$, giving a larger CV error. Hence, cross-validation is likely to choose $\lambda = 0$ (or some sufficiently small value of λ). The following theorem shows that cross-validation is consistent in estimating x^* under the only-bias setting.

Theorem 4.9. *Suppose the assumptions (A1), (A2) and (A3) hold. Consider any $x^* \in \mathbb{R}^d$. Suppose there is no noise, or equivalently suppose $\eta = 0$ in (A1). Suppose $c \in (0, 1)$ is a constant. Suppose the partial ordering is either:*

(a) any group ordering satisfying the all c-fraction assumption, or

(b) any total ordering with $d = 2$.

Let $0 \in \Lambda$. Then for any $\delta > 0$ and $\epsilon > 0$, there exists some integer n_0 (dependent on $\epsilon, \delta, c, d, \sigma$), such that for every $n \geq n_0$ and every partial ordering satisfying (a) or (b):

$$\mathbb{P}\Big(\|\widehat{x}^{(\lambda_{\mathrm{cv}})} - x^*\|_2 < \epsilon\Big) \geq 1 - \delta.$$

The proof of this result is provided in Section 11.5. From Theorem 4.5 we have that the estimator $\widehat{x}^{(0)}$ (at $\lambda = 0$) is also consistent under the only-bias setting. Combining Theorem 4.5 with Theorem 4.9, we have $\widehat{x}^{(\lambda_{\mathrm{cv}})}$ approaches $\widehat{x}^{(0)}$. Formally, under the conditions of Theorem 4.9, we have

$$\mathbb{P}\Big(\|\widehat{x}^{(\lambda_{\mathrm{cv}})} - \widehat{x}^{(0)}\|_2 < \epsilon\Big) \geq 1 - \delta.$$

The next result considers the case when there is only noise and no bias, and we show that cross-validation obtains a solution that is close to the solution using a fixed choice of $\lambda = \infty$ (sample mean). Intuitively, at small values of λ the estimator still tries to estimate a non-trivial amount of the interpolated bias $\widetilde{B}^{(\lambda)}$. However, any such non-trivial interpolated bias is erroneous since there is no bias in the observations to start with, increasing the CV error $\|Y - \widehat{x}^{(\lambda)}\mathbf{1}^T - \widetilde{B}^{(\lambda)}\|_{\Omega^{\mathrm{v}}}^2$ by doing a wrong bias "correction". On the other hand, at $\lambda = \infty$ (or some λ that is sufficiently large), the interpolated bias $\widetilde{B}^{(\lambda)}$ is zero (or close to zero), which is the right thing to do and hence gives a smaller CV error. The following theorem shows that cross-validation is consistent in estimating x^* under the only-noise setting.

Theorem 4.10. *Suppose the assumptions (A1), (A2) and (A3) hold. Consider any $x^* \in \mathbb{R}^d$. Suppose there is no bias, or equivalently assume $\sigma = 0$ in (A2). Suppose $c_1, c_2 \in (0, 1)$ are constants. Suppose the partial ordering is either:*

(a) any group ordering satisfying the single c_1-fraction assumption, or

(b) any total ordering satisfying the c_2-fraction interleaving assumption with $d = 2$.

Let $\infty \in \Lambda$. Then for any $\delta > 0$ and $\epsilon > 0$, there exists some integer n_0 (dependent on $\epsilon, \delta, c_1, c_2, d, \eta$), such that for every $n \geq n_0$ and every partial ordering satisfying (a) or (b):

$$\mathbb{P}\Big(\|\widehat{x}^{(\lambda_{\mathrm{cv}})} - x^*\|_2 < \epsilon\Big) \geq 1 - \delta.$$

The proof of this result is provided in Section 11.6. By the consistency of $\widehat{x}^{(\infty)}$ implied from Proposition 4.8 under the only-noise setting, this result implies that the estimator $\widehat{x}^{(\lambda_{\mathrm{cv}})}$ approaches $\widehat{x}^{(\infty)}$. Formally, under the conditions of Theorem 4.10, we have

$$\mathbb{P}\Big(\|\widehat{x}^{(\lambda_{\mathrm{cv}})} - \widehat{x}^{(\infty)}\|_2 < \epsilon\Big) \geq 1 - \delta.$$

Recall that the sample mean estimator is commonly used and minimax-optimal in the absence of bias. This theorem suggests that our cross-validation algorithm, by adapting the amount of regularization in a data-dependent manner, recovers the sample mean estimator under the setting when sample mean is suitable (under only noise and no bias).

These two theorems, in conjunction to the properties of the estimator at $\lambda = 0$ and $\lambda = \infty$ given in Sections 4.4.2 and 4.4.3 respectively, indicate that our proposed cross-validation algorithm achieves our desired goal in the two extremal cases. The main intuition underlying these two results is that if the magnitude of the estimated bias from the training set aligns with the true amount of bias, the interpolated bias from the validation set also aligns with the true amount of bias and hence gives a small CV error. Extending this intuition to the general case where there is both bias and noise, one may expect cross-validation to still able to identify an appropriate value of λ.

4.5 Experiments

We now conduct experiments to evaluate our estimator and our cross-validation algorithm under various settings. We consider the metric of the squared ℓ_2 error. To estimate the qualities using our cross-validation algorithm, we first use Algorithm 2 to obtain a value of the hyperparameter λ_{cv}; we then compute the estimate $\widehat{x}^{(\lambda_{\mathrm{cv}})}$ as the solution to (4.2) at $\lambda = \lambda_{\mathrm{cv}}$ (that is, we solve (4.2) on the entire data combining the training set and the validation set).[4] Implementation details for the cross-validation algorithm (Algorithm 2) are provided in Section 4.7.1. Throughout the experiments, we use $\Lambda = \{2^i : -9 \leq i \leq 5, i \in \mathbb{Z}\} \cup \{0, \infty\}$. We also plot the error incurred by the best fixed choice of $\lambda \in \Lambda$, where for each point in the plots, we pick the value of $\lambda \in \Lambda$ which minimizes the empirical ℓ_2 error over all fixed choices in Λ. Note that this best fixed choice is not realizable in practice since we cannot know the actual value of the ℓ_2 error.

We compare our cross-validation algorithm with the mean, median, and also the reweighted mean estimator introduced in Section 4.6.2. The mean estimator is the sample mean for each course (same as our estimator at $\lambda = \infty$) defined as $[\widehat{x}_{\mathrm{mean}}]_i = \frac{1}{n}\sum_{j \in [n]} y_{ij}$ for each $i \in [d]$, and the median estimator is defined as $[\widehat{x}_{\mathrm{med}}]_i = \mathrm{median}(y_{i1}, \ldots, y_{in})$ for each $i \in [d]$. The reweighted mean estimator is not applicable to total orderings.

[4] Note that this is different from the theoretical results in Section 4.4.4, where we solve (4.2) at $\lambda = \lambda_{\mathrm{cv}}$ only on the training set.

In the model (4.1), we assume that the noise terms $\{z_{ij}\}_{i\in[d],j\in[n]}$ and the bias terms $\{b_{ij}\}_{i\in[d],j\in[n]}$ follow the assumptions (A1) and (A2) respectively for our theoretical results in Section 4.4.1. In our simulations, we consider three cases for the amounts of bias and noise: only bias ($\sigma = 1, \eta = 0$), only noise ($\sigma = 0, \eta = 1$), and both bias and noise ($\sigma = 0.5, \eta = 0.5$). Throughout the experiments we use $x^* = 0$, and as explained in Proposition 11.5 in Section 11.2.1, the results remain the same for any value of x^*.

Each point in all the plots is computed as the empirical mean over 250 runs. Error bars in all the plots represent the standard error of the mean.

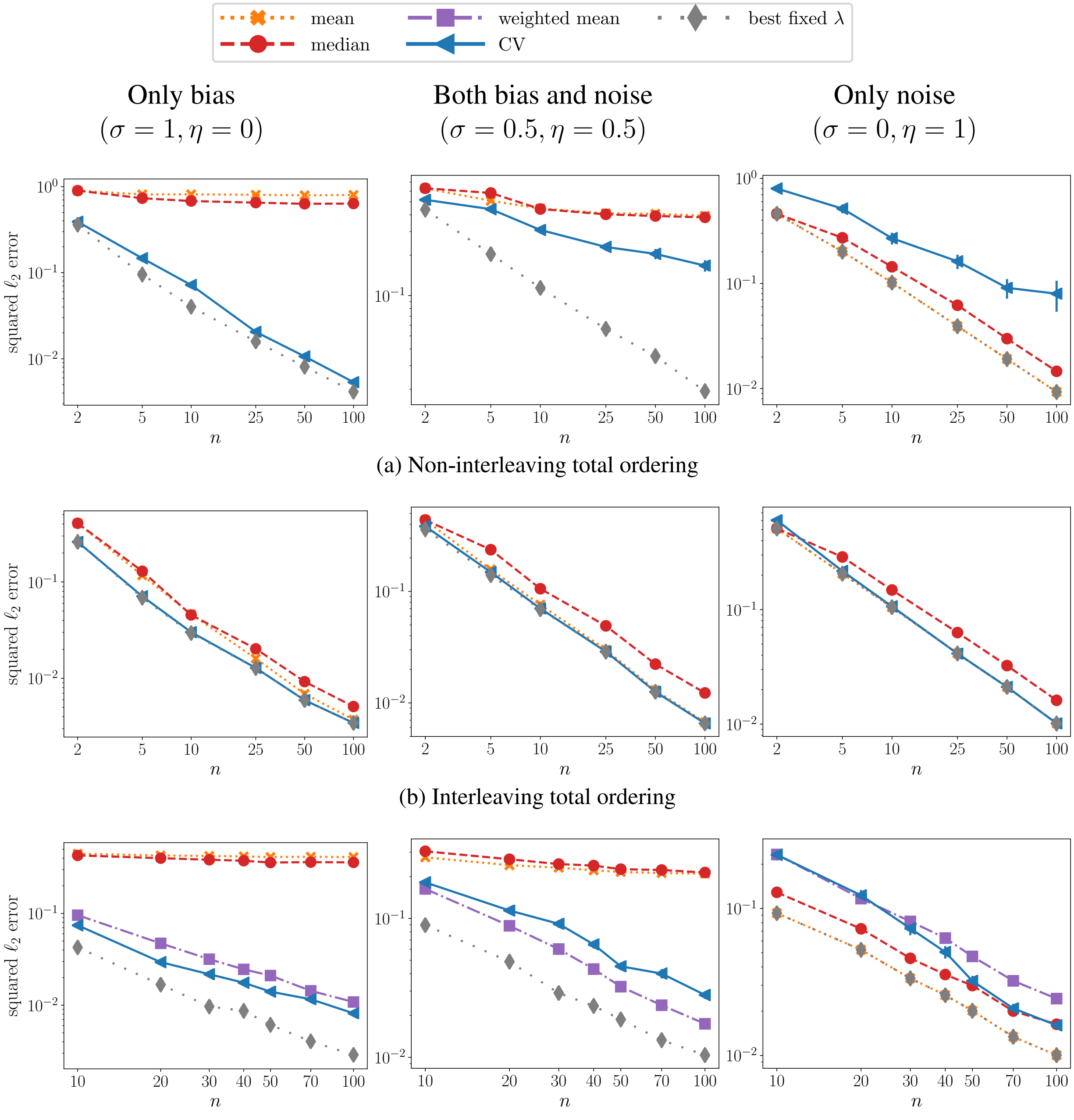
4.5.1 Dependence on n

We first focus on group orderings. We evaluate the performance of our estimator under different values of n, under the following types of group orderings.

- **Non-interleaving total ordering:** We call a total ordering a "non-interleaving" total ordering, if the total ordering is $b_{11} \leq \ldots \leq b_{1n} \leq b_{21} \leq \ldots \leq b_{2n} \leq \ldots \leq b_{d1} \leq \ldots b_{dn}$. In the non-interleaving total ordering, the values of the bias terms vary quite significantly across courses. Our goal is to evaluate whether our estimator provides good estimates under such imbalanced bias.
- **Interleaving total ordering:** We call a total ordering an "interleaving" total ordering, if the total ordering is $b_{11} \leq b_{21} \leq \ldots \leq b_{d1} \leq b_{12} \leq \ldots \leq b_{d2} \leq b_{1n} \leq \ldots \leq b_{dn}$. In contrast to the non-interleaving total ordering, in the interleaving total ordering the bias terms are more balanced across different courses, and we expect the mean and the median baselines to work well in this setting. Our goal is to evaluate whether the cross-validation algorithm deviates much from the baselines when the baselines work well.
- **Binary ordering:** We call a group ordering a "binary" ordering, if there are $r = 2$ groups. Specifically, we consider a group distribution where $(\ell_{i1}, \ell_{i2}) = (0.9n, 0.1n)$ for half of the courses i, and $(\ell_{i1}, \ell_{i2}) = (0.1n, 0.9n)$ for the other half of the courses i.

We consider $d = 3$ courses for the non-interleaving and interleaving total orderings, and consider $d = 4$ for the binary ordering. The results are shown in Fig. 4.1. In the non-interleaving case (Fig. 4.1a) and the binary case (Fig. 4.1c) where the distribution of the bias is quite imbalanced, our estimator performs better than the mean and median baselines when there is bias (with or without noise). The improvement is the most significant in the case when there is only bias and no noise. In the case where there is only noise, our estimator still performs reasonably as compared to the the baselines – the performance of our estimator is worse, but this is not unexpected, because while our algorithm tries to compensate for possible bias, the mean and median baselines do not. Indeed, as the theory (Proposition 4.8) suggests, the mean estimator is ideal for the only-noise setting, but in practice we do not know whether we operate in this only-noise setting a priori. In the interleaving case where the bias is more balanced (Fig. 4.1b), our estimator performs on par with the baselines, and is still able to correct the small amount of bias in the only-bias case.

We also compare our estimator with the reweighted mean estimator in the binary case. Recall that the reweighted mean estimator is more specialized and not applicable to total orderings or more general partial orderings. Our estimator performs slightly better than the reweighted mean



(a) Non-interleaving total ordering

(b) Interleaving total ordering

(c) Binary ordering

Figure 4.1: The performance of our estimator (with cross-validation and with the best fixed λ) for various values of n, compared to the mean, median and reweighted mean estimators.

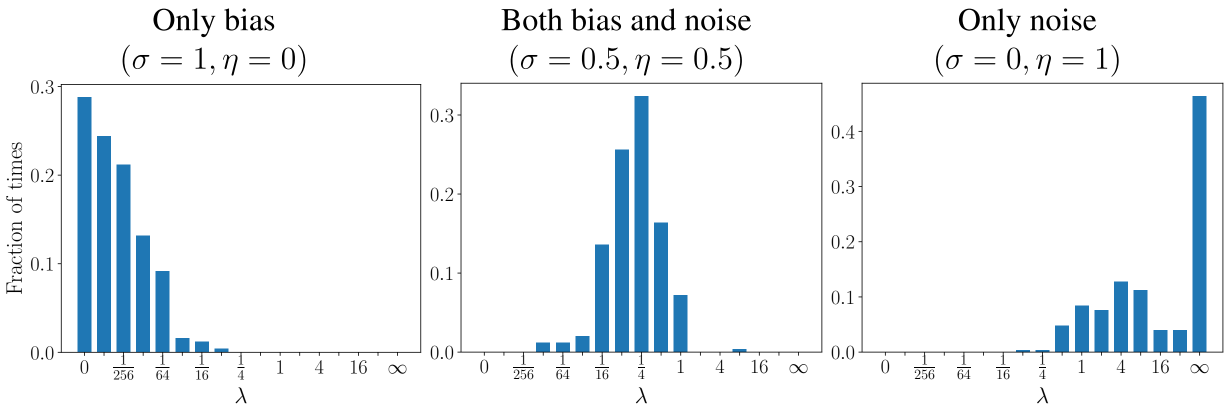


Figure 4.2: The histogram on the fraction of times each value of λ is chosen by cross-validation. Cross-validation is able to choose the value of λ adaptive to different amounts of bias and noise.

estimator in the two extremal (only-bias and only-noise) cases. In the noisy case, the best fixed λ is better than the reweighted mean estimator but the cross-validation algorithm is worse. In general, we observe that there remains a non-trivial gap between the best fixed λ and cross-validation in the noisy case (also see the non-interleaving total ordering in the noisy case). If prior knowledge about the relative amounts of bias and noise is given, we may be able to achieve better performance with our estimator by setting the value of λ manually.

4.5.2 Choices of λ by cross-validation

We inspect the choices of the hyperparameter λ made by our cross-validation algorithm. We use the binary setting from Section 4.5.1, with $n = 50$. The histograms in Fig. 4.2 plot the fraction of times that each value of $\lambda \in \Lambda$ is chosen by cross-validation. When there is only bias, the chosen value of λ is small (with $\lambda = 0$ as the most chosen); when there is only noise, the chosen value of λ is large (with $\lambda = \infty$ as the most chosen). When there is both bias and noise, the value of λ lies in the middle of the two extremal cases. These trends align with our intuition and theoretical results about cross-validation in Section 4.4.4, and show that cross-validation is indeed able to adapt to different amounts of bias and noise present in the data.

4.5.3 The regime of $d > n$

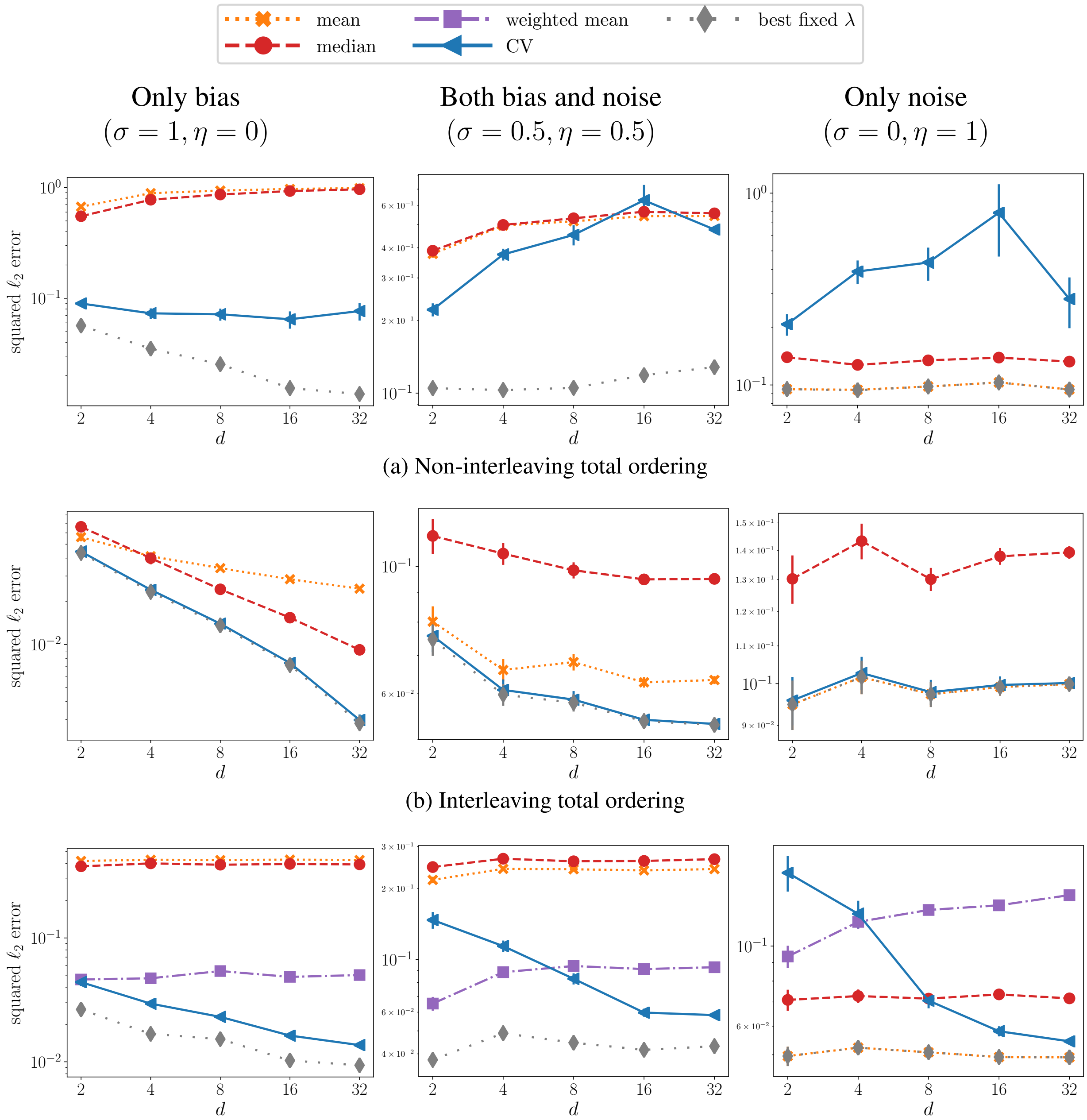
In our theoretical results from Section 4.4, we restricted our attention to the case where the number of courses d is a fixed constant. We now evaluate the regime where the number of courses d becomes large compared to the number of students n, in order to test the general applicability of our estimator. We again consider the three types of group orderings from Section 4.5.1. We set $n = 10$ for the non-interleaving and interleaving total orderings, and $n = 20$ for the binary ordering.

The results with different choices of d are shown in Fig. 4.3. The mean baseline has a flat curve (except for the small sample-size regime of small values of d) and converges to some non-zero constant in all of the settings. The flat curves come from the fact that the number of parameters (i.e., the number of courses d) grows linearly in the number of observations. The median baseline also has a relatively flat curve, with the exception that in the only-bias case for the interleaving ordering, the error decreases rapidly for small values of d, and eventually converges to a very small constant (not shown), because the median observations across courses have very close bias due to the interleaving ordering). Again, our estimator performs better than the mean and median baselines when there is bias. In the binary case, our estimator also performs better than the reweighted mean estimator for large values of d. One notable setting where our estimator does not perform as well is the only-noise case for the non-interleaving ordering. Note that this is a case not covered by the theory in Theorem 4.10(b) because the non-interleaving ordering does not satisfy the constant-fraction interleaving assumption. In this case, our estimator at $\lambda = 0$ (or small values of λ) incurs a large error. Therefore, despite the fact that we empirically observe that cross-validation still chooses large values of λ for a large fraction of times, due to the very large error when small values of λ are chosen, the overall error is still large. The reason that our estimator at $\lambda = 0$ (or small values of λ) gives a large error is that our estimator attempts to explain the data (that has no bias and only noise) as much as possible by the bias. Since in the non-interleaving ordering, course i has smaller bias than course $(i+1)$, our estimator at $\lambda = 0$ mistakenly estimates that $\widehat{x}_i$ is about a constant larger than $\widehat{x}_{i+1}$ for each $i \in [d-1]$, incurring a large error.

4.5.4 General partial orderings

In our theoretical results from Section 4.4, we restricted our attention to group orderings. While group orderings cover a large range of common cases in practice, there may exist other types of partial orderings. We now consider the following two types of general partial orderings that are not group orderings to test the general applicability of our estimator.

- **Total binary tree:** We consider a binary tree, and denote the number of levels (depth) of the tree as ℓ. Each node in the tree represents a single element from the observations. Each node has a direct edge to both of its children, and the partial ordering is the set of all directed edges. Specifically, we consider $d = 2$ courses. In this case, the total number of observations dn is even. Therefore, we construct a binary tree with one (arbitrary) leaf node removed. We assign all the $2^{\ell-1} - 1$ nodes from levels 1 to $(\ell - 1)$ to the first course, and assign all the $2^{\ell-1} - 1$ nodes from level ℓ (leaf nodes) to the second course. This construction is conceptually similar to total orderings in group orderings, where each element takes a distinct role in the partial ordering. In this construction we have the relation $dn = 2^{\ell} - 2$.
- **Binary tree of 3 levels:** We consider a binary tree of 3 levels and therefore 7 nodes in total. Each node contains k elements. There is an ordering constraint between two elements if and only if there is an edge between the corresponding nodes they belong to. We have the relation $dn = 7k$. We consider $d = 3$, and therefore we have $n = \frac{7}{3}k$. The three courses have the following assignment, where the elements in each level are sampled uniformly at random from all elements in this level:
    - Course 1: all k elements from level 1; k elements from level 2; $\frac{k}{3}$ elements from level 3,



(a) Non-interleaving total ordering

(b) Interleaving total ordering

(c) Binary ordering

Figure 4.3: The performance of our estimator (with cross-validation and with the best fixed λ) for various values of d, compared to the mean, median, and reweighted mean estimators.

- Course 2: k elements from level 2; $\frac{4}{3}k$ elements from level 3,
- Course 3: $\frac{7}{3}k$ elements from level 3.

This construction is conceptually similar to a group ordering with a constant number of groups.

We evaluate our estimator under these two types of tree partial orderings for various values of n (setting the values of ℓ and k accordingly). Given that the reweighted mean estimator is defined only for group orderings, we also consider its two extensions that are tailored to tree orderings, termed "reweighted mean (node)" and "reweighted mean (level)" as explained in Section 4.7.2. Similar to the case of group orderings, these two reweighted mean estimators are applicable to the binary tree of 3 levels but not the total binary tree.

The results are shown in Fig. 4.4. Again, when there is noise, we observe that our estimator performs better than the mean and median baselines in both of these two tree orderings. In the binary tree of 3 levels, the construction procedure specifies the number of elements in each course from each level, but there is randomness in which nodes in the level these elements from belong to. Due to this randomness, the reweighted mean (node) estimator is not always applicable, and we use hollow squares to indicate these settings and only compute the error across the runs where the estimator is applicable. We observe that our cross-validation algorithm performs better than the two reweighted mean estimators in the only-bias case. When there is noise (with or without bias), our cross-validation algorithm performs on par while the best fixed λ performs better than the reweighted mean estimators.

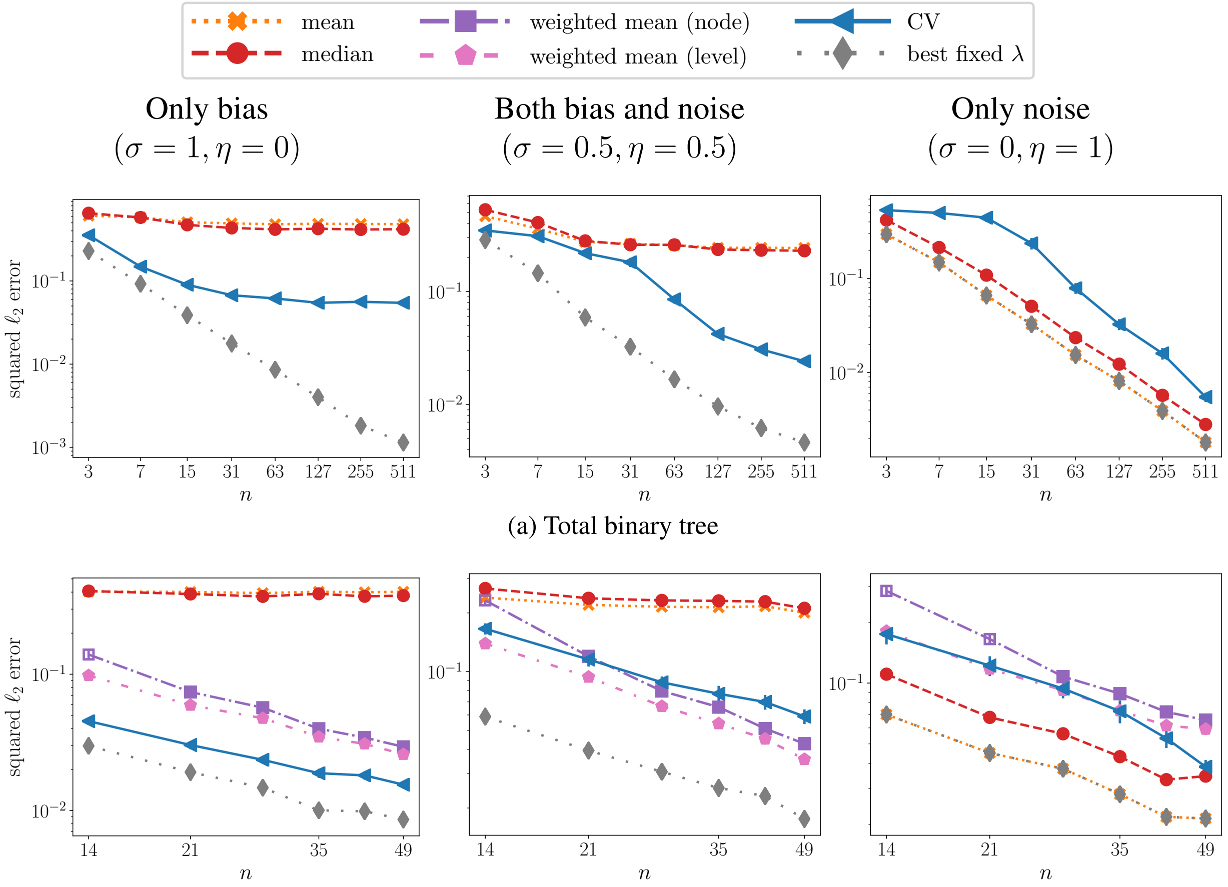
4.5.5 Semi-synthetic grading data

In this section we conduct a semi-synthetic experiment using real grading statistics. We use the grading data from Indiana University Bloomington [90], where the possible grades that students receive are A+ through D-, and F. We consider three ways to construct the group orderings:

- **Fine grades:** The 13 groups correspond to the grades of A+ through D-, and F.
- **Coarse grades:** The fine grades are merged to 5 groups of A, B, C, D and F, where grades in $\{$A+, A, A-$\}$ are all considered A, etc.
- **Binary grades:** The grades are further merged to 2 groups of P and F (meaning pass and fail), where all grades except F are considered P. According to the university's policies, D- is the lowest passing grade.

We use the grading data from the course "Business Statistics" from Spring 2020. This course consists of 10 sessions taught by multiple instructors. The average number of students per session is 50. We choose this course because this course has multiple sessions, so that the grading distributions across different sessions are more balanced. Therefore, many common grades (A+ through B) appear in all sessions, allowing the reweighted mean estimator to use more observations and perform well. Instead, if we consider all 31 statistics courses taught in the semester, then the only grade appearing in all courses is A, and the reweighted mean estimator has to discard the data from all other grades.

We use the number of students and the grade distribution from this course, and synthesize the observations using our model (4.1) under the Gaussian assumptions (A2) and (A1). The



(a) Total binary tree

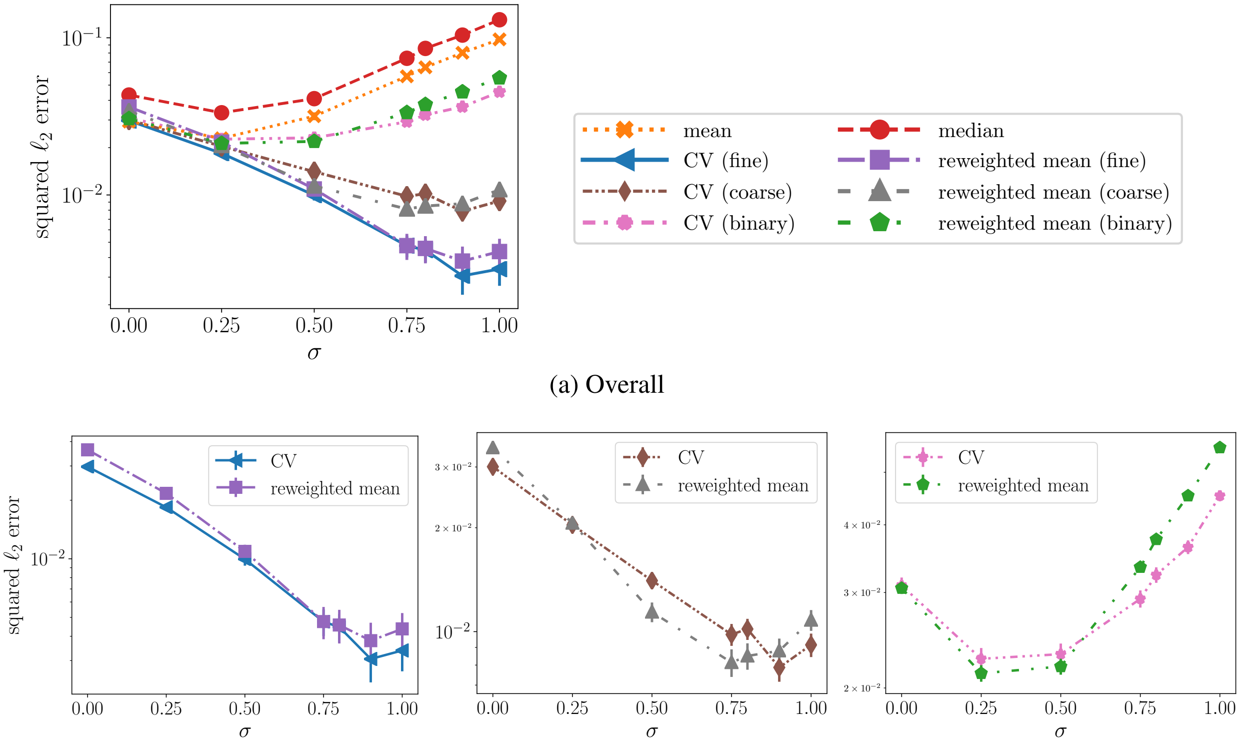
(b) Binary tree of 3 levels

Figure 4.4: The performance of our estimator (with cross-validation and with the best fixed λ) compared to the mean, median, and two reweighted mean estimators, under two types of partial orderings that are not group orderings.

true quality is set as $x^* = 0$ (again the results are independent from the value of x^*); the bias is generated according to the group ordering induced by the fine grades, with a marginal distribution of $\mathcal{N}(0, \sigma^2)$, and the noise is generated i.i.d. from $\mathcal{N}(0, \eta^2)$. We set $\eta = 1 - \sigma$, and consider different choices of σ. The estimators are given one of the three group orderings listed above.

Note that the number of students is unequal in different sessions of the course. The mean and median baselines are still defined as taking the mean and median of each course respectively. The precise definitions of the reweighted mean estimator and our estimator are in Section 4.7.3. We estimate the quality of the 10 sessions of the course individually, even if some sessions are taught by the same instructor.

The results are shown in Fig 4.5. As in previous simulations, the mean and median baselines do not perform well when there is considerable bias (corresponding to a large value of σ). As the number of groups increases from the binary grades to coarse grades and then to the fine grades, the performance of both our estimator and the reweighted mean estimator improves, because the finer orderings provide more information about the bias. Our estimator performs slightly better than the reweighted mean estimator for the fine grades (Fig. 4.5b), and slightly better on a subset



(a) Overall

(b) Fine grades

(c) Coarse grades

(d) Binary grades

Figure 4.5: The performance of our estimator (with cross-validation) on semi-synthetic grading data, compared to the mean, median and reweighted mean estimators.

of values of σ for the coarse grades (Fig. 4.5c). For the binary grades, the error of both our estimator and the reweighted mean estimator increases as the relative amount of bias increases (Fig. 4.5d). This increase is likely due to the model mismatch as the data is generated from fine grades. In this case our estimator performs better than the reweighted mean estimator for large values of σ.

4.6 Auxiliary results

In this section, we present auxiliary theoretical results on comparing our estimator with the mean estimator (Section 4.6.1) and a reweighted mean estimator that we introduce (Section 4.6.2).

4.6.1 Comparison with the mean estimator

Recall from Section 4.5 that the mean estimator for estimating x^* is defined as $[\widehat{x}_{\text{mean}}]_i = \frac{1}{n}\sum_{j\in[n]} y_{ij}$ for each class $i \in [d]$. Taking the mean ignores the bias, and hence it is natural to expect that this estimator does not perform well when the bias in the data is distributed unequally across classes. Intuitively, let us consider two classes of different quality. If students in a stronger class receive lower grades than students in a weaker class, then the bias induced by this distribution of grades may result in the mean estimator ranking the classes incorrectly. The following proposition formalizes this intuition and shows that the mean estimator indeed fails to compare the qualities of courses in the only-bias setting.

Proposition 4.11. *Suppose the assumptions (A1), (A2) and (A3) hold and there is no noise, or equivalently $\eta = 0$ in (A1). Suppose the partial ordering satisfies any one of the conditions in Theorem 4.5:*

(a) any group ordering of r groups with all c-fractions, where $c \in (0, \frac{1}{r})$ is a constant, or

(b) any group ordering with $d = 2$ courses and $r = 2$ groups, or

(c) any total ordering.

Then there exist a partial ordering that satisfies any one of the conditions (a) (with any number of groups $r \geq 2$), (b) or (c), true qualities $x^* \in \mathbb{R}^d$, a pair of courses $i, i' \in [d]$, and an integer n_0 (dependent on the standard parameter σ of the distribution of the bias and the number of groups r in condition (a)), such that for all $n \geq n_0$, we have

$$\mathbb{P}\Big(\operatorname{sign}\left([\widehat{x}_{\text{mean}}]_i - [\widehat{x}_{\text{mean}}]_{i'}\right) = \operatorname{sign}(x^*_i - x^*_{i'}) \Big) < 0.01.$$

The proof of this result is provided in Section 11.7. Note that in condition (a) we require $c \neq \frac{1}{r}$. This requirement is necessary because if $c = \frac{1}{r}$, then the number of students in any course $i \in [d]$ and any group $k \in [r]$ has to be exactly cn. In this case, the bias is evenly distributed across all courses, and in this case the mean estimator is consistent. This negative result on comparing pairs of courses (combined with the fact that both model (4.1) and the mean estimator are shift invariant) implies the following negative result on estimation – the mean estimator $\widehat{x}_{\text{mean}}$ does not converge to the true x^* in probability.

Corollary 4.12. *Suppose the assumptions (A1), (A2) and (A3) hold and there is no noise, or equivalently $\eta = 0$ in (A1). Consider any $x^* \in \mathbb{R}^d$. Suppose the partial ordering satisfies Then there exist a partial ordering that satisfies any one of the conditions (a), (b) or (c), and there exists a constant $\epsilon > 0$ such that for all $n \geq 1$ we have*

$$\mathbb{P}\Big(\|\widehat{x}_{\text{mean}} - x^*\|_2^2 < \epsilon \Big) < 0.01.$$

Recall that our estimator at $\lambda = 0$ is consistent in both comparing the quality of any pair of courses (Corollary 4.6) and estimating the qualities (Theorem 4.5). In contrast, the negative results in Proposition 4.11 and Corollary 4.12 show that the mean estimator is not consistent in comparison or estimation. Moreover, these negative results are stronger, in that they show the probability of correct comparison or estimation not only does not converge to 1, but also can be arbitrarily small. The negative results on the mean estimator stem from the fact that the mean estimator completely ignores the fact that the bias is not evenly distributed across different courses. We remedy this issue by proposing a second baseline – termed a reweighted mean estimator in the following subsection.

4.6.2 A reweighted mean estimator

The second baseline, defined on group orderings only, re-weighs the observations to make the bias evenly distributed across courses, allowing to then take the mean. For each group $k \in [r]$, denote $\ell_{k,\min} := \min_{i \in [d]} \ell_{ik}$ as the minimum number of students in group k among all courses.

Denote $R = \{k \in [r] : \ell_{k,\min} > 0\}$ as the set of groups that appear in all courses. The reweighted mean estimator consists of the following two steps.

Reweighting step The estimator computes a weighted mean of each course $i \in [d]$ as

$$[\widehat{x}_{\text{rw}}]_i = \sum_{k \in R} \frac{\ell_{k,\min}}{\sum_{k' \in R} \ell_{k',\min}} \sum_{j:(i,j) \in Gk} \frac{y_{ij}}{\ell_{ik}}. \tag{4.4}$$

Intuitively, the observations are reweighted in a way such that the bias distribution is balanced among courses. Specifically, for each course $i \in [d]$ and each group $k \in [r]$, this reweighted mean estimator computes its group mean $\sum_{j:(i,j)\in Gk} \frac{y_{ij}}{\ell_{ik}}$, and weighs the contribution of this group mean to the overall mean by the factor of $\frac{\ell_{k,\min}}{\sum_{k' \in R} \ell_{k',\min}}$. This reweighting can bee seen as the expected version of a sampling procedure, where for each course $i \in [d]$ and each group $k \in [r]$, we sample $\ell_{k,\min}$ out of ℓ_{ik} observations so that the number of observations in group k is equal across all courses, and then take the mean on the sampled observations. Note that there are an infinite number choices for the weights to balance the biases, and the choice in (4.4) motivated by sampling is quite natural. It has the property that if all courses have the same group distribution, then the reweighted mean reduces to sample mean.

Recentering step We use the assumption that the bias and noise are centered, that is, $\sum_{i \in [d] j \in [n]} \mathbb{E}[b_{ij}] = 0$ and $\sum_{i \in [d], j \in [n]} \mathbb{E}[z_{ij}] = 0$. Under this assumption, we have

$$\frac{1}{n} \sum_{i \in [d], j \in [n]} \mathbb{E}[y_{ij}] = \frac{1}{n} \sum_{i \in [d], j \in [n]} \mathbb{E}[x_i^* + b_{ij} + z_{ij}] = \sum_{i \in [d]} x_i^*. \tag{4.5}$$

Hence, we shift $\widehat{x}_{\text{rw}}$ by a constant such that the empirical version of (4.5) holds, that is, $\sum_{i \in d} [\widehat{x}_{\text{rw}}]_i = \frac{1}{n} \sum_{i \in [d], j \in [n]} y_{ij}$.

$$\widehat{x}_{\text{rw}} \leftarrow \widehat{x}_{\text{rw}} + \left(-\frac{1}{d} \sum_{i \in [d]} [\widehat{x}_{\text{rw}}]_i + \frac{1}{dn} \sum_{i \in [d], j \in [n]} y_{ij} \right) \mathbf{1} \tag{4.6}$$

This recentering step is necessary, because the expected mean of the bias over all courses after the reweighting step may not be 0, as the reweighting step only aligns the bias across courses, but not necessarily to 0. From (11.16b) in Lemma 11.4, our estimator also satisfies $\sum_{i \in [d]} \widehat{x}_i = \frac{1}{n} \sum_{i \in [d], j \in [n]} y_{ij}$ for all $\lambda \in [0, \infty]$, so this recentering also ensures a fair comparison with our estimator. Empirically we observe that the reweighted mean estimator always performs better after the recentering step.

Note that reweighted mean is undefined for total orderings. For group orderings with all constant fractions, reweighted mean is also consistent. In this case, we present a simple example below, where our estimator at $\lambda = 0$ still performs better than reweighted mean by a constant factor (uniform bias is assumed for analytical tractability).

Proposition 4.13. *Suppose the number of courses is $d = 2$. Suppose the number of groups is $r = 2$, with a grade distribution of $(\ell_{11}, \ell_{12}) = ((rn, (1-r)n)$ and $(\ell_{21}, \ell_{22}) = ((1-r)n, rn)$ for some $r \in (0,1)$. Suppose there is no noise. Suppose bias in group 1 is generated i.i.d. from Unif$[-1, 0]$, and bias in group 2 is generated i.i.d. from Unif$[0,1]$. Then the squared ℓ_2-risk for the reweighted mean estimator is $\widehat{x}_{\mathrm{rw}}$ and for our estimator $\widehat{x}^{(0)}$ at $\lambda = 0$ is respectively*

$$\frac{1}{2}\mathbb{E}\|\widehat{x}_{\mathrm{rw}} - x^*\|_2^2 = \frac{1}{24n} + \frac{1}{96r(1-r)n} \geq \frac{1}{12n}$$

$$\frac{1}{2}\mathbb{E}\|\widehat{x}^{(0)} - x^*\|_2^2 = \frac{1}{24n} + O\left(\frac{1}{n^2}\right).$$

The proof of this result is provided in Section 11.8. Note that the risk of our estimator is at most half of the error of reweighted mean, if ignoring the higher-order term $O\left(\frac{1}{n^2}\right)$.

4.7 Additional experimental details

In this section, we provide additional details for the experiments in Section 4.5.

4.7.1 Implementation

We now discuss the implementation of our estimator.

Solving the optimization (Line 10 in Algorithm 2): We describe the implementation of solving the optimization (4.2) depending on the value of λ.

- $\boldsymbol{\lambda = \infty}$**:** The estimator is computed as taking the mean of each course according to Proposition 4.7.
- $\boldsymbol{\lambda \in (0, \infty)}$**:** In the proof of Proposition 11.1 we show that the objective 4.1 is strictly convex in (x, B) on a convex domain. Hence, the problem is a QP with a unique solution. We solve for the QP using the CVXPY package.
- $\boldsymbol{\lambda = 0}$**:** It can be shown that the objective (4.1) is still convex, but there may exist multiple solutions before the tie-breaking. We first obtain one solution of the QP using CVXPY, denoted (x_0, b_0). The optimization (4.2) only has the first term, which is an ℓ_2-projection from y to the convex domain $\{x\mathbf{1}^T + b : x \in \mathbb{R}^d, b \in \mathbb{R}^{d\times n}, b \text{ satisfies } \mathcal{O}\}$. Hence, the value of $(x\mathbf{1}^T + b)$ is unique among all solutions (x, b), and the set of solutions can be written as $\{(x, b) : x = x_0 + u, b = b_0 - u\mathbf{1}^T, u \in \mathbb{R}^d\}$. We implement the tie-breaking by solving u using CVXPY, minimizing $\|b\|_F^2 = \|b_0 - u\mathbf{1}^T\|_F^2$ subject to the ordering constraints on $b = b_0 - u\mathbf{1}^T$.

Finally, we discuss a speed-up technique for solving the QP. For total orderings, the number of constraints in $\mathcal{O}$ is linear in the number of samples, whereas for general group orderings, the number of constraints in $\mathcal{O}$ can become quadratic, making the QP solver slow. To speed up the optimization, it can be shown that for all elements within any course and any group, the ordering of the estimated bias $\widehat{B}$ at these elements is the same as the ordering of the observations

Y at these elements. Therefore, among the constraints in $\mathcal{O}$ involving these elements, we only keep the constraints that involve the maximum and the minimum elements in this course and this group. Then we add the ordering of Y at these elements to the partial ordering $\mathcal{O}$. This replacement reduces the number of constraints in $\mathcal{O}$ and speeds up the QP solver.

Sampling a total ordering from the partial ordering $\mathcal{O}$ (Line 2 in Algorithm 2): When $\mathcal{O}$ is a group ordering, sampling a total ordering uniformly at random is implemented by first sorting the elements according to their group, and then permuting the them uniformly at random within each group.

When $\mathcal{O}$ is a tree or a group tree, we sample a total ordering using the following procedure. We first take all elements at the root of the tree, and place them in the total ordering as the lowest-ranked elements (if there are multiple elements at the root, then permute them uniformly at random in the total ordering). Consider each sub-tree consisting of a child node of the root and all its descendants. For the remaining positions in the total ordering, we assign these positions to the sub-trees uniformly at random. Then we proceed recursively to sample a total ordering for each sub-tree, and fill them back to their positions in the total ordering.

Interpolation (Line 15 in Algorithm 2): We sample 100 total orderings to approximate the interpolation.

4.7.2 Extending the reweighted mean estimator to tree orderings

We introduce the definitions of the two reweighted mean estimators on tree orderings used in the simulation in Section 4.5.4. Note that the reweighted mean estimator defined in Section 4.6.2 is with respect to the groups $\{Gk\}_{k\in[r]}$. We replace the groups in the reweighted mean estimator by the following two partitions of the elements.

Reweighted mean (node): Each subset in the partition consists of all elements in the same node of the tree.

Reweighted mean (level): Each subset in the partition consists of all elements on the same level of the tree.

4.7.3 Extending our estimator and the reweighted mean estimator to an unequal number of students per course

In the semi-synthetic experiment in Section 4.5.5, the number of students is unequal in different courses. We describe a natural extension of the reweighted mean estimator and our estimator to this case.

First, we explain how to format the observations back to a matrix form. Denote n_i as the number of students in course $i \in [d]$. Let $n = \max_{i\in[d]} n_i$. Construct a matrix $Y \in \mathbb{R}^{d\times n}$, where the first n_i elements in each row $i \in [d]$ correspond to the observations in this course, and the values of the remaining elements are set arbitrarily. Construct the set of observations $\Omega \in [d] \times [n]$, where the first n_i elements in each row $i \in [d]$ are in Ω. Estimation under an unequal number of students per course is equivalent to estimation given Y (and its corresponding partial ordering $\mathcal{O}$) restricted to the set Ω. It remains to define the reweighted mean estimator and our estimator restricted to any set $\Omega \in [d] \times [n]$.

The reweighted mean estimator: In the definition of the the reweighted mean estimator in Section 4.6.2, the reweighting step is the same (only using the observations in Ω). The recentering step restricted to Ω is defined as:

$$\widehat{x}_{\text{rw}} \leftarrow \widehat{x}_{\text{rw}} + \left(-\sum_{i\in[d]} \frac{n_i}{|\Omega|} [\widehat{x}_{\text{rw}}]_i + \frac{1}{|\Omega|} \sum_{i\in[d], j\in[n]} y_{ij} \right) \mathbf{1}$$

Similar to Section 4.6.2, after this recentering step, the reweighted mean estimator satisfies the empirical version of an equality (Eq. (11.15b) in Section 11.2.1) that our estimator also satisfies.

Our estimator: We extend Algorithm 2 naturally to being restricted to a set Ω as follows. In the data-splitting step, in Line 2, we replace the number of elements from dn to $\sum_{i\in[d]} n_i$; in Lines 4-7, we replace the number of students from n to n_i, and only find the sub-ordering of the n_i elements in Ω. The validation step remains the same.

4.8 Discussion

Evaluations given by participants in various applications are often spuriously biased by the evaluations received by the participant. We formulate the problem of correcting such outcome-induced bias, and propose an estimator and a cross-validation algorithm to address it. The cross-validation algorithm adapts to data without prior knowledge of the relative extents of bias and noise. Access to any such prior knowledge can be challenging in practice, and hence not requiring such prior knowledge provides our approach more flexibility.

Open problems There are a number of open questions of interest resulting out of this work. An interesting and important set of open questions pertains to extending our theoretical analysis of our estimator and cross-validation algorithm to more general settings: in the regime where there is both bias and noise, under other types of partial orderings, in a non-asymptotic regime, and in a high-dimensional regime with $d \gg n$. In addition, while our work aims to correct biases that already exist in the data, it is also helpful to mitigate such biases during data elicitation itself. This may be done from a mechanism design perspective where we align the users with proper incentives to report unbiased data, or from a user-experience perspective where we design multitude of questions that jointly reveal the nature of any bias.

Limitations There are several caveats that need to be kept in mind when interpreting or using our work. First, our work only claims to address biases obeying the user-provided information such as biases associated with the grading practice of the instructor (which follow the ordering constraints), and does *not* address biases associated with aspects such as the demographics of the instructor (which may not align with the ordering constraints). Second, the user should be careful in supplying the appropriate ordering constraints to the algorithm, ensuring these constraints have been validated separately. Third, our theoretical guarantees hold under specific shape assumptions of the bias and the noise. Our algorithm is designed distribution-free, and we speculate similar guarantees to hold under other reasonable, well-behaved shape assumptions; however, formal guarantees under more general models remain open. Our algorithm consequently may be appropriate for use as an assistive tool along with other existing practices (e.g., sample mean) when making decisions, particularly in any high-stakes scenario. Aligned results between our algorithm and other practices give us more confidence that the result is correct; different results between our algorithm and other practices suggests need for additional information or deliberation before drawing a conclusion.

Acknowledgments

This work was supported in part by NSF CAREER award 1942124, NSF CIF-1763734, CCF-2007911 and DMS-2015447.

Part II

Estimation Bias

Chapter 5

Bias Reduction in Estimation from Pairwise Comparisons

A number of applications (e.g., AI bot tournaments, sports, peer grading, crowdsourcing) use pairwise comparison data and the Bradley-Terry-Luce (BTL) model to evaluate a given collection of items (e.g., bots, teams, students, search results). Past work has shown that under the BTL model, the widely-used maximum-likelihood estimator (MLE) is minimax-optimal in estimating the item parameters, in terms of the mean squared error. However, another important desideratum for designing estimators is fairness. In this work, we consider fairness modeled by the notion of bias in statistics. We show that the MLE incurs a suboptimal rate in terms of bias. We then propose a simple modification to the MLE, which "stretches" the bounding box of the maximum-likelihood optimizer by a small constant factor from the underlying ground truth domain. We show that this simple modification leads to an improved rate in bias, while maintaining minimax-optimality in the mean squared error. In this manner, our proposed class of estimators provably improves fairness represented by bias without loss in accuracy.

5.1 Introduction

A number of applications involve data in the form of pairwise comparisons among a collection of items, and entail an evaluation of the individual items from this data. An application gaining increasing popularity is competition between pairs of AI bots (e.g., [131]). Here a number of AI bots compete with each other in pairwise matchups for a certain task, where each bot plays every other bot a certain number of times in a *round robin* fashion, with the goal of evaluating the quality of each bot. A second example is the evaluation of self-play of AI algorithms in their training phase [164], where again, different copies of an AI bot play against each other a number of times. Applications involving humans include sports and online games such as the English Premier League of football [3, 98] (unofficial ratings) and official world rankings for chess (e.g., FIDE [1] and USCF [69] ratings). The influence of scientific journals has also been analyzed in this manner, where citations from one journal to another are modeled by pairwise comparisons [169].

A common method of evaluating the items based on pairwise comparisons is to assume that the probability of an item beating another equals the logistic function of the difference in the true quality of the two items, and then infer the true quality from the observed outcomes of the comparisons (e.g., the Elo rating system). Various applications employ such an approach to rating from pairwise comparisons, with some modifications tailored to that specific application. Our goal is not to study the application-specific versions, but the foundational underpinnings of such rating systems.

In this chapter, we study the pairwise-comparison model that underlies [5, 70] these rating systems, namely the Bradley-Terry-Luce (BTL) model [24, 109]. The BTL model assumes that each item is associated to an unknown real-valued parameter representing the quality of that item, and assumes that the probability of an item beating another is the logistic function applied to the difference of the parameters of these two items. The BTL model is also employed in the applications of peer grading [104, 155] (where the grades of the students are set as the BTL parameters to be estimated), crowdsourcing [37, 137], and understanding consumer choice in marketing [74].

5.1.1 BTL model and maximum likelihood estimation

Now we present a formal definition of the BTL model. Let $d \geq 2$ denote the number of items. The d items are associated to an unknown parameter vector $\theta^* \in \mathbb{R}^d$ whose i^{th} entry represents the underlying quality of item $i \in [d]$. When any item $i \in [d]$ is compared with any item $j \in [d]$ in the BTL model, the item i beats item j with probability

$$\frac{1}{1 + e^{-(\theta_i^* - \theta_j^*)}}, \tag{5.1}$$

independent of all other comparisons. The probability of item j beating i is one minus the expression (5.1) above. We consider the "league format" [5] of comparisons where every pair of items is compared k times.

We follow the usual assumption [79, 156] under the BTL model that the true parameter vector θ^* lies in the set Θ_B parameterized by a *constant* $B > 0$ and satisfy:

$$\Theta_B = \{\theta \in \mathbb{R}^d \mid \|\theta\|_\infty \leq B \text{ and } \sum_{i=1}^d \theta_i = 0\}. \tag{5.2}$$

The first constraint requires that the magnitude of the parameters is bounded by some constant B. We call this constraint the "box constraint". A box constraint is necessary, because otherwise the estimation error can diverge to infinity [156, Appendix G]. The second constraint requires the parameters to sum to 0. This is without loss of generality due to the shift-invariance property of the BTL model.

A large amount of both theoretical [79, 89, 127, 156, 173] and applied [37, 137, 163, 169] literature focuses on the goal of estimating the parameter vector θ^* of the BTL model. A standard and widely-studied estimator is the maximum-likelihood estimator (MLE):

$$\widehat{\theta}^{(B)} = \operatorname*{argmin}_{\theta \in \Theta_B} \ell(\theta), \tag{5.3}$$

where ℓ is the negative log-likelihood function. Letting W_{ij} denote a random variable representing the number of times that item $i \in [d]$ beats item $j \in [d]$, the log-likelihood function ℓ is given by:

$$\ell(\theta) := \ell(\{W_{ij}\}; \theta) = -\sum_{1 \leq i < j \leq d} \left[W_{ij} \log\left(\frac{1}{1+e^{-(\theta_i - \theta_j)}}\right) + W_{ji} \log\left(\frac{1}{1+e^{-(\theta_j - \theta_i)}}\right) \right].$$

5.1.2 Metrics

Accuracy. A common metric used in the literature on estimating the BTL model is the *accuracy* of the estimate, measured in terms of the mean squared error. Formally, the accuracy of any estimator $\widehat{\theta}$ is defined as:

$$\alpha(\widehat{\theta}) := \sup_{\theta^* \in \Theta_B} \mathbb{E}[\|\widehat{\theta} - \theta^*\|_2^2].$$

Importantly, past work [79, 156] has shown that the MLE (5.3) has the appealing property of being minimax-optimal in terms of the accuracy.

Bias. Another important desideratum for designing and evaluating estimators is fairness. For example, in sports or online games, we do not want to assign scores in such a way that it systematically gives certain players higher scores than their true quality, but at the same time gives certain other players lower scores than their true quality. In this chapter, we use the standard definition of *bias* in statistics as the notion of fairness. For any estimator, the bias incurred by this estimator on a parameter is defined as the difference between the expected value of the estimator and the true value of the parameter. Since our parameters are a vector, we consider the worst-case bias, that is, the maximum magnitude of the bias across all items. Formally, the bias of any estimator $\widehat{\theta}$ is defined as:

$$\beta(\widehat{\theta}) := \sup_{\theta^* \in \Theta_B} \|\mathbb{E}[\widehat{\theta}] - \theta^*\|_\infty.$$

With this background, we now provide an overview of the contributions of this chapter.

5.1.3 Contribution I: performance of MLE

Our first contribution is to analyze the widely-used MLE (5.3) in terms of its bias. Let us begin with a visual illustration through simulation. Consider $d = 25$ items with parameter values equally spaced in the interval $[-1, 1]$, where $k = 5$ pairwise comparisons are observed between each pair of items under the BTL model. We estimate the parameters using the MLE, and plot the bias on each item across 5000 iterations of the simulation in Figure 5.1 (striped red). The MLE shows a systematic bias: it induces a negative bias (under-estimation) on the large positive parameters, and a positive bias (over-estimation) on the large negative parameters. In the applications of interest, the MLE thus systematically underestimates the abilities of the top players/students/items and overestimates the abilities of those at the bottom.

In this chapter, we theoretically quantify the bias incurred by the MLE.

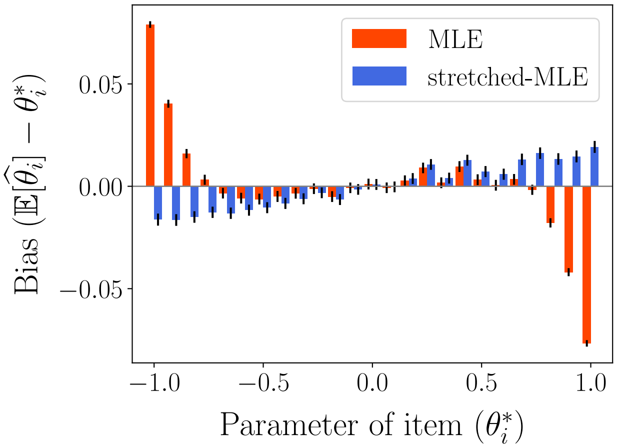


Figure 5.1: Biases on items of different parameters, induced by the MLE and our stretched-MLE (with $A = 2$). Our estimator significantly reduces the maximum magnitude of the bias across the items. Note that this figure plots the bias including its sign: A positive bias means over-estimation of the parameter, and a negative bias means under-estimation of the parameter. Each bar is a mean over 5000 iterations.

Theorem 5.1 (MLE bias lower bound; Informal)**.** *The MLE* (5.3) *incurs a bias* $\beta(\widehat{\theta}^{(B)})$ *lower bounded as* $\Omega(\frac{1}{\sqrt{dk}})$.

As shown by our results to follow, this bias is suboptimal. Our proof for this result indicates that the bias is incurred because the MLE operates under the accurately specified model with the box constraint at B. That is, the MLE "clips" the estimate to lie within the set Θ_B. This issue is visible in the simulation of Figure 5.1 where the bias is the largest when the true values of the parameters are near the boundaries $\pm B$. For example, consider a true parameter whose value equals B. The estimate of this parameter sometimes equals the largest allowed value B (due to the box constraint), and sometimes is smaller than B (due to the randomness of the data). Therefore, in expectation, the estimate of this parameter incurs a negative bias. An analogous argument explains the positive bias when the true parameter equals or is close to $-B$.

5.1.4 Contribution II: proposed stretched estimator and its theoretical guarantees

Our goal is to design an estimator with a lower bias while maintaining high accuracy. Since the MLE (5.3) is already widely studied and used, it is also desirable from a practical and computational standpoint that the new estimator is a simple modification of the MLE (5.3). With this motivation in mind, an intuitive approach is to consider the MLE but without the box constraint "$\|\theta\|_\infty \leq B$". We call the estimator without the box constraint as the "unconstrained MLE", and denote it by $\widehat{\theta}^{(\infty)}$, because removing the box constraint is equivalent to setting the box constraint to ∞:

$$\widehat{\theta}^{(\infty)} = \operatorname*{argmin}_{\theta \in \Theta_\infty} \ell(\theta), \tag{5.4}$$

Estimator	Bias	Mean squared error
Standard MLE $\widehat{\theta}^{(B)}$	$\Omega(\frac{1}{\sqrt{dk}})$ (Thm. 5.4(a))	$\mathcal{O}(\frac{1}{k})$ minimax-optimal [79, 156]
Unconstrained MLE $\widehat{\theta}^{(\infty)}$	Undefined	∞
Stretched MLE $\widehat{\theta}^{(A)}$	$\widetilde{\mathcal{O}}(\frac{1}{dk})$ (Thm. 5.4(b))	$\mathcal{O}(\frac{1}{k})$ minimax-optimal (Thm. 5.5(b))

Table 5.1: Theoretical guarantees for the MLE $\widehat{\theta}^{(B)}$, unconstrained MLE $\widehat{\theta}^{(\infty)}$ and the proposed stretched-MLE $\widehat{\theta}^{(A)}$ (with a constant A such that $A > B$). The proposed stretched-MLE achieves a better rate on bias, while retaining minimax optimality in terms of accuracy. Recall that d denotes the number of items and k denotes the number of comparisons per pair.

where $\Theta_\infty := \{\theta \in \mathbb{R}^d \mid \sum_{i=1}^d \theta_i = 0\}$. The unconstrained MLE $\widehat{\theta}^{(\infty)}$ incurs an unbounded error in terms of accuracy. This is because with non-zero probability an item beats all others, in which case the unconstrained MLE estimates the parameter of this item as ∞, thereby inducing an unbounded mean squared error.

Consequently, in this work, we propose the following simple modification to the MLE which is a middle ground between the standard MLE (5.3) and the unconstrained MLE. Specifically, we consider a “stretched-MLE”, which is associated to a parameter A such that $A > B$. Given the parameter A, the stretched-MLE is identical to (5.3) but “stretches” the box constraint to A:

$$\widehat{\theta}^{(A)} = \operatorname*{argmin}_{\theta \in \Theta_A} \ell(\theta), \tag{5.5}$$

where $\Theta_A := \{\theta \in \mathbb{R}^d \mid \|\theta\|_\infty \leq A \text{ and } \sum_{i=1}^d \theta_i = 0\}$. That is, Θ_A simply replaces the box constraint $\|\theta\|_\infty \leq B$ in (5.2) by the “stretched” box constraint $\|\theta\|_\infty \leq A$.

The bias induced by the stretched-MLE (with $A = 2$) in the previous experiment is also shown in Figure 5.1 (solid blue). Observe that the maximum bias (incurred at the leftmost item with the largest negative parameter, or the rightmost item with the largest positive parameter) is significantly reduced compared to the MLE. Moreover, the bias induced by the stretched-MLE looks qualitatively more evened out across the items.

Our second main theoretical result proves that the stretched-MLE indeed incurs a significantly lower bias.

Theorem 5.2 (Stretched-MLE bias upper bound; Informal)**.** *The stretched-MLE* (5.5) *with* $A = 2$ *incurs a bias* $\beta(\widehat{\theta}^{(A)})$ *upper bounded as* $\widetilde{\mathcal{O}}(\frac{1}{dk})$.

Given the significant bias reduction by our estimator, a natural question is about the accuracy of the stretched-MLE, particularly given the unbounded error incurred by the unconstrained MLE. We prove that our stretched-MLE is able to maintain the same minimax-optimal rate on the mean squared error as the standard MLE.

Theorem 5.3 (Stretched-MLE accuracy upper bound; Informal)**.** *The stretched-MLE* (5.5) *with* $A = 2$ *incurs a mean squared error* $\alpha(\widehat{\theta}^{(A)})$ *upper bounded as* $\mathcal{O}(\frac{1}{k})$*, which is minimax-optimal.*

This result shows **a win-win** by our stretched-MLE: *reducing the bias while retaining the accuracy guarantee*. The comparison of the MLE and the stretched-MLE in terms of accuracy and bias is summarized in Table 5.1. Another attractive feature of our result is that the proposed stretched-MLE is a simple modification of the standard MLE, which can easily be incorporated in any existing implementation. It is important to note that while our modification to the estimator is simple to implement, our theoretical analyses and the proofs are non-trivial.

5.1.5 Related work

The logistic nature (5.1) of the BTL model relates our work to studies of logistic regression (e.g., [58, 83, 138, 172]), among which the paper [172] is the most closely related to ours. The paper [172] considers an unconstrained MLE in logistic regression, and shows its bias in the opposite direction as compared to our results on the standard MLE (constrained) in the BTL model. Specifically, the paper [172] shows that the large positive coefficients are overestimated, and the large negative coefficients are underestimated. There are several additional key differences between the results in [172] as compared to the present chapter. The paper [172] studies the asymptotic bias of the unconstrained MLE, showing that the unconstrained MLE is not consistent. On the other hand, we operate in a regime where the MLE is still consistent, and study finite-sample bounds. Moreover, the paper [172] assumes that the predictor variables are i.i.d. Gaussian. On the other hand, in the BTL model the probability that item i beats item j can be written as $\frac{1}{1+e^{-x_{ij}^T\theta^*}}$, where each predictor variable $x_{ij} \in \mathbb{R}^d$ has entry i equal to 1, entry j equal to -1, and the remaining entries equal to 0.

A common way to achieve bias reduction is to employ finite-sample correction, such as Jackknife [140] and other methods [8, 48, 60] to the MLE (or other estimators). These methods operate in a low-dimensional regime (small d) where the MLE is asymptotically unbiased. Informally, these methods use a Taylor expansion and write the expression for the bias as an infinite sum $\frac{f_1(\theta^*)}{n} + \frac{f_2(\theta^*)}{n^2} + \ldots$, where n is the number samples, for some functions $f_1, f_2, \ldots$. These works then modify the estimator in a variety of ways to eliminate the lower-order terms in this bias expression. However, since the expression is an infinite sum, eliminating the first term does not guarantee a low rate of the bias. Moreover, since the functions f_i are implicit functions of θ^*, eliminating lower-order terms does not directly translate to explicit worst-case guarantees.

Returning to the pairwise-comparison setting, in addition to the mean squared error, some past work has also considered accuracy in terms of the ℓ_1 norm error [4] and the ℓ_∞ norm error [40, 41, 92]. The ℓ_∞ bound for a regularized MLE is analyzed in [41]. Our proof for bounding the bias of the standard MLE (unregularized) relies on a high-probability ℓ_∞ bound for the unconstrained MLE (unregularized). It is important to note that the bound for regularized MLE from [41] does not carry to unregularized MLE, because the proof from [41] relies on the strong convexity of the regularizer. On the other hand, *our intermediate result provides a partial answer to the open question in [41]* about the ℓ_∞ norm for the unregularized MLE (Lemma 12.5 in Section 12.1): We establish an ℓ_∞ bound for unregularized MLE when $p_{obs} = 1$, which has the same rate as that of the regularized MLE in [41].

Another common occurrence of bias is the phenomenon of regression towards the mean [170]. Regression towards the mean refers to the phenomenon that random variables taking large (or small) values in one measurement are likely to take more moderate (closer to average) values in subsequent measurements. On the contrary, we consider items whose indices are fixed (and are not order statistics). For fixed indices, our results suggest that under the BTL model, the bias (under-estimation of large true values) is in the opposite direction as that in regression towards the mean (over-estimation of large observed values).

Finally, the paper [98] models the notion of fairness in Elo ratings in terms of the "variance", where an estimator is considered fair if the estimator is not much affected by the underlying randomness of the pairwise-comparison outcomes. The paper [98] empirically evaluates this notion of fairness on the English Premier League data, but presents no theoretical results.

5.2 Main results

In this section, we formally provide our main theoretical results on bias and on the mean squared error.

5.2.1 Bias

Recall that d denotes the number of items and k denotes the number of comparisons per pair of items. The true parameter vector is $\theta^* \in \Theta_B$ for some pre-specified constant $B > 0$. The following theorem provides bounds on the bias of the standard MLE $\widehat{\theta}^{(B)}$ and that of our stretched-MLE $\widehat{\theta}^{(A)}$ with parameter A. In particular, it shows that if A is a finite constant strictly greater than B, then our stretched-MLE has a much smaller bias than the MLE when d and k are sufficiently large.

Theorem 5.4. *(a) There exists a constant $c > 0$ that depends only on the constant B, such that*

$$\beta(\widehat{\theta}^{(B)}) \geq \frac{c}{\sqrt{dk}}, \tag{5.6a}$$

for all $d \geq d_0$ and all $k \geq k_0$, where d_0 and k_0 are constants that depend only on the constant B.

(b) Let A be any finite constant such that $A > B$. There exists a constant $c > 0$ that depends only on the constants A and B, such that

$$\beta(\widehat{\theta}^{(A)}) \leq c\frac{\log d + \log k}{dk}, \tag{5.6b}$$

for all $d \geq d_0$ and all $k \geq k_0$, where d_0 and k_0 are constants that depend only on the constants A and B.

We note that in Theorem 5.4(b), we allow A to be any positive constant as long as $A > B$. Therefore, the difference between A and B can be any arbitrarily small constant. It is perhaps surprising that stretching the box constraint only by a small constant yields such a significant improvement in the bias. We provide intuition behind this result in Section 5.2.1.

We devote the remainder of this section to providing a sketch of the proof of Theorem 5.4. We first prove Theorem 5.4(b) and then Theorem 5.4(a), because the proof of Theorem 5.4(a) depends on the proof of Theorem 5.4(b). The complete proof is provided in Section 12.1.

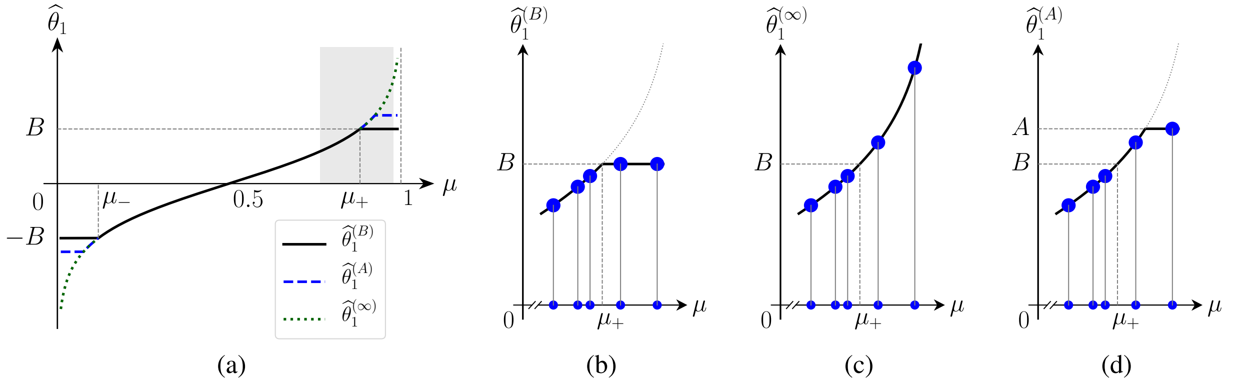


Figure 5.2: Intuition on the sources of bias. (a) The estimators standard MLE $\widehat{\theta}^{(B)}$, stretched-MLE $\widehat{\theta}^{(A)}$ and unconstrained MLE $\widehat{\theta}^{(\infty)}$ (on item 1), as a function of μ when there are $d = 2$ items. We consider $\theta^* = [B, -B]$, under which the true probability that item 1 beats item 2 is μ_+. We zoom in to the region around $\mu = \mu_+$ indicated by the grey box. (b) The standard MLE $\widehat{\theta}^{(B)}$ incurs a negative bias, because the estimate is required to be at most B. (c) The unconstrained MLE $\widehat{\theta}^{(\infty)}$ incurs a positive bias by Jensen's inequality, because the estimator function is convex on $\mu \in (0.5, 1)$. (d) Our estimator balances out the negative bias and the positive bias.

For Theorem 5.4(b), we first analyze the unconstrained MLE $\widehat{\theta}^{(\infty)}$. By plugging $\widehat{\theta}^{(\infty)}$ into the first-order optimality condition of the negative log-likelihood function and using concentration on the comparison outcomes, we prove an ℓ_∞ bound of the form $\|\widehat{\theta}^{(\infty)} - \theta^*\|_\infty = \widetilde{\mathcal{O}}(\frac{1}{\sqrt{dk}})$ with sufficiently high probability (which partially resolves the open problem from [41], in the regime where $p_{obs} = 1$). Next, using a second-order mean value theorem on the first-order optimality condition and taking an expectation, we show a result of the form $\|\mathbb{E}[\widehat{\theta}^{(\infty)} \mid \mathcal{E}] - \theta^*\|_\infty \approx \|\widehat{\theta}^{(\infty)} - \theta^*\|_\infty^2 = \widetilde{\mathcal{O}}(\frac{1}{dk})$, where $\mathcal{E}$ is some high-probability event (recall from Table 5.1 that for unconstrained MLE, the bias $\|\mathbb{E}[\widehat{\theta}^{(\infty)}] - \theta^*\|_\infty$ without conditioning on $\mathcal{E}$ is undefined). Finally, we show that the unconstrained MLE $\widehat{\theta}^{(\infty)}$ and the stretched-MLE $\widehat{\theta}^{(A)}$ are identical with high probability for sufficiently large d and k, and perform some algebraic manipulations to finally arrive at the claim (5.6b).

For Theorem 5.4(a), we first prove a bound on the order of $\frac{1}{\sqrt{d}}$ when there are $d = 2$ items. Then for general d, we consider the bias on item 1 under the true parameter vector $\theta^* = [B, -\frac{B}{d-1}, \ldots, -\frac{B}{d-1}]$. We construct an "oracle" MLE, such that analyzing the bias of the "oracle" MLE can be reduced to the proof of the 2-item case, and thereby prove a bias on the order of $\frac{1}{\sqrt{dk}}$ for the oracle MLE. Finally, we show that the difference between the oracle MLE and the standard MLE is small, by repeating arguments from the proof of Theorem 5.4(b).

Intuition for Theorem 5.4

In this section, we provide intuition why stretching the box constraint from B to A significantly reduces the bias. Specifically, we consider a simplified setting with $d = 2$ items. Due to the centering constraint, we have $\theta_2^* = -\theta_1^*$ for the true parameters, and we have $\widehat{\theta}_2 = -\widehat{\theta}_1$ for any estimator $\widehat{\theta}$ that satisfies the centering constraint. Therefore, it suffices to focus only on item 1.

Denote μ as the random variable representing the fraction of times that item 1 beats item 2, and denote the true probability that item 1 beats item 2 as $\mu^* := \frac{1}{1+e^{-(\theta_1^*-\theta_2^*)}}$. We consider the true parameter of item 1 as $\theta_1^* \in [-B, B]$. Then we have $\mu^* \in [\mu_-, \mu_+]$, where $\mu_- = \frac{1}{1+e^{2B}}$ and $\mu_+ = \frac{1}{1+e^{-2B}}$. The standard MLE $\widehat{\theta}^{(B)}$, the stretched-MLE $\widehat{\theta}^{(A)}$ and the unconstrained MLE $\widehat{\theta}^{(\infty)}$ can be solved in closed form:

$$\widehat{\theta}_1^{(B)}(\mu) = \begin{cases} -B & \text{if } \mu \in [0, \mu_-] \\ -\frac{1}{2}\log\left(\frac{1}{\mu}-1\right) & \text{if } \mu \in (\mu_-, \mu_+) \\ B & \text{if } \mu \in [\mu_+, 1]. \end{cases}$$

$$\widehat{\theta}_1^{(A)}(\mu) = \begin{cases} -A & \text{if } \mu \in \left[0, \frac{1}{1+e^{2A}}\right] \\ -\frac{1}{2}\log\left(\frac{1}{\mu}-1\right) & \text{if } \mu \in \left(\frac{1}{1+e^{2A}}, \frac{1}{1+e^{-2A}}\right) \\ A & \text{if } \mu \in \left[\frac{1}{1+e^{-2A}}, 1\right]. \end{cases}$$

$$\widehat{\theta}_1^{(\infty)}(\mu) = -\frac{1}{2}\log\left(\frac{1}{\mu}-1\right).$$

See Fig. 5.2a for a comparison of these three estimators.

Now we consider the bias incurred by these three estimators. For intuition, let us consider the case $\theta_1^* = B$, which incurs the largest bias in our simulation of Fig. 5.1. If the observation μ were noiseless (and thus equals the true probability μ_+), then all three estimators would output the true parameter B. However, the observation μ is noisy, and only concentrates around μ_+. To investigate how these three estimators behave differently under this noise, we zoom in to the region around $\mu = \mu_+$ indicated by the grey box in Fig. 5.2a. (Note that the observation μ can lie outside the grey box, but for intuition we ignore this low-probability event due to concentration.)

The behaviors of the three estimators in the grey box are shown in Fig. 5.2b, Fig. 5.2c and Fig. 5.2d, respectively. For each of these estimators, the blue dots on the x-axis denotes the noisy observation of μ across different iterations, and the blue dots on the estimator function denotes the corresponding noisy estimates. The expected value of the estimator is a mean over the blue dots on the estimator function. For the standard MLE $\widehat{\theta}^{(B)}$ (Fig. 5.2b), the box constraint requires that the estimate shall never exceed B. We call this phenomenon the "clipping" effect, which introduces a negative bias. For the unconstrained MLE $\widehat{\theta}^{(\infty)}$ (Fig. 5.2c), since the estimator function is convex, by Jensen's inequality, the unconstrained MLE $\widehat{\theta}^{(\infty)}$ introduces a positive bias. Our proposed stretched-MLE $\widehat{\theta}^{(A)}$ (Fig. 5.2d) lies in the middle between the standard MLE and the unconstrained MLE. Therefore, the stretched-MLE balances out the negative bias from the "clipping" effect and the positive bias from the convexity of the estimator function, thereby yielding a smaller bias on the item parameter. In practice, one can numerically tune the parameter A to minimize the bias across all possible parameter vector $\theta^* \in \Theta_B$. Simulation results on different values of A are included in Section 5.3.

5.2.2 Accuracy

Given the result of Theorem 5.4 on the bias reduction of the estimator $\widehat{\theta}^{(A)}$, we revisit the mean squared error. Past work [79, 156] has shown that the standard MLE $\widehat{\theta}^{(B)}$ is minimax-optimal in terms of the mean squared error. The following theorem shows that this minimax-optimality also holds for our proposed stretched-MLE $\widehat{\theta}^{(A)}$, where A is any constant such that $A > B$. The theorem statement and its proof follows Theorem 2 from [156], after some modification to accommodate the bounding box parameter A.

Theorem 5.5. *(a) [Theorem 2(a) from [156]] There exists a constant $c > 0$ that depends only on the constant B, such that any estimator $\widehat{\theta}$ has a mean squared error lower bounded as*

$$\alpha(\widehat{\theta}) \geq \frac{c}{k}, \tag{5.7a}$$

for all $k \geq k_0$, where k_0 is a constant that depends only on the constant B.

(b) Let A be any finite constant such that $A > B$. There exists a constant $c > 0$ that depends only on the constants A and B, such that

$$\alpha(\widehat{\theta}^{(A)}) \leq \frac{c}{k}. \tag{5.7b}$$

Theorem 5.5 shows that using the estimator $\widehat{\theta}^{(A)}$ retains the minimax-optimality achieved by $\widehat{\theta}^{(B)}$ in terms of the mean squared error. Combining Theorem 5.4 and Theorem 5.5 shows the Pareto improvement of our estimator $\widehat{\theta}^{(A)}$: the estimator $\widehat{\theta}^{(A)}$ decreases the rate of the bias, while still performing optimally on the mean squared error.

The proof of Theorem 5.5 closely mimics the proof of Theorem 2(b) from [156], replacing the steps involving the domain Θ_B by the stretched domain Θ_A. The details are provided in Section 12.2.

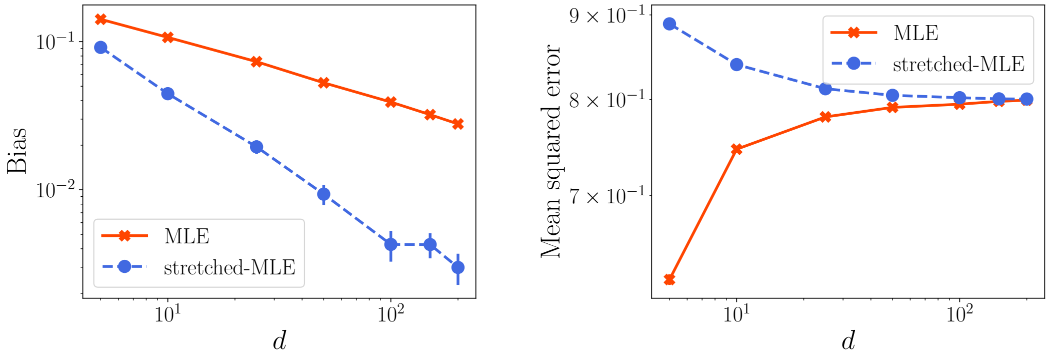
5.3 Simulations

In this section, we explore our problem space and compare the standard MLE and our proposed stretched-MLE by simulations. In what follows, we set $B = 1$, and unless specified otherwise we set $A = 2$ and $\theta^* = [1, -\frac{1}{d-1}, -\frac{1}{d-1}, \ldots, -\frac{1}{d-1}]$. We also evaluate the performance of other values of θ^* subsequently. Error bars in all the plots represent the standard error of the mean.

(i) **Dependence on d:** We vary the number of items d, while fixing $k = 5$. The results are shown in Fig. 5.3. Observe that the stretched-MLE has a significantly smaller bias, and performs on par with the MLE in terms of the mean squared error when d is large. Moreover, the simulations also suggest the rate of bias as of order $\frac{1}{\sqrt{d}}$ for the MLE and $\frac{1}{d}$ for the stretched-MLE, as predicted by our theoretical results.

(ii) **Dependence on k:** We vary the number of comparisons k per pair of items, while fixing $d = 10$. The results are shown in Fig. 5.4. As in the simulation (i) with varying d, we observe that the stretched-MLE has a significantly smaller bias, and performs on par with the MLE in terms of the mean squared error. Moreover, the simulations also suggest the rate of bias as of order $\frac{1}{\sqrt{k}}$ for the MLE and $\frac{1}{k}$ for the stretched-MLE, as predicted by our theoretical results.

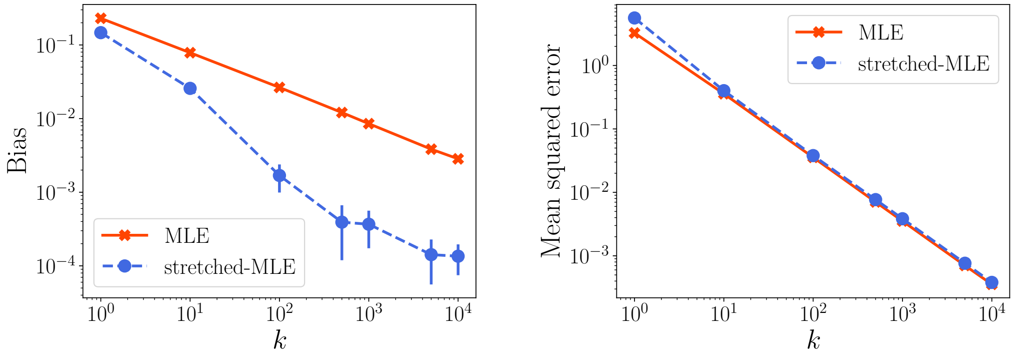
(iii) **Different values of A:** In our theoretical analysis, we proved bounds that hold for all constant A such that $A > B$. In this simulation, we empirically compare the performance



(a) Bias

(b) Mean squared error

Figure 5.3: Performance of estimators for various values of d, with $k = 5$ and $A = 2$. Each point is a mean over 10000 iterations.



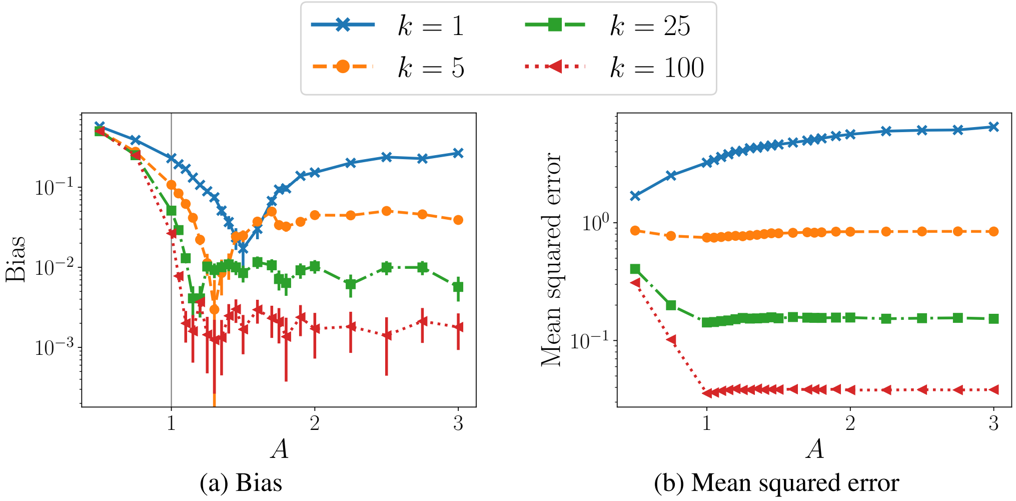
(a) Bias

(b) Mean squared error

Figure 5.4: Performance of estimators for various values of k, with $d = 10$ and $A = 2$. Each point is a mean over 10000 iterations.

of the stretched-MLE for different values of A (note that setting $A = 1$ is equivalent to the standard MLE). We fix $d = 10$, varying $A \in [0.5, 3]$ and k from 1 to 100. The results are shown in Fig. 5.5. For the bias, we observe that the bias keeps decreasing in the range of $A \in [0.5, 1]$. This is because as we increase A to 1, the negative bias introduced by the "clipping" effect is reduced. The optimal value of A for all settings of k is always greater than 1. Moreover, the optimal A seems to be closer to 1 when we increase k. This agrees with the intuition in Section 5.2.1. When k is larger, the estimate becomes more concentrated around the true parameter. Then the "clipping" effect becomes smaller and can be accommodated by a smaller A. The mean squared error is insensitive to the choice of A as long as $A \geq 1$.

(iv) **Different settings of the true parameter θ^*:** Our theoretical result considers the worst-case bias and accuracy. In this simulation, we empirically compare the performance of the stretched-MLE under different settings of the true parameter vector θ^* (again, recall that setting $A = 1$ is equivalent to the standard MLE). Specifically, we consider the following



(a) Bias

(b) Mean squared error

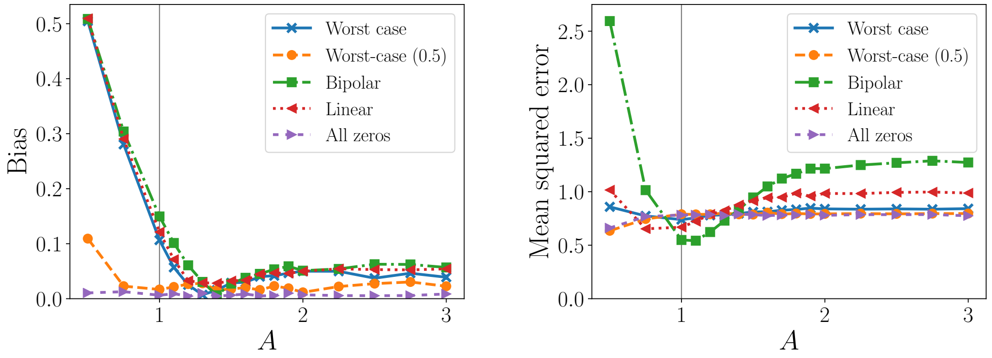
Figure 5.5: Performance of estimators for various values of A and k, with $d = 10$. Setting $A = 1$ is equivalent to the standard MLE. Each point is a mean over 5000 iterations.

values of θ^*:

- *Worst case:* $\theta^* = [1, -\frac{1}{d-1}, \ldots, -\frac{1}{d-1}]$.
- *Worst case (0.5):* $\theta^* = [0.5, -\frac{0.5}{d-1}, \ldots, -\frac{0.5}{d-1}]$.
- *Bipolar:* half of the values are 1, and the other half are -1.
- *Linear:* the values are equally spaced in the interval $[-1, 1]$.
- *All zeros:* all parameters are 0.

We fix $d = 10$ and $k = 5$, varying $A \in [0.5, 3]$ under different settings of the true parameter vector θ^*. The results are shown in Fig. 5.6. Two high-level takeaways from the empirical evaluations are that the bias generally reduces with an increase in A till past B, and that the mean squared error remains relatively constant beyond $A = 1$ in the plotted range. In more detail, for the bias, we observe that the performance primarily depends on the largest magnitude of the items (that is, $\|\theta^*\|_\infty$). For the settings *worst case, bipolar* and *linear* (where $\|\theta^*\|_\infty = 1$), the bias keeps decreasing when A is past $B = 1$. For the setting *worst-case (0.5)* (where $\|\theta^*\|_\infty = 0.5$), the bias keeps decreasing when A is past 0.5. This makes sense since in this case we effectively have $B = 0.5$ (although the algorithm would not know this in practice). The bias for the setting *all zeros* stays small across values of A. For the mean squared error, the increase when A is past 1 is relatively small under most of the settings of the true parameter vector θ^*. The *bipoloar* setting has the largest increase in the mean squared error. Under this setting, all parameters θ_i^* take values at the boundaries $\pm B$, and therefore the estimates of all parameters are affected by the box constraint.

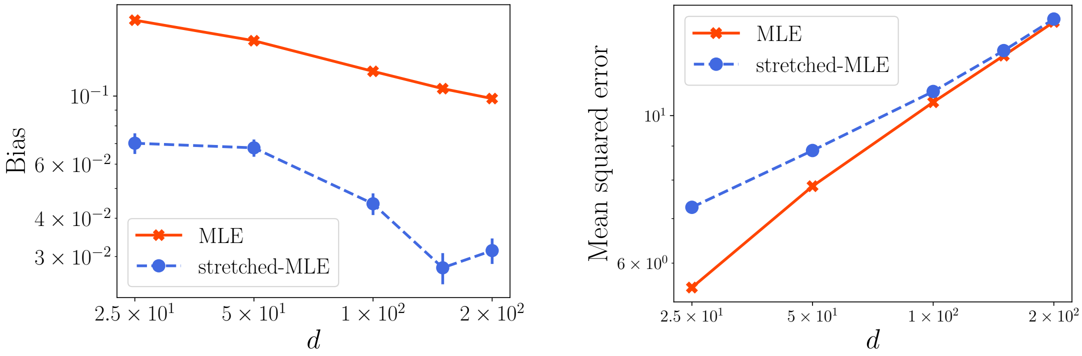
(v) **Sparse observations:** So far we have considered a league format where k comparisons are observed between any pair of items. Now we consider a random-design setup, where k comparisons are observed between any pair of items independently with probability $p_{obs} \in (0, 1)$, and none otherwise [41, 127]. In our simulations, we set $p_{obs} = \frac{1}{\sqrt{d}}$ and



(a) Bias

(b) Mean squared error

Figure 5.6: Performance of estimators for various values of A and various settings of θ^*, with $d = 10$ and $k = 5$. Setting $A = 1$ is equivalent to the standard MLE. Each point is a mean over 5000 iterations.



(a) Bias

(b) Mean squared error

Figure 5.7: Performance of estimators for various values of d under sparse observations, with $A = 2$. A number of $k = 5$ comparisons are observed between any pair independently with probability $p_{obs} = \frac{1}{\sqrt{d}}$ and none otherwise. Each point is a mean over 10000 iterations.

$k = 5$. We discard an iteration if the graph is not connected, since the problem is not identifiable under such a graph. The results are shown in Figure 5.7. We observe that the stretched-MLE continues to outperform MLE in terms of bias, and perform on par in terms of the mean squared error.

5.4 Conclusion and discussion

In this work, we show that the widely-used MLE is suboptimal in terms of bias, and propose a class of estimators called the "stretched-MLE", which provably reduces the bias while maintaining the minimax-optimality in terms of accuracy. These results on the performance of the MLE and the stretched-MLE are of both theoretical and practical interest. From the theoretical point of view, our analysis and proofs provide insights on the cause of the bias, explain why stretching the box alleviates this cause, and prove theoretical guarantees in bias reduction by stretching the box. Our results on the benefits of the stretched-MLE thus suggest theoreticians to consider the stretched-MLE for analysis instead of the standard MLE.

From the practical point of view, the constant B is often unknown, and practitioners oten estimate the value of B by fitting the data or from past experience. Our results thus suggest that one should estimate B leniently, as an estimation smaller than or equal to the true B causes significant bias. Moreover, our proposed estimator is a simple modification to the MLE, which can be incorporated into any existing implementation at ease.

Our results lead to several open problems. First, it is of interest to extend our theoretical analysis to settings where the observations are sparse. For example, one may consider a random-design setup, where k comparisons are observed between any pair independently with probability p_{obs} and none otherwise [41, 127] (also see simulation (v) in Section 5.3). In terms of the bias under this random-design setup, we think that the lower-bound for MLE and the upper-bound for our stretched-MLE also depend on d and k as $\Omega(\frac{1}{\sqrt{dk}})$ and $\widetilde{\mathcal{O}}(\frac{1}{dk})$ respectively; we also think that the dependence of the stretched-MLE on p_{obs} is no worse than that of the standard MLE. Second, it is of interest to extend our results to other parametric models such as the Thurstone model [177], and we envisage similar results to hold across a variety of such models. Finally, the ideas and techniques developed in this chapter may also help in improving the Pareto efficiency on other learning and estimation problems, in terms of the bias-accuracy tradeoff.

Acknowledgments

This was supported in part by NSF grants 1755656 and 1763734, and 1527032.

Part III

Policy Bias

Chapter 6

Allocation Schemes in Distributed Evaluation: Evaluate a Full Application or a Single Attribute?

Many applications such as hiring and educational admissions involve evaluation and selection of applicants. The number of applicants is often large, thereby making it infeasible for a single reviewer to evaluate all applications. The common practice is to assign the evaluation task to multiple reviewers in a distributed fashion. Specifically, each reviewer is assigned a subset of the applications, and asked to assess all relevant information for their assigned group of applicants. However, such a selection process is subject to problems such as miscalibration (reviewers see only a small fraction of applicants and may not get a good sense of relative quality) and of bias and discrimination (irrelevant attributes of applicants influence the evaluation).

We propose an alternative "segmented" approach to assigning candidates to reviewers. Our approach requires each reviewer to evaluate more candidates but fewer attributes per candidate. We compare our proposed approach to the traditional aforementioned approach on several dimensions via theoretical and experimental methods. We establish various tradeoffs between these two approaches, and identify conditions under which the segmented approach results in more accurate evaluations.

6.1 Introduction

In applications such as hiring and educational admissions, the goal is to evaluate the quality of the candidates, and select a subset of the candidates of the highest perceived quality (according to some criteria). Notably, in both cases, the number of candidates is large (particularly in admissions, which can be on a scale of hundreds or thousands), and thus it is unrealistic for a single reviewer to evaluate all applications. The common practice is to assign the evaluation task to many reviewers in a distributed fashion. Specifically, each reviewer is assigned a subset of the applications, and asked to assess all relevant information for their assigned group of applicants. We term this the *holistic* approach to evaluation.

It is well-documented that in many settings, the holistic selection process is subject to various

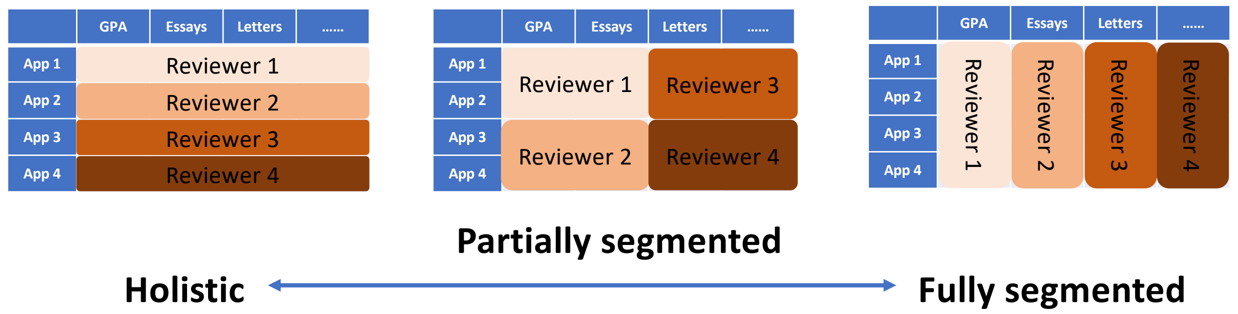


Figure 6.1: An illustration of the spectrum of holistic vs different segmented schemes.

errors. One common source of error is miscalibration: reviewers see only a small fraction of applicants, therefore do not have the same sense of the relative quality, and thus some judge applicants in their set more harshly and others more leniently, leading some candidates to get more or less favorable evaluations than their qualifications merit [159]. Another source of error is bias and discrimination, where irrelevant characteristics such as race or gender influence the evaluation on relevant attributes leading to systematic over- or under-estimation of ability for some groups of applications [43].

In this work, we propose an alternative approach to task assignment. In our proposed approach, each reviewer is assigned a subset of the attributes for evaluation, but a greater number of applicants. We term this a *segmented* approach. A pictorial illustration contrasting the holistic versus segmented approaches is in Figure 6.1. Here each application is represented by a row, and each column represents an attribute. Review assignments are characterized by how regions of the table are allocated to different reviewers. In the holistic approach (left), rectangles of different reviewers are tiled vertically. In the "fully segmented" approach (right), rectangles are tiled horizontally. We use the term "segmented": to refer to any assignment where each reviewer evaluates a (strict) subset of the attributes, i.e., each rectangle does not cover the entire row. Between holistic and fully segmented, there exist schemes where the rectangles are of different aspect ratios to allocate rows and columns to reviewers, termed "partially segmented." Hence, "segmented" includes both fully segmented and partially segmented, forming a spectrum. For the rest of the chapter, we also call the task assigned to a reviewer as a "tile," and each attribute of a single application as a "cell" (of the table depicted in Figure 6.1. Note that for simplicity, we do not explicitly consider the case where different rectangle sizes are simultaneously used, or the case where the there is overlap in the assignment of attributes to reviewers. However, our analysis does not explicitly exclude these cases, and we expect that our general conclusions will apply to such cases as well.

We also note two important boundary conditions based on our assumptions and the contexts to which we expect to generalize. First, we focus on evaluation processes that involve judgment of attributes which are separable. For example, when applying to college, students submit essays, grades, test scores, recommendation letters, etc. which could be assigned to the same or different reviewers. In contrast, when a manuscript is evaluated for publication, the attributes under consideration are less easily separable. Even if the reviewers are only asked to rate a subset of the attributes of the paper (e.g., writing, novelty, impact) it will typically be necessary for them to read the entire paper. A second boundary condition is our focus on settings where the evaluation involves at least some attributes that have a significant subjectivity component, requiring expert judgement. This would stand in contrast to evaluations based almost entirely on attributes involving demonstrably correct answers or universally accepted standards for identifying higher quality performance, such as tests of math ability or physical evaluations of speed or strength. These situations can still be affected by bias, in that the consideration of or weighting of such inputs in final decisions could be influenced by irrelevant characteristics (i.e. for some groups we look at math scores, for others we ignore them) which would be more likely to occur in a holistic process. However these are not the types of influences we consider in the present study.

In what follows, we first discuss two considerations that motivate the proposed segmented approach, namely the potential improvements on *calibration* (Section 6.1.1) and *fairness* (Section 6.1.2) by the segmented approach. Calibration is the estimation of quality from the given applicant samples where the segmented approach enjoys an advantage of size. As alluded to above, fairness is improved in a segmented scheme when irrelevant attributes are hidden from the evaluator. To present a fair comparison, we also discuss potential benefits of the holistic approach – aggregation (Section 6.1.3) and efficiency (Section 6.1.4). Specifically, aggregation in the holistic approach is the ability of the reviewers to jointly consider all attributes account making a final determination about each candidate. Efficiency in the holistic approach refers to the synergies in reviewing two different attributes simultaneously (such as reading a manuscript to evaluate different aspects) as well as the reviewers' flexibility in adapting time and effort allocation on the basis of judged candidate quality. For example, when multiple objective indicators all point to a particular candidate being below threshold, a holistic reviewer could decide to move on to another and allocate more effort to applications above average. Our theoretical and experimental results in the subsequent sections analyze these aspects quantitatively, describe tradeoffs between them, and characterize the regimes where the segmented approach is desirable.

In the next sections, we first present a brief review of relevant literature (Section 6.2). We then describe the general form of our model for evaluation (Section 6.3), followed by theoretical and experimental results (Section 6.4) focused on the four aspects, namely calibration (Section 6.4.1), fairness (Section 6.4.2), aggregation (Section 6.4.3) and efficiency (Section 6.4.4). Finally, we conclude and discuss practical considerations for implementing the proposed segmented approach and future directions (Section 6.5).

6.1.1 Advantages of the segmented approach in reviewer calibration

In the context of evaluation, calibration is the ability of reviewers to evaluate candidates in a manner that is consistent and accurately reflects the candidate's quality relative to the entire pool of candidates candidates. Note that if a reviewer is able to perfectly identify the placement of each candidate with respect to all other candidates, then a perfect ranking of all candidates is derived. However, in most real world evaluation situations, reviewers lack information about the full set of candidates and thus are not able to calibrate perfectly.

Problems with reviewer calibration comprise one of the main disadvantages of the holistic approach to evaluation. In the holistic approach, each reviewer reviews all attributes for each candidate, and with the exception of very small applicant pools, this necessitates that each reviewer sees on a small subset of the entire pool. Pictorially, in Figure 6.1, if we assume each rectangle is of a fixed area (representing a fixed workload of each reviewer), then the holistic approach entails rectangles of the longest width (number of attributes), and therefore the smallest height (number of applications). Due to the limited scope (rows of the matrix in Figure 6.1) the reviewers see, their evaluation of each candidate is heavily dependent on the quality of candidates they review, and incurs an error if their assigned subset is not representative of the entire pool (due to the randomness in the partition of subsets). Consequently, when scores for candidates are combined from different reviewers, the resulting decisions will be less accurate as a result of this miscalibration.

In contrast, we argue that a segmented approach is likely to reduce errors by improving calibration. In a segmented approach, if we hold workload constant, each reviewer evaluates one or a few attributes for a larger number of applications (such as in the right panel of Figure 1). Consequently, reviewers have better information about the distribution of quality in the pool on their assigned attribute given their exposure to a much bigger sample if not the entire set of scores.

Preview of results related to calibration In Section 6.4.1, we present a formal definition of calibration, and subsequently present experimental results that validate this intuition, demonstrating the benefit of the segmented approach in improving reviewer calibration.

6.1.2 Advantages of the segmented approach in reducing prejudice

A second drawback of the holistic approach relates to prejudice. Extensive literature (e.g., [20]) has shown that many judgments are biased by a variety of characteristics that are irrelevant to the content of the evaluation (such as race or gender). Many of these biases operate on a subconscious level [76] and thus affect evaluations even when the reviewers intend to be fair. Even if some reviewers are biased and some are not, in the holistic approach when scores from the reviewers are combined, the result will still be inaccurate with scope for potential prejudice given the lack of consistency across reviewers.

Therefore, a second potential benefit of the segmented approach to evaluation is the reduction of prejudice as a result of bias. This can occur for two reasons.

(F1) **Reducing the impact of biased reviewers:** Reviewers may be biased to different extents. With the holistic approach, each candidate is assigned to only one reviewer, which is either biased or unbiased. In contrast, with the segmented approach, since each reviewer is only assigned a subset of the attributes, candidates can be assigned a mixture of biased and unbiased reviewers. The probability that the evaluation of all attributes of a candidate will be affected by bias is reduced in the segmented evaluation regime compared to the holistic regime as is the probability of prejudice against certain groups of candidates.

(F2) **Restricting access to biasing information:** In the segmented approach, each reviewer only evaluates a subset of the attributes for their assigned candidates. Therefore, the reviewers only need to be provided with information about the candidates that are relevant to the evaluated attributes. For example, the reviewer evaluating the research statement of graduate school applicants does not need to access the biographical information of the candidates or their recommendation letters. When the reviewers are unaware of potentially biasing information, the resulting evaluations are less likely to be influenced by prejudice.

Preview of results related to reducing discrimination We focus on studying the first pathway for reducing prejudice (F1). In Section 6.4.2, we present a multi-attribute model inspired by [100], and provide theoretical support for the conditions under which the segmented approach results in more accurate and less biased evaluations than the holistic approach. Our model supports our argument that a key means by which the segmented approach reduces bias is by redistributing and reducing the impact of biased reviewers.

6.1.3 Advantages of the holistic approach in better aggregation

Despite the advantages of the segmented approach we have described, we note that many settings still structure selection processes using the holistic approach. There are two potential advantages of the holistic approach which relate to the aggregation of judgments, and the efficiency of conducting evaluations.

First, a potential benefit of the holistic approach is that reviewers may more accurately combine information from different attributes than would be the case if each were evaluated by different reviewers independently. This might take one of the following forms.

(A1) **Weighing attributes:** Reviewers conducting holistic evaluations may intuitively weight attributes more effectively than might be true of the aggregation that would occur with segmented evaluation. We model the weighting of attributes by letting each reviewer report a score for each application they review, aggregated along the attributes they review. By comparison, for the segmented approach, we expect an aggregation error, because now different reviewers individually aggregate their assigned subset of attributes, losing the ability to adjust the relative emphasis on different attributes across the entire available set.

(A2) **Jointly reasoning about attributes:** There may be benefits to utilizing information about one attribute when interpreting another. This might be the case when there is a lot of ”noise” in the signal of certain attributes or when the evaluation of an attribute is done more accurately with knowledge of another. For example, standard test scores and essays may be positively correlated, as better writing skills can improve both of them. In this case, the evaluation of writing skills would be improved by knowledge of a candidate's test scores and essay evaluation. Reviewers under the segmented approach may not have access to such cross-attribute reasoning.

On the other hand, it is also possible that reviewers erroneously assume non-existing correlations, or erroneously estimate the amount of the correlation. For example, if a candidate has good test scores, the reviewers may form a good overall impression about the candidate, and consequently be lenient when scoring the other attributes of the candidate, resulting in an inflated evaluation. This type of ”halo effect” can be modeled by an overestimation on the positive correlation between attributes.

Preview of results related to aggregation We focus on studying the aspect (A1). In Section 6.4.3, we present simulation results comparing the performance of holistic vs segmented approaches. The results suggest a trade-off between reducing errors related to calibration (where the segmented approach performs better) vs errors related to aggregation (where the holistic approach performs better). We conclude that the segmented approach is better when the calibration error is significant and when the attributes are correlated, and the holistic approach performs better under other conditions.

6.1.4 Efficiency advantages of the holistic approach

Another potential benefit of the holistic approach is related to the efficiency of conducting evaluations. We consider two reasons why the holistic approach may be more efficient:

(E1) **Adaptively allocating effort:** Recall that our goal is to select the best subset of applications. If an application is clearly below widely recognized threshold on a subset of the attributes, a reviewer may conclude that the application will not be selected, without scrutinizing the remaining attributes or giving a precise score to the application. This is also often the case in practice. For example, in admissions, standardized test scores and GPAs are often used as preliminary filters to eliminate some applications from further consideration. In terms of reviewers, if a reviewer sees red flags in the recommendation letters, the reviewer may only skim through the rest of the material given that context. The segmented approach, by contrast, lacks the ability to perform such adaptive reasoning, because the segmented approach assigns the attributes to the reviewers in parallel. One potential way to alleviate this drawback is to employ a filtering rule (such as test score cut-offs) to reduce the pool *before* segmenting on remaining attributes. However, it would be important to insure that filtering rules did not inadvertently exclude qualified candidates or build in other sources of error.

(E2) **Automation:** Many organizations have tried to automate the evaluation of the entire applications in hiring. However, these attempts have been fraught with biases [51, 174], and eliminating these biases from evaluation of the entire application is a hard problem. Our segmented approach can allow to reap the benefits of automation in a careful manner. Specifically, the segmented approach allows an organization to first identify a the set of attributes which they think can be evaluated in an automated manner without bias, and these attributes can then be separately evaluated via automation and not assign them to a human reviewer.

(E3) **Switching cost:** In the holistic approach, the reviewers primarily switch between different attributes; in the segmented approach, the reviewers primarily switch between candidates. Depending on the attribute being evaluated, switching between candidates may involve greater effort, because the reviewers need to access information that is likely stored in a manner (i.e. different directories with restricted access) that makes navigation among files intentionally more difficult for security reason. Under such conditions, the holistic approach incurs lower switching costs. On the other hand, switching between different attributes of the same candidate may involve greater cognitive effort if the evaluation process is very different for each, such as reviewing transcripts versus reading essays. Under these conditions, the segmented approach may incur lower switching costs. Consequently, whether the holistic or seqmented evaluation regime results in greater switching costs is highly application-dependent.

Preview of results related to efficiency We focus on studying adaptive allocation of effort (E1). In Section 6.4.4, we consider a setting where there are only two attributes. We implement a simple heuristic where the reviewers evaluate the second attribute only if the first attribute passes a certain threshold. The simulation results quantitatively show that the the holistic approach enjoys notable savings especially if there is high (positive) correlation between the attributes, without significant loss on evaluation accuracy. This result points to a trade-off between the holistic vs segmented approaches and highlights a condition in which holistic would have an advantage.

6.2 Related work

In this section, we present an overview of related literature.

Inaccuracy of human evaluation Evaluation of creative work or portfolios is inherently a subjective task. As an example, the NeurIPS 2014 experiment involved a fraction of papers that were assigned to two different halves of the PC. The resulting inconsistency in the results led the writer of this blog post to conclude that Computer Science conference acceptances seem to be more random than was previously assumed. A recent dissertation highlights various aspects of subjectivity in hiring. An economic perspective of algorithmic fairness [47] also highlights how human inconsistency in evaluations might be benchmarked with that introduced by algorithms. In a field experiment, Cowgill [46] describes how machine learning can lead to better outcomes than human evaluators by reducing the inconsistency under certain conditions.

Controversy of automated evaluation Algorithmic hiring is not a panacea against inconsistent human judgement either. The famous lapses of the AI-inspired hiring tool in Amazon documented in [51, 174] highlights many of the challenges in the deployment of AI in hiring decisions. In our work, we suggest a natural point of efficiency in using algorithmic strategies for evaluation in the form of segmenting the work. Our goal is to delineate conditions under which this efficiency can be realized, while also highlighting under which settings aspects of fairness that may be compromised.

Applications using a segmented approach "Segmented" approaches are frequently used in other applications. One key example is grading by multiple teaching assistants. The grading task is usually split by having each grader grade a specific subset of questions across all students' homeworks, rather than each grader grading the entire homework of a subset of students (which would mirror holistic evaluators).

People work on different parts of complicated tasks in [143]. Here a complex task is broken down, where different crowdsourcing workers work on different parts, and then their work is computationally (or sometimes manually) put together.

Calibration There is ample evidence in the literature that human evaluators are miscalibrated, that is, different evaluators have different evaluation scales [7, 64, 123]. Some evaluators may be strict, some lenient, some extreme, some conservative, etc. Furthermore, this miscalibration is known to be quite complex [28].

Our definition of using the relative scale for calibration is inspired by applications such as admissions and hiring, where one often marks an applicant as being in a certain relative bin. In peer review, various conferences ask reviewers to rate papers according to the relative position among all submitted papers [159].

Fairness The problems in the Amazon automated resume screening system [174] provide a clear example of the importance of enforcing fairness in evaluation tasks. When unfair outcomes arise as a result of reviewers taking into account information from protected categories, segmenting evaluations offers a natural advantage by potentially avoiding revealing irrelevant information (such as protected categories) for the task of evaluating the attributed assigned. Kleinberg and Raghavan [100] have highlighted the relative advantages of rules promoting diversity (such as the Rooney rule) in avoiding bias in the selection of a best candidate from a pool. In a similar vein, we will highlight the conditions under which segmented evaluations retain an edge over holistic ones for finding the best candidate in Section 6.4.2.

Hierarchical decision making The process for evaluating candidates for employment has been debated throughout the history of industrial psychology [153]. There has been a strong tendency among those reviewing candidates toward relying on intuition and informal processes, such as unstructured interviews, despite the wealth of data pointing to the benefits of a more structured approach [183]. Meta-analyses of selection methods have demonstrated that integrating standardized measures such as cognitive ability testing, work samples, and structured interviews greatly increased the reliability and predictive validity of selection processes [152] and that they could be used in combination as each added incremental validity to the overall prediction of future performance [30]. Following a hierarchical decision process [101], the best selection systems do a thorough job task analysis to identify the strongest predictors of future performance, then identify the measurement approach that achieves the best balance of high reliability vs. cost.

Halo effects during aggregation Despite the abundant evidence of the effectiveness of more structured approaches to selection [152] many organizations continue to utilize unstructured and informal methods [183]. However, these approaches pose a multitude of problems. First, they tend to take on the "holistic" form mentioned previously, in that reviewers provide an overall rating or ranking on the set of candidates they review. This raises the chances of a variety of biases to reduce the quality of decisions. One of the oldest of these, known as the "halo effect" [176], refers to the extent to which a rater's evaluations of another individual on a series of individual traits are influenced by the rater's overall liking (or disliking) for the individual. Bias can occur for a variety of reasons; implicit bias (such as the halo effect) can occur as a result of a variety of physical attributes associated with stereotypes including attractiveness, height, race, gender, age, weight, physical disability or similarity to the prototype for the role [18]. They can also occur as a result of similarity to the reviewer [144], as well as unusual scores or performance on some dimensions of the evaluation that alter perceptions of other dimensions. These biases can work in a candidate's favor or against them, but in either case they serve to increase the error in selection decisions in a manner that is often non-random and problematic, as certain attributes such as race, gender, or socioeconomic status elicit biases in a consistent enough manner that unstructured selection processes predictably result in systematic discrimination [128]. A segmented approach to evaluation offers the possibility of mitigating such biases, as such a system could be designed so that reviewers evaluate candidates on particular attributes while remaining unaware of their status on other, potentially biasing, attributes.

Aggregation in social choice and voting; subjectivity There is a rich literature on aggregation of people's preferences and evaluations over various domains. The field of social choice theory concerns itself with aggregating subjective opinions of voters on multiple candidates [26]. A different type of aggregation – more closely related to our work – is that of aggregating over multiple criteria (or features). Specifically, there may be multiple criteria on which an item is evaluated, and the overall evaluation of the item depends on how the evaluations of the individual criteria are combined [106, 130]. The aggregation of criteria is frequently modeled using linear models [129, 151], and we also make this assumption in parts of our work.

6.3 Model

In this section, we introduce our proposed model for the evaluation procedure.

Notations We assume there are n applications, and each application has d attributes. We let $X \in \mathbb{R}^{n \times d}$ be a matrix whose $(i, j)^{\text{th}}$ entry x_{ij} represents the "true" value of application $i \in [n]$ on attribute $j \in [d]$. A greater value represents higher quality. We divide the task into tiles as shown in Figure 6.1, where each reviewer evaluates a $n_0 \times d_0$ tile of n_0 applications on a subset of d_0 attributes. For simplicity, we assume each attribute of each application is evaluated once. Then the number of reviewers is defined as $R := \frac{nd}{n_0 d_0}$. For each reviewer $k \in [R]$, denote $A_k \subseteq [n]$ as the set of applications assigned to this reviewer, and $T_k \subseteq [d]$ as the set of attributes assigned to this reviewer. For any reviewer k, we let $\{x\}_{A_k, T_k}$ denote the submatrix of X comprising the rows A_k and columns T_k; this is the data seen and evaluated by reviewer k.

Ground-truth We assume there exists a true ranking of the applications. To derive the true ranking, we define true scores of the applications as a weighted linear combination on a mapping of the attributes. That is, consider a vector $w \in \mathbb{R}^d$, and a function $f^* : \mathbb{R} \to \mathbb{R}$. Then the true score for each application $i \in [n]$ is defined as $y_i^* := \sum_{j \in [d]} w_j f^*(x_{ij})$. The true ranking is the ranking induced by the scores $\{y_i^*\}_{i \in [n]}$. We use the function f^* to model the non-linear preference on each individual attribute. In some of our subsequent results, f^* as the percentile function mapping from the values to their percentage (when the values are drawn from a distribution and the number of applications is large, then this percentile function becomes the inverse c.d.f. of the distribution). For the weights w, for simplicity we later consider w as the all-one vector. Then the linear combination reduces to taking the mean of all attributes.

Metric In applications such as admissions, the goal is to choose the subset of the maximum quality given a fixed size. Hence, the natural metric is the top-K error in ranking. For simplicity, we consider the top-1 accuracy in ranking. That is, the accuracy is 1 if the estimated ranking correctly identifies the maximum value in the true ranking, and 0 otherwise[1].

Evaluation process The evaluation process is described in Algorithm 3. We assume each reviewer $k \in [R]$ has a calibration function $f_k : \mathbb{R} \to \mathbb{R}$ on each individual attribute, and an aggregation function $g_k : \mathbb{R} \to \mathbb{R}$ for aggregating their assigned subset of attributes. The realizations of the functions f_k and g_k are discussed later. The algorithm consists of two steps. In Step 1 (Line 1-6), each reviewer evaluates their assigned subset of attributes (Line 3), and reports a score $y_i^{(k)} \in \mathbb{R}$ for each application they are assigned (Line 5). In Step 2 (Line 7), we compute a final score for each application $i \in [n]$, by aggregating the scores reported by all the reviewers assigned this application (on a subset of attributes). These scores from reviewers are aggregated exogenously, by using the weight vector $w^{(k)} \in \mathbb{R}^{\binom{d}{d_0}}$. When the d-dimensional weight vector w is all-ones (i.e. taking the mean), it means all the attributes are treated with equal importance. In this case, it is natural to define $w^{(k)}$ as the all-ones vector too, taking the mean of all reviewers assigned to each application.

Algorithm 3: The evaluation procedure.

```
   Input: Applications $X \in \mathbb{R}^{n \times d}$, the number $n_0$ of applications and the number of
          attributes $d_0$ assigned to each reviewer, the weight vector $w \in \mathbb{R}^d$
   Output: Application scores $y \in \mathbb{R}^n$
   /* Step 1:  each individual reviewer evaluates their assigned tile */
1  foreach reviewer $k \in [R]$ do
2      foreach application $i \in A_k$ do
3          Compute the attribute score $y_{ij} \in \mathbb{R}$ as $y_{ij} = f_k(x_{ij}; \{x\}_{A_k,T_k})$        // Compute
            each cell
4      end
5      Compute the reviewer-aggregated score $y_i^{(k)} = g_k(\sum_{j \in T_k} w_j y_{ij}; \{x\}_{A_k,T_k}, w)$.
        // Aggregate within each tile
6  end
   /* Step 2:  aggregate across reviewers */
7  Compute the aggregated scores $y \in \mathbb{R}^n$ as $y_i = \sum_{k \in [R]} w^{(k)} y_i^{(k)}$. // Aggregate across
    tiles
```

[1] For simulation and experiments, we make sure that there are no ties in the maximum application in the true ranking. If there are ties in the estimated ranking, the accuracy is computed as 1 / (number of maximum applications) if the true top-1 application is one of the maximum application in the tie, and 0 otherwise.

6.4 Main results

In this section we present main results. In what follows, we present our results on the four aspects: calibration (Section 6.4.1), fairness (Section 6.4.2), aggregation (Section 6.4.3) and efficiency (Section 6.4.4). The error bars presented in all the plots represent the standard error of the mean.

6.4.1 Calibration

In this section, we present empirical results that validate the improvement on calibration by the segmented approach via a study in a crowdsourcing platform. We start by giving our precise definition of calibration, and then discuss our experimental results.

Definition of calibration Formally, we define calibration as the reviewers' accuracy of estimating the ranking (or percentile) of each candidate with respect to the entire pool of all candidates. We define calibration on this relative scale for three reasons: (1) the selection problem is intrinsically relative in nature, that is, we aim to select the top candidates compared to the entire pool; (2) in many applications, the evaluation data that the reviewers are asked to report is relative. For example, reviewers may be asked to give scores on a scale of 1-5, where the criteria define the score of 1 as the candidate being the bottom 20% among all candidates, and 2 as being 20%-40% among all candidates, etc.; (3) people's reasoning involves a relative nature. For example, being a "top" candidate is perceived as simply being significantly better than the rest of the candidates.

Hypothesis Recall that the potential improvement in calibration provided by the segmented approach is given by the larger number of applications that the reviewer has access to in this approach. Intuitively, in the segmented approach, the reviewers acquire more knowledge about the pool for each attribute they examine, and hence have better calibration. The goal of our experiment is to verify this hypothesis. For simplicity, we focus on a single attribute to isolate it from the effect of aggregation across attributes. While this relationship is intuitive, it is not immediately clear that it is realized in reality. For example, a counter-argument may say that reviewers have a fixed short-term memory, so viewing more samples may have a minimal effect on calibration.

Experimental setup We designed the experiment focused on a single attribute. We recruited 52 crowdsourcing workers on the Prolific platform. The workers were introduced to a hiring context and asked to evaluate scores of candidates. Specifically, they were told that there are 1000 candidates with scores that are integers in the range of 200-300 (without any distributional information about the scores). Then the workers were presented some numbers in between 200-300, and are asked to estimate the percentile of the scores. For each score, the workers answered a multiple-choice question with 5 choices to estimate the percentile of the score with respect to the population: 0-20%, 20-$40\%, \ldots, 80$-100%. We chose to ask the reviewers to report in 5 quantized bins instead of directly reporting a number for percentile, because extensive studies have shown that workers are not able to perceive fine numbers accurately and have higher accuracy if the elicitation format is quantized. We also conducted a preliminary study comparing soliciting data with 5 bins and 10 bins, and confirm this trend. At the end of the study, the workers are also asked to describe in text their strategy in deriving answers to these questions.

Question grouping The workers were divided into two groups. In the first group, each worker was presented 5 scores (termed "5Q-group"). In the second group, each worker was presented 20 scores (termed "20Q-group"). The workers are presented with 5 questions per page. That is, for the 20Q-group, the 20 questions are distributed across 4 pages. Neither group of workers was told the number of the scores they will be presented before starting the task. The workers were required to answer all questions on a page before proceeding to the next page. The workers were allowed to go back to previous pages at anytime. In both groups, the workers were allowed to edit their answers to any question at any time before submission. We chose the design of grouping 5 questions per page and not informing the workers the number of total questions a priori, because in a preliminary study where workers from both groups were presented their respective number of questions on one page, we observes that the results from the workers in the 20Q-group were of lower accuracy in estimating the the first five questions compared to the 5Q-group. We hypothesize that this is due to the fact the workers in the 20Q-group are aware in the beginning that they have a lot of questions to answer, and accordingly reduce their effort in answering each question. Henceforth, we restricted each page to have 5 questions, and did not inform the workers the number of questions in total, to eliminate this confounding factor.

Values of the scores The scores are generated at random from $\mathcal{N}(230, 25)$, truncated to the range of $[200, 300]$. We choose this distribution for generating the scores, because in a preliminary study where the workers were presented scores in the range of $[0, 100]$, we observed that the workers appeared to have a strong uniform prior, mapping values in $[0, 20]$ to percentile 0-20%, etc. We therefore chose a range that is not $[0, 100]$ so that the workers do not rely on such priors.

Results We record the reviewer calibration measured by their accuracy in estimating the percentiles. We compute the ℓ_1 error between the workers' reported percentile against the true percentile. Formally, we number the percentile 0-$20\%, 20$-$40\%, 40$-$60\%, 60$-80% and 80-100% as the bins $1, 2, 3, 4$ and 5 respectively. Then the error is computed as the absolute difference between the workers' reported bin and the true bin of the value, followed by taking the mean over the questions and the workers. For the 20Q-group, we separately compute the error restricted to each page of 5 questions (i.e., Q1-5, Q6-10, Q11-15, Q16-20).

The results are shown in Figure 6.2. We make the following observations. First, the workers in the 5Q-group and 20Q-group have comparable accuracy in answering the first 5 questions – this matches what we expect from the experimental set up. Second, the error in the 20Q-group decreases significantly as the workers see more numbers in the later pages. This validates our hypothesis that the reviewers have better calibration when they see more candidates.

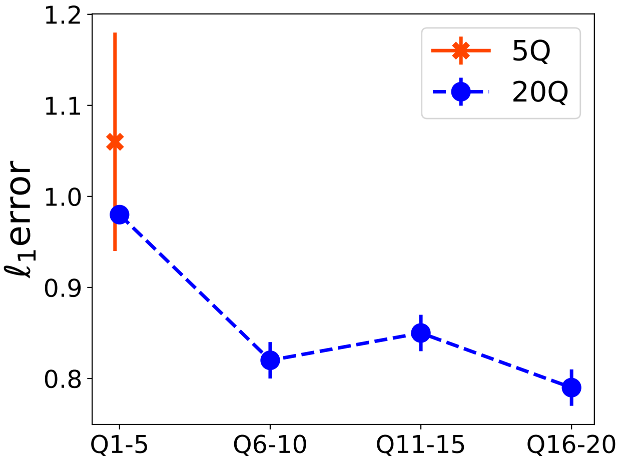


Figure 6.2: The ℓ_1 error in estimating percentile, for workers in the 5Q-group (representing holistic) and the 20Q-group (representing segmented), computed individually for each page of 5 questions.

6.4.2 Fairness

In this section, we formulate a simple model to highlight the advantage of the segmented over the holistic approach in terms of avoiding biases or unfairness that creeps in from evaluating other attributes. As in [100], we consider the problem of finding the best candidate in a pool of applicants, and present theoretical guarantees that characterize a wide range of regimes where the segmented approach yields higher accuracy than the holistic approach.

Formulation We follow [100] and consider "absolute" values instead of relative values for this section. The marginal distribution of each attribute is generated from a continuous distribution $\mathcal{D}$ whose support only contains non-negative values. For the reviewers, we let f^* be the identity function. We also let g be the identity function. A fraction of α applicants are from the disadvantaged group, and a fraction λ of the attributes are "protected". Each reviewer has i.i.d. $Bernoulli(\gamma)$ probability of being biased in the following sense: an unbiased reviewer reports the (noiseless) true values of the assigned tile for any applicant and any attribute, while a biased reviewer applies a multiplicative discount factor $\beta \in [0, 1)$ to the protected attributes of the disadvantaged applicants, and reports the true value otherwise. In other words, for unbiased reviewer k, its function f_k is the identity function. For a biased reviewer k evaluating attribute j for applicant i (of value x_{ij}), we have

$$f_k(x_{ij}) = \begin{cases} \beta x_{ij} & \text{if } j \text{ is a protected attribute and } i \text{ is a disadvantaged applicant} \\ x_{ij} & \text{otherwise.} \end{cases}$$

Note that here the function f_k no longer represents calibration but a discrimination of the biased reviewers.

Notation: Denote the underlying distribution as $\mathcal{D}$. To illustrate the extreme situation, in what follows, we consider the case where attributes are perfectly correlated (i.e., the attributes of an applicant have identical values), and we call this the "quality" of the applicant or candidate. Denote the quality of the disadvantaged candidates as $X_1, \ldots, X_{\alpha n} \sim \mathcal{D}$, and the advantaged candidates as $Y_1, \ldots, Y_{(1-\alpha)n} \sim \mathcal{D}$. Define the maximum of the disadvantaged candidate as $X^{\max} := \max_{i \in [\alpha n]} X_i$ and $Y^{\max}$ similarly. Denote the expected top-1 error incurred by the mean estimator under holistic and segmented approaches as e_{hol} and e_{seg} respectively. The power law distribution with parameter δ is defined as $\mathbb{P}[Z \geq t] = t^{-(1+\delta)}$ supported on $[1, \infty)$.

Proposition 6.1. *Let the number of attributes be $d = 2$. Let $\alpha = 0.5$ (that is, half of the applicants are disadvantaged). Let the two attributes be perfectly correlated. Consider two reviewers, under holistic (i.e., each reviewer sees half of the applicants, where the split is uniformly at random) or segmented (i.e., each reviewer sees one attribute of all applicants) approaches.*

(a) Let $\lambda = 0.5$ (that is, one of the two attributes is protected). Consider any $\beta \in [0, 1)$. Then for any $\gamma \in (0, 1)$, the segmented approach is better than the holistic approach, i.e. $e_{seg} \leq e_{hol}$.

(b) Let $\lambda = 1$ (that is, both attributes are protected). Let $\beta = 0$ (i.e., extreme downward bias on disadvantaged applicants). Then

$$e_{hol} - e_{seg} = \frac{\gamma(1-\gamma)}{2} \left[4 \cdot \mathbb{P}\Big(X^{\max} > 2Y^{\max}\Big) - 1\right]. \tag{6.1}$$

Hence, for any $\gamma \in (0, 1)$, the segmented approach is better than the holistic approach, if and only if

$$\mathbb{P}\Big(X^{\max} > 2Y^{\max}\Big) > 0.25. \tag{6.2}$$

This condition (6.2) is dependent on the number of applicants n and and the distribution $\mathcal{D}$, and independent of the other problem parameters. As an example, for $PowerLaw(\delta)$ with parameter δ, for sufficiently large n, the segmented approach is better than the holistic approach if and only if

$$\delta < \frac{\log(3)}{\log(2)} - 1 \approx 0.58. \tag{6.3}$$

This proposition reveals that segmented is better than holistic in terms of accuracy over a large range of parameters modeling the presence of mild levels of discrimination among the reviewers based on attributes are protected. Despite its extreme assumptions, it readily illustrates the advantages of segmented in terms of avoiding bias creeping in from signals from other irrelevant attributes.

The proof of this proposition is in Chapter 13. Part (a) is intuitive: While holistic is led astray on any disadvantaged applicant on the protected attribute, segmented ensures that one reviewer out of the two only sees the attribute that is not protected and hence adds more information compared to the holistic case. The intuition for part (b) is as follows. The difference in holistic vs. segmented lies in the case where one reviewer is biased and the other is unbiased. For segmented, one attribute of the best disadvantaged applicant is assigned the biased reviewer. For holistic, the best disadvantaged applicant is assigned the biased reviewer with probability 0.5 (this is worse than in the segmented case – giving error probability 1 when $\beta = 0$), and the unbiased reviewer with probability 0.5 (better than segmented – giving error probability 0 when $\beta = 1$). The probability in (6.2) characterizes the success probability of segmented. For power law, with smaller δ, the distribution of the max, $X^{\max}$ and $Y^{\max}$ (identical in distribution when $\alpha = 0.5$), has a heavier tail by definition of the power law distribution, so it's more likely for segmented to succeed in this regime of δ.

Generalizing the underlying advantage of segmented in the above proposition to other cases (such as a general d and any tile shape for the segmented work) is an interesting open problem.

6.4.3 Aggregation (and its tradeoff with calibration)

In this section, we analyze the aggregation aspect, that is, how each reviewer derives an aggregated score for each application by combining the attributes they review. Recall that we reason that the holistic approach may have the benefit of better aggregation, thereby resulting in a tradeoff between aggregation and calibration in the comparison of the holistic versus the segmented approaches. We first describe the choices of reviewer functions f and g (see Section 6.3 and Algorithm 3) that incorporate both the calibration and the aggregation aspects. Then we present simulation results that describe this tradeoff on the spectrum from the holistic to the segmented approaches.

Model of the true calibration function f^* Following the definition of calibration in Section 6.4.1, we define the true calibration function f^* as the true ranking (or percentile) of each attribute (among all applications).

Model of reviewer calibration function f We describe the choices of the reviewer calibration function f and the aggregation function g. For the calibration function f, recall that we define calibration as comparing a candidate to the entire pool. Intuitively, the more knowledge a reviewer has about the entire population, the better their calibration is. Such knowledge about the pool comes from two sources: (1) the applications that the reviewer is assigned to; (2) reviewer's prior observations or belief about the distribution of the pool. Recall from Section 6.3 that each reviewer $k \in [R]$ has a calibration function f that is a mapping on each individual attribute. We define the function f as follows. Recall that n_0 denotes the number of applications that each reviewer is assigned. We characterize the prior information by a set of past applications, whose size is denoted as N. Then the calibration function f maps each application to its (normalized) ranking among the $(n_0 + N)$ current and past applications. We denote this function as $f_{n_0,N}$. We allow the prior applications to be different for different reviewers, so both the current and past applications depend on the reviewer $k \in [R]$, and thus the function $f_{n_0,N}$ implicitly depends on the reviewer $k \in [R]$.

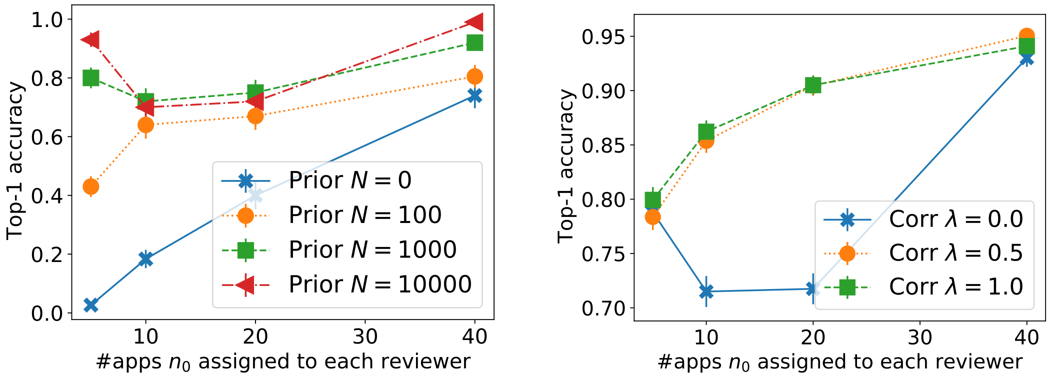
Intuition about the calibration function f Note that in the special case of $N = 0$, the function reduces to ranking among current applications. On the other hand, in the special case of $N \to \infty$, the function converges to the true inverse c.d.f. With $N \to \infty$, the holistic scheme approaches the true ranking. In this context, the calibration experiment in Section 6.4.1 verified that $f_{n_0,N}$ gets better when n_0 increases. This model also incorporates the miscalibration of lenient and strict reviewers – when the reviewer sees a skewed subset of good applications (from the current applications or from the prior), the reviewer will have strict calibration. Likewise when the reviewer sees a skewed subset of low-quality applications, the reviewer will have lenient calibration.

Model of aggregation functions g Similar to the calibration functions f, we define the aggregation function g on a relative scale. Specifically, each reviewer aggregates their assigned attributes using the weight vector w, and derives a score for each assigned application (including the ones in the prior). Then the aggregation function g maps each application to its (noramlized) ranking among the $(n_0 + N)$ scores of the current and past applications. We note that the aggregation is performed only within the attributes that a reviewer is assigned to. Hence, we expect that the aggregation function g introduces error in the segmented approach. On the other hand, as suggested by Section 6.4.1, the segmented approach improves calibration. In what follows, we conduct synthetic simulation to investigate this tradeoff between calibration and aggregation, and the regimes where the holistic or the segmented approach is better.

Setting We let the number of applications be $n = 200$ and the number of attributes be $d = 8$. We assume each reviewer has a fixed workload of reviewing 40 cells in total. We vary the number of applications $n_0 d_0 = 40$ assigned to each reviewer, where $n_0 = 5$ corresponds to the holistic approach and $n_0 = 40$ corresponds to the (fully-)segmented approach. We generate all attribute values (for both the current and past applications) from the power law distribution with parameter $\delta = 1$. To model the correlation between attributes, we order the attributes by the ranking induced by a d-dimensional Normal distribution $\mathcal{N}(0_d, \Sigma)$, where we set the covariance matrix as $\Sigma = (1-\lambda)I_d + \lambda \mathbf{1}\mathbf{1}^T$. With this definition of the covariance matrix Σ, we have $\lambda = 0$ correspond to i.i.d. attributes, and λ correspond to all attributes of an application having the same ranking (perfect correlation). We consider the top-1 ranking accuracy, and report simulation results over 200 runs.

Results First, we examine the performance over the holistic vs. segmented spectrum for different amounts of prior knowledge. We consider $\lambda = 0$ for independent attributes. The amount of prior knowledge is controlled by the parameter N. The result is shown in Fig 6.3(a). We observe that when the prior knowledge is limited (when $N = 0$), the top-1 error is dominated by the calibration error, so holistic does not perform well. When the prior knowledge is large (when $N = 1000$), the calibration error decreases and the aggregation error induced by g increases. In this case, the holistic approach is better than the partially-segmented approaches.

Second, we examine the calibration-aggregation tradeoff under different amount of correlation between the attributes. We set the prior size to be $N = 1000$. The result is shown in Figure 6.3(b). We expect that the aggregation error decreases when the (positive) correlation between attributes increases. This is because all attributes give the same ranking, and the role of aggregation becomes less important. Figure 6.3(b) confirms this trend that more correlation between the attributes benefits the segmented approach.



(a) Different sizes of the prior knowledge

(b) Different amounts of the correlation between attributes

Figure 6.3: A comparison of the spectrum of the holistic vs. segmented approaches (x-axis) in terms of the calibration-aggregation tradeoff.

Caveats in interpreting the results This simulation seems to suggest that the fully-segmented approach is always the best. This is because in this case, the aggregation function g of individual reviewers become trivial, so there is no aggregation error. Then combining zero aggregation error with the best calibration yields the best performance for the fully-segmented approach. However, we should be careful about drawing the conclusion that fully-segmented is always optimal, due to the following caveats: (1) we assume the exogenous aggregation for the segmented approach is perfect, which may not be the case in reality (especially if true w is not the all-one vector); (2) it makes sense for other reasons to group attributes together, giving less preference to the fully-segmented approach. For example, the TOEFL and GRE scores of a graduate applicant are correlated, and it makes sense to assign a single reviewer to evaluate both; (3) the fully-segmented approach does not allow adaptive decision to save evaluation effort (see more in the subsequent Section 6.4.4).

6.4.4 Efficiency from adaptive decision-making

Another potential benefit of the holistic approach is its efficiency in making evaluation adaptively. Specifically, when a reviewer evaluates multiple attributes, if an application has a very low value on one attribute, we may be confident to conclude that it is unlikely for the application to have high values on other attributes due to correlation, or even if the values on other attributes are high, since we combine the attributes such as taking their mean, it is still unlikely for this application to be one of the best applications. We now conduct synthetic simulation to quantitatively analyze such saving given by the holistic approach.

Setting We consider $n = 200$ applications. For simplicity, we consider $d = 2$ attributes, and two reviewers. Then each holistic reviewer evaluates both attributes on half of the applications; each segmented reviewer evaluates one attribute on all applications. We assume both reviewers always evaluate the first attribute. Then the reviewers only proceed to evaluate the second

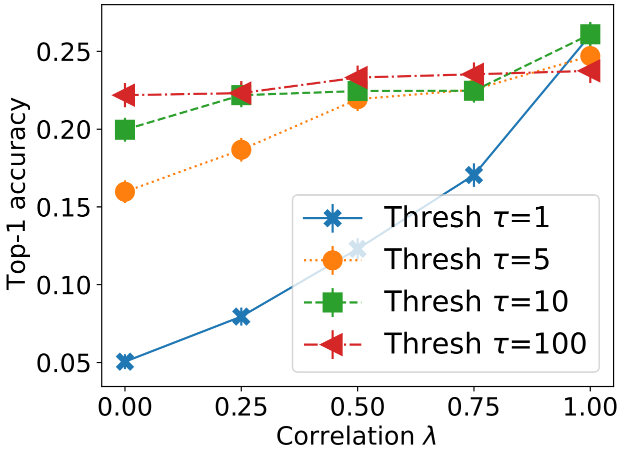


Figure 6.4: The top-1 accuracy for different threshold values τ, for various values of the correlation (x axis) of the attributes.

attribute of an application, if the value of its first attribute is the top-τ. Finally, the top-1 application is selected only from applications whose both attributes are evaluated. The simulation results are reported over 1000 runs.

Results We compute the top-1 accuracy for different threshold values τ, under different correlations of the attributes. The result is shown in Figure 6.4. We observe that with high correlation between the attributes, the accuracy only decreases marginally when less evaluation is performed (smaller value of τ). However, the threshold introduces a significant amount of saving in terms of the number of cells that are evaluated. Hence, we conclude that a higher correlation between the attributes allows more saving in the holistic approach.

In contrast, such saving is not possible in the segmented approach because the evaluation tasks are typically allocated in parallel to the reviewers. Alternatively, choosing or reducing the evaluation task can be made possible by decomposing the evaluation into multiple rounds. However, having multiple rounds adds the complication to the implementation of the evaluation procedure, and may also require more time needed for the evaluation process. This comparison again suggests a tradeoff between the holistic and the segmented approaches, on efficiency versus calibration.

6.5 Discussion

In this work, we propose using the segmented approach as an alternative to the conventional holistic approach, for applications such as hiring and admissions. We provide a detailed discussion comparing the two approaches, and provide theoretical and experimental results focused on four aspects: calibration, fairness, aggregation and efficiency. These results indicate the potential improvement by the segmented approach on calibration and fairness, while also suggesting that the holistic approach has potential benefits on aggregation and efficiency. These results together suggest a trade-off between the holistic and the segmented approaches (and the spectrum in between). The trade-off and the choice of which scheme to use depends on the specific application and desiderata. There are various considerations, as well as open problems, to this end:

- The segmented approach requires grouping of attributes, and the system designer needs to do this grouping appropriately. For example, in the case of admissions, one may group test scores and GPAs as one attribute called "scholarly performance". In addition, in order to provide appropriate context to reviewers, one may need to provide some attributes to multiple reviewers.
- We assumed for simplicity that we aggregate the attributes by taking their mean. In practice, we may want to use different weights for different attributes, or even use non-linear functions. This gives rise to two interesting open problems. First, if the aggregation function is non-linear, then it requires grouping the attributes in a manner that allows for the best possible approximate aggregation after segmentation. Second, in some cases, the aggregation function may not be precisely provided by the system designer, but needs to be learnt from past data. This problem of learning the aggregation function from evaluations has been studied in the specific context of peer review of scientific papers [130], and it is of interest to extend such results to more general applications.
- This work discusses a spectrum of choices on how to tile the attributes and applications in terms of the spectrum between holistic and segmented. An open problem of interest is to establish the optimal point(s) on this spectrum theoretically and/or practically for any given specification of application and desiderata.

Chapter 7

Reducing Bias from Alphabetcial-Ordering Authorship in Scientific Publications

7.1 Introduction

It is common in some academic fields such as theoretical computer science to order the authors of a paper according to the alphabetical order of their last names. Alphabetical ordering is also employed in other contexts like listing of names of people on the web, for instance, to order the participant list and pictures on the ITA conference website[1].

Although alphabetical ordering mitigates some issues with other ordering approaches (e.g., possible conflicts among authors under contribution-based ordering), it causes its own biases. These biases form the focus of this post. What are these biases? A number of papers have empirically studied the effects of the convention of alphabetically-ordered authorship, which reveal biases associated to this convention. Here is an excerpt from the study [57] by Einav and Yariv:

> *"We begin our analysis with data on faculty in all top 35 U.S. economics departments. Faculty with earlier surname initials are significantly more likely to receive tenure at top ten economics departments, are significantly more likely to become fellows of the Econometric Society, and, to a lesser extent, are more likely to receive the Clark Medal and the Nobel Prize. These statistically significant differences remain the same even after we control for country of origin, ethnicity, religion or departmental fixed effects. All these effects gradually fade as we increase the sample to include our entire set of top 35 departments.*
>
> *We suspect the 'alphabetical discrimination' reported in this paper is linked to the norm in the economics profession prescribing alphabetical ordering of credits on coauthored publications. As a test, we replicate our analysis for faculty in the top 35 U.S. psychology departments, for which coauthorships are not normatively ordered alphabetically. We find no relationship between alphabetical placement and tenure status in psychology."*

Various other studies make similar observations and draw similar conclusions (e.g., see [85, 184] and references therein).

[1] `http://ita.ucsd.edu/workshop/18/?year=18#participants`.

Conference	Number of papers	Number of papers using “First author et al.” in its text
STOC 2017	99	70
STOC 2016	79	59
FOCS 2017	79	48
FOCS 2016	73	43
EC 2017	75	48
EC 2016	99	87

Table 7.1: A significant portion of the papers accepted at theory conferences in recent years uses the “First author et al.” citation format.

What is the source of these biases? There are at least two types of bias effects.

- **Implicit bias – Primacy effects:** Primacy effects describe the human cognitive bias that people are more likely to remember and choose items showing up earlier in a list than items later in the list – in short, “first is best” [32]. Primacy effects have been widely studied in psychology, and observed in many laboratory and field settings, e.g., people are more likely to recall words earlier in a list [96]; people are more likely to choose the first candidate on the ballot for an election [38]. In the context of author ordering, primacy effect suggests that authors whose names show up earlier in the author list are likely to receive more attention from the reader.
- **Explicit bias – “First author et al.”:** A more conspicuous bias arises when papers use a “First author et al.” format in its text to refer to other papers. Now, it may be argued that communities which use alphabetical-ordering conventions do not use the “First author et al.” format. So we put this hypothesis to the test. Publication venues in computer science that primarily follow alphabetical orderings include STOC, FOCS and EC. We performed a search on Google Scholar, and Tab. 7.1 shows the number of papers in these conferences which use the “First author et al.” format in their own text. We observe that a significant portion of the papers accepted at theory conferences in recent years uses the “First author et al.” citation format.

What are alternative solutions? For ordering authors in papers, a contribution-based arrangement is a popular alternative. However, this manner of ordering can cause conflicts between authors regarding their contributions. An alternative is to employ a technique that computer scientists use extensively in their research – randomization. Under such a randomized arrangement, authors could be ordered uniformly at random. Or otherwise the authors could be arranged as a combination of contribution-based and randomized methods, where contributions can determine a partial order and then a total order is selected uniformly at random from among all total orders consistent with the partial order. In this case, symbols or footnotes can be used to distinguish authors whose orders are contribution-based and whose orders are random. See, for instance, the paper [142] for a more detailed discussion on randomized author ordering.

Likewise for lists of names on the web, one could randomize the order whenever feasible. This randomization could be dynamic (a new ordering whenever the page is loaded) or static (permute once and fix the permutation). Now, if we were dealing with listing names in some printed material, searching for any particular individual would have been difficult. But on the browser, one can always use Ctrl/Cmd+F to search.

7.2 Main results

We reached out to a number of organizations regarding mitigating the bias from alphabetical-ordering authorship in scientific publications. We hope that these efforts constitute the important first steps in reducing such bias:

- We reached out to the program chairs of ACM EC 2019, Nicole Immorlica and Ramesh Johari. They kindly agreed to change the submission style file with numbered references as default from the "First author et al." format, and also keep numbered references in the camera ready versions.
- Taking cognizance of these biases, starting October 24, 2019, the Machine Learning Department at CMU has randomized the ordering of students and faculty on its webpages[2]. One concern was that users may get confused since the standard practice is to order alphabetically. To this end, we put a small bar on top of the page indicating these biases and a link to this post for details. Our webmaster tells us that the user experience has been same as before (along with a lot of positive feedback that this was the right thing to do). The CMU Theory group website also uses dynamic orderings now[3].
- We reached out to Virginia Vassilevska Williams, the program chair of STOC 2021. Taking cognizance of these issues, the call for papers for the conference added:

  > **Recommended best practices for citations:** Authors are asked to avoid "et al." in citations in favor of an equal mention of all authors' surnames (unless the number of authors is very large, and if it is large, consider just using `\cite{}` with no "et al.").

[2] `https://www.ml.cmu.edu/people/`

[3] `http://theory.cs.cmu.edu/`

Chapter 8

Gender Bias in Conference Awards

In this chapter, we evaluate the gender distribution of best paper awardees in various top CS conferences. The inspiration to do so comes from an interview of Andrea Goldsmith in the IEEE Information Theory Society (ITSoc) newsletter [2] in which she lays out some interesting statistics:

> "Going by the names of authors, it seems that of the 64 papers that have won the ITSoc paper award, not a single one has a female author. Similarly, it appears that not a single female student has won the ISIT student paper award. Only five women have been elevated to IEEE Fellow through ITSoc, which is quite a small number given that approximately 3–5 members have been elevated annually to Fellow through ITSoc going back many decades. Finally, only one of the nine Padovani lecturers, who are selected as role models for current ITSoc graduate students, has been female. In my own experience serving on the ITSoc awards committees and Fellows committee, I rarely see women nominated for society awards and honors. When they are my sense is that their research, achievements, and impact are judged more harshly than that of the men. Perhaps that is why women are not well represented among the recipients of ITSoc's highest honors and awards."

In this section, we ask a similar question for other venues. Specifically, we compile data on the fraction of women authors in award-winning papers in a number of top conferences in Computer Science in this decade (2010-2018). Fig. 8.1 plot shows the percentage of women authors among all authors in award-winning papers. Fig. 8.2 shows the percentage of award-winning papers with at least one woman author. For comparison, all except two award-winning papers (one from ACL and one from FOCS, both single-authored) have at least one male author.

The numbers are quite striking especially in venues on the left side of these plots. It is also important to note that the conferences which have a healthier distribution in a relative sense (those towards the right of the plots), still have the number of award-winning women authors less than 20%. Also interestingly, comparing conferences on similar topics, we see that STOC and FOCS are almost identical in these plots; on the other hand NeurIPS and ICML are significantly different ($p < 0.05$ for in percentage of women authors in award-winning papers; Fisher's two-sided exact test).

We do hope that the compilation of this data in the post will at least spur some conversation about this topic. Two specific topics of discussion are:

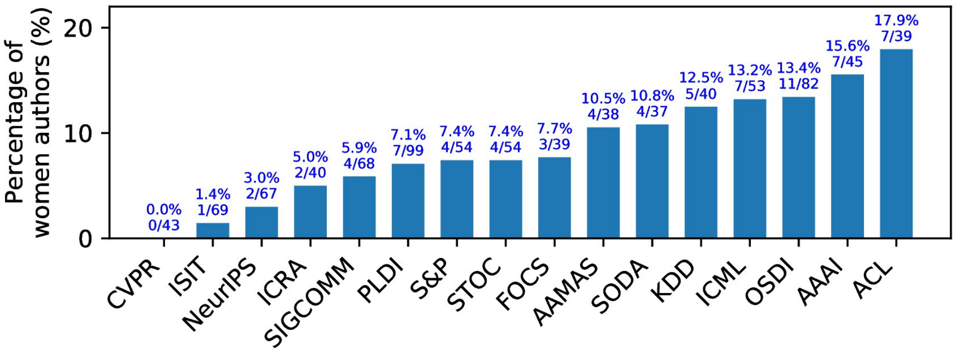


Figure 8.1: The percentage of women authors among all authors in award-winning papers.

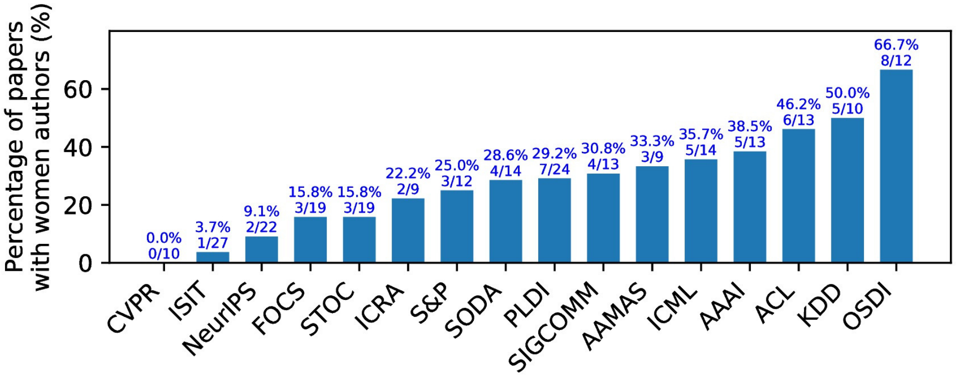


Figure 8.2: The percentage of award-winning papers with at leaset one women author.

1. Evaluation: It would be of interest to compare with the distribution of genders among the submitted papers, accepted papers, and papers nominated for the awards at these venues. Can we identify the main source(s) of this discrepancy in the peer-review pipeline — whether it is in the submissions itself, paper acceptance decisions in reviews, nominations for the awards, or in the final award decisions?
2. Transparency: The process of determining the awards is often not clear. For instance, in the conferences which adopt a double-blind policy, are the author identities visible to the award committee? How is the committee determined, and what criteria do they use? Finally, if these conferences release some details on why a certain paper was chosen for the award, it will not only provide some criteria for the award but also help motivate and guide budding researchers. For example, the award selection process along with remarks on the award-winning papers for ACL 2017 is described in the program chairs' blog [14].

Part IV

Proofs

Chapter 9

Proofs of Chapter 2

In this section, we present the proofs of our theoretical results.

For notational simplicity, we use "$1 \prec 2$" to denote that item 1 has a smaller value than item 2. Since the items have distinct values, we have $1 \prec 2$ if and only if $2 \succ 1$. For the 0-1 loss $L(\pi^*, \widehat{\pi}) = \mathbb{1}\{\widehat{\pi} \neq \pi^*\}$, we call the expected loss $\mathbb{E}[L(\pi^*, \widehat{\pi})] = \mathbb{P}(\widehat{\pi} \neq \pi^*)$ as the "probability of error" of any estimator $\widehat{\pi}$, and $\mathbb{P}(\widehat{\pi} = \pi^*)$ as the "probability of success". For the canonical setting and A/B testing, the probability of success of random guessing is 0.5. To show that some estimator $\widehat{\pi}$ strictly uniformly dominates random guessing for the canonical setting or A/B testing, we only need to show that the probability of success of this estimator is strictly higher than 0.5, or equivalently, the probability of error of of this estimator is strictly lower than 0.5.

9.1 Proof of Theorem 2.2

We prove that no deterministic cardinal estimator can strictly uniformly dominate the random-guessing estimator $\widehat{\pi}_{\text{can}}$, which implies the negative result for any deterministic ordinal estimator.

Recall the notation $\widehat{i}^{(1)} = \operatorname{argmax}_{i \in \{1,2\}} y_i$ as the item receiving the higher score (with ties broken uniformly at random), and the notation $\widehat{i}^{(2)}$ as the remaining item. First, we consider a deterministic estimator that always outputs $\widehat{i}^{(1)}$ as the item whose value is greater. We call this estimator the "sign estimator", denoted $\widehat{\pi}_{\text{sign}}$:

$$\widehat{\pi}_{\text{sign}}(A, y_1, y_2) = (\widehat{i}^{(1)} \succ \widehat{i}^{(2)}).$$

The proof consists of two steps. (1) We show that the sign estimator does not strictly uniformly dominate random guess. (2) Building on top of (1), we show that more generally, no deterministic estimator strictly uniformly dominates random guess.

Step 1: The sign estimator does not strictly uniformly dominate random guess.

We construct the following counterexample such that the probability of error of the sign estimator is 0.5. We construct reviewer calibration functions such that their ranges are disjoint, that is, one reviewer always gives a higher score than the other reviewer, regardless of the items they are assigned. Then the relative ordering of the two scores does not convey any information about the relative ordering of the two items, and we show that in this case, the sign estimator has a probability of error of 0.5. Concretely, let the item values be bounded as $x_1, x_2 \in (0, 1)$, and let the calibration functions be $f_1(x) = x$ and $f_2(x) = x + 1$. Then the score given by reviewer 2 is higher than the score given by reviewer 1 regardless of the item values they are assigned. The sign estimator always observes $y_1 < y_2$, and outputs the item assigned to reviewer 2 as the larger item. The assignment is either $A = (S_1 = 1, S_2 = 2)$ or $(S_1 = 2, S_2 = 1)$ with probability 0.5 each. Under assignment $(S_1 = 1, S_2 = 2)$, the sign estimator outputs $1 \prec 2$. Under assignment $(S_1 = 2, S_2 = 1)$, the sign estimator outputs $1 \succ 2$. Under one (and exactly one) of the two assignments, the output of the sign estimator is correct. Hence, the probability of error of the sign estimator is 0.5.

Step 2: No deterministic estimator strictly uniformly dominates random guess.

Let $\mathcal{A}$ be the set of the two assignments, $\mathcal{A} = \{(S_1 = 1, S_2 = 2), (S_1 = 2, S_2 = 1)\}$. A deterministic estimator $\widehat{\pi}_{\text{det}} : \mathcal{A} \times \mathbb{R} \times \mathbb{R} \to \{1 \succ 2, 1 \prec 2\}$ is a deterministic function that takes as input the assignment and the scores for the two items, and outputs the relative ordering between the two items. Step 1 has shown that the sign estimator does not strictly uniformly dominate random guess. Hence, we only need to prove that any deterministic estimator $\widehat{\pi}_{\text{det}}$ that is different from the sign estimator does not strictly uniformly dominate random guess. For this deterministic estimator $\widehat{\pi}_{\text{det}}$, there exist some input values $(a, \widetilde{y}_1, \widetilde{y}_2)$ such that the output of this deterministic estimator differs from the sign estimator. If the two estimators $\widehat{\pi}_{\text{sign}}$ and $\widehat{\pi}_{\text{det}}$ only differ at points where $\widetilde{y}_1 = \widetilde{y}_2$, then we can use the same counterexample in Step 1 to show that the probability of error of this deterministic estimator is 0.5. It remains to consider the case when $\widetilde{y}_1 \neq \widetilde{y}_2$. Without loss of generality, assume $\widetilde{y}_1 > \widetilde{y}_2$. Then consider the following counterexample. Let $x_1 > x_2$. Let f_1, f_2 be strictly-increasing functions such that $f_1(x_1) = f_2(x_1) = \widetilde{y}_1$, $f_1(x_2) = f_2(x_2) = \widetilde{y}_2$. Regardless of the reviewer assignment, the score y_1 for item 1 is $\widetilde{y}_1$, and the score y_2 for item 2 is $\widetilde{y}_2$. The item receiving a higher score is always $\widehat{i}^{(1)} = \operatorname{argmax}_{i \in \{1,2\}} y_i = 1$, so the sign estimator $\widehat{\pi}_{\text{sign}}$ always outputs $1 \succ 2$. Under assignment a, the deterministic estimator differs from the sign estimator, so the deterministic estimator gives the incorrect output $(1 \prec 2)$. The assignment a happens with probability 0.5, so the probability of error of this deterministic estimator is at least 0.5.

The two steps above complete the proof that there exists no deterministic estimator that strictly uniformly dominates random guess.

9.2 Proof of Theorem 2.3

In what follows, we prove that the probability of success of our estimator is strictly greater than 0.5 under arbitrary item values x_1, x_2 and arbitrary calibration functions f_1, f_2. We start with re-writing our estimator in (2.2) into an alternative and equivalent expression, and then prove the result on this new expression of our estimator.

Recall that $\widehat{i}^{(1)} = \operatorname{argmax}_{i \in \{1,2\}} y_i$ denotes the item receiving the higher score, and $\widehat{i}^{(2)}$ denotes the remaining item (with ties broken uniformly). Depending on the relative ordering of y_1 and y_2, we can split (2.2) into the following three cases:

$$\widetilde{\pi}^{\text{our}}_{\text{can}}(A, y_1, y_2 \mid y_1 > y_2) = \begin{cases} 1 \succ 2 & \text{with probability } \frac{1+w(y_1-y_2)}{2} \\ 2 \succ 1 & \text{otherwise,} \end{cases} \tag{9.1a}$$

$$\widetilde{\pi}^{\text{our}}_{\text{can}}(A, y_1, y_2 \mid y_1 < y_2) = \begin{cases} 1 \succ 2 & \text{with probability } \frac{1-w(y_2-y_1)}{2} \\ 2 \succ 1 & \text{otherwise,} \end{cases} \tag{9.1b}$$

$$\widetilde{\pi}^{\text{our}}_{\text{can}}(A, y_1, y_2 \mid y_1 = y_2) = \begin{cases} 1 \succ 2 & \text{with probability } \frac{1}{2} \\ 2 \succ 1 & \text{otherwise.} \end{cases} \tag{9.1c}$$

Recall that the function w is from $[0, \infty)$ to $[0, 1)$. Now we define the following auxiliary function $\widetilde{w} : \mathbb{R} \to (0, 1)$:

$$\widetilde{w}(x) = \begin{cases} \frac{1+w(x)}{2} & \text{if } x > 0 \\ \frac{1}{2} & \text{if } x = 0 \\ \frac{1-w(-x)}{2} & \text{otherwise.} \end{cases} \tag{9.2}$$

Combining (9.1) and (9.2), we have

$$\widetilde{\pi}^{\text{our}}_{\text{can}}(A, y_1, y_2) = \begin{cases} 1 \succ 2 & \text{with probability } \widetilde{w}(y_1 - y_2) \\ 2 \succ 1 & \text{otherwise.} \end{cases} \tag{9.3}$$

Without loss of generality, assume $x_1 > x_2$. The assignment is either $a := (S_1 = 1, S_2 = 2)$ or $a' := (S_1 = 2, S_2 = 1)$ with probability 0.5 each. Thus, the estimator observes $\{y_1 = f_1(x_1), y_2 = f_2(x_2)\}$ under assignment a, or $\{y_1 = f_2(x_1), y_2 = f_1(x_2)\}$ under assignment a'. The probability of success of our estimator $\widetilde{\pi}^{\text{our}}_{\text{can}}$ is:

$$\begin{aligned} \mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{can}} = \pi^*) &= \sum_{\widetilde{a} \in \{a, a'\}} \mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{can}} = \pi^* \mid A = \widetilde{a}) \mathbb{P}(A = \widetilde{a}) \\ &\overset{\text{(i)}}{=} \frac{1}{2}\widetilde{w}(f_1(x_1) - f_2(x_2)) + \frac{1}{2}\widetilde{w}(f_2(x_1) - f_1(x_2)) \\ &= \frac{1}{2}\left[\widetilde{w}(f_1(x_1) - f_2(x_2)) + \widetilde{w}(f_2(x_1) - f_1(x_2))\right] \\ &\overset{\text{(ii)}}{=} \frac{1}{2}\left[1 + \widetilde{w}(f_1(x_1) - f_2(x_2)) - \widetilde{w}(f_1(x_2) - f_2(x_1))\right], \end{aligned} \tag{9.4}$$

where step (i) is true by plugging in (9.3), and step (ii) is true because $\widetilde{w}(x) + \widetilde{w}(-x) = 1$ by the definition of the function $\widetilde{w}$ in (9.2).

By the monotonicity of the functions f_1 and f_2, and by the assumption that $x_1 > x_2$, we have $f_1(x_1) + f_2(x_1) > f_1(x_2) + f_2(x_2)$, and therefore $f_1(x_1) - f_2(x_2) > f_1(x_2) - f_2(x_1)$. Since $w(0) \geq 0$ and w is monotonically increasing on $[0, \infty)$, it is straightforward to verify that $\widetilde{w}$ is monotonically increasing on $\mathbb{R}$. Hence, we have

$$\widetilde{w}(f_1(x_1) - f_2(x_2)) > \widetilde{w}(f_1(x_2) - f_2(x_1)). \tag{9.5}$$

Combining (9.4) and (9.5), we have

$$\mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{can}} = \pi^*) > \frac{1}{2}.$$

9.3 Proof of Theorem 2.4

We construct a counterexample on which the mean, median and sign estimators have a probability of error of 0.5. In this counterexample, let the item values be bounded as $x_1, x_2 \in (0, 1)$, and let the m reviewer calibration functions be as follows:

$$f_j(x) = \begin{cases} x + (j-1) & \text{if } 1 \leq j \leq m-1 \\ x + \dfrac{m(m-1)}{2} & \text{if } j = m. \end{cases} \tag{9.6}$$

In these calibration functions, the score provided by each reviewer is the sum of the true value of the item assigned to this reviewer, and a bias term specific to this reviewer. The analysis is performed separately for the three estimators. At a high level, the analysis for the mean estimator uses the fact that one reviewer (specifically, reviewer m) has a significantly greater bias than the rest of the reviewers. The analysis for the median and the sign estimators uses the fact that the ranges of these calibration functions are disjoint.

Mean estimator: Recall that each reviewer is assigned one of the two items. Given any assignment, consider the item assigned to reviewer m. Trivially, the sum of the scores for this item must be strictly greater than $f_m(0) = \frac{m(m-1)}{2}$. Now consider the remaining item (not assigned to reviewer m). The sum of the scores for the remaining item can be at most $\sum_{j=1}^{m-1} f_j(1) = \sum_{j=1}^{m-1} j = \frac{m(m-1)}{2}$.

From these two bounds on the sum of the scores, an item has a greater sum of scores if and only if reviewer m is assigned to this item. By symmetry of the assignment, reviewer m is assigned to either item with probability 0.5. With the true ranking being either $1 \succ 2$ or $1 \prec 2$, the mean estimator makes an error in one of the two assignments, and this assignment happens with probability 0.5. Hence, the mean estimator makes an error with probability 0.5.

Median estimator and sign estimator: For the median estimator and the sign estimator, we first present an alternative view on the assignment, which is used for the analysis of both estimators. Recall that the assignment specifies $m/2$ reviewers to evaluate item 1, drawn uniformly at random without replacement, and the remaining $m/2$ reviewers to item 2. Equivalently, we can view this assignment as comprising the following two steps. (1) We sample uniformly at random a permutation of the m reviewers, denoted as a list $(j_1, \ldots, j_m)$. Define R and R' as the first half and second half of the reviewers in the list, $R = (j_1, \ldots, j_{\frac{m}{2}})$ and $R' = (j_{\frac{m}{2}+1}, \ldots, j_m)$. (2) We draw uniformly at random one of the two items, and assign the list R of reviewers to this item. Then assign the list R' of reviewers to the remaining item. For each $k \in [m/2]$, call reviewers $\{j_k, j_{\frac{m}{2}+k}\}$ as the k^{th} pair of reviewers.

For the median estimator and the sign estimator, we prove that given any arbitrary lists of reviewers R and R' in Step (1) of the assignment, the randomness in Step (2) yields the probability of error of the two estimators as 0.5.

Recall that the item values are bounded as $x_1, x_2 \in (0, 1)$. Since the biases of any two reviewers differ by at least 1 in Eq. (9.6), any reviewer j gives a higher score than any other reviewer j' if and only if $j < j'$, independent of the item values and the assignment. Formally, for any $x, x' \in (0, 1)$, and any $j, j' \in [m]$, we have

$$f_j(x) < f_{j'}(x') \quad \text{if and only if} \quad j < j'. \tag{9.7}$$

The remaining analysis is performed separately for the median estimator and the sign estimator.

Median estimator: Denote j_1^{med} and j_2^{med} as the indices of the reviewers providing the (upper) median scores in the set R_1 and R_2, respectively. From (9.7), we have

$$\begin{aligned} j_1^{\text{med}} &= \text{median}(j_1, \ldots, j_{\frac{m}{2}}) \\ j_2^{\text{med}} &= \text{median}(j_{\frac{m}{2}+1}, \ldots, j_m). \end{aligned} \tag{9.8}$$

Also from (9.7), the higher score in the two scores given by reviewer j_1^{med} and j_2^{med} is the reviewer with the larger index, $\max\{j_1^{\text{med}}, j_2^{\text{med}}\}$. In Step (2) of the assignment, reviewer j_1^{med} is assigned to item 1 or item 2 with equal probability. Hence, the probability of error of the median estimator is 0.5. This proves the claim that the (upper) median estimator does not strictly uniformly dominates random guess.

We now comment on using the median function defined as the lower median, or as the mean of the two middle values. For the lower median, the same argument as above applies. Now consider the median defined as the mean of the two middle values. When $m/2$ is odd, Eq. (9.8) still holds, and the argument as above still applies. When $m/2$ is even, the median value may not be equal to any scores from the reviewers. We construct a counterexample where the item values are still bounded as $x_1, x_2 \in (0, 1)$, and the calibration functions as follows:

$$f_j(x) = x + 2^j \qquad \text{for every } j \in [m].$$

With these calibration functions, for any $x, x', x'', x''' \in (0, 1)$, and any $j, j', j'', j''' \in [m]$, we have

$$f_j(x) + f_{j'}(x') < f_{j''}(x'') + f_{j'''}(x''') \quad \text{if and only if} \quad \max\{j, j'\} < \max\{j'', j'''\}.$$

Using this fact, we can show that the output of this median estimator only depends on reviewer indices and the realization of Step (2), independent of the item values. The probability of error of this median estimator is also 0.5.

Sign estimator: Denote a as the assignment that reviewers in R are assigned to item 1, and denote a' as the assignment that reviewers in R are assigned to item 2. For each $k \in [m/2]$, define $v_k \in \{0, 1\}$ as the binary value of whether the higher score in the k^{th} pair of scores comes from item 1, under assignment a. Set $v_k = 1$ if the higher score comes from item 1 and $v_k = 0$ otherwise. Define $v'_k \in \{0, 1\}$ similarly under assignment a'. Set $v'_k = 1$ if the higher score comes from item 1, and $v'_k = 0$ otherwise. Inequality (9.7) implies that $v_k + v'_k = 1$ for any $k \in [m/2]$. Define $v = \sum_{k=1}^{m/2} v_k$ as the count of pairwise wins for item 1 under assignment a, and define v' similarly. Then we have

$$v + v' = \frac{m}{2}. \tag{9.9}$$

The sign estimator outputs the item with more pairwise wins. That is, the sign estimator outputs item 1 under assignment a if $v > m/4$, outputs item 1 under assignment a' if $v' > m/4$, and outputs one of the two items uniformly at random if $v = m/4$ or $v' = m/4$. When $v = v' = m/4$, then under either assignment, the sign estimator has a tie, and hence outputs one of the two items uniformly at random. The probability of error of the sign estimator is 0.5. Otherwise, we have $v \neq m/4$. By (9.9), we have either $v > m/4 > v'$ or $v' > m/4 > v$. The sign estimator gives different outputs under the two assignments, out of which one and only one output is correct. The probability of error of the sign estimator is 0.5.

9.4 Proof of Theorem 2.5

Recall that a subset of $m/2$ reviewers, drawn uniformly at random without replacement, are assigned to item 1, and the remaining $m/2$ reviewers are assigned to item 2. We provide an alternative and equivalent view of the assignment as the following two steps:

(1) We sample two reviewers, uniformly at random without replacement, as the first pair of reviewers for the two items, and call them $\{j_1, j'_1\}$. Then sample two reviewers, uniformly at random without replacement, from the remaining $(m-2)$ reviewers as the second pair of reviewers for the two items, and call them $\{j_2, j'_2\}$. Continue until all m reviewers are exhausted, and call the subsequent pairs of reviewers $\{j_3, j'_3\}, \ldots, \{j_{m/2}, j'_{m/2}\}$.

(2) Within each pair, assign the pair of reviewers to the two items uniformly at random. That is, for each $k \in [m/2]$, assign reviewer j_k to one of the two items uniformly at random, and assign reviewer j'_k to the remaining item. The assignments are independent across pairs.

Consider any arbitrary values of items $x_1, x_2 \in \mathbb{R}$. Given any arbitrary realization of Step (1) of the assignment procedure described above, we apply Theorem 2.3 and show that on each pair of reviewers, the canonical estimator gives the correct output with probability strictly greater than 0.5. Then we show that combining the $m/2$ pairs by majority-voting yields probability of success strictly greater than 0.5.

Denote $\lambda(x_1, x_2, \{f, f'\})$ as the probability that our canonical estimator in Eq. (2.2) gives the correct output comparing items of values x_1, x_2 under reviewer calibration functions f, f'. In Step (2) of the assignment procedure described above, for any $k \in [m/2]$, consider the k^{th} pair of reviewers, $\{j_k, j'_k\}$. Suppose that the calibration functions of these two reviewers are denoted as $\{f, f'\}$. By Theorem 2.3, since the two reviewers are assigned to the two items uniformly at random, we have

$$\lambda\left(x_1, x_2, \{f, f'\}\right) > \frac{1}{2} \qquad \text{for all permissible } f, f'. \tag{9.10}$$

Let λ_{min} denote the probability of success of our canonical estimator when run on the worst pair of calibration functions among all pairs of reviewers

$$\lambda_{\text{min}} = \min_{f,f' \in \{f_1,\ldots,f_m\}} \lambda(x_1, x_2, \{f, f'\}) \overset{\text{(i)}}{>} \frac{1}{2},$$

where inequality (i) is true because of Eq. (9.10), and because the number of reviewers m is finite.

Now assume that we are given any arbitrary realization of Step (1) of the assignment. For each $k \in [m/2]$, define $V_k \in \{0, 1\}$ as the indicator variable of the correctness of our canonical estimator on the k^{th} pair of scores. We set $V_k = 1$ if the canonical estimator gives the correct output on the k^{th} pair, and 0 otherwise. Then V_k is a Bernoulli random variable with mean $\lambda(x_1, x_2, \{f_{j_k}, f_{j'_k}\}) \geq \lambda_{\text{min}}$. Moreover, since Step (2) of the assignment is performed independently across all pairs, the variables $\{V_j\}_{j=1}^{k}$ are independent given the item values and Step (1) of the assignment.

Let $V = \sum_{j=1}^{m/2} V_j$ be the number of pairs for which the canonical estimator $\widetilde{\pi}_{\text{can}}^{\text{our}}$ gives the correct output. Define a binomial random variable B with k trials and the success probability parameter λ_{min}. Then the random variable V stochastically dominates the random variable B. Recall that our estimator breaks ties uniformly at random. The probability of success of our estimator with the majority-voting procedure is then bounded as

$$\begin{aligned}
\mathbb{P}[V > \frac{k}{2}] + \frac{1}{2}\mathbb{P}[V = \frac{k}{2}] =& \frac{1}{2}\left(\mathbb{P}[V > \frac{k}{2}] + \mathbb{P}[V \geq \frac{k}{2}]\right) \\
\geq& \frac{1}{2}\left(\mathbb{P}[B > \frac{k}{2}] + \mathbb{P}[B \geq \frac{k}{2}]\right) \\
=& \mathbb{P}[B > \frac{k}{2}] + \frac{1}{2}\mathbb{P}[B = \frac{k}{2}] \\
\overset{\text{(i)}}{>}& \frac{1}{2},
\end{aligned}$$

where inequality (i) is true because the success probability parameter λ_{min} of the binomial variable is strictly greater than $\frac{1}{2}$.

We complete the proof that the probability of success of our estimator is strictly greater than 0.5 uniformly on any item values x_1, x_2 and any permissible calibration functions $\{f_j\}_{j=1}^{m}$.

9.5 Proof of Theorem 2.6

We first provide a high-level description of the proof. We call a pair of items "flippable", if Algorithm 1 uses the canonical estimator to decide the relative ordering of this pair (that is, the if-condition in Line 6 in Algorithm 1 is true). Note that a "flippable" pair may or may not be flipped by the algorithm, as the outcome depends on the output of the canonical estimator. In Theorem 2.3, we show that our canonical estimator $\widetilde{\pi}_{\text{can}}^{\text{our}}$ predicts the relative ordering of a pair of items correctly with probability strictly greater than 0.5. The main idea of the proof is to apply Theorem 2.3 to each flippable pair. Then we show that an improvement on the probability of correctness on these flippable pairs translates to an improvement on the probability of success of exact recovery.

Theorem 2.3 requires that within each pair, the two reviewers are assigned the two items uniformly at random. To be able to apply this theorem, we separate the different sources of randomness in the joint procedure of the assignment and the algorithm. We derive an equivalent algorithm by re-ordering the steps of Algorithm 1, so that in this equivalent algorithm, given any flippable pair of items and two reviewers evaluating this pair, the last sources of randomness comes from the random assignment of the two reviewers to the two items within this pair.

We introduce some additional notation for our re-ordered algorithm. Recall the notation of $A = (S_1, \ldots, S_m)$ for the reviewer assignment, where S_j is a pair of items assigned to reviewer j for each $j \in [m]$. Denote $\mathcal{Q} = \{\widetilde{S}_j\}_{j=1}^m$ as the same m pairs of items, but the reviewer assigned to each pair $\widetilde{S}_j$ is unspecified. Now we present an equivalent joint procedure of the assignment and the cardinal estimator $\widetilde{\pi}^{\text{our}}_{\text{rank}}$ in Algorithm 4. In what follows, we provide a high-level summary of Algorithm 4:

(1) *Line 1-2:* We sample m pairwise comparisons of the items, drawn uniformly at random without replacement from the $\binom{n}{2}$ pairs. Obtain an initial estimate $\widehat{\pi}$ of the ranking, by computing a topological ordering on the graph $\mathcal{G}(\mathcal{B})$.

(2) *Line 3-18:* Store the positions of all flippable pairs (if any) determined by Algorithm 1. If an item is included in some flippable pair, then this item is matched to a distinct pairwise comparison in $\mathcal{Q}$. Store the matching between the items in flippable pairs and the pairwise comparisons.

(3) *Line 19:* For the two pairwise comparisons associated with each pair of flippable items, sample two reviewers uniformly at random without replacement to evaluate the two comparisons.

(4a) *Line 20-21:* Within each flippable pair, assign the two reviewers to the two items uniformly at random.

(4b) *Line 22-28:* Run the canonical estimator on each flippable pair, and flip the pair if the canonical estimator decides to do so (Line 23-26). After all flippable pairs are examined, output the final ranking $\widehat{\pi}$.

We now briefly discuss the equivalence of Algorithm 4 to Algorithm 1. We first discuss the equivalence of the assignment procedures in the two algorithms, and then the estimation aspect in the next paragraph. The assignment consists of Steps (1), (3) and (4a). Recall that the assignment in Algorithm 1 samples m pairwise comparisons, uniformly at random without replacement, to assign to the m reviewers. In Algorithm 4, this assignment is decomposed into the choice of pairwise comparisons, the choice of a pair of reviewers to two pairwise comparisons in each flippable pair, and the assignment within each flippable pair, corresponding to Steps (1), (3) and (4a), respectively. Note that only the selected pairwise comparison for each item within some flippable pair is used for Algorithm 4, so we do not need to specify the assignment of the reviewers for the rest of the comparisons. This re-ordering of the assignment is equivalent to Algorithm 1.

The cardinal ranking estimator consists of the rest of the steps, namely Steps (2) and (4b). In the original presentation of the estimator in Algorithm 1, the estimator scans through the items, identifies flippable pairs, calls the canonical estimator on each flippable pair, and flips the pairs accordingly. Note that the identification of flippable pairs does not need the assignment of reviewers or the scores from the reviewers, so Algorithm 4 first scans through the items and identifies all flippable pairs, without using the choice of the reviewers in the assignment or using the scores from the reviewers. Then Algorithm 4 calls the canonical estimator on each flippable pair once the choice of the reviewers and the scores are determined, and flips each pair based on the corresponding output from the canonical estimator. Note that when checking for a flippable pair (the if-condition in Line 6 in Algorithm 1 and Line 9 in Algorithm 4), Algorithm 1 checks whether the flipped ranking $\widehat{\pi}_{\text{flip}}$ is a topological ordering, where the previous flippable pairs in $\widehat{\pi}_{\text{flip}}$ may have already been flipped. In Algorithm 4, the previous flippable pairs are identified but are not flipped. However, whether the flipped ranking $\widehat{\pi}_{\text{flip}}$ is a topological ordering is independent of whether the previous flippable pairs in $\widehat{\pi}_{\text{flip}}$ are flipped. Hence, the identification of the flippalbe pairs is equivalent in the two algorithms. The re-ordering of the steps of the cardinal estimator $\widetilde{\pi}^{\text{our}}_{\text{rank}}$ is valid.

Having now established the equivalence of Algorithm 4 to Algorithm 1, we now prove Theorem 2.6 with respect to Algorithm 4. Let us denote $\widetilde{\pi}^{\text{eq}}_{\text{rank}}$ as the cardinal estimator in Algorithm 4. Denote $\text{topo}(\mathcal{B})$ as the set of all topological orderings on the directed graph $\mathcal{G}(\mathcal{B})$ induced by the set of ordinal observations $\mathcal{B}$. We denote a random variable $T(\mathcal{B}) := |\text{topo}(\mathcal{B})|$ as the number of such topological orderings. Note that the definition of flippable pairs carries over from Algorithm 1 to Algorithm 4. We denote a random variable L as the number of flippable pairs in Algorithm 4.

Let us first consider the probability of success of the ordinal estimator. The following lemma describes the posterior distribution of the true ranking conditioned on the set of ordinal observations $\mathcal{B}$. Using this posterior distribution, the optimal ordinal estimators and their probability of success are derived.

Lemma 9.1. *(a) Given any possible set of ordinal observations β, the posterior distribution of the true ranking π^* is uniformly distributed over the $T(\beta)$ topological orderings:*

$$\mathbb{P}(\pi^* = \pi \mid \mathcal{B} = \beta) = \begin{cases} \frac{1}{T(\beta)} & \textit{if } \pi \in \text{topo}(\beta) \\ 0 & \textit{otherwise}. \end{cases} \tag{9.11}$$

(b) Any ordinal estimator $\widehat{\pi}^{opt}_{rank}$ is optimal for the 0-1 loss, if and only if given any set of ordinal observations β, the output of this ordinal estimator belongs to the $T(\beta)$ topological orderings with probability 1, that is, if and only if

$$\mathbb{P}(\widehat{\pi}^{opt}_{rank}(\beta) \in \text{topo}(\beta) \mid \mathcal{B} = \beta) = 1 \qquad \textit{for all possible set of ordinal observations } \beta. \tag{9.12}$$

Moreover, conditioned on the set of ordinal observations β, the probability of success of any optimal ordinal estimator $\widehat{\pi}^{opt}_{rank}$ is

$$\mathbb{P}(\widehat{\pi}^{opt}_{rank} = \pi^* \mid \mathcal{B} = \beta) = \frac{1}{T(\beta)}. \tag{9.13}$$

See Section 9.5.1 for the proof of the lemma.

Now consider the probability of success of our cardinal estimator $\widetilde{\pi}^{\text{eq}}_{\text{rank}}$ from Algorithm 4. We write the probability of success of our cardinal estimator as

$$\mathbb{P}(\widetilde{\pi}^{\text{eq}}_{\text{rank}} = \pi^*) = \sum_{\beta} \sum_{\ell} \mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{rank}} = \pi^* \mid \mathcal{B} = \beta, L = \ell)\mathbb{P}(\mathcal{B} = \beta, L = \ell), \tag{9.14}$$

where β is summed over all possible sets of ordinal observations, and ℓ is summed from 0 to $\lfloor n/2 \rfloor$.

We consider each term $\mathbb{P}(\widetilde{\pi}^{\text{eq}}_{\text{rank}} = \pi^* \mid \mathcal{B} = \beta, L = \ell)$ separately for each β and ℓ. We prove that for any β and any ℓ, the probability of success of our cardinal estimator is greater than or equal to the probability of success of any optimal ordinal estimator $\widehat{\pi}^{\text{opt}}_{\text{rank}}$. We also show that the probability of success is strictly greater for some β and ℓ. We discuss the following two cases separately, depending on the number of flippable pairs.

Case 1: $\ell = 0$.

We have the number of flippable pairs $L = 0$ either if there is a unique topological ordering on the graph $\mathcal{G}(\mathcal{B})$, or if in each pair of adjacent items that can be flipped without violating pairwise comparisons, at least one item in this pair does not have any score. Note that these two conditions are fully determined by the set of ordinal observations. Hence, conditioned on the set of ordinal observations $\mathcal{B} = \beta$, the event of $L = 0$ is fully determined, and is independent of everything else given $\mathcal{B}$.

The initial estimated ranking of the cardinal estimator is a topological ordering (Line 2 of Algorithm 4). Since there is no flippable pair, the cardinal estimator simply outputs this topological ordering. For any set of ordinal observations β such that $\mathbb{P}(\mathcal{B} = \beta, L = 0) > 0$, we have

$$\begin{aligned}\mathbb{P}(\widetilde{\pi}^{\text{eq}}_{\text{rank}} = \pi^* \mid \mathcal{B} = \beta, L = 0) &\overset{\text{(i)}}{=} \mathbb{P}(\widetilde{\pi}^{\text{eq}}_{\text{rank}} = \pi^* \mid \mathcal{B} = \beta)\\ &\overset{\text{(ii)}}{=} \mathbb{P}(\widehat{\pi}^{\text{opt}}_{\text{rank}} = \pi^* \mid \mathcal{B} = \beta),\end{aligned} \tag{9.15}$$

where $\widehat{\pi}^{\text{opt}}_{\text{rank}}$ denotes any optimal ordinal estimator. Here in (9.15), equality (i) is true because the event $L = 0$ is fully determined by $\mathcal{B}$, and equality (ii) is true because this cardinal estimator that simply outputs a topological ordering is equivalent to an ordinal estimator that outputs the same topological ordering. From (9.12), this ordinal estimator is one optimal ordinal estimator.

Case 2: $\ell > 0$.

In this case, Algorithm 4 identifies at least one flippable pair. The probability of success of our cardinal estimator is

$$\begin{aligned}\mathbb{P}(\widetilde{\pi}^{\text{eq}}_{\text{rank}} = \pi^* \mid \mathcal{B} = \beta, L = \ell) &= \sum_{\pi \in \Pi} \mathbb{P}(\widetilde{\pi}^{\text{eq}}_{\text{rank}} = \pi \mid \pi^* = \pi, \mathcal{B} = \beta, L = \ell)\mathbb{P}(\pi^* = \pi \mid \mathcal{B} = \beta, L = \ell)\\ &\overset{\text{(i)}}{=} \sum_{\pi \in \Pi} \mathbb{P}(\widetilde{\pi}^{\text{eq}}_{\text{rank}} = \pi \mid \pi^* = \pi, \mathcal{B} = \beta, L = \ell)\mathbb{P}(\pi^* = \pi \mid \mathcal{B} = \beta)\\ &\overset{\text{(ii)}}{=} \frac{1}{T(\beta)} \sum_{\pi \in \text{topo}(\beta)} \mathbb{P}(\widetilde{\pi}^{\text{eq}}_{\text{rank}} = \pi \mid \pi^* = \pi, \mathcal{B} = \beta, L = \ell),\end{aligned} \tag{9.16}$$

where equality (i) is true because L is independent of π^* conditioned on $\mathcal{B}$. Equality (ii) is true by plugging in (9.11).

In Algorithm 4, Lines 1-19 fully determine the number of the flippable pairs, their positions, and the two reviewers evaluating each flippable pair. In Lines 20-28, within each flippable pair, the algorithm first assigns uniformly at random one reviewer to one item and the remaining reviewer to the remaining item, and then calls the canonical estimator to output the relative ordering of this pair. Conditioned on the randomness in Lines 1-19 of Algorithm 4, we now apply Theorem 2.3 to each flippable pair. Since the reviewer assignment within each flippable pair (Line 21) is uniformly at random, by Thoerem 2.3, the probability that the canonical estimator outputs the correct relative ordering of each flippable pair is strictly greater than $\frac{1}{2}$. Since the assignment within each flippable pair is independent across pairs, the probability that the canonical estimator outputs the correct relative ordering of all ℓ flippable pairs is strictly greater than $\frac{1}{2^\ell}$.

Recall that the initial estimated ranking of our cardinal estimator is a topological ordering. Consider all total rankings that are identical to this initial ranking, except for (possibly) the relative ordering of the ℓ flippable pairs. Since the items in the flippable pairs are disjoint, there are 2^ℓ such total rankings. By definition, a pair is flippable only if the total ranking after this pair is flipped is still a topological ordering. Hence, all these 2^ℓ total rankings are topological orderings on the graph $\mathcal{G}(\mathcal{B})$. In (9.16), the summation of π is over all topological orderings. In particular, this summation includes these 2^ℓ total rankings. On each of these 2^ℓ total rankings, the output of our ranking estimator $\widetilde{\pi}^{\text{eq}}_{\text{rank}}$ is correct if and only if the canonical estimator outputs the correct relative orderings for the ℓ flippable pairs, which happens with probability strictly greater than $\frac{1}{2^\ell}$. Hence, we bound (9.16) as

$$\begin{aligned}\mathbb{P}(\widetilde{\pi}^{\text{eq}}_{\text{rank}} = \pi^* \mid \mathcal{B} = \beta, L = \ell) > \frac{1}{T(\beta)} \cdot 2^\ell \cdot \frac{1}{2^\ell} =& \frac{1}{T(\beta)}\\ \overset{\text{(i)}}{=}& \mathbb{P}(\widehat{\pi}^{\text{opt}}_{\text{rank}} = \pi^* \mid \mathcal{B} = \beta),\end{aligned} \tag{9.17}$$

where $\widehat{\pi}^{\text{opt}}_{\text{rank}}$ denotes any optimal ordinal estimator. Equality (i) is true because of (9.13) in Lemma 9.1.

Plugging (9.15) and (9.17) into (9.14), we have

$$\mathbb{P}(\widetilde{\pi}^{\text{eq}}_{\text{rank}} = \pi^*) \geq \mathbb{P}(\widehat{\pi}^{\text{opt}}_{\text{rank}} = \pi^*) \qquad \text{for any optimal ordinal estimator } \widehat{\pi}^{\text{opt}}_{\text{rank}}, \tag{9.18}$$

and a strict inequality holds in (9.18) if there exists some β and some $\ell > 0$, such that

$$\mathbb{P}(\mathcal{B} = \beta, L = \ell) > 0. \tag{9.19}$$

It remains to find some β and some $\ell > 0$ such that (9.19) is true. We construct such β and $\ell > 0$ as follows. Consider the true ranking $1 \succ 2 \succ \cdots \succ n$, which happens with a strictly positive probability as the prior distribution of the true ranking is uniform. Conditioned on this true ranking, consider the event that the sampled pairwise comparisons in $\mathcal{Q}$ do not include a direct comparison between items 1 and 2, but both item 1 and item 2 have at least one score each (from comparisons with at least one of the remaining $(n-2)$ items). Recall that the number of pairs satisfies $1 < m < \binom{n}{2}$, so such a set $\mathcal{Q}$ of pairwise comparisons arises with a strictly positive probability. Let β be the set of ordinal observations derived from the true ranking and the set $\mathcal{Q}$ of pairwise comparisons described as above. With this β, item 1 and item 2 are the first two items in the initial ranking of the topological ordering, they can be flipped, and they both have some scores. Hence, item 1 and item 2 form a flippable pair, and we have $L > 0$. Hence, with this construction of β, we have

$$\sum_{\ell=1}^{\lfloor n/2 \rfloor} \mathbb{P}(\mathcal{B} = \beta, L = \ell) > 0.$$

Thus there exists some $\ell > 0$ such that $\mathbb{P}(\mathcal{B} = \beta, L = \ell) > 0$. Hence, Equation (9.19) is true, implying the strictly inequality in (9.18). Consequently, the probability of success of our cardinal ranking estimator $\widetilde{\pi}^{\text{eq}}_{\text{rank}}$ is strictly uniformly greater than the probability of success of any optimal ordinal estimator. By the equivalence of Algorithm 4 and Algorithm 1, this result also holds for the original cardinal estimator $\widetilde{\pi}^{\text{our}}_{\text{rank}}$, completing the proof.

9.5.1 Proof of Lemma 9.1

We first prove part (a) of the lemma. By Bayes rule, for any ranking $\pi \in \Pi$ and any possible set of ordinal observations β, we have

$$\mathbb{P}(\pi^* = \pi \mid \mathcal{B} = \beta) = \frac{\mathbb{P}(\mathcal{B} = \beta \mid \pi^* = \pi)\mathbb{P}(\pi^* = \pi)}{\mathbb{P}(\mathcal{B} = \beta)}. \tag{9.20}$$

Given the set of ordinal observations β, the denominator in (9.20) is independent of π. Since the prior of the true ranking is uniform, in the numerator we have $\mathbb{P}(\pi^* = \pi) = \frac{1}{n!}$, independent of π. Now it remains to consider the term $\mathbb{P}(\mathcal{B} = \beta \mid \pi^* = \pi)$ in the numerator. Recall the notation of the random variable $\mathcal{Q}$ as the set of pairwise comparisons in the ordinal observations (but $\mathcal{Q}$ does not include the results of the relative orderings of these pairs). We write the term $\mathbb{P}(\mathcal{B} = \beta \mid \pi^* = \pi)$ as

$$\begin{aligned}\mathbb{P}(\mathcal{B} = \beta \mid \pi^* = \pi) =& \sum_q \mathbb{P}(\mathcal{B} = \beta \mid \mathcal{Q} = q, \pi^* = \pi)\mathbb{P}(\mathcal{Q} = q \mid \pi^* = \pi)\\ \overset{\text{(i)}}{=}& \sum_q \mathbb{P}(\mathcal{B} = \beta \mid \mathcal{Q} = q, \pi^* = \pi)\mathbb{P}(\mathcal{Q} = q),\end{aligned} \tag{9.21}$$

where q is summed over all possible sets of m pairwise comparisons. Equality (i) is true because the sampling of the set of pairwise comparisons $\mathcal{Q}$ is independent of the true ranking π^*.

Recall that the set of ordinal observations β includes the pairwise comparisons and results of the relative orderings of these pairwise comparisons, whereas q only includes the pairwise comparisons themselves, so β fully determines q. For this term to be non-zero, the set of pairwise comparisons indicated by β and the set of pairwise comparisons indicated by q need to be identical. Hence, there is only one q in the summation of (9.21) consistent with β, and we denote $\widetilde{q}$ as the set of pairs consistent with β. Then (9.21) reduces to

$$\mathbb{P}(\mathcal{B} = \beta \mid \pi^* = \pi) =\mathbb{P}(\mathcal{B} = \beta \mid \mathcal{Q} = \widetilde{q}, \pi^* = \pi)\mathbb{P}(\mathcal{Q} = \widetilde{q}), \tag{9.22}$$

In (9.22), the second term $\mathbb{P}(\mathcal{Q} = \widetilde{q})$ is independent of π. Now consider the first term $\mathbb{P}(\mathcal{B} = \beta \mid \mathcal{Q} = \widetilde{q}, \pi^* = \pi)$. If π is a topological ordering on $\mathcal{G}(\beta)$, then by definition, the relative orderings on the pairs $\widetilde{q}$ induced by the ranking π is the set of ordinal observations β. If π is not a topological ordering, then by definition, the relative orderings induced by the ranking π violates at least one relative ordering in β. Hence,

$$\mathbb{P}(\mathcal{B} = \beta \mid \mathcal{Q} = \widetilde{q}, \pi^* = \pi) = \begin{cases} 1 & \text{if } \pi \in \text{topo}(\beta) \\ 0 & \text{otherwise.} \end{cases} \tag{9.23}$$

Combining the law of total probability with (9.20), (9.22) and (9.23), the posterior distribution of the true ranking is

$$\mathbb{P}(\pi^* = \pi \mid \mathcal{B} = \beta) = \begin{cases} \frac{1}{T(\beta)} & \text{if } \pi \in \text{topo}(\beta) \\ 0 & \text{otherwise.} \end{cases} \tag{9.24}$$

Conditioned on the set of ordinal observations β, the posterior distribution of the true ranking is uniform over all topological ordering on the graph $\mathcal{G}(\beta)$. This completes the proof for part (a) of the lemma.

For part (b) of the lemma, we condition on any possible set of ordinal observations β. On the input β, the probability of success of any (possibly-randomized) ordinal estimator $\widehat{\pi}_{\text{rank}}$ is:

$$\begin{aligned}\mathbb{P}(\widehat{\pi}_{\text{rank}}(\beta) = \pi^* \mid \mathcal{B} = \beta) &= \sum_{\pi \in \Pi} \mathbb{P}(\widehat{\pi}_{\text{rank}}(\beta) = \pi \mid \pi^* = \pi, \mathcal{B} = \beta) \mathbb{P}(\pi^* = \pi \mid \mathcal{B} = \beta) \\ &\overset{\text{(i)}}{=} \frac{1}{T(\beta)} \sum_{\pi \in \text{topo}(\beta)} \mathbb{P}(\widehat{\pi}_{\text{rank}}(\beta) = \pi \mid \pi^* = \pi, \mathcal{B} = \beta) \\ &\overset{\text{(ii)}}{=} \frac{1}{T(\beta)} \sum_{\pi \in \text{topo}(\beta)} \mathbb{P}(\widehat{\pi}_{\text{rank}}(\beta) = \pi) \\ &\overset{\text{(iii)}}{\leq} \frac{1}{T(\beta)},\end{aligned} \tag{9.25}$$

where equality (i) is true by plugging in (9.24). Equality (ii) is true because the output of the ordinal estimator $\widehat{\pi}_{\text{rank}}(\beta)$ on the input β only depends on its internal randomness, and hence independent of the π^* and $\mathcal{B}$. Inequality (iii) is true by the law of total probability. In particular, the equality sign in (iii) holds if and only if the output of the ordinal estimator is always a topological ordering consistent with β, that is, if and only if

$$\mathbb{P}(\widehat{\pi}_{\text{rank}}(\beta) \in \text{topo}(\beta) \mid \mathcal{B} = \beta) = 1. \tag{9.26}$$

Taking an expectation over all possible ordinal observations β, we have

$$\mathbb{P}(\widehat{\pi}_{\text{rank}}(\mathcal{B}) = \pi^*) = \sum_{\beta} \mathbb{P}(\widehat{\pi}_{\text{rank}}(\beta) = \pi^* \mid \mathcal{B} = \beta) \mathbb{P}(\mathcal{B} = \beta). \tag{9.27}$$

Combining (9.27) with the condition (9.26) for the equality sign in (9.25), an ordinal estimator is optimal if and only if Eq. (9.26) holds on all possible ordinal observations β where $\mathbb{P}(\mathcal{B} = \beta) > 0$. This completes the proof for part (b) of the lemma.

9.6 Proof of Theorem 2.7

The proof is a slight modification to the proof of Theorem 2.3, so we only highlight the difference.

Recall that ϵ_{ij} denotes the noise in the reported score of reviewer $j \in \{1, 2\}$ for item $i \in \{1, 2\}$. In Eq. (9.4) from the proof of Theorem 2.3, we replace all the noiseless terms $f_j(x_i)$ by the noisy terms $f_j(x_i) + \epsilon_{ij}$ for each $i \in \{1, 2\}$ and $j \in \{1, 2\}$. Using the fact that the noise terms are independent of everything else, and taking an expectation over all the noise terms, we have

$$\begin{aligned}\mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{can}} = \pi^*) =& \frac{1}{2}\mathbb{E}_{\epsilon_{11},\epsilon_{12},\epsilon_{21},\epsilon_{22}}\left[1 + \widetilde{w}((f_1(x_1) + \epsilon_{11}) - (f_2(x_2) + \epsilon_{22})) - \widetilde{w}((f_1(x_2) + \epsilon_{21}) - (f_2(x_1) + \epsilon_{12}))\right] \\ =& \frac{1}{2}\mathbb{E}_{\epsilon_{11},\epsilon_{12},\epsilon_{21},\epsilon_{22}}\left[1 + \widetilde{w}(f_1(x_1) - f_2(x_2) + \epsilon_{11} - \epsilon_{22}) - \widetilde{w}(f_1(x_2) - f_2(x_1) + \epsilon_{21} - \epsilon_{12})\right] \\ \overset{\text{(i)}}{=}& \frac{1}{2}\mathbb{E}_{\epsilon_1,\epsilon_2}\left[1 + \widetilde{w}(f_1(x_1) - f_2(x_2) + \epsilon_1 - \epsilon_2) - \widetilde{w}(f_1(x_2) - f_2(x_1) + \epsilon_1 - \epsilon_2)\right],\end{aligned} \tag{9.28}$$

where step (i) uses linearity of expectation with a change of variable names, as the noise terms ϵ_{ij} are i.i.d.

Without loss of generality, assume $x_1 > x_2$. Recall from the proof of Theorem 2.3 that $f_1(x_1) - f_2(x_2) > f_1(x_2) - f_2(x_1)$, and therefore we have the deterministic inequality

$$f_1(x_1) - f_2(x_2) + \epsilon_1 - \epsilon_2 > f_1(x_2) - f_2(x_1) + \epsilon_1 - \epsilon_2, \quad \text{for any } \epsilon_1, \epsilon_2 \in \mathbb{R}.$$

Using the monotonicity of $\widetilde{w}$, we have

$$\widetilde{w}(f_1(x_1) - f_2(x_2) + \epsilon_1 - \epsilon_2)) > \widetilde{w}(f_1(x_2) - f_2(x_1) + \epsilon_1 - \epsilon_2). \tag{9.29}$$

Taking an expectation over ϵ_1 and ϵ_2 in (9.29) and combining with (9.28) gives

$$\mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{can}} = \pi^*) > \frac{1}{2}.$$

9.7 Proof of Theorem 2.8

We first present the construction of a cardinal estimator $\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}$, which has access to one call to any arbitrary ordinal estimator $\widehat{\sigma}_{\text{rank}}$. For any $i, i' \in [n]$ with $i \neq i'$, we call the pair of items (i, i') "topologically-identical" under the set of ordinal observations $\mathcal{B}$, if the following conditions are met. There is no direct comparison between item i and item i' in $\mathcal{B}$. For any item $\ell \notin \{i, i'\}$, the set $\mathcal{B}$ includes a comparison between item i and item ℓ, if and only if the set $\mathcal{B}$ includes a comparison between item i' and item ℓ. Moreover, if two comparisons (i, ℓ) and (i', ℓ) are in the set $\mathcal{B}$, their comparison results are the same, that is, $\mathbb{1}\{i \succ \ell\} = \mathbb{1}\{i' \succ \ell\}$. Note that it is possible that item i is compared to no item in $\mathcal{B}$ (and hence item i' is also compared to no item).

For any item $i \in [n]$ and any possible set of ordinal observations $\mathcal{B}$, we define the following sets:

$$\begin{aligned}V^+(i, \mathcal{B}) =& \{\ell \in [n], \ell \neq i \mid \text{there exists a directed path from item } \ell \text{ to item } i \text{ in the graph } \mathcal{G}(\mathcal{B})\} \\ V^-(i, \mathcal{B}) =& \{\ell \in [n], \ell \neq i \mid \text{there exists a directed path from item } i \text{ to item } \ell \text{ in the graph } \mathcal{G}(\mathcal{B})\}.\end{aligned}$$

In words, $V^+(i, \mathcal{B})$ is the set of items that are ranked higher than item i according to the set of ordinal observations $\mathcal{B}$, and $V^-(i, \mathcal{B})$ is the set of items that are ranked lower than item i. For any topologically-identical pair (i, i'), we have $V^+(i, \mathcal{B}) = V^+(i', \mathcal{B})$ and $V^-(i, \mathcal{B}) = V^-(i', \mathcal{B})$, so we denote $V^+(i, i', \mathcal{B}) := V^+(i, \mathcal{B})$ and $V^-(i, i', \mathcal{B}) := V^-(i, \mathcal{B})$. Now we present a cardinal estimator $\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}$ in Algorithm 5.

In words, Algorithm 5 obtains an initial estimated ranking $\widehat{\sigma}_{\text{init}}$ by making one call to the given ordinal estimator $\widehat{\sigma}_{\text{rank}}$. Then Algorithm 5 identifies two items that are topologically-identical. If such a topologically-identical pair (i, i') exists, we perform the following two steps on this topologically-identical pair:

(1) *Line 5-13:* Using the set of ordinal observations $\mathcal{B}$, we obtain a new ranking $\widehat{\sigma}$ by re-arranging the items in the initial estimated ranking $\widehat{\sigma}_{\text{init}}$. In this new ranking $\widehat{\sigma}$, we keep all items outside $V^+ \cup V^- \cup \{i, i'\}$ in the same positions as they are in $\widehat{\sigma}_{\text{init}}$. We re-arrange the items in $V^+ \cup V^- \cup \{i, i'\}$, so that they occupy the remaining positions; the set V^+ is ranked higher than items $\{i, i'\}$, and the set V^- is ranked lower than items $\{i, i'\}$. Moreover, the relative ranking of items within each set (V^+, V^- or $\{i, i'\}$) is preserved. That is, if $\ell, \ell' \in V$ with some $V \in \{V^+, V^-, \{i, i'\}\}$, we have $\widehat{\sigma}(\ell) < \widehat{\sigma}(\ell')$ if and only if $\widehat{\sigma}_{\text{init}}(\ell) < \widehat{\sigma}_{\text{init}}(\ell')$.

(2) *Line 14-18:* We sample uniformly at random a score for each item in the topologically-identical pair (i, i'). Based on this pair of scores, we call the canonical estimator to decide the relative ordering of these two items. Depending on the outcome of the canonical estimator, we keep the relative ordering of these two items unchanged, or flip the two items accordingly.

This completes the description of the cardinal estimator $\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}$.

We now show that the cardinal estimator $\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}$ takes polynomial time in the number of items n, in addition to the time taken by one call to the given ordinal estimator $\widehat{\sigma}_{\text{rank}}$. To check if a pair of items (i, i') is topologically-identical, it takes polynomial time to go through the pairwise comparisons in $\mathcal{B}$. Hence, it takes polynomial time to identify a topologically-identical pair (or determine that such a pair does not exist). For any topologically-identical pair, in the re-arranging step, the set $V^-(i, i', \mathcal{B})$ can be found by a graph traversal from node i. The set $V^+(i, i', \mathcal{B})$ can be found by a graph traversal from node i on the graph $\mathcal{G}(\mathcal{B})$ but with all edges reversed. Both traversals take polynomial time. Hence, Algorithm 5 takes polynomial time, in addition to one call to the ordinal estimator $\widehat{\sigma}_{\text{rank}}$.

We now present the proof for the uniform strict dominance of the cardinal estimator $\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}$ over the given ordinal estimator $\widehat{\sigma}_{\text{rank}}$. Given any two rankings σ_1, σ_2 and any two items (i, i'), we denote $\alpha(\sigma_1, \sigma_2, i, i') := \mathbb{1}\{\mathbb{1}\{\sigma_1(i) > \sigma_1(i')\} \neq \mathbb{1}\{\sigma_2(i) > \sigma_2(i')\}\}$ as Kendall-tau distance between σ_1 and σ_2 contributed by the pair of items (i, i'). Then we can write Kendall-tau distance between σ_1, σ_2 as

$$\begin{aligned} L_{\text{KT}}(\sigma_1, \sigma_2) &= \sum_{\substack{i \in [n], i' \in [n]: \\ \sigma_1(i) < \sigma_2(i')}} \mathbb{1}\{\sigma_2(i) > \sigma_2(i')\} \\ &= \sum_{1 \leq i < i' \leq n} \mathbb{1}\{\mathbb{1}\{\sigma_1(i) > \sigma_1(i')\} \neq \mathbb{1}\{\sigma_2(i) > \sigma_2(i')\}\} \\ &= \sum_{1 \leq i < i' \leq n} \alpha(\sigma_1, \sigma_2, i, i'). \end{aligned} \tag{9.30}$$

For Spearman's footrule dsitance, for each item $i \in [n]$, we call the term $|\sigma_1(i) - \sigma_2(i)|$ as Spearman's footrule distance between σ_1 and σ_2 contributed by item i.

We analyze Step (1) of re-arranging the items and Step (2) of evoking the canonical estimator separately. The following rearrangement inequality is used for analyzing both steps. For any $a_1, a_2, b_1, b_2 \in \mathbb{R}$ where $a_1 < a_2$ and $b_1 < b_2$, it is straightforward to verify that

$$|a_1 - b_2| + |a_2 - b_1| \geq |a_1 - b_1| + |a_2 - b_2|. \tag{9.31}$$

We first analyze the re-arranging step in Line 5-13 of Algorithm 5. We denote the random variable $\widehat{\sigma}_{\text{re}}$ as the estimated ranking after the re-arranging step (that is, the value of the quantity $\widehat{\sigma}$ after Line 13 of Algorithm 5). The re-arranged ranking $\widehat{\sigma}_{\text{re}}$ is a deterministic function of the initial ranking $\widehat{\sigma}_{\text{init}}$. The following lemma proves a deterministic result about this re-arranging step.

Lemma 9.2. *For any true ranking σ^*, any set of ordinal observations $\mathcal{B}$ consistent with the true ranking, and any initial estimated ranking $\widehat{\sigma}_{init}$, the re-arranged ranking $\widehat{\sigma}_{re}$ yields smaller or equal loss compared to the initial ranking $\widehat{\sigma}_{init}$, regarding Kendall-tau distance and Spearman's footrule distance. That is,*

$$L_{KT}(\widehat{\sigma}_{re}, \sigma^*) \leq L_{KT}(\widehat{\sigma}_{init}, \sigma^*) \tag{9.32a}$$

$$L_{SF}(\widehat{\sigma}_{re}, \sigma^*) \leq L_{SF}(\widehat{\sigma}_{init}, \sigma^*). \tag{9.32b}$$

The lemma is proved at the end of this section.

Now we turn to analyze the second step of calling the canonical estimator on the topologically-identical pair. This step starts from the re-arranged ranking $\widehat{\sigma}_{\text{re}}$. Denote E as the event that Algorithm 5 identifies some topologically-identical pair (that is, Line 5-19 of Algorithm 5 is executed). Then E^c denotes the event that no topologically-identical pair is found. If there exists no topologically-identical pairs, then the second step in Line 14-18 of Algorithm 5 is never executed. Trivially, the final output $\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}$ is identical to the re-arranged ranking $\widehat{\sigma}_{\text{re}}$. We have

$$\mathbb{E}[L_{\text{KT}}(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*) \mid E^c] = \mathbb{E}[L_{\text{KT}}(\widehat{\sigma}_{\text{re}}, \sigma^*) \mid E^c] \tag{9.33a}$$

$$\mathbb{E}[L_{\text{SF}}(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*) \mid E^c] = \mathbb{E}[L_{\text{SF}}(\widehat{\sigma}_{\text{re}}, \sigma^*) \mid E^c]. \tag{9.33b}$$

It remains to consider the case when the event E is true. We start by showing that the event E happens with non-zero probability. Consider any arbitrary true ranking σ^*. Under this true ranking, denote the top item as $i^{(1)}$, and denote the second-ranked item as $i^{(2)}$. Conditioned on this true ranking, consider the set of pairwise comparisons $\mathcal{Q}$ such that the set $\mathcal{Q}$ includes comparisons between item $i^{(1)}$ and a subset of $\min\{\lfloor m/2 \rfloor, n-2\}$ items from $[n] \setminus \{i^{(1)}, i^{(2)}\}$. Assume that $\mathcal{Q}$ also includes comparisons between item $i^{(2)}$ and the identical subset of items from $[n] \setminus \{i^{(1)}, i^{(2)}\}$. The rest of the comparisons can be arbitrary between the $(n-2)$ items in $[n] \setminus \{i^{(1)}, i^{(2)}\}$. Recall that $1 < m < \binom{n}{2}$, so such a set $\mathcal{Q}$ arises with non-zero probability. Hence, the event E happens with non-zero probability.

Note that the set of ordinal observations $\mathcal{B}$ fully determines the topologically-identical pair (if any) selected by Algorithm 5. Since the event E happens with non-zero probability, there exists β such that $\mathbb{P}(\mathcal{B} = \beta, E) > 0$. We condition on the event E and any set of ordinal observations β such that $\mathbb{P}(\mathcal{B} = \beta, E) > 0$. We denote the two items in the topologically-identical pair selected by the algorithm as items $(i(\beta), i'(\beta))$ (or items (i, i') in short). In what follows, we consider Kendall-tau distance and Spearman's footrule separately.

Kendall-tau distance: For each $\ell, \ell' \in [n]$ with $\ell \neq \ell'$, we consider Kendall-tau distance contributed by the pair (ℓ, ℓ') . Recall that conditioned on the event event E and the set of ordinal observations β, the only pair that can be flipped by Algorithm 5 is $(i(\beta), i'(\beta))$. We only need to consider the pairs (ℓ, ℓ') such that the relative ordering of (ℓ, ℓ') can be potentially changed by flipping the pair (i, i'). We consider the following two cases separately.

Case 1: We consider Kendall-tau distance contributed by the pair (i, i') itself. That is, $\{\ell, \ell'\} = \{i, i'\}$.

Consider the ranking $\widehat{\sigma}_{\text{re}}$ from the re-arranging step. We have

$$\begin{aligned}
\mathbb{E}[\alpha(\widehat{\sigma}_{\text{re}}, \sigma^*, i, i') \mid \mathcal{B} = \beta, E] &= \sum_{\sigma} \mathbb{E}[\alpha(\widehat{\sigma}_{\text{re}}, \sigma^*, i, i') \mid \sigma^* = \sigma, \mathcal{B} = \beta, E] \cdot \mathbb{P}(\sigma^* = \sigma \mid \mathcal{B} = \beta, E) \\
&\overset{\text{(i)}}{=} \sum_{\sigma} \mathbb{E}[\alpha(\widehat{\sigma}_{\text{re}}, \sigma, i, i') \mid \sigma^* = \sigma, \mathcal{B} = \beta, E] \cdot \mathbb{P}(\sigma^* = \sigma \mid \mathcal{B} = \beta) \\
&\overset{\text{(ii)}}{=} \frac{1}{T(\beta)} \sum_{\sigma \in \text{topo}(\beta)} \mathbb{E}[\alpha(\widehat{\sigma}_{\text{re}}, \sigma, i, i') \mid \sigma^* = \sigma, \mathcal{B} = \beta, E]. \qquad (9.34)
\end{aligned}$$

where equality (i) is true because σ^* is independent of E conditioned on $\mathcal{B}$. Equality (ii) is true because of (9.11) in Lemma 9.1.

Recall that the initial ranking $\widehat{\sigma}_{\text{init}}$ is obtained by calling the (possibly randomized) ordinal estimator $\widehat{\sigma}_{\text{rank}}$ taking input $\mathcal{B}$, and the re-arranged ranking $\widehat{\sigma}_{\text{re}}$ is fully determined by $\widehat{\sigma}_{\text{init}}$. Hence, we further write (9.34) as

$$\begin{aligned}
&\mathbb{E}[\alpha(\widehat{\sigma}_{\text{re}}, \sigma^*, i, i') \mid \mathcal{B} = \beta, E] \\
&\quad = \frac{1}{T(\beta)} \sum_{\widehat{\sigma}} \sum_{\sigma \in \text{topo}(\beta)} \mathbb{E}[\alpha(\widehat{\sigma}, \sigma, i, i') \mid \widehat{\sigma}_{\text{re}} = \widehat{\sigma}, \sigma^* = \sigma, \mathcal{B} = \beta, E] \cdot \mathbb{P}(\widehat{\sigma}_{\text{re}} = \widehat{\sigma} \mid \sigma^* = \sigma, \mathcal{B} = \beta, E) \\
&\quad \overset{\text{(i)}}{=} \frac{1}{T(\beta)} \sum_{\widehat{\sigma}} \sum_{\sigma \in \text{topo}(\beta)} \mathbb{E}[\alpha(\widehat{\sigma}, \sigma, i, i') \mid \widehat{\sigma}_{\text{re}} = \widehat{\sigma}, \sigma^* = \sigma, \mathcal{B} = \beta, E] \cdot \mathbb{P}(\widehat{\sigma}_{\text{re}} = \widehat{\sigma} \mid \mathcal{B} = \beta),
\end{aligned}$$

(9.35)

where equality (i) is true, because $\widehat{\sigma}_{\text{rank}}$ is independent of the true ranking σ^* and the event E conditioned on $\mathcal{B}$. Hence, $\widehat{\sigma}_{\text{re}}$ is independent of the true ranking π^* and the event E conditioned on $\mathcal{B}$.

Define the set $\Omega_{i \succ i'} \subseteq \text{topo}(\beta)$ as the collection of topological orderings where i is ranked higher than i'. Define the set $\Omega_{i \prec i'} \subseteq \text{topo}(\beta)$ as the collection of topological orderings where i is ranked lower than i. Then $\{\Omega_{i \succ i'}, \Omega_{i \prec i'}\}$ is a partition of the collection of all topological orderings, $\text{topo}(\beta)$. Given that the pair (i, i') is topologically-identical, for any ranking $\sigma \in \text{topo}(\beta)$, we can flip items (i, i'), and the flipped ranking is still a topological ordering. Flipping the items (i, i') defines a bijection between the set $\Omega_{i \succ i'}, \Omega_{i \prec i'}$, so we have $|\Omega_{i \prec i'}| = |\Omega_{i \prec i'}|$. Any ranking $\widehat{\sigma}_{\text{re}}$ is correct on one and only one of the sets $\Omega_{i \succ i'}$ and $\Omega_{i \prec i'}$, and hence the re-arranged ranking $\widehat{\sigma}_{\text{re}}$ is correct on exactly half of the topological orderings. For any $\widehat{\sigma}$, we have

$$\sum_{\sigma \in \text{topo}(\beta)} \mathbb{E}[\alpha(\sigma, \pi, i, i') \mid \widehat{\sigma}_{\text{re}} = \widehat{\sigma}, \sigma^* = \sigma, \mathcal{B} = \beta, E] = \frac{1}{2}. \tag{9.36}$$

Combining (9.36) with (9.35) yields

$$\mathbb{E}[\alpha(\widehat{\sigma}_{\text{re}}, \sigma^*, i, i') \mid \mathcal{B} = \beta, E] = \frac{1}{2}.$$

Now consider the cardinal estimator. Similar to the proof of Theorem 2.6, we have

$$\mathbb{E}[\alpha(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*, i, i') \mid \mathcal{B} = \beta, E] < \frac{1}{2}.$$

Consequently, in Case 1, we have

$$\mathbb{E}[\alpha(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*, i, i') \mid \mathcal{B} = \beta, E] < \mathbb{E}[\alpha(\widehat{\sigma}_{\text{re}}, \sigma^*, i, i') \mid \mathcal{B} = \beta, E]. \tag{9.37}$$

Case 2: Consider any pair (ℓ, ℓ') that is not identical to the pair (i, i'). Since the relative ordering of the pair (ℓ, ℓ') is changed by flipping the pair (i, i'), then one item has to be either i or i'. Without loss of generality, assume $\ell \notin \{i, i'\}$ and $\ell' \in \{i, i'\}$. We consider pairs in the form of (ℓ, i) and (ℓ, i').

If the position of ℓ is not in between item i and item i' in the ranking $\widehat{\sigma}_{\text{re}}$ (that is, if $\widehat{\sigma}_{\text{re}}(\ell) < \min\{\widehat{\sigma}_{\text{re}}(i), \widehat{\sigma}_{\text{re}}(i')\}$ or $\widehat{\sigma}_{\text{re}}(\ell) > \max\{\widehat{\sigma}_{\text{re}}(i), \widehat{\sigma}_{\text{re}}(i')\}$), then flipping the pair (i, i') does not change the relative ordering of the pair (ℓ, i) or (ℓ, i'). Now we restrict our attention to item ℓ ranked in between item i and item i' in the ranking $\widehat{\sigma}_{\text{re}}$. Moreover, if the position of ℓ is not in between the positions of item i and item i' in the true ranking (that is, if $\sigma^*(\ell) < \min\{\sigma^*(i), \sigma^*(i')\}$ or $\sigma^*(\ell) > \max\{\sigma^*(i), \sigma^*(i')\}$), then whether flipping the pair (i, i') or not, one and only one of the two comparisons (ℓ, i) and (ℓ', i) is correct. Hence, we only need to consider each item ℓ ranked between the two items i and i', in both the re-arranged ranking $\widehat{\sigma}_{\text{re}}$ and the true ranking σ^*. For each such item ℓ, for any re-arranged ranking $\widehat{\sigma}_{\text{re}}$, we have the determinisitc equality

$$\begin{aligned} &\alpha(\widehat{\sigma}_{\text{re}}, \sigma^*, \ell, i) + \alpha(\widehat{\sigma}_{\text{re}}, \sigma^*, \ell, i') = 2\alpha(\widehat{\sigma}_{\text{re}}, \sigma^*, i, i') \\ &\alpha(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*, \ell, i) + \alpha(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*, \ell, i') = 2\alpha(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*, i, i') \end{aligned} \tag{9.38}$$

Combining (9.38) and (9.37), for each item ℓ ranked in between item i and item i' in both the re-arranged ranking $\widehat{\sigma}_{\text{re}}$ and the true ranking σ^*, we have

$$\begin{aligned}\mathbb{E}[\alpha(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*, \ell, i) + \alpha(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*, \ell, i') \mid \mathcal{B} = \beta, E]&\\ < \mathbb{E}[\alpha(\widehat{\sigma}_{\text{re}}, \sigma^*, \ell, i) + \alpha(\widehat{\sigma}_{\text{re}}, \sigma^*, \ell, i') \mid \mathcal{B} = \beta, E].&\end{aligned} \tag{9.39}$$

Combining the expression of Kendall-tau distance in (9.30) with the two cases in (9.37) and (9.39) of which the relative ordering of some pair (ℓ, ℓ') is changed, we have

$$\mathbb{E}[L_{\text{KT}}(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*) \mid \mathcal{B} = \beta, E] < \mathbb{E}[L_{\text{KT}}(\widehat{\sigma}_{\text{re}}, \sigma^*) \mid \mathcal{B} = \beta, E].$$

Recall that $\mathbb{P}(\mathcal{B} = \beta, E) > 0$ for some β. Taking an expectation over $\mathcal{B}$ yields

$$\mathbb{E}[L_{\text{KT}}(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*) \mid E] < \mathbb{E}[L_{\text{KT}}(\widehat{\sigma}_{\text{re}}, \sigma^*) \mid E]. \tag{9.40}$$

Combining (9.40) and (9.33a) yields

$$\mathbb{E}[L_{\text{KT}}(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*)] < \mathbb{E}[L_{\text{KT}}(\widehat{\sigma}_{\text{re}}, \sigma^*)]. \tag{9.41}$$

Finally, combining (9.41) with inequality (9.32a) for the re-arranging step completes the proof for Kendall-tau distance.

Spearman's footrule distance: We condition on the event E and any set of ordinal observations β such that $\mathbb{P}(\mathcal{B} = \beta, E) > 0$. Since only one pair $(i(\beta), i'(\beta))$ can be flipped by Algorithm 5, we only need to consider Spearman's footrule distance contributed by these two items. Consider any ranking $\sigma_{i \succ i'} \in \Omega_{i \succ i'}$. Let $\sigma_{i \prec i'}$ be the ranking obtained by flipping items (i, i') in $\sigma_{i \succ i'}$. Then we have $\sigma_{i \prec i'} \in \Omega_{i \prec i'}$. For any such pair $\{\sigma_{i \succ i'}, \sigma_{i \prec i'}\}$, we have

$$\begin{aligned}\mathbb{P}(\sigma^* \in \{\sigma_{i \succ i'}, \sigma_{i \prec i'}\}, \mathcal{B} = \beta, E) =&\mathbb{P}(\sigma^* \in \{\sigma_{i \succ i'}, \sigma_{i \prec i'}\} \mid \mathcal{B} = \beta, E) \cdot \mathbb{P}(\mathcal{B} = \beta, E)\\ =&\mathbb{P}(\sigma^* \in \{\sigma_{i \succ i'}, \sigma_{i \prec i'}\} \mid \mathcal{B} = \beta, E) \cdot \mathbb{P}(\mathcal{B} = \beta, E)\\ =&\mathbb{P}(\sigma^* \in \{\sigma_{i \succ i'}, \sigma_{i \prec i'}\} \mid \mathcal{B} = \beta) \cdot \mathbb{P}(\mathcal{B} = \beta, E) \qquad (9.42)\\ \overset{\text{(i)}}{>}&0, \qquad (9.43)\end{aligned}$$

where inequality (i) is true, because the two terms in (9.42) are both non-zero. The first term in (9.42) is non-zero by the fact that $\sigma_{i \succ i'}, \sigma_{i \prec i'}$ are topological orderings, and by (9.11) in Lemma 9.1. The second term in (9.42) is non-zero, because by construction we find β such that the second term $\mathbb{P}(\mathcal{B} = \beta, E) > 0$.

Now we analyze the Spearman's footrule distance conditioned on the event $\sigma^* \in \{\sigma_{i \succ i'}, \sigma_{i \prec i'}\}$. Using the argument deriving (9.37), we can further derive

$$\begin{aligned}\mathbb{E}[\alpha(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*, i, i') \mid \sigma^* \in \{\sigma_{i \succ i'}, \sigma_{i \prec i'}\}, \mathcal{B} = \beta, E]&\\ <\mathbb{E}[\alpha(\widehat{\sigma}_{\text{re}}, \sigma^*, i, i') \mid \sigma^* \in \{\sigma_{i \succ i'}, \sigma_{i \prec i'}\}, \mathcal{B} = \beta, E].&\end{aligned} \tag{9.44}$$

By the rearrangement inequality (9.31), if the relative ordering of the pair (i, i') is correct, then Spearman's footrule distance does not increase compared to the ranking with the relative ordering of (i, i') incorrect. Eq. (9.44) implies that conditioned on β, the event E and the event of $\sigma^* \in \{\sigma_{i \succ i'}, \sigma_{i \prec i'}\}$, the probability that the cardinal estimator $\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}$ gives the correct relative ordering of the pair (i, i') is higher than the probability that $\widehat{\sigma}_{\text{re}}$ gives the correct relative ordering. Hence, for any set of ordinal observations β and any pair $\{\sigma_{i \succ i'}, \sigma_{i \prec i'}\}$ of the true rankings, we have

$$\mathbb{E}[L_{\text{SF}}(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*) \mid \sigma^* \in \{\sigma_{i \succ i'}, \sigma_{i \prec i'}\}, \mathcal{B} = \beta, E] \leq \mathbb{E}[L_{\text{SF}}(\widehat{\sigma}_{\text{re}}, \sigma^*) \mid \sigma^* \in \{\sigma_{i \succ i'}, \sigma_{i \prec i'}\}, \mathcal{B} = \beta, E]. \tag{9.45}$$

Note that directly applying the re-arrangement inequality does not translate the strict inequality from (9.37) to (9.45). The reason is that correctly ordering a topologically-identical pair does not guarantee strictly smaller Spearman's footrule distance. For example, if item i and item i' are the top-2 items in the true ranking, but are the bottom-2 items in $\widehat{\sigma}_{\text{re}}$. Then the relative ordering of the pair (i, i') does not change the Spearman's footrule distance. In the rearrangement inequality (9.31), strictly inequality holds if $a_1 \leq \{b_1, b_2\} \leq a_2$. Hence, we find one pair of true rankings $\{\sigma^*_{i \succ i'}, \sigma^*_{i \prec i'}\}$ such that one of the following is true:

$$\begin{aligned} & \sigma^*_{i \succ i'}(i) \leq \{\widehat{\sigma}_{\text{re}}(i), \widehat{\sigma}_{\text{re}}(i')\} \leq \sigma^*_{i \succ i'}(i') \\ \text{or} \qquad & \sigma^*_{i \succ i'}(i') \leq \{\widehat{\sigma}_{\text{re}}(i), \widehat{\sigma}_{\text{re}}(i')\} \leq \sigma^*_{i \succ i'}(i). \end{aligned} \tag{9.46}$$

Then strictly inequality in (9.44) holds on the pair $\{\sigma^*_{i \succ i'}, \sigma^*_{i \prec i'}\}$. Now we provide the construction of this pair $\{\sigma^*_{i \succ i'}, \sigma^*_{i \prec i'}\}$.

We start by constructing a topological ordering $\sigma(i, i', \beta)$ (or σ in short) as follows. We topologically sort the items in $V^+ := V(i, i', \beta)$ and place them as the top $|V^+|$ items in σ. We topologically sort the items in $V^- := V^-(i, i, \beta)$ and place them as the bottom $|V^-|$ items. Arbitrarily choose one item from $\{i, i'\}$ and place it at the position $(|V^+| + 1)$, and place the remaining item from the pair $\{i, i'\}$ at the position $(n - |V^-|)$. Topologically sort the rest of the items, and place them in the remaining positions in σ.

We now prove that the ranking σ is a valid topological ordering. Assume for contradiction that σ is not a valid topological ordering. Then there exists a pair (ℓ, ℓ') that violates some pairwise comparison in $\mathcal{B}$. Denote $V^c = [n] \setminus (V^+ \cup V^- \cup \{i, i'\})$. Within each set V^+, V^- or V^c, the items are ordered by a topological ordering. Moreover, there is no direct comparison between item i and item i', so items $\{i, i'\}$ can be ranked with either $i \succ i'$ or $i \prec i'$. Hence, ℓ and ℓ' cannot belong to the same set of V^+, V, V^c or $\{i, i'\}$. By the definition of the sets V^+ and V^-, in the true ranking V^+ should be ranked higher than $\{i, i'\}$, and V^- should be ranked lower than $\{i, i'\}$. In our ranking σ, we also rank V^+ higher than $\{i, i'\}$, and V^- lower than $\{i, i'\}$. Hence, if both item ℓ and item ℓ' are in $V^+ \cup V^- \cup \{i, i'\}$, the relative ordering between (ℓ, ℓ') must be consistent with $\mathcal{B}$. Then at least one item from the pair (ℓ, ℓ') must be in V^c. Without loss of generality, assume $\ell' \in V^c$. Since ℓ and ℓ' cannot belong to the same set, we have $\ell \notin V^c$. If $\ell \in \{i, i'\}$, since the pair (ℓ, ℓ') violates some pairwise comparison, the items (ℓ, ℓ') are compared in $\mathcal{B}$, that is, ℓ' is compared to either i or i'. By the definition of the sets V^+ and V^-, it must be true that $\ell' \in V^+$ or $\ell' \in V^-$, contradicting the assumption that $\ell' \in V^c$. If $\ell \in V^+$, by construction the ranking σ ranks ℓ higher than ℓ'. Since the pair (ℓ, ℓ') violates some pairwise comparison, the set $\mathcal{B}$ must include the pairwise comparison $\ell' \succ \ell$. By the definition of V^+, since $\ell \in V^+$, there exists a path from ℓ to i. Concatenating the pairwise comparison $\ell' \succ \ell$ with the path from ℓ to i, we have a path from ℓ' to i. Hence, $\ell' \in V^+$, contradicting the assumption that $\ell' \in V^c$. Similarly, $\ell \in V^-$ gives a contradiction. Hence, in the ranking σ there exists no pair of items violating pairwise comparisons in $\mathcal{B}$. By definition, the ranking σ is a topological ordering. The ranking σ places items $\{i, i'\}$ at positions $\{|V^+| + 1, n - |V^-|\}$. In Algorithm 5, the re-arranged ranking $\widehat{\sigma}_{\text{re}}$ places the set V^+ before items $\{i, i'\}$, and the set V^- after items $\{i, i'\}$. Hence, we have either $\sigma(i) \leq \{\widehat{\sigma}_{\text{re}}(i), \widehat{\sigma}_{\text{re}}(i')\} \leq \sigma(i')$ or $\sigma(i') \leq \{\widehat{\sigma}_{\text{re}}(i), \widehat{\sigma}_{\text{re}}(i')\} \leq \sigma(i)$.

Recall that when constructing σ, we arbitrarily place an item from the set $\{i, i'\}$ at position $(|V^+| + 1)$, and the remaining item from $\{i, i'\}$ at position $(n - |V^-|)$. Denote $\sigma^*_{i \succ i'}$ as the topological ordering with item i in position $(|V^+| + 1)$. Denote $\sigma^*_{i \prec i'}$ as the topological ordering with item i' in position $(|V^+| + 1)$. For any possible $\widehat{\sigma}_{\text{re}}$, one of the conditions in (9.46) holds on the pair $\{\sigma^*_{i \succ i'}, \sigma^*_{i \prec i'}\}$, and hence strict inequality in (9.45) holds for the pair $\{\sigma^*_{i \succ i'}, \sigma^*_{i \prec i'}\}$.

Eq. (9.43) implies that the event $\sigma^* \in \{\sigma^*_{i \succ i'}, \sigma^*_{i \prec i'}\}$ arises with non-zero probability. Taking an expectation over all possible pairs $\{\sigma_{i \succ i'}, \sigma_{i \prec i'}\}$ in (9.45), and using the strict inequality for the pair $\{\sigma^*_{i \succ i'}, \sigma^*_{i \prec i'}\}$ yields

$$\mathbb{E}[L_{\text{SF}}(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*) \mid \mathcal{B} = \beta, E] < \mathbb{E}[L_{\text{SF}}(\widehat{\sigma}_{\text{re}}, \sigma^*) \mid \mathcal{B} = \beta, E].$$

Taking an expectation over the set of ordinal observations $\mathcal{B}$ yields

$$\mathbb{E}[L_{\text{SF}}(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*) \mid E] < \mathbb{E}[L_{\text{SF}}(\widehat{\sigma}_{\text{re}}, \sigma^*) \mid E]. \tag{9.47}$$

Combining (9.47) with inequality (9.33b) for the re-arranging step yields

$$\mathbb{E}[L_{\text{SF}}(\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}, \sigma^*)] < \mathbb{E}[L_{\text{SF}}(\widehat{\sigma}_{\text{re}}, \sigma^*)]. \tag{9.48}$$

Finally, combining (9.48) with inequality (9.32b) for the re-arranging step completes the proof for Spearman's footrule.

We make a comment about having multiple topologically-identical pairs. Notice that in Algorithm 5, we only find one topologically-identical pair, and then break out of the for-loops. Alternatively, we can identify and flip multiple disjoint topologically-identical pairs in a similar fashion as in Algorithm 1. This is still a valid algorithm, because each step of processing one topologically-identical pair does not increase Kendall-tau distance or Spearman's footrule distance.

It remains to prove Lemma 9.2.

9.8 Proof of Lemma 9.2

Consider any two items $\ell, \ell' \in [n]$, such that $\ell \succ \ell'$ in the true ranking σ^*. Let $\widehat{\sigma}_1$ be an arbitrary ranking. Let $\widehat{\sigma}_2$ be a ranking where all items are ranked the same as in $\widehat{\sigma}_1$, except that the positions of items ℓ and ℓ' are flipped as compared to $\widehat{\sigma}_1$. The remainder of the proof is broken into two parts.

Part 1: If the relative ordering of a pair is inconsistent with the relative ordering indicated by the true ranking, then flipping this pair does not increase Kendall-tau distance or Spearman's footrule distance.

Specifically, we claim that if $\ell \prec \ell'$ in $\widehat{\pi}_1$, then $\widehat{\sigma}_2$ has a smaller or equal loss than $\widehat{\sigma}_1$, with respect to Kendall-tau distance and Spearman's footrule distance. We discuss the two distance metrics separately.

Kendall-tau distance: First, consider Kendall-tau distance contributed by the pair (ℓ, ℓ'). We have $\ell \prec \ell'$ in $\widehat{\sigma}_1$ and $\ell \succ \ell'$ in $\widehat{\sigma}_2$. Since we have $\ell \succ \ell'$ in the true ranking, the relative ordering of this pair is correct in $\widehat{\sigma}_2$, and incorrect in $\widehat{\sigma}_1$. Hence,

$$0 = \alpha(\widehat{\sigma}_2, \sigma^*, \ell, \ell') < \alpha(\widehat{\sigma}_1, \sigma^*, \ell, \ell') = 1. \tag{9.49}$$

Denote ℓ_{mid} as any item ranked in between ℓ and ℓ' in $\widehat{\sigma}_1$ (or equivalently, in $\widehat{\sigma}_2$). In the rest of the pairs that are not (ℓ, ℓ'), the flip only changes the relative ordering of each pair of the form $(\ell, \ell_{\text{mid}})$ or $(\ell', \ell_{\text{mid}})$. If in the true ranking σ^*, item ℓ_{mid} is ranked higher than both (ℓ, ℓ'), or ranked lower than both (ℓ, ℓ'), then the sum of the contributions to Kendall-tau distance by the pair $(\ell, \ell_{\text{mid}})$ and the pair $(\ell', \ell_{\text{mid}})$ is the same in $\widehat{\sigma}_1$ and $\widehat{\sigma}_2$:

$$\alpha(\widehat{\sigma}_2, \sigma^*, \ell, \ell_{\text{mid}}) + \alpha(\widehat{\sigma}_2, \sigma^*, \ell', \ell_{\text{mid}}) = 1 = \alpha(\widehat{\sigma}_1, \sigma^*, \ell, \ell_{\text{mid}}) + \alpha(\widehat{\sigma}_1, \sigma^*, \ell', \ell_{\text{mid}}). \tag{9.50}$$

Otherwise ℓ_{mid} is ranked in between ℓ and ℓ' in the true ranking σ^*, then we have

$$0 = \alpha(\widehat{\sigma}_2, \sigma^*, \ell, \ell_{\text{mid}}) + \alpha(\widehat{\sigma}_2, \sigma^*, \ell', \ell_{\text{mid}}) < \alpha(\widehat{\sigma}_1, \sigma^*, \ell, \ell_{\text{mid}}) + \alpha(\widehat{\sigma}_1, \sigma^*, \ell', \ell_{\text{mid}}) = 2. \tag{9.51}$$

Combining the expression (9.30) of Kendall-tau distance with (9.49), (9.50) and (9.51) yields

$$L_{\text{KT}}(\widehat{\sigma}_2, \sigma^*) < L_{\text{KT}}(\widehat{\sigma}_1, \sigma^*).$$

Spearman's footrule distance: By flipping the positions of the items (ℓ, ℓ'), only Spearman's footrule distance contributed by these two items has changed. Recall that the condition for flipping the pair (ℓ, ℓ') requires $\ell \prec \ell'$ in $\widehat{\pi}_1$ and $\ell \succ \ell'$ in σ^*. Applying the rearrangement inequality (9.31) with $a_1 = \widehat{\sigma}_1(\ell'), a_2 = \widehat{\sigma}_1(\ell), b_1 = \sigma^*(\ell), b_2 = \sigma^*(\ell')$, we have

$$\begin{aligned} |\widehat{\sigma}_1(\ell') - \sigma^*(\ell')| + |\widehat{\sigma}_1(\ell) - \sigma^*(\ell)| \geq & |\widehat{\sigma}_1(\ell') - \sigma^*(\ell)| + |\widehat{\sigma}_1(\ell) - \sigma^*(\ell')| \\ = & |\widehat{\sigma}_2(\ell) - \sigma^*(\ell)| + |\widehat{\sigma}_2(\ell') - \sigma^*(\ell')|. \end{aligned} \tag{9.52}$$

Combining (9.52) with the definition of Spearman's footrule distance yields

$$L_{\text{SF}}(\widehat{\sigma}_2, \sigma^*) \leq L_{\text{SF}}(\widehat{\sigma}_1, \sigma^*).$$

This completes Part 1 of the proof.

Part 2: The re-arranging step in Algorithm 5 is equivalent to a sequence of pair flips.

With Part 1 in place, we now explain the rest of the proof. For any arbitrary topologically-identical pair of items (i, i') and any arbitrary set of ordinal observations $\mathcal{B}$, denote the sets

$V_1 := V^+(i, i', \mathcal{B}), V_2 := \{i, i'\}, V_3 := V^-(i, i', \mathcal{B})$. We consider the following procedure. We start by setting $\widehat{\sigma}_1$ as the initial estimated ranking $\widehat{\sigma}_{\text{init}}$. We identify one pair (ℓ, ℓ') (if any) such that the following three conditions are met. First, we have $\ell \in V_j, \ell' \in V_{j'}$ with $j < j'$. Second, we have $\ell \prec \ell'$ in $\widehat{\sigma}_1$. Third, there is no item in $V_1 \cup V_2 \cup V_3$, whose position is in between ℓ and ℓ' in the ranking $\widehat{\sigma}_1$. If such a pair is found, we flip the positions of ℓ and ℓ', and update $\widehat{\sigma}_1$ to be this new ranking. Repeat this procedure until no such pair can be found.

Now we show that this procedure is equivalent to the re-arranging step in Line 5-13 of Algorithm 5. This procedure properly terminates, because each pair of items (ℓ, ℓ') can be swapped at most once, and there is a finite number of pairs. When the procedure terminates, the ranking is identical to the re-arranged ranking $\widehat{\sigma}$ after Line 13 of Algorithm 5. To see this claim, we first note that this procedure has never moved items outside $V_1 \cup V_2 \cup V_3$, so we only need to concern about the items in $V_1 \cup V_2 \cup V_3$ and their positions. For each pair (ℓ, ℓ') to be flipped, the procedure specifies that ℓ and ℓ' belong to two different sets from V_1, V_2 and V_3. Moreover, by the condition on the pair (ℓ, ℓ'), there cannot be any item in $V_1 \cup V_2 \cup V_3$ that is ranked in between ℓ and ℓ'. Hence, the relative ordering of the items within each set of V_1, V_2 or V_3 is unchanged, consistent with the ranking specified in Line 10 and Line 12 of Algorithm 5. Moreover, the re-arranging step in Algorithm 5 ranks all items in V_1 before all items in V_2, and all items in V_2 before all items in V_3. Assume that the final output of the procedure is a different ranking from the re-arranging step in Algorithm 5, then we can find a pair (ℓ, ℓ') that can be flipped, contradicting the fact that no such pairs can be found at the termination of the procedure. Hence, the procedure and the re-arranging step in Algorithm 5 are equivalent. Applying Part 1 to each flip in this procedure completes the proof of the lemma.

9.9 Proof of Theorem 2.9

The proof is a slight modification to the proof of Theorem 2.6, so we only highlight the difference. First, we consider the probability of success of the optimal ordinal estimator $\widehat{\pi}_{\text{rank-unif}}$ that outputs one of the topological orderings uniformly at random:

$$\begin{aligned}\mathbb{P}(\widehat{\pi}_{\text{rank-unif}}(\beta) = \pi^* \mid \mathcal{B} = \beta) &= \sum_{\pi \in \text{topo}(\beta)} \mathbb{P}(\pi = \pi^* \mid \widehat{\pi}_{\text{rank-unif}} = \pi, \mathcal{B} = \beta)\mathbb{P}(\widehat{\pi}_{\text{rank-unif}} = \pi \mid \mathcal{B} = \beta) \\ &\overset{\text{(i)}}{=} \frac{1}{T(\beta)} \sum_{\pi \in \text{topo}(\beta)} \mathbb{P}(\pi = \pi^* \mid \widehat{\pi}_{\text{rank-unif}} = \pi, \mathcal{B} = \beta),\end{aligned} \tag{9.53}$$

where equality (i) is true because the ordinal estimator $\widehat{\pi}_{\text{rank-unif}}$ uniformly at random outputs one of the topological orderings consistent with β.

Now we consider each term $\mathbb{P}(\pi = \pi^* \mid \widehat{\pi}_{\text{rank-unif}} = \pi, \mathcal{B} = \beta)$ in (9.53). The quantities π^* and π are both deterministic. Trivially, we have

$$\mathbb{P}(\pi = \pi^* \mid \widehat{\pi}_{\text{rank}} = \pi, \mathcal{B} = \beta) = \begin{cases} 1 & \text{if } \pi = \pi^* \\ 0 & \text{otherwise.} \end{cases} \tag{9.54}$$

Combining (9.53) and (9.54) with the fact that the true ranking π^* must be a topological ordering consistent with β, we have

$$\mathbb{P}(\widehat{\pi}_{\text{rank-unif}}(\beta) = \pi^* \mid \mathcal{B} = \beta) = \frac{1}{T(\beta)}. \tag{9.55}$$

Now consider the cardinal estimator $\widetilde{\pi}^{\text{our}}_{\text{rank-unif}}$. When the number of flippable pairs is zero, the cardinal estimator behaves equivalently as the ordinal estimator $\widehat{\pi}_{\text{rank-unif}}$. Following a similar argument as Case 1 in the proof of Theorem 2.6, for any set of ordinal observations β, we have (cf. Equation (9.15) in the proof of Theorem 2.6):

$$\mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{rank-unif}} \mid \mathcal{B} = \beta, L = 0) = \mathbb{P}(\widehat{\pi}_{\text{rank-unif}} = \pi^* \mid \mathcal{B} = \beta). \tag{9.56}$$

Denote $\widehat{\pi}_{\text{init}}$ as the initial estimated ranking obtained by calling the ordinal estimator $\widehat{\pi}_{\text{rank-unif}}$. When the number of flippable pairs is $L = \ell > 0$, the probability of success of the cardinal estimator is

$$\begin{aligned}
&\mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{rank-unif}} = \pi^* \mid \mathcal{B} = b, L = \ell) \\
&\qquad = \sum_{\pi \in \text{topo}(\beta)} \mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{rank-unif}} = \pi^* \mid \mathcal{B} = \beta, L = \ell, \widehat{\pi}_{\text{init}} = \pi)\mathbb{P}(\widehat{\pi}_{\text{init}} = \pi \mid \mathcal{B} = \beta, L = \ell) \\
&\qquad \overset{\text{(i)}}{=} \frac{1}{T(\beta)} \sum_{\pi \in \text{topo}(\beta)} \mathbb{P}(\widetilde{\pi}^{\text{our}}_{\text{rank-unif}} = \pi^* \mid \mathcal{B} = \beta, L = \ell, \widehat{\pi}_{\text{init}} = \pi),
\end{aligned} \tag{9.57}$$

where equality (i) is true because the ordinal estimator $\widehat{\pi}_{\text{rank-unif}}$ outputs a topological ordering uniformly at random.

The remaining argument is similar to Case 2 in the proof of Theorem 2.6, so we only outline the main steps. Consider all total rankings that are identical to the true ranking π^*, except for (possibly) the relative ordering of the ℓ flippable pairs. There are 2^ℓ such total rankings, and all these 2^ℓ total rankings are topological orderings on the graph $\mathcal{G}(\mathcal{B})$. In (9.57), the summation of π is over all topological orderings. In particular, this summation includes these 2^ℓ total rankings. Recall that the cardinal estimator $\widetilde{\pi}^{\text{our}}_{\text{rank-unif}}$ is obtained by replacing Line 2 of Algorithm 1 by calling the ordinal estimator $\widehat{\pi}_{\text{rank-unif}}$. To be able to apply Theorem 2.3, we obtain a cardinal estimator $\widetilde{\pi}^{\text{eq}}_{\text{rank-unif}}$ by replacing Line 2 of Algorithm 4 by calling the ordinal estimator $\widehat{\pi}_{\text{rank-unif}}$. This estimator $\widetilde{\pi}^{\text{eq}}_{\text{rank-unif}}$ is equivalent to the original estimator $\widetilde{\pi}^{\text{our}}_{\text{rank-unif}}$. When the initial estimated ranking $\widehat{\pi}_{\text{init}}$ is any of the 2^ℓ total rankings, the probability that the cardinal estimator $\widetilde{\pi}^{\text{eq}}_{\text{rank-unif}}$ gives the correct output is strictly greater than $\frac{1}{2^\ell}$. Hence, we bound (9.57) as (cf. Equation (9.17) in the proof of Theorem 2.6):

$$\mathbb{P}(\widetilde{\pi}^{\text{eq}}_{\text{rank-unif}} = \pi^* \mid \mathcal{B} = b, L = \ell) > \frac{1}{T(\beta)} \cdot 2^\ell \cdot \frac{1}{2^\ell} = \frac{1}{T(\beta)} \overset{\text{(i)}}{=} \mathbb{P}(\widehat{\pi}_{\text{rank-unif}} = \pi^* \mid \mathcal{B} = \beta). \tag{9.58}$$

where equality (i) is true from (9.55).

Having established (9.56) and (9.58), the rest of the argument follows the proof of Theorem 2.6.

Algorithm 4: An equivalent joint procedure of the assignment A, the evaluation $\mathcal{Y}$, and the execution of our cardinal ranking estimator $\widetilde{\pi}^{\text{our}}_{\text{rank}}(A, \mathcal{Y})$ in Algorithm 1.

```
Sample pairwise comparisons $\mathcal{Q} = \{\widetilde{S}_j\}_{j=1}^{m}$ uniformly at random from all $\binom{n}{2}$ pairs.
  Obtain the ordinal comparisons $\mathcal{B}$.
Compute a topological ordering $\widehat{\pi}$ on the graph $\mathcal{G}(\mathcal{B})$, with ties broken in order of the
  indices of the items.
$t \leftarrow 1$.
$\mathcal{Q}_{\text{avail}} \leftarrow \mathcal{Q}$.
flippable_positions $\leftarrow [\,]$.
reviewer_indices $\leftarrow [\,]$.
while $t < n$ do
    Let $\widehat{\pi}_{\text{flip}}$ be the ranking obtained by flipping the positions of the $t^{th}$ and the $(t+1)^{th}$
      items in $\widehat{\pi}$.
    if $\widehat{\pi}_{\mathit{flip}}$ is a topological ordering on $\mathcal{G}(\mathcal{B})$, and both the $t^{th}$ and $(t+1)^{th}$ items are
      each included in at least one pairwise comparison in $\mathcal{Q}_{\mathit{avail}}$ then
        From all of the pairwise comparisons in $\mathcal{Q}_{\text{avail}}$ including the $t^{th}$ item, sample one
          uniformly at random and denote it as $\widetilde{S}_t$. Likewise denote $\widetilde{S}_{t+1}$ as a randomly
          chosen pairwise comparison including the $(t+1)^{th}$ item from $\mathcal{Q}_{\text{avail}}$.
        Append $t$ to flippable_positions.
        Append the pair $[\widetilde{S}_t, \widetilde{S}_{t+1}]$ to reviewer_indices.
        Remove $\widetilde{S}_t$ and $\widetilde{S}_{t+1}$ from $\mathcal{Q}_{\text{avail}}$.
        $t \leftarrow t + 2$.
    else
        $t \leftarrow t + 1$.
    end
end
For each pair $[\widetilde{S}_t, \widetilde{S}_{t+1}]$ in reviewer_indices, sample uniformly at random without
  replacement a pair of reviewers $\{j_t, j_{t+1}\}$.
foreach $t \in$ flippable_positions do
    Assign reviewer $j_t$ to one of the two pairs $\widetilde{S}_t$ or $\widetilde{S}_{t+1}$, uniformly at random. Assign
      reviewer $j_{t+1}$ to the remaining pair. Obtain the scores from these two reviewers for
      their corresponding pair.
    Denote $y_{\widehat{\pi}(t)}$ as the score for the $t^{th}$ item in $\widetilde{S}_t$. Likewise denote $y_{\widehat{\pi}(t+1)}$ as the score
      for the $(t+1)^{th}$ item in $\widetilde{S}_{t+1}$.
    if $\widetilde{\pi}^{\mathit{our}}_{\mathit{can}}(y_{\widehat{\pi}(t)}, y_{\widehat{\pi}(t+1)})$ outputs $\widehat{\pi}(t+1) \succ \widehat{\pi}(t)$ then
        Let $\widehat{\pi}_{\text{flip}}$ be the ranking obtained by flipping the positions of the $t^{th}$ and the
          $(t+1)^{th}$ items in $\widehat{\pi}$.
        $\widehat{\pi} \leftarrow \widehat{\pi}_{\text{flip}}$.
    end
end
Output $\widehat{\pi}$.
```

Algorithm 5: Our cardinal ranking estimator $\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}(A, \mathcal{Y})$ concerning Kendall-tau distance and Spearman's footrule distance.

```
1  Deduce the ordinal observations $\mathcal{B}$ from the cardinal observations $\mathcal{Y}$. Compute an initial
     estimated ranking $\widehat{\sigma}_{\text{init}} = \widehat{\sigma}_{\text{rank}}(\mathcal{B})$.;
2  for $i = 1, \ldots, n$ do
3      for $i' = (i+1), \ldots, n$ do
4          if the pair $(i, i')$ is topologically-identical, and both items $i$ and $i'$ have at least
             one score each from $\mathcal{Y}$ then
5              Compute $V^+ := V^+(i, i', \mathcal{B})$. Denote the items in $V^+$ as $i^+_1 \succ \cdots \succ i^+_{|V^+|}$
                 under the ranking $\widehat{\pi}_{\text{init}}$.;
6              Compute $V^- := V^-(i, i', \mathcal{B})$. Denote the items in $V^-$ as $i^-_1 \succ \cdots \succ i^-_{|V^-|}$
                 under the ranking $\widehat{\pi}_{\text{init}}$.;
7              positions $= \{\ell \in V^+ \cup V^- \cup \{i, i'\} \mid \widehat{\sigma}_{\text{init}}(\ell)\}$. ;
8              $\widehat{\sigma} \leftarrow \widehat{\sigma}_{\text{init}}$.;
9              if $i \succ i'$ under $\widehat{\sigma}_{init}$ then
10                 Re-arrange items in $V^+ \cup V^- \cup \{i, i'\}$ in $\widehat{\sigma}$, such that they still occupy
                     positions, and $i^+_1 \succ \cdots \succ i^+_{|V^+|} \succ i \succ i' \succ i^-_1 \succ \cdots \succ i^-_{|V^-|}$.;
11             else
12                 Re-arrange items in $V^+ \cup V^- \cup \{i, i'\}$ in $\widehat{\sigma}$, such that they still occupy
                     positions, and $i^+_1 \succ \cdots \succ i^+_{|V^+|} \succ i' \succ i \succ i^-_1 \succ \cdots \succ i^-_{|V^-|}$.;
13             end
14             From all of the scores of item $i$ in $\mathcal{Y}$, sample one uniformly at random and
                 denote it as $y_i$. Likewise denote $y_{i'}$ as a randomly chosen score of item $i'$
                 from $\mathcal{Y}$.;
15             if $\widetilde{\pi}^{our}_{can}(y_i, y_{i'})$ indicates a relative ordering of the pair $(i, i')$ different from $\widehat{\sigma}$
                 then
16                 Let $\widehat{\sigma}_{\text{flip}}$ be the ranking obtained by flipping items $i$ and $i'$ in $\widehat{\sigma}$.;
17                 $\widehat{\sigma} \leftarrow \widehat{\sigma}_{\text{flip}}$.;
18             end
19             break from both for-loops and go to Line 23.
20         end
21     end
22 end
23 Output $\widetilde{\sigma}^{\text{our}}_{\text{rank-metric}}(A, \mathcal{Y}) = \widehat{\sigma}$.;
```

Chapter 10

Proofs of Chapter 3

In this chapter, we present the proofs of all theoretical results.

10.1 Preliminary results

In this section, we present preliminary results that are used in the proofs. For the regularizer R, it can be verified that we have the symmetry

$$R_{ii'jj'} = R_{ii'j'j} = R_{i'ijj'} = R_{i'ij'j}.$$

We say that an entry (i, j) does not contribute to the regularizer if $R_{ii'jj'} = 0$ for all $i' \in [n]$ and $j' \in [d]$. We say that a row/column does not contribute to the regularizer if none of the entries in the row/column contributes to the regularizer. We say that (i, i', j, j') is a "conflicting quadruple" if we have $(A_{ij} - A_{ij'})(A_{i'j} - A_{i'j'}) < 0$. By the definition (3.7) of the regularizer, an entry (i, j) does not contribute to the regularizer if and only if the quadruple (i, i', j, j') is not a conflicting quadruple for each $i' \in [n]$ and $j' \in [d]$.

10.1.1 Derivative of the objective

We compute the derivative of the regularizer term $R_{ii'jj'}$ as

$$\frac{\partial R_{ii'jj'}}{\partial A_{ij}} = \begin{cases} 0 & \text{if } (A_{ij} - A_{ij'})(A_{i'j} - A_{i'j'}) \geq 0 \\ 2(A_{ij} - A_{ij'})(A_{i'j} - A_{i'j'})^2 & \text{otherwise.} \end{cases} \tag{10.1}$$

Hence, we have

$$\operatorname{sign}\left(\frac{\partial R_{ii'jj'}}{\partial A_{ij}}\right) = \operatorname{sign}(A_{ij} - A_{ij'}), \quad \text{if } (A_{ij} - A_{ij'})(A_{i'j} - A_{i'j'}) < 0. \tag{10.2}$$

It can be verified that we have the symmetry

$$\frac{\partial R_{ii'jj'}}{\partial A_{ij}} = \frac{\partial R_{ii'j'j}}{\partial A_{ij}} = \frac{\partial R_{i'ijj'}}{\partial A_{ij}} = \frac{\partial R_{i'ij'j}}{\partial A_{ij}}. \tag{10.3}$$

Combining (10.3) with the expression (3.6) of R, we have

$$\frac{\partial R}{\partial A_{ij}} = 4 \sum_{i' \in [n], j' \in [d]} \frac{\partial R_{ii'jj'}}{\partial A_{ij}}. \tag{10.4}$$

The derivative of the objective L is computed as

$$\nabla L(A) = 2(A - Y)_\Omega + \lambda \nabla R(A). \tag{10.5}$$

We have the partial derviative

$$\frac{\partial L}{\partial A_{ij}} = 2(A_{ij} - Y_{ij}) \cdot \mathbb{1}\{(i,j) \in \Omega\} + \lambda \frac{\partial R}{\partial A_{ij}} \tag{10.6}$$

$$\overset{\text{(i)}}{=} 2(A_{ij} - Y_{ij}) \cdot \mathbb{1}\{(i,j) \in \Omega\} + 4\lambda \sum_{i' \in [n], j' \in [d]} \frac{\partial R_{ii'jj'}}{\partial A_{ij}}, \tag{10.7}$$

where (i) is true by plugging in (10.4).

10.1.2 Additional preliminary results

For notational simplicity, we denote the projection step (3.10b) as $\mathcal{P}_{[0,1]}A := \min\{1, \max\{0, A\}\}$ for any $A \in \mathbb{R}^{d \times n}$. The following lemma states that the objective L does not increase after a projection step.

Lemma 10.1. *Consider any $Y \in [0,1]^{n \times d}$. Then for any $A \in \mathbb{R}^{n \times d}$, we have $L(\mathcal{P}_{[0,1]}A) \leq L(A)$.*

Proof of Lemma 10.1 We consider the two terms in the objective (3.8). For the first term $\|A - Y\|_\Omega^2$, it is straightforward to verify that

$$\|Y - \mathcal{P}_{[0,1]}(A)\|_\Omega \leq \|Y - A\|_\Omega, \qquad \forall Y \in [0,1]^{n \times d}. \tag{10.8}$$

For the second term, we consider $R_{ii'jj'}$ for each quadruple (i, i', j, j'). Note that for any scalar values $a, b \in \mathbb{R}$, the term $(\mathcal{P}_{[0,1]}(a) - \mathcal{P}_{[0,1]}(b))$ either has the same sign as $(a - b)$ or has a value of 0. Now we discuss the following two cases depending on the sign of each quadruple (i, i', j, j').

Case 1: $(A_{ij} - A_{ij'})(A_{i'j} - A_{i'j'}) \geq 0$.

In this case, we have $(\mathcal{P}_{[0,1]}A_{ij'} - \mathcal{P}_{[0,1]}A_{i'j})(\mathcal{P}_{[0,1]}A_{i'j} - \mathcal{P}_{[0,1]}A_{i'j'}) \geq 0$. Hence, by the definition of the function $R_{ii'jj'}$, we have

$$0 = R_{ii'jj'}(A) = R_{ii'jj'}(\mathcal{P}_{[0,1]}A) \tag{10.9}$$

Case 2: $(A_{ij} - A_{ij'})(A_{i'j} - A_{i'j'}) < 0$.

In this case, we have $(\mathcal{P}_{[0,1]}A_{ij'} - \mathcal{P}_{[0,1]}A_{i'j})(\mathcal{P}_{[0,1]}A_{i'j} - \mathcal{P}_{[0,1]}A_{i'j'}) \leq 0$. Moreover, due to the projection we have

$$\left|\mathcal{P}_{[0,1]}A_{ij} - \mathcal{P}_{[0,1]}A_{ij'}\right| \leq |A_{ij} - A_{ij'}|$$
$$\left|\mathcal{P}_{[0,1]}A_{i'j} - \mathcal{P}_{[0,1]}A_{i'j'}\right| \leq |A_{i'j} - A_{i'j'}|.$$

By the definition of the function $R_{ii'jj'}$, it can be verified that

$$R_{ii'jj'}(A) \geq R_{ii'jj'}(\mathcal{P}_{[0,1]}A). \tag{10.10}$$

Combining (10.9) and (10.10) from the two cases, we have

$$R(A) \geq R(\mathcal{P}_{[0,1]}A), \quad \forall (i, i', j, j'), \forall A \in [0,1]^{n\times d}. \tag{10.11}$$

Finally, combining the two terms (10.8) and (10.11) of the objective L, we have

$$L(A) \geq L(\mathcal{P}_{[0,1]}A),$$

completing the proof.

Now we analyze the local optima of the objective. Standard results suggest that any local optimum in the interior of the domain satisfies the first-order optimality condition, namely having a gradient of 0. The following lemma suggests that any local optimum on the boundary of the domain also satisfies the first-order optimality condition. We define $\nabla L(A)$ as the gradient on $\mathbb{R}$, without restricting to the domain $[0,1]^{n\times d}$.

Lemma 10.2. *For any local optimum A of the objective* (3.8) *defined on the domain* $[0,1]^{d\times n}$*, we have* $\nabla L(A) = 0$.

Proof of Lemma 10.2 If any local optimum A is in the interior, then standard first-order optimality condition [16, Theorem 2.6] yields $\nabla L(A) = 0$. It remains to consider the case where A is on the boundary of the domain.

Assume for contradiction that there exists a local optimum A on the boundary with $\nabla L(A) \neq 0$. Without loss of generality we assume $\frac{\partial L(A)}{\partial A_{11}} \neq 0$. By definition of the local optimum, there exists some $\delta > 0$, such that $L(A') \geq L(A)$ for all $A' \in [0,1]^{n\times d}$ with $\|A' - A\|_F < \delta$. On the other hand, let E_{11} denote the matrix whose $(1,1)$-entry is 1 and all other entries are 0. By definition of the partial derivative, there exists some $\delta' \in (0, \delta)$ such that $L(A + \delta' E_{11}) < L(A)$. Now consider the point $\mathcal{P}_{[0,1]}(A + \delta' E_{11})$. By Lemma 10.1, we have

$$L(\mathcal{P}_{[0,1]}(A + \delta' E_{11})) \leq L(A + \delta' E_{11}) < L(A). \tag{10.12}$$

Since $[0,1]^{n\times d}$ is a convex set and $A \in [0,1]^{n\times d}$, by Lemma 10.1 we have

$$\|\mathcal{P}_{[0,1]}(A + \delta' E_{11}) - A\|_F \leq \|A + \delta' E_{11} - A\|_F = \delta' < \delta. \tag{10.13}$$

Combining (10.12) and (10.13), the point $\mathcal{P}_{[0,1]}(A + \delta' E_{11})$ yields a contradiction to the local optimality of A.

10.2 Proof of Theorem 3.1

The proof consists of two steps. First, we show that our objective L has a Lipschitz gradient. Second, we incorporate the projected step straightforwardly into standard analysis of gradient descent for functions with Lipschitz gradient.

Step 1: Bound the magnitude of the gradient $\|\nabla L\|_F$ and the Lipschitz constant

As a general definition, consider any $d \geq 1$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to have a Lipschitz gradient with constant K on domain $D \subseteq \mathbb{R}^d$ if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq K\|x - y\|_2, \text{ for all } x, y \in D.$$

For projected gradient descent, the gradient step (3.10a) may give solutions outside the domain $[0, 1]^{n\times d}$, so we bound the gradient on an enlarged domain, namely $[-1, 2]^{n\times d}$. For any $A \in [-1, 2]^{n\times d}$, its partial derivative is given by (10.7) as:

$$\frac{\partial L}{\partial A_{ij}} = 2(A_{ij} - Y_{ij}) \cdot \mathbb{1}\{(i, j) \in \Omega\} + 4\lambda \sum_{i' \in [n], j' \in [d]} \frac{\partial R_{ii'jj'}}{\partial A_{ij}}. \tag{10.14}$$

Consider the term $\frac{\partial R_{ii'jj'}}{\partial A_{ij}}$ in (10.14). For each $i' \in [n]$ and $j' \in [d]$, we have

$$\left|\frac{\partial R_{ii'jj'}}{\partial A_{ij}}\right| \leq 2|A_{ij} - A_{ij'}| \cdot (A_{i'j} - A_{i'j'})^2 \leq 54. \tag{10.15}$$

Combining (10.15) and (10.14), we have

$$\left|\frac{\partial L}{\partial A_{ij}}\right| \leq 6 + 216\lambda nd, \tag{10.16}$$

and hence

$$\|\nabla L(A)\|_F \leq \sqrt{nd}(6 + 216\lambda nd). \tag{10.17}$$

Now we bound the Lipschitz constant of the objective L. Let $A, B \in [-1, 2]^{n\times d}$ be any two matrices. Using (10.17), we have:

$$\begin{aligned}\|\nabla L(A) - \nabla L(B)\|_F^2 &\leq 4(nd)(6 + 216\lambda nd)^2 \\ &\overset{\text{(i)}}{\leq} 4(2 + 72\lambda nd)^2\|A - B\|_F^2,\end{aligned}$$

where (i) holds because $A, B \in [-1, 2]^{n\times d}$ Hence, L has a Lipschitz gradient with $K = K(n, d, \lambda) := 4 + 144\lambda nd$ on $[-1, 2]^{n\times d}$.

Step 2: Incorporate the projection step into standard analysis of gradient descent

The following standard result states that a gradient descent step with a sufficiently small stepsize decreases the objective.

Lemma 10.3 (Sufficient Decrease Lemma; Lemma 4.23 and Lemma 4.24 of [16])**.** *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ has Lipschitz gradient with constant K. Then for any $x \in \mathbb{R}^d$ and $\gamma > 0$, we have*

$$f(x) - f(x - \gamma\nabla f(x)) \geq \left(1 - \frac{K\gamma}{2}\right) \|\nabla f(x)\|_2^2. \tag{10.18}$$

Now denote $\{A_t^{\text{grad}}\}_{t\geq 0}$ as the sequence after the gradient step (3.10a) in each iteration, and denote $\{A_t\}_{t\geq 0}$ as the sequence after the projection step (3.10b) in each iteration. We set the stepsize γ such that $\gamma \in (0, \frac{1}{4K})$. Due to the projection we have $A_t \in [0,1]^{n\times d}$ for all $t \geq 0$. Then for the gradient step, using (10.16) it can be verified that

$$A_t^{\text{grad}} = A_{t-1} - \gamma \nabla L(A_{t-1}) \in [-1, 2]^{n\times d}.$$

By Lemma 10.3 we have

$$L(A_{t-1}) - L(A_t^{\text{grad}}) \geq \left(1 - \frac{K\gamma}{2}\right) \|\nabla L(A_{t-1})\|_2^2 \geq 0. \tag{10.19}$$

For the projection step, by Lemma 10.1 we have

$$L(A_t^{\text{grad}}) - L(A_t) \geq 0. \tag{10.20}$$

Combining (10.19) and (10.20), we have

$$L(A_{t-1}) - L(A_t) \geq \left(1 - \frac{K\gamma}{2}\right) \|\nabla L(A_{t-1})\|_F^2 \geq 0. \tag{10.21}$$

Hence, the sequence $\{L(A_t)\}_{t\geq 0}$ is non-increasing. Furthermore, it is straightforward to verify that L is bounded below by 0. Since the sequence $\{L(A_t)\}_{t\geq 0}$ is non-increasing and bounded below by 0, we have

$$\lim_{t\to\infty} L(A_{t-1}) - L(A_t) = 0. \tag{10.22}$$

Plugging (10.22) into (10.21), we have $\lim_{t\to\infty} \|\nabla A_t\|_F = 0$, completing the proof.

10.3 Proof of Theorem 3.2

Since $Y \in \mathcal{M}$, we have A^* is a global minimum if and only if

$$A^*_\Omega = Y_\Omega$$
$$\text{and } A^* \in \mathcal{M}.$$

By Lemma 10.2 any local optima (on the boundary) is a stationary point, so we only consider stationary points for the proof. To show that any stationary point is the global optimum, we separately discuss the three cases: $d = 2$, $d = 3$ and $n = 2$. In each case, we show that any stationary point A satisfies $A \in \mathcal{M}$. Since we have $\nabla L(A) = 0$ for any $A \in \mathcal{M}$, setting the derivative (10.5) to 0 gives $A_\Omega = Y_\Omega$.

10.3.1 $d = 2$

Consider any stationary point A. With $d = 2$, the matrix A has two columns. Assume for contradiction that $A \notin \mathcal{M}$. Denote the sets

$$I_+ := \{i \in [n] : A_{i1} - A_{i2} > 0\} \tag{10.23a}$$

$$I_- := \{i \in [n] : A_{i1} - A_{i2} < 0\}. \tag{10.23b}$$

By the assumption that $A \notin \mathcal{M}$, we have $I_+ \neq \emptyset$ and $I_- \neq \emptyset$.

For each $i \in I_+$, we have

$$\operatorname{sign}\left(\frac{\partial R}{\partial A_{i1}}\right) = \operatorname{sign}\left(\sum_{i' \in I_-} \frac{\partial R_{i,i',1,2}}{\partial A_{i1}}\right) \overset{\text{(i)}}{=} \operatorname{sign}\left(A_{i1} - A_{i2}\right) \overset{\text{(ii)}}{=} 1 \tag{10.24a}$$

$$\operatorname{sign}\left(\frac{\partial R}{\partial A_{i2}}\right) = \operatorname{sign}\left(\lambda \sum_{i' \in I_-} \frac{\partial R_{ii',2,1}}{\partial A_{i2}}\right) \overset{\text{(i)}}{=} \operatorname{sign}\left(A_{i2} - A_{i1}\right) \overset{\text{(ii)}}{=} -1, \tag{10.24b}$$

where the steps (i) are true due to (10.2), and the steps (ii) are true due to the definition (10.23a) of I_+. Likewise for each $i \in I_-$, we have

$$\operatorname{sign}\left(\frac{\partial R}{\partial A_{i1}}\right) = \operatorname{sign}(A_{i1} - A_{i2}) = -1 \tag{10.24c}$$

$$\operatorname{sign}\left(\frac{\partial R}{\partial A_{i2}}\right) = \operatorname{sign}(A_{i2} - A_{i1}) = 1. \tag{10.24d}$$

Case 1: If any entry (i, j) in the rows $I_+ \cup I_-$ is not observed (i.e., not in Ω), then by the gradient expression (10.6) we have

$$\frac{\partial L}{\partial A_{ij}} = \frac{\partial R}{\partial A_{ij}} \neq 0,$$

where the inequality holds due to (10.24). Contradiction to the assumption that A is a stationary point with $\nabla L(A) = 0$.

Case 2: All the entries in the rows $I_+ \cup I_-$ are observed.

Now consider any $i \in I_+$. Setting the gradient expression (10.6) to 0, we have

$$A_{i1} - Y_{i1} + \frac{\partial R}{\partial A_{i1}} = 0$$

$$Y_{i1} = A_{i1} + \frac{\partial R}{\partial A_{i1}}. \tag{10.25a}$$

Likewise, we have

$$Y_{i2} = A_{i2} + \frac{\partial R}{\partial A_{i2}}. \tag{10.25b}$$

Subtracting (10.25b) from (10.25a), we have

$$Y_{i1} - Y_{i2} = A_{i1} - A_{i2} + \left(\frac{\partial R}{\partial A_{i1}} - \frac{\partial R}{\partial A_{i2}} \right) > 0, \tag{10.26a}$$

where the last inequality holds because $(A_{i1} - A_{i2}) > 0$ by the definition (10.23a) of I_+, and because $\frac{\partial R}{\partial A_{i1}} - \frac{\partial R}{\partial A_{i2}} > 0$ due to (10.24a) and (10.24b). Likewise, for each $i \in I_-$,

$$Y_{i1} - Y_{i2} = A_{i1} - A_{i2} + \left(\frac{\partial R}{\partial A_{i1}} - \frac{\partial R}{\partial A_{i2}} \right) < 0. \tag{10.26b}$$

Combining (10.26) contradicts the assumption that $Y \in \mathcal{M}$.

10.3.2 $n = 2$

With $n = 2$, the matrix A has two rows. We prove by induction on the number of columns d. For $d = 1$, we trivially have $A \in \mathcal{M}$. For $d = 2$, the proof in Section 10.3.1 yields the claimed result. Now suppose the claim holds for all $2 \times d$ matrices. We now consider any $2 \times (d+1)$ matrix.

Let $A \in \mathbb{R}^{2\times(d+1)}$ be a stationary point given the observations $Y \in \mathbb{R}^{2\times(d+1)}$. Without loss of generality, we re-index the columns such that $A_{11} \leq A_{12} \leq \ldots \leq A_{1,d+1}$. Now consider the maximum entry in the second row of A.

Case 1: The entry $A_{2,d+1}$ is the maximum in the second row of A.

In this case, column $(d+1)$ contains the maximum for both rows. That is, we have $A_{i,d+1} \geq A_{ij}$ for each $i \in \{1, 2\}$ and each $j \in [d]$. It can be verified that this column $(d+1)$ of the matrix, namely the column $\begin{bmatrix} A_{1,d+1} \\ A_{2,d+1} \end{bmatrix}$ does not contribute to the regularizer R. Hence, the gradient of the submatrix $\{A_{ij}\}_{i\in\{1,2\},j\in[d]}$ remains the same if the last column is removed. That is, for each $i \in \{1, 2\}$ and $j \in [d]$, we have

$$\frac{\partial L(\{A_{ij}\}_{i\in\{1,2\},j\in[d]})}{\partial A_{ij}} = \frac{\partial L(A)}{\partial A_{ij}}.$$

Applying the induction hypothesis on the submatrix $\{A_{ij}\}_{i\in\{1,2\},j\in[d]}$, we have $\{A_{ij}\}_{i\in\{1,2\},j\in[d]} \in \mathcal{M}$. Since the last column $\begin{bmatrix} A_{1,d+1} \\ A_{2,d+1} \end{bmatrix}$ has the maximum entries in both rows, we have $A \in \mathcal{M}$.

Case 2: The entry $A_{2,d+1}$ is not a maximum in the second row.

Assume that a maximum in the second row is A_{2j^*} for some $1 \leq j^* < d$. Then we have $A_{2j^*} > A_{2,d+1}$.

Now consider the entry A_{1j^*}. By assumption we have $A_{1j^*} \leq A_{1,d+1}$. If $A_{1j^*} = A_{1,d+1}$, then the two entries in column j^* are both the maximum in their respective rows. Applying a similar inductive argument as in Case 1 to the submatrix $\{A_{ij}\}_{i\in\{1,2\},j\in[d+1]\setminus\{j^*\}}$ yields $A \in \mathcal{M}$. It remains to consider the case of $A_{1j^*} < A_{1,d+1}$.

We first analyze row 2. Using (10.2) combined with the fact that A_{2j^*} is the maximum entry in row 2, we have $\frac{\partial R}{\partial A_{2j^*}} \geq 0$. Moreover, since $A_{1j^*} < A_{1,d+1}$ and $A_{2j^*} > A_{2,d+1}$, the quadruple

$(1, 2, j, d+1)$ is a conflicting quadruple, and hence we have the strict inequality

$$\frac{\partial R}{\partial A_{2j^*}} > 0. \tag{10.27}$$

On the other hand, we have $\frac{\partial R}{\partial A_{2,d+1}} \leq 0$, because for any conflicting quadruple $(2, d+1, 1, j)$ for some $j \in [d]$ that contributes to the derivative $\frac{\partial R}{\partial A_{2j}}$, we have

$$\operatorname{sign}\left(\frac{\partial R_{2,1,d+1,j}}{\partial A_{2j}}\right) \overset{\text{(i)}}{=} \operatorname{sign}(A_{2,d+1} - A_{2,j}) \overset{\text{(ii)}}{=} -\operatorname{sign}(A_{1,d+1} - A_{1,j}) \overset{\text{(iii)}}{=} -1,$$

where step (i) is true due to (10.2); step (ii) is true because $(2, 1, d+1, j)$ is assumed to be a conflicting quadruple and hence $(A_{2,d+1} - A_{2,j})(A_{1,d+1} - A_{1,j}) < 0$; step (ii) is true because by assumption $A_{1,d+1}$ is the maximum entry in the first row. Furthermore, the quadruple $(1, 2, j^*, d+1)$ is a conflicting quadruple, so we have strict inequality

$$\frac{\partial R}{\partial A_{2,d+1}} < 0. \tag{10.28}$$

Now consider whether the entries A_{2,j^*} and $A_{2,d+1}$ are observed. If either A_{2,j^*} or $A_{2,d+1}$ is not observed, then combining the gradient expression (10.6) with the strict inequalities (10.27) and (10.28), we have $\frac{\partial L}{\partial A_{2,d+1}} \neq 0$ or $\frac{\partial L}{\partial A_{2,j^*}} \neq 0$, contradicting the assumption that A is a stationary point. Hence, both A_{2,j^*} and $A_{2,d+1}$ are observed. Setting the gradient expression (10.6) to 0 respectively for the two entries $(2, j^*)$ and $(2, d+1)$, we have

$$Y_{2j^*} - Y_{2,d+1} = (A_{2j^*} - A_{2,d+1}) + \frac{\partial R}{\partial A_{2j^*}} - \frac{\partial R}{\partial A_{2,d+1}} > 0, \tag{10.29a}$$

where the inequality holds because $(A_{2j^*} - A_{2,d+1}) > 0$ as A_{2j^*} is the maximum entry in the second row, and because of (10.27) and (10.28).

Now we analyze row 1. Using a similar argument as in row 2, we have $\frac{\partial R}{\partial A_{1,d+1}} > 0$ because $A_{1,d+1}$ is the maximum entry in row 1, and strict inequality holds due to the existence of the conflicting quadruple $(1, 2, j^*, d+1)$. Moreover, we have $\frac{\partial R}{\partial A_{1j^*}} < 0$, because A_{2j^*} is the maximum entry in row 2 and the same conflicting quadruple $(1, 2, j^*, d+1)$. Similar to the analysis of row 2, we derive that both entries $(1, j^*)$ and $(1, d+1)$ are observed. Therefore,

$$Y_{1j^*} - Y_{1,d+1} = (A_{1j^*} - A_{1,d+1}) + \frac{\partial R}{\partial A_{1j^*}} - \frac{\partial R}{\partial A_{1,d+1}} < 0. \tag{10.29b}$$

Combining (10.29) contradicts the assumption that $Y \in \mathcal{M}$. Therefore, the entry $A_{2,d+1}$ is the maximum in row 2. From Case 1 we have $A \in \mathcal{M}$ for any $2 \times (d+1)$ matrices, completing the inductive step.

10.3.3 $d = 3$

With $d = 3$, the matrix has 3 columns. We consider the maximum entry in each row of the matrix. If a row has multiple maxima, one is chosen arbitrarily unless otherwise specified.

Case 1: The maxima in all the n rows of the matrix lie in the same column.

Without loss of generality, assume that the column containing all the maxima is column 3. It can be verified that all entries in column 3 do not contribute to the regularizer R. Applying the proof of the $d = 2$ case in Appendix 10.3.1 to the submatrix $\{A_{ij}\}_{i\in[n],j\in\{1,2\}}$ yields $\{A_{ij}\}_{i\in[n],j\in\{1,2\}} \in \mathcal{M}$. Since column 3 contains the maximum of each row, we have $A \in \mathcal{M}$.

Case 2: The maxima of the n rows lie in two different columns.

If the 3 entries within each row are identical, then we have $A \in \mathcal{M}$, so it remains to consider the case where there exists a row whose values are not all identical. Without loss of generality, we assume that the entries are not all identical in row 1. We re-index the columns such that the first row is non-decreasing. Hence, we have $A_{11} < A_{13}$. We also re-index the rows, so that rows whose maxima are in the same column are grouped together. Then the matrix A is in one of the two following forms:

$$\left[\begin{array}{ccc} \text{min} & * & \text{max} \\ \vdots & \vdots & \vdots \\ * & * & \text{max} \\ \hline * & \text{max} & * \\ \vdots & \vdots & \vdots \\ * & \text{max} & * \end{array}\right] \tag{10.30a}$$

or

$$\left[\begin{array}{ccc} \text{min} & * & \text{max} \\ \vdots & \vdots & \vdots \\ * & * & \text{max} \\ \hline \text{max} & * & * \\ \vdots & \vdots & \vdots \\ \text{max} & * & * \end{array}\right], \tag{10.30b}$$

where we use “min” and “max” to indicate that the matrix entry is respectively a minimum or a maximum of its row (allowing ties). We use $*$ to indicate a general matrix entry, and use the horizontal line to indicate that the matrix structure decomposes into two blocks of rows. We denote the upper block and the lower block of the matrix as A_{U} and A_{L}, respectively, so that the matrix is also written as $\left[\begin{array}{c} A_{\text{U}} \\ \hline A_{\text{L}} \end{array}\right]$. We denote the row indices of the upper block and the lower block as $I_{\text{U}}, I_{\text{L}} \subseteq [n]$, respectively. By the assumption of the case, we have $I_{\text{U}}, I_{\text{L}} \neq \emptyset$.

Case 2.1: We consider the matrix form (10.30a).

We assume that in the lower block A_{L}, the entries in column 2 are strictly greater than the entries in column 3 within each row. That is, we assume $A_{i2} > A_{i3}$ for each $i \in I_{\text{L}}$. This assumption is without loss of generality, because otherwise we have $A_{i2} = A_{i3}$, so that one can move row i to the upper block of the matrix.

Case 2.1.1: There exists a strict min-entry in column 2 in some row of the upper block. That is, there exists $i^* \in I_{\text{U}}$ such that $A_{i^*1} > A_{i^*2}$. Since column 3 contains the maximum for all rows in the upper block, we have the strict inequality $A_{i^*2} < A_{i^*3}$.

Using (10.2), it can be verified that

$$\frac{\partial R}{\partial A_{i^*2}} < 0 \tag{10.31a}$$

$$\frac{\partial R}{\partial A_{i^*3}} > 0, \tag{10.31b}$$

where strict inequalities hold because the quadruple $(i^*, i', 2, 3)$ is a conflicting quadruple for each $i' \in I_{\mathrm{L}}$. Setting the gradient (10.6) for the stationary point A and combining with (10.31), we have the entries $(i^*, 1)$ and $(i^*, 2)$ must both be observed. Subtracting the gradient expression (10.6) on the entries $(i^*, 1)$ and $(i^*, 2)$, we have

$$Y_{i^*,2} - Y_{i^*,3} = (A_{i^*,2} - A_{i^*,3}) + \left(\frac{\partial R}{\partial A_{i^*,2}} - \frac{\partial R}{\partial A_{i^*,3}}\right) < 0,$$

where the last inequality holds by combining the fact of $A_{i^*2} < A_{i^*3}$ with inequalities (10.31). Hence, we have

$$Y_{i^*2} < Y_{i^*3}. \tag{10.32}$$

Now consider the case where there exists a min-entry in column 3 in the lower block, and denote this row as $i_{\mathrm{L}} \in I_{\mathrm{L}}$. Since we assume $A_{i2} > A_{i3}$ for each $i \in I_{\mathrm{L}}$ for Case 2.1, we have $(i_{\mathrm{L}}, 3)$ is a strict min-entry. Note that $(i^*, i_{\mathrm{L}}, 2, 3)$ is a conflicting quadruple. Using an argument similar to the derivation of (10.32), we have

$$Y_{i_{\mathrm{L}},2} > Y_{i_{\mathrm{L}},3}. \tag{10.33}$$

Combining (10.32) and (10.33) contradicts the assumption that $Y \in \mathcal{M}$. Hence, there does not exist any min-entry in column 3 in the lower block. Hence, the min-entry must lie in column 1 in the lower block, and all such min-entries are strict. Now the matrix A can be written in the form

$$\left[\begin{array}{ccc} \text{min} & * & \text{max} \\ \vdots & \vdots & \vdots \\ * & * & \text{max} \\ \hline \text{min} & \text{max} & * \\ \vdots & \vdots & \vdots \\ \text{min} & \text{max} & * \end{array}\right].$$

Now consider any row $i_{\mathrm{L}} \in I_{\mathrm{L}}$. We have $\frac{\partial R}{\partial A_{i_{\mathrm{L}},2}} > 0$ because column 2 contains a max-entry, and strict inequality holds due to the conflicting quadruple $(i^*, i_{\mathrm{L}}, 2, 3)$. On the other hand, we have $\frac{\partial R}{\partial A_{i_{\mathrm{L}},3}} \leq 0$, because no quadruple within the lower block contributes to the regularizer, and in the upper block column 3 contains the max-entry. Moreover, we have the strict inequality $\frac{\partial R}{\partial A_{i_{\mathrm{L}},3}} < 0$ due to the conflicting quadruple $(i^*, i_{\mathrm{L}}, 2, 3)$ again. Setting the gradient expression (10.6) for the stationary point A, we have that both entries $(i_{\mathrm{L}}, 2)$ and $(i_{\mathrm{L}}, 3)$ are observed. Subtracting the two gradient expression, we have

$$Y_{i_{\mathrm{L}},2} > Y_{i_{\mathrm{L}},3}. \tag{10.33'}$$

Combining (10.32) and (10.33') yields a contradiction to the assumption of $Y \in \mathcal{M}$, completing the proof of Case 2.1.1.

Case 2.1.2: There does not exist a min-entry in column 2 in the upper block.

In this case, the matrix is in the form

$$\left[\begin{array}{ccc} \text{min} & * & \text{max} \\ \vdots & \vdots & \vdots \\ \text{min} & * & \text{max} \\ \hline * & \text{max} & * \\ \vdots & \vdots & \vdots \\ * & \text{max} & * \end{array}\right].$$

We consider column 2 in the upper block. If $A_{i2} = A_{i3}$ for all $i \in I_{\mathrm{U}}$, then column 2 of the entire matrix only contains max-entries, and we apply the proof of Case 1 to column 2. It remains to consider the case where there exists some $i \in I_{\mathrm{U}}$ such that $A_{i_{\mathrm{U}},2} < A_{i_{\mathrm{U}},3}$. We have $\frac{\partial R}{\partial A_{i_{\mathrm{U}},3}} > 0$, where strict inequality holds due to the conflicting quadruple $(I_{\mathrm{U}}, I_{\mathrm{L}}, 2, 3)$ for any $i_{\mathrm{L}} \in I_{\mathrm{L}}$. Moreover, we have $\frac{\partial R}{\partial A_{i_{\mathrm{U}},2}} \leq 0$, because no quadruple within the upper block contributes to the regularizer, and in the lower block column 2 contains the max-entries. We have the strict inequality $\frac{\partial R}{\partial A_{i_{\mathrm{U}},2}} < 0$ due to the conflicting quadruple $(I_{\mathrm{U}}, I_{\mathrm{L}}, 2, 3)$ for any $i_{\mathrm{L}} \in I_{\mathrm{L}}$. Using the gradient expression (10.6), both entries $(i_{\mathrm{U}}, 2)$ and $(i_{\mathrm{U}}, 3)$ are observed, and we have

$$Y_{i_{\mathrm{U}},2} < Y_{i_{\mathrm{U}},3}. \tag{10.34a}$$

Now consider column 3 in the lower block. If for any row $i_{\mathrm{L}} \in I_{\mathrm{L}}$, column 3 contains the min-entry. Then due to the quadruple $(i_{\mathrm{U}}, i_{\mathrm{L}}, 2, 3)$ we have

$$Y_{i_{\mathrm{L}},2} < Y_{i_{\mathrm{L}},3}. \tag{10.34b}$$

Combining (10.34) yields a contradiction to the assumption that $Y \in \mathcal{M}$. Hence, column 3 does not contain any min-entry in the lower block. That is, the matrix can be written in the form

$$\left[\begin{array}{ccc} \text{min} & * & \text{max} \\ \vdots & \vdots & \vdots \\ \text{min} & * & \text{max} \\ \hline \text{min} & \text{max} & * \\ \vdots & \vdots & \vdots \\ \text{min} & \text{max} & * \end{array}\right].$$

Note that column 1 of the entire matrix only contains min-entries. Applying Case 1 to the minima (instead of the maxima) completes the proof of Case 2.1.2.

Case 2.2: We consider the form (10.30b).

Without loss of generality, we assume strict inequality $A_{i_{\mathrm{L}},1} > A_{i_{\mathrm{L}},3}$ for all $i_{\mathrm{L}} \in I_{\mathrm{L}}$. Otherwise, we have $A_{i_{\mathrm{L}},1} = A_{i_{\mathrm{L}},3}$ and one can move row i_{L} to the upper block. Assume that column

3 in the lower block contains a min-entry for some row $i_{\mathrm{L}} \in I_{\mathrm{L}}$. Combining row i_{L} with row 1 gives a conflicting quadruple $(1, i_{\mathrm{L}}, 1, 3)$. Using an argument similar to Case 2.1, we have

$$\begin{aligned} &Y_{11} < Y_{13} \\ &Y_{i_{\mathrm{L}},1} > Y_{i_{\mathrm{L}},3}, \end{aligned}$$

contradicting to the assumption $Y \in \mathcal{M}$. Hence, column 3 in the lower block does not contain any min-entry. Therefore, the matrix can be written as

$$\left[\begin{array}{ccc} \text{min} & * & \text{max} \\ \vdots & \vdots & \vdots \\ * & * & \text{max} \\ \hline \text{max} & \text{min} & * \\ \vdots & \vdots & \vdots \\ \text{max} & \text{min} & * \end{array}\right].$$

For any i_{L}, the quadruple $(1, i_{\mathrm{L}}, 1, 3)$ is again a conflicting quadruple. We have

$$\begin{aligned} &Y_{11} < Y_{13} \\ &Y_{i_{\mathrm{L}},1} > Y_{i_{\mathrm{L}},3}, \end{aligned}$$

contradicting to the assumption $Y \in \mathcal{M}$, completing the proof of Case 2.2.

Case 3: The maxima of the n rows span all the 3 columns. That is, the matrix can be written in the form:

$$\left[\begin{array}{ccc} \text{min} & * & \text{max} \\ \vdots & \vdots & \vdots \\ * & * & \text{max} \\ \hline * & \text{max} & * \\ \vdots & \vdots & \vdots \\ * & \text{max} & * \\ \hline \text{max} & * & * \\ \vdots & \vdots & \vdots \\ \text{max} & * & * \end{array}\right].$$

Denote the three blocks in the matrix as $A_{\mathrm{U}}, A_{\mathrm{M}}$ and A_{L} respectively, so that the matrix is also written as $\left[\begin{array}{c} A_{\mathrm{U}} \\ \hline A_{\mathrm{M}} \\ \hline A_{\mathrm{L}} \end{array}\right]$. Denote the corresponding sets of row indices as $I_{\mathrm{U}}, I_{\mathrm{M}}$ and I_{L}, respectively. Without loss of generality, we assume

$$\begin{aligned} &A_{i2} > A_{i3} \qquad \forall i \in I_{\mathrm{M}} \\ &A_{i1} > \{A_{i2}, A_{i3}\} \qquad \forall i \in I_{\mathrm{L}}. \end{aligned}$$

Otherwise, we may move the rows in the middle block to the upper block, and move the rows in the lower block to the upper or middle blocks.

Now consider the lower block. Assume that there exists some min-entry in column 3 of the lower block. That is, assume that there exists some $i_{\mathrm{L}} \in I_{\mathrm{L}}$, such that $A_{i_{\mathrm{L}},3}$ is a min-entry. Then the quadruple $(1, i_{\mathrm{L}}, 1, 3)$ is a conflicting quadruple. Hence, we have

$$\begin{aligned} &Y_{11} < Y_{13} \\ &Y_{i_{\mathrm{L}},1} > Y_{i_{\mathrm{L}},3}, \end{aligned}$$

contradicting with the assumption that $Y \in \mathcal{M}$. Hence, there does not exist any min-entry in column 3 of the lower block. Then the matrix can be written in the form:

$$\left[\begin{array}{ccc} \text{min} & * & \text{max} \\ \vdots & \vdots & \vdots \\ * & * & \text{max} \\ \hline * & \text{max} & * \\ \vdots & \vdots & \vdots \\ * & \text{max} & * \\ \hline \text{max} & \text{min} & * \\ \vdots & \vdots & \vdots \\ \text{max} & \text{min} & * \end{array}\right].$$

Now consider row 1. The quadruple $(1, i_{\mathrm{L}}, 1, 3)$ is a conflicting quadruple for each row $i_{\mathrm{L}} \in I_{\mathrm{L}}$ in the lower block. Hence, we have

$$Y_{11} < Y_{13}. \tag{10.35}$$

Assume without loss of generality that there exists some $i_{\mathrm{M}} \in I_{\mathrm{M}}$, such that $A_{i_{\mathrm{M}},1} < A_{i_{\mathrm{M}},2}$. Otherwise, the first column in the middle block contains all max-entries, and the matrix reduces to Case 2.2. Now consider any row $i_{\mathrm{L}} \in I_{\mathrm{L}}$. The quadruple $(i_{\mathrm{M}}, i_{\mathrm{L}}, 1, 2)$ is a conflicting quadruple. Hence, we have

$$Y_{i_{\mathrm{L}},1} > Y_{i_{\mathrm{L}},2}. \tag{10.36}$$

Combining (10.35) and (10.36) along with the assumption that $Y \in \mathcal{M}$, we have

$$Y_{i2} \leq Y_{i1} \leq Y_{i3}, \qquad \forall i \in [n]. \tag{10.37}$$

Now consider row i_{M} again in the middle block. Assume $A_{i_{\mathrm{M}},1}$ is the min-entry in row i_{M}. The quadruple $(i_{\mathrm{M}}, i_{\mathrm{L}}, 1, 2)$ is a conflicting quadruple for any $i_{\mathrm{L}} \in I_{\mathrm{L}}$. Hence, we have

$$Y_{i_{\mathrm{M}},2} > Y_{i_{\mathrm{M}},1},$$

contradicting (10.37). Hence, it must be the case that $A_{i_{\mathrm{M}},3}$ is the min-entry. Then we have $A_{I_{\mathrm{M}},2} > A_{I_{\mathrm{M}},1} \geq A_{i_{\mathrm{M}},3}$. Now again consider any row $i_{\mathrm{L}} \in I_{\mathrm{L}}$. Recall that we have established that $A_{i_{\mathrm{L}},3}$ cannot be a min-entry, so we have $A_{i_{\mathrm{L}},3} > A_{i_{\mathrm{L}},2}$. Then the quadruple $(i_{\mathrm{M}}, i_{\mathrm{L}}, 2, 3)$ is a conflicting quadruple. Hence, we have

$$Y_{i_{\mathrm{M}},2} > Y_{i_{\mathrm{M}},3},$$

again contradicting (10.37), completing the proof of Case 3.

Finally, combining the three cases yields the claimed result.

10.4 Proof of Proposition 3.3

Without loss of generality, we consider any $j, j' \in [d]$ such that

$$A_{1,j} < A_{1,j'} \tag{10.38a}$$
$$A_{2,j} > A_{2,j'}, \tag{10.38b}$$

and prove that

$$Y_{1,j} < Y_{1,j'}, \quad \text{if } (1,j),(1,j') \in \Omega.$$

First, consider the quadruple $(1, 2, j, j')$. By (10.38), it is a conflicting quadruple. By (10.2), we have

$$\frac{\partial R_{1,2,j,j'}}{\partial A_{1,j}} < 0 \tag{10.39a}$$
$$\frac{\partial R_{1,2,j,j'}}{\partial A_{1,j'}} > 0. \tag{10.39b}$$

Now consider quadruples involving any other column $k \in [d] \setminus \{j, j'\}$. We consider all possible orderings of the entries in column k relative to the columns j and j' as follows (we bold the entries in column k for better readability).

Case 1:

$$\begin{matrix} \mathbf{A_{1,k}} \le A_{1,j} < A_{1,j'} \\ \mathbf{A_{2,k}} \le A_{2,j'} < A_{2,j} \end{matrix}, \quad \text{or} \quad \begin{matrix} A_{1,j} < A_{1,j'} \le \mathbf{A_{1,k}} \\ A_{2,j'} < A_{2,j} \le \mathbf{A_{2,k}} \end{matrix}$$

It can be verified that column k does not form conflicting quadruples with columns j or j'. Hence, column k does not contribute to the gradient of the regularizer with respect to A_{1j} or $A_{1j'}$:

$$\frac{\partial R_{1,2,j,k}}{\partial A_{1j}} = \frac{\partial R_{1,2,j',k}}{\partial A_{1j'}} = 0.$$

Case 2:

$$\begin{matrix} A_{1,j} < \mathbf{A_{1,k}} \le A_{1,j'} \\ \mathbf{A_{2,k}} \le A_{2,j'} < A_{2,j} \end{matrix}, \quad \text{or} \quad \begin{matrix} A_{1,j} < A_{1,j'} \le \mathbf{A_{1,k}} \\ A_{2,j'} \le \mathbf{A_{2,k}} < A_{2,j} \end{matrix}$$

It can be verified that column k contributes a negative gradient to the regularizer with respect to A_{1j}, and no gradient to the regularizer with respect to $A_{1j'}$:

$$\frac{\partial R_{12jk}}{\partial A_{1j}} < 0 = \frac{\partial R_{12j'k}}{\partial A_{1j'}}.$$

Case 3:

$$\begin{array}{c} A_{1,j} \leq \mathbf{A}_{\mathbf{1},\mathbf{k}} < A_{1,j'} \\ A_{2,j'} < A_{2,j} \leq \mathbf{A}_{\mathbf{2},\mathbf{k}} \end{array}, \quad \text{or} \quad \begin{array}{c} \mathbf{A}_{\mathbf{1},\mathbf{k}} \leq A_{1,j} < A_{1,j'} \\ A_{2,j'} < \mathbf{A}_{\mathbf{2},\mathbf{k}} \leq A_{2,j} \end{array}$$

It can be verified that column k contributes no gradient to the regularizer with respect to $A_{1,j}$, and a positive gradient to the regularizer with respect to $A_{1j'}$:

$$\frac{\partial R_{1j2k}}{\partial A_{1j}} = 0 < \frac{\partial R_{1j'2k}}{\partial A_{1j'}}.$$

Case 4:

$$\begin{array}{c} A_{1,j} < \mathbf{A}_{\mathbf{1},\mathbf{k}} < A_{1,j'} \\ A_{2,j'} < \mathbf{A}_{\mathbf{2},\mathbf{k}} < A_{2,j}, \end{array}$$

It can be verified that column k contributes a negative gradient to the regularizer with respect to A_{1j}, and a positive gradient to the regularizer with respect to $A_{1j'}$:

$$\frac{\partial R_{1j2k}}{\partial A_{1j}} < 0 < \frac{\partial R_{1j'2k}}{\partial A_{1j'}}.$$

Case 5:

$$\begin{array}{c} \mathbf{A}_{\mathbf{1},\mathbf{k}} < A_{1,j} < A_{1,j'} \\ A_{2,j'} < A_{2,j} < \mathbf{A}_{\mathbf{2},\mathbf{k}}, \end{array}$$

It can be verified that column k contributes positive gradients to the regularizer with respect to both A_{1j} and $A_{1j'}$. By (10.1), we have

$$\begin{aligned} \frac{\partial R_{1j2k}}{\partial A_{1j}} &= 2(A_{1j} - A_{1k})(A_{2j} - A_{2k})^2 \\ \frac{\partial R_{1j'2k}}{\partial A_{1j'}} &= 2(A_{1j'} - A_{1k})(A_{2j'} - A_{2k})^2, \end{aligned}$$

and hence

$$0 < \frac{\partial R_{1j2k}}{\partial A_{1j}} < \frac{\partial R_{1j'2k}}{\partial A_{1j'}}.$$

Finally, combining all the 5 cases, it can be verified that they cover all possible orderings of the entries in column k relative to columns j and j'. Moreover, we have

$$\frac{\partial R_{1j2k}}{\partial A_{1j}} < \frac{\partial R_{1j'2k}}{\partial A_{1j'}} \qquad \forall k \in [d] \setminus \{j, j'\}. \tag{10.40}$$

Plugging (10.39) and (10.40) to (10.4), we have

$$\begin{aligned}\frac{\partial R}{\partial A_{1j}} &= 4\lambda\left(\frac{\partial R_{1,2,j,j'}}{\partial A_{1j}} + \sum_{k\in[d]\backslash\{j,j'\}}\frac{\partial R_{1j2k}}{\partial A_{1j}}\right)\\ &< 4\lambda\left(\frac{\partial R_{1,2,j',j'}}{\partial A_{1j'}} + \sum_{k\in[d]\backslash\{j,j'\}}\frac{\partial R_{1j'2k}}{\partial A_{1j'}}\right) = \frac{\partial R}{\partial A_{1j'}} \end{aligned}\tag{10.41}$$

Since we assume $(1, j), (1, j') \in \Omega$, using the gradient expression (10.6), we have

$$Y_{1j'} - Y_{1j} = (A_{1j} - A_{1j'}) + \left(\frac{\partial R}{\partial A_{1j}} - \frac{\partial R}{\partial A_{1j'}}\right) < 0,$$

where the inequality holds due to (10.38a) and (10.41), completing the proof.

10.5 Proof of Theorem 3.4

To present the main ideas of the proof, we first prove the following lemma under a simplified setting of Theorem 3.4, where the partition includes two subsets, $[d] = S \cup \overline{S}$ under full observations $\Omega = [n] \times [d]$. Then we present how to generalize Lemma 10.4 to any partition and partial observations.

Lemma 10.4. *Consider any matrix $Y \in \mathbb{R}^{n\times d}$, and full observations $\Omega = [n] \times [d]$. Consider $n = 2$. Assume there exists a partition of columns $[d] = S \cup \overline{S}$, such that any column in $\overline{S}$ dominates any column in S. That is, for any $j \in S$ and $j' \in \overline{S}$, we have*

$$Y_{i,j} < Y_{i,j'} \qquad \forall i \in \{1, 2\}. \tag{10.42}$$

Then we have the same relation for any stationary point A. That is,

$$A_{i,j} < A_{i,j'} \qquad \forall i \in \{1, 2\}, \forall j \in S \text{ and } j' \in \overline{S}. \tag{10.43}$$

10.5.1 Proof of Lemma 10.4

We decompose the proof into the following steps.

Step 1: Show that conflicting quadruples cannot lie across $(S, \overline{S})$

Assume for contradiction that there exists a conflicting quadruple across $(S, \overline{S})$. That is, assume that there exists $j \in S$ and $j' \in \overline{S}$ such that $(A_{1,j} - A_{1,j'})(A_{2j} - A_{2j'}) < 0$. Applying Proposition 3.3, we have $(Y_{1,j} - Y_{1,j'})(Y_{2j} - Y_{2j'}) < 0$, contradicting the dominance assumption (10.42). Hence, all conflicting quadruples must lie within S, or within $\overline{S}$. Formally, for any $j, j' \in [d]$ such that $(1, 2, j, j')$ is a conflicting quadruple, we have either $j, j' \in S$ or $j, j' \in \overline{S}$.

Step 2: Partition columns into blocks We partition the columns into blocks $[d] = B_1 \cup B_2 \cup \ldots \cup B_K$ for some $K \geq 2$, such that the following conditions are satisfied:

(a) For $k \in [K]$, the block B_k includes columns only from S, or only from $\overline{S}$. That is, for each $k \in [K]$ we have $B_k \subseteq S$ or $B_k \subseteq \overline{S}$.

(b) For each $k \in [K-1]$, the blocks B_k and B_{k+1} are in different sets of the partition $(S, \overline{S})$. That is, for each $k \in [K-1]$, we have either $B_k \subseteq S$ and $B_{k+1} \subseteq \overline{S}$, or $B_k \subseteq \overline{S}$ and $B_{k+1} \subseteq S$.

(c) For each $k \in [K-1]$, the columns in B_{k+1} dominates the columns in B_k. That is,

$$A_{ij} \leq A_{ij'} \qquad \forall i \in \{1,2\}, \forall k \in [K], \forall j \in B_k, \text{ and } \forall j' \in B_{k+1}.$$

Due to Step 1, all conflicting quadruples lie within S or $\overline{S}$, so it can be verified that a partition of blocks with $K \geq 2$ satisfying (a)-(c) exists.

Step 3: Show that A satisfies the claimed dominance relation (10.43).

We define

$$k_{\mathrm{H}} := \max\{k \in [K] : B_k \subseteq S\} \tag{10.44a}$$

$$k_{\mathrm{L}} := \min\{k \in [K] : B_k \subseteq \overline{S}\}, \tag{10.44b}$$

where ties are broken arbitrary. That is, $B_{k_{\mathrm{L}}}$ is the block that is ordered the lowest among all blocks consisting of columns in $\overline{S}$, and $B_{k_{\mathrm{H}}}$ is the block that is ordered the highest among all blocks consisting of columns in S. Furthermore, we define

$$j_{\mathrm{H}} := \operatorname*{argmax}_{j \in B_{k_{\mathrm{H}}}} A_{1j} \tag{10.45a}$$

$$j_{\mathrm{L}} := \operatorname*{argmin}_{j \in B_{k_{\mathrm{L}}}} A_{1,j}, \tag{10.45b}$$

where ties are broken arbitrarily. That is, $(1, j_{\mathrm{H}})$ is the the maximum entry of A in row 1 among columns $B_{k_{\mathrm{H}}}$, and $(1, j_{\mathrm{L}})$ is the minimum entry of A in row 1 among columns $B_{k_{\mathrm{L}}}$.

Case 1: $A_{1,j_{\mathrm{H}}} < A_{1,j_{\mathrm{L}}}$

By condition (c) of the construction, we have $k_{\mathrm{L}} > k_{\mathrm{H}}$. Hence, for all $j \in S$ and $j' \in \overline{S}$, we have

$$A_{1j} \overset{\text{(i)}}{\leq} A_{1j_{\mathrm{H}}} < A_{1j_{\mathrm{L}}} \overset{\text{(ii)}}{\leq} A_{1j'},$$

where steps (i) and (ii) are true due to the definitions (10.44) and (10.45) along with the fact that $k_{\mathrm{L}} > k_{\mathrm{H}}$. This completes Case 1.

Case 2: $A_{1,j_{\mathrm{H}}} \geq A_{1,j_{\mathrm{L}}}$

If any conflicting quadruple includes the entry $A_{1,j_{\mathrm{H}}}$, then from Step 1 we have that all such conflicting quadruples are within S. By the definition (10.44a) of k_{H} and the definition (10.45a) of j_{H}, the entry $A_{1,j_{\mathrm{H}}}$ is the maximum entry among all entries in row 1 among column S. Hence, we have

$$\frac{\partial R}{\partial A_{1,j_{\mathrm{H}}}} \geq 0 \tag{10.46a}$$

and likewise

$$\frac{\partial R}{\partial A_{1,j_{\mathrm{L}}}} \leq 0. \tag{10.46b}$$

Using the gradient expression (10.6), we have

$$Y_{1,j_{\mathrm{L}}} - Y_{1,j_{\mathrm{H}}} = (A_{1,j_{\mathrm{L}}} - A_{1,j_{\mathrm{H}}}) + \lambda \left(\frac{\partial R}{\partial A_{1,j_{\mathrm{L}}}} - \frac{\partial R}{\partial A_{1,j_{\mathrm{H}}}} \right) \leq 0,$$

where the last inequality is true due to (10.46) along with the assumption of the case. This contradicts the assumption (10.42) that the columns $\overline{S}$ dominates the columns S, completing Case 2.

Combining the two cases completes the proof.

10.5.2 Proof of Theorem 3.4

Now we extend Lemma 10.4 to partial observations, stated as follows.

Lemma 10.5. *Consider any matrix $Y \in [0,1]^{n\times d}$, and partial observations $\Omega \subseteq [n] \times [d]$. Consider $n = 2$. Assume there exists a partition of columns $[d] = S \cup \overline{S}$, such that any column in $\overline{S}$ dominates any column in S. That is, we have*

$$Y_{i,j} < Y_{i,j'} \qquad \forall i \in \{1,2\}, \forall j \in S \textit{ and} \, \forall j' \in \overline{S}. \tag{10.47}$$

Moreover, we assume that for each $j \in S, j' \in \overline{S}$, we have

$$\exists i \in \{1,2\} \textit{ such that } (i,j),(i,j') \in \Omega. \tag{10.48}$$

Then for any stationary point A, we have

$$A_{i,j} < A_{i,j'} \qquad \forall i \in \{1,2\}, \forall j \in S \textit{ and } j' \in \overline{S}. \tag{10.49}$$

We first use Lemma 10.5 to prove Theorem 3.4, and then prove Lemma 10.5. To prove Theorem 3.4, applying Lemma 10.5 with

$$\begin{aligned} S &= \cup_{r=1}^{k} S_r \\ \overline{S} &= \cup_{r=k+1}^{m} S_r \end{aligned}$$

with every $k \in [m-1]$ gives

$$A_{ij} < A_{ij'} \qquad \forall i \in \{1,2\}, \forall j \in S_k, \text{ and } \forall j' \in S_{k+1},$$

completes the proof of Theorem 3.4. It now remains to prove Lemma 10.5.

Proof of Lemma 10.5 We extend the three steps in the proof of Lemma 10.4 to partial observations as follows.

Step 1: Show that conflicting quadruples cannot lie across $(S, \overline{S})$

Assume for contradiction that $(1,2,j,j')$ is a conflicting quadruple with $j \in S$ and $j' \in \overline{S}$. Assume without loss of generality that

$$A_{1j} < A_{1j'} \tag{10.50a}$$

$$A_{2j} > A_{2j'}. \tag{10.50b}$$

If all the 4 entries in this quadruple are observed, then applying Proposition 3.3 yields a contradiction. By assumption (10.48), one pair in the quadruple is observed. If the pair $A_{2j} > A_{2j'}$ is observed, then applying Proposition 3.3 gives $Y_{2j} > Y_{2j'}$, yielding a contradiction to (10.47). Hence, it remains to consider the case that the pair $A_{1j} < A_{1j'}$ is observed.

We first show that one entry in the pair $A_{2j} > A_{2j'}$ must be observed. Using the same argument as in the proof of Proposition 3.3, we have

$$\frac{\partial R}{\partial A_{2j}} > \frac{\partial R}{\partial A_{2j'}}. \tag{10.51}$$

If both entries in this pair are unobserved, then combining (10.51) with the gradient expression (10.6), we have

$$\frac{\partial L}{\partial A_{2j}} = \lambda \frac{\partial R}{\partial A_{2j}} > \lambda \frac{\partial R}{\partial A_{2j'}} = \frac{\partial L}{\partial A_{2j'}},$$

contradicting the assumption that A is a stationary point with a gradient of 0, and hence $\frac{\partial L}{\partial A_{2j}} = \frac{\partial L}{\partial A_{2j'}} = 0$. Hence, one entry in the pair $A_{2j} > A_{2j'}$ is observed. We now separately discuss the two cases depending on which entry in this pair is observed.

Case 1: A_{2j} is observed and $A_{2j'}$ is unobserved.

Since $A_{2j'}$ is unobserved, we have

$$\frac{\partial L}{\partial A_{2j'}} = \frac{\partial R}{\partial A_{2j'}} = 0. \tag{10.52}$$

Since $(1, 2, j, j')$ is a conflicting quadruple, we have

$$\frac{\partial R_{1,2,jj'}}{\partial A_{2j'}} < 0. \tag{10.53}$$

Combining (10.52) and (10.53), there must exist some $k \in [d]$ such that $\frac{\partial R_{1,2,j'k}}{\partial A_{2j'}} > 0$. That is, we have

$$A_{1j'} < A_{1k} \tag{10.54a}$$
$$A_{2j'} > A_{2k}. \tag{10.54b}$$

If $k \in S$, then $(1, 2, j', k)$ is a quadruple across the partition $(S, \overline{S})$. Recall by the assumption of the case that $A_{2j'}$ is unobserved, by condition (10.48), the pair $A_{1j'} < A_{1k}$ must be observed. Applying Proposition 3.3 yields $Y_{1j'} < Y_{1k}$, contradicting the dominance assumption (10.47).

It now remains to consider $k \in \overline{S}$. Recall by the assumption of the case that $A_{2j'}$ is unobserved. If A_{2k} is also unobserved, then the applying the arguments in Proposition 3.3 to the conflicting quadruple $(1, 2, j, k)$, we have

$$\frac{\partial R}{\partial A_{2j'}} < \frac{\partial R}{\partial A_{2k}},$$

and hence

$$\frac{\partial L}{\partial A_{2j'}} = \frac{\partial R}{\partial A_{2j'}} < \frac{\partial R}{\partial A_{2k}} = \frac{\partial L}{\partial A_{2k}},$$

contradicting the assumption that A is a stationary point with a gradient of 0. Hence, A_{2k} is observed. Combining (10.50) and (10.54), we have

$$A_{1j} < A_{1k}$$
$$A_{2j} > A_{2k}.$$

That is, $(1, 2, j, k)$ is a conflicting quadruple. Note that all the 4 entries in this conflicting quadruple are observed. Note that by the assumption that $j \in S$ and $k \in \overline{S}$, this conflicting quadruple is across the partition $(S, \overline{S})$. Applying Proposition 3.3 yields a contradiction with the dominance relation (10.47) of the partition $(S, \overline{S})$.
Case 2: A_{2j} is unobserved and $A_{2j'}$ is observed. A similar argument as in Case 1 applies.

Combining the two cases completes Step 1.
Step 2: Parition columns into blocks
We use the same construction of the blocks described in Step 2 of the proof of Lemma 10.4, and obtain the blocks $[d] = B_1 \cup \ldots \cup B_K$.

Step 3: Show that A satisfies the claimed dominance relation (10.49)
We follow Step 3 of the proof of Lemma 10.4, and use the same definition of $k_{\mathrm{H}}, k_{\mathrm{L}}$ from (10.44), and the definition of $(j_{\mathrm{H}}, j_{\mathrm{L}})$ from (10.45). Again assume for contradiction that the dominance relation (10.49) does not hold on A. We separately discuss the following cases depending on whether the entries $(1, j_{\mathrm{H}})$ and $(1, j_{\mathrm{L}})$ are observed.
Case 1: Both $(1, j_{\mathrm{H}})$ and $(1, j_{\mathrm{L}})$ are observed. Then Step 3 of Lemma 10.4 can be applied directly.
Case 2: Both $(1, j_{\mathrm{H}})$ and $(1, j_{\mathrm{L}})$ are unobserved. Due to the definitions (10.44a) and (10.45a), the entry $(1, j_{\mathrm{H}})$ is the maximum entry of A in row 1 among columns S. If the entry $(1, j_{\mathrm{H}})$ is involved in any conflicting quadruple, then due to Step 1, all such conflicting quadruples must lie within S. Hence, all conflicting quadruples contribute a positive gradient to $\frac{\partial R}{\partial A_{1,j_{\mathrm{H}}}}$. Since $(1, j_{\mathrm{H}})$ is unobserved, setting the gradient expression (10.6) to 0 for the stationary point A, we have

$$\frac{\partial L}{\partial A_{1,j_{\mathrm{H}}}} = \frac{\partial R}{\partial A_{1,j_{\mathrm{H}}}} = 0.$$

Hence, the entry $(1, j_{\mathrm{H}})$ cannot be in any conflicting quadruples. Therefore, $(2, j_{\mathrm{H}})$ is the maximum entry in row 2 among columns S. Likewise $(1, j_{\mathrm{L}})$ cannot be in any conflicting quadruples, and $(2, j_{\mathrm{L}})$ is the minimum entry in row 2 among columns $\overline{S}$. By the assumption (10.48), both $(2, j_{\mathrm{L}})$ and $(2, j_{\mathrm{L}})$ are observed. Applying the arguments in Case 1 to the pair of $(2, j_{\mathrm{H}})$ and $(2, j_{\mathrm{L}})$ completes Case 2.
Case 3: $(1, j_{\mathrm{L}})$ is observed and $(1, j_{\mathrm{H}})$ is unobserved.
Denote $(2, j'_{\mathrm{L}})$ as the minimum entry in row 2 among columns $\overline{S}$. If $(2, j'_{\mathrm{L}})$ is unobserved, then as in Case 2, the entry $(2, j'_{\mathrm{L}})$ cannot be in any conflicting quadruples, and hence $j'_{\mathrm{L}} = j_{\mathrm{L}}$.

We have $(1, j_{\mathrm{H}})$ and $(2, j_{\mathrm{L}})$ both unobserved, contradicting (10.48). Hence, $(2, j'_{\mathrm{L}})$ must be observed, and likewise $(2, j'_{\mathrm{H}})$ must be observed, where $(2, j'_{\mathrm{H}})$ is the maximum entry in row 2 among columns S. Applying Case 1 to the pair of $(2, j'_{\mathrm{L}})$ and $(2, j'_{\mathrm{H}})$ completes the proof.
Case 4: $(1, j_{\mathrm{L}})$ is unobserved and $(1, j_{\mathrm{H}})$ is observed. By symmetry, a similar argument as in Case 3 applies.

Finally, combining the 4 cases completes the proof.

Chapter 11

Proof of Chapter 4

In this chapter, we provide proofs for all the theoretical claims in Chapter 4. We begin by introducing some additional notation in Section 11.1 which is used throughout the proofs. In Section 11.2, we then provide certain preliminaries that are useful for the proofs. We then present the proofs in subsequent subsections.

For ease of notation, we ignore rounding throughout the proofs as it does not affect the claimed results.

11.1 Notation

Training-validation split $(\Omega^{\mathrm{t}}, \Omega^{\mathrm{v}})$**:** By Algorithm 2, the number of elements restricted to the set Ω^{t} or Ω^{v} is the same for each course i. Hence, we denote n^{t} and n^{v} as the number of students per course in Ω^{t} and Ω^{v} respectively. Throughout the proofs, for simplicity we assume that n is *even*. In this case we have

$$n^{\mathrm{t}} = n^{\mathrm{v}} = \frac{n}{2}. \tag{11.1}$$

All the proofs extend to the case where n is odd under minor modifications.

We define the elements in each course $i \in [d]$ restricted to Ω^{t} or Ω^{v} as:

$$\begin{aligned} \Omega_i^{\mathrm{t}} &:= \{(i,j) \in \Omega^{\mathrm{t}}\} \\ \Omega_i^{\mathrm{v}} &:= \{(i,j) \in \Omega^{\mathrm{v}}\}. \end{aligned}$$

We slightly abuse the notation and say $j \in \Omega_i^{\mathrm{t}}$ if $(i,j) \in \Omega_i^{\mathrm{t}}$. Likewise for Ω_i^{v}.

Group orderings: Recall that from Definition 4.1 that G_k denotes the set of elements in group $k \in [r]$. We define

$$\begin{aligned} G_k^{\mathrm{t}} &:= G_k \cap \Omega^{\mathrm{t}} \\ G_k^{\mathrm{v}} &:= G_k \cap \Omega^{\mathrm{v}}. \end{aligned}$$

We denote the elements of group $k \in [r]$ in course $i \in [d]$ restricted to Ω^{v} as:

$$G_{ik} := G_k \cap \Omega_i.$$

Furthermore, we define the elements of G_{ik} restricted to Ω^{v} as

$$G^{\text{t}}_{ik} := G^{\text{t}}_k \cap \Omega^{\text{t}}_i \qquad G^{\text{v}}_{ik} := G^{\text{v}}_k \cap \Omega^{\text{v}}_i.$$

Again, we slightly abuse the notation and say $j \in G^{\text{v}}_{ik}$ if $(i, j) \in G^{\text{v}}_{ik}$.

We define ℓ_{ik} as the the number of students of group $k \in [r]$ in course $i \in [d]$. We define ℓ_k as the number of students of group $k \in [r]$. We denote $\ell_{-i,k}$ as the number of students of group $k \in [r]$ and not in course i. Namely,

$$\ell_{ik} := |G_{ik}| \tag{11.2a}$$

$$\ell_k := |G_k| = \sum_{i \in [d]} \ell_{ik} \tag{11.2b}$$

$$\ell_{-i,k} := |G_k \setminus G_{ik}| = \sum_{i' \neq i} \ell_{i'k}. \tag{11.2c}$$

Furthermore, we define

$$\ell^{\text{t}}_k := \left|G^{\text{t}}_k\right| \qquad \ell^{\text{v}}_k := |G^{\text{v}}_k|, \tag{11.3a}$$

$$\ell^{\text{t}}_{ik} := \left|G^{\text{t}}_{ik}\right| \qquad \ell^{\text{v}}_{ik} := |G^{\text{v}}_{ik}|. \tag{11.3b}$$

Total ordering: Consider the dn elements. We say that the element (i, j) is of rank $t \in [dn]$ if (i, j) is the t^{th}-smallest element in among the dn elements.

We denote t_{ij} as the rank of each element $(i, j) \in [d] \times [n]$. We denote (i_t, j_t) as the element of rank $t \in [dn]$.

Observations Y and bias B: Denote the mean of all observations as

$$\overline{y} = \frac{1}{dn} \sum_{i \in [d], j \in [n]} y_{ij}. \tag{11.4}$$

Denote the mean of the observations in any course $i \in [d]$ as

$$\overline{y}_i = \frac{1}{n} \sum_{j=1}^{n} y_{ij}. \tag{11.5}$$

Likewise we denote the mean of the bias in any course $i \in [d]$ as $\overline{b}_i$. We denote the mean of the bias of any course $i \in [d]$ as

$$\overline{b}_{G_k} = \frac{1}{\ell_k} \sum_{(i,j) \in G_k} b_{ij}.$$

Now restrict to group orderings. For any course $i \in [d]$ and any group $k \in [r]$, denote the smallest and the largest observation in course i and group k as

$$y_{ik,\max} := \max_{j:(i,j)\in G_k} y_{ij} \tag{11.6a}$$

$$y_{ik,\min} := \min_{j:(i,j)\in G_k} y_{ij} \tag{11.6b}$$

We define $b_{ik,\max}$ and $b_{ik,\min}$ likewise. In addition, we define the smallest and the bias of any group $k \in [r]$ as

$$\begin{aligned} b_{k,\min} &= \min_{(i,j)\in G_k} b_{ij} \\ b_{k,\max} &= \max_{(i,j)\in G_k} b_{ij}. \end{aligned} \tag{11.7}$$

Statistics: We g as the p.d.f. of $\mathcal{N}(0,1)$. Denote G and G^{-1} as the corresponding c.d.f., and the inverse c.d.f., respectively. We slightly abuse notation and write $\mathbb{P}(X)$ as the p.d.f. of any continuous variable X.

For a set of i.i.d. random variables $X_1, \ldots, X_n$, we denote $X^{(k)}$ as the k^{th} order statistics of $\{X_i\}_{i=1}^n$. We use the notation $X^{(k:n)}$ when we emphasize the sample size n.

Let $d \geq 2$ be any integer, and let π be a total ordering of size d. We denote the monotonic cone with respect to π as $\mathcal{M} := \left\{\theta \in \mathbb{R}^d : \theta_{\pi(1)} \leq \ldots \leq \theta_{\pi(d)}\right\}$. For any vector $x \in \mathbb{R}^d$, we denote the isotonic projection of x as

$$\Pi_{\mathcal{M}}(x) := \operatorname*{argmin}_{u\in\mathcal{M}_\pi} \|x-u\|_2^2. \tag{11.8}$$

We denote $\mathcal{M}$ as the monotonic cone with respect to the identity ordering.

Our estimator and the cross-validation algorithm: Recall from Line 10 of Algorithm 2 that our estimator restricted to any set of elements $\Omega \subseteq [d] \times [n]$ is defined as the solution to:

$$\operatorname*{argmin}_{x\in\mathbb{R}^d} \min_{\substack{B\in\mathbb{R}^{d\times n} \\ B \text{ satisfies } \mathcal{O}}} \left\|Y - x\mathbf{1}^T - B\right\|_\Omega^2 + \lambda\|B\|_\Omega^2, \tag{11.9}$$

with the ties broken by minimizing $\|B\|_F^2$.

We use the shorthand notation $(\widehat{x}, \widehat{B})$ to denote the solution $(\widehat{x}^{(\lambda)}, \widehat{b}^{(\lambda)})$ to (11.9) when the value λ is clear from the context. Likewise we use the shorthand notation $\widetilde{B}^{(\lambda)}$ to denote the interpolated bias $\widetilde{B}^{(\lambda)}$ obtained in Line 15 of Algorithm 2.

Recall from Line 13 in Algorithm 2 that we find the element $(i^\pi, j^\pi) \in \Omega^{\text{t}}$ (or two elements $(i_1^\pi, j_1^\pi), (i_2^\pi, j_2^\pi) \in \Omega^{\text{t}}$) that is close to the considered element $(i,j) \in \Omega^{\text{v}}$ in any total ordering π. We call these one or two elements from Ω^{t} as the "nearest-neighbor" of (i,j) with respect to π, denoted $\text{NN}(i,j;\pi)$. Recall from Line 17 in Algorithm 2 that $e^{(\lambda)}$ denotes the CV error at λ.

Define the random variable Λ_ϵ as the set

$$\Lambda_\epsilon := \{\lambda \in [0,\infty] : \|\widehat{x}^{(\lambda)}\|_2 > \epsilon\}. \tag{11.10}$$

Under $x^* = 0$, the set Λ_ϵ consists of the "bad" choices of λ whose estimate $\widehat{x}^{(\lambda)}$ incurs a large squared ℓ_2-error.

Taking the limit of $n \to \infty$: For ease of notation, we define the limit of taking $n \to \infty$ as follows. For example, in the statement of Theorem 4.5(a), we consider any fixed $\epsilon > 0$. Then the notation

$$\lim_{n\to\infty} \mathbb{P}\Big(\|\widehat{x}^{(0)} - x^*\|_2 < \epsilon\Big) = 1 \tag{11.11}$$

is considered equivalent to the original statement of Theorem 4.5(a) that for any $\delta > 0$, there exists an integer n_0, such that for every $n \geq n_0$ and every partial ordering satisfying the condition (a) we have

$$\mathbb{P}\Big(\|\widehat{x}^{(0)} - x^*\|_2 < \epsilon\Big) = 1.$$

The notation (11.11) has the alternative interpretation as follows. We construct a sequence of partial orderings $\{\mathcal{O}_n\}_{n=1}^{\infty}$, where the partial ordering $\mathcal{O}_n$ is on d courses and n students and satisfies the condition (a). With n students, the estimator $\widehat{x}^{(0)}$ is provided the partial ordering $\mathcal{O}_n$. We consider any such fixed sequence $\{\mathcal{O}_n\}_{n=1}^{\infty}$. Then the limit of $n \to \infty$ in (11.11) is well-defined.

11.2 Preliminaries

In this section we present preliminary results that are used in the subsequent proofs. Some of the preliminary results are defined based on a set of elements $\Omega \subseteq [d] \times [n]$. We define the elements in each course $i \in [d]$ as

$$\Omega_i := \{(i, j) \in \Omega\}.$$

Again we say $j \in \Omega_i$ if $(i, j) \in \Omega_i$. We define the number of elements in each course $i \in [d]$ as $n_i := |\Omega_i|$.

Throughout the proofs, whenever a set $\Omega \subseteq [d] \times [n]$ is considered, *we assume the set Ω satisfies $n_i > 0$ for each $i \in [d]$* to avoid pathological cases. For ease of presentation, the order of the preliminary results does not exactly follow the sequential order that they are proved.

11.2.1 Properties of the estimator

In this section we present a list of properties of our estimator. We start with the following proposition. This proposition shows the existence and uniqueness of the solution to our estimator (11.9) under its tie-breaking rule for any $\lambda \in [0, \infty)$. That is, the estimator is well-defined on $\lambda \in [0, \infty)$.

Proposition 11.1 (Existence of the estimator at $\lambda \in [0, \infty)$)**.** *For any $\lambda \in [0, \infty)$ and any $\Omega \subseteq [d] \times [n]$, there exists a unique solution to our estimator* (4.2) *under the tie-breaking rule, given any inputs $Y \in \mathbb{R}^{d\times n}$ and any partial ordering $\mathcal{O}$.*

The proof of this result is provided in Appendix 11.9.1. Recall that the solution to (11.9) at $\lambda = \infty$ is defined by taking the limit of $\lambda \to \infty$ as:

$$\widehat{x}^{(\infty)} := \lim_{\lambda \to \infty} \widehat{x}^{(\lambda)} \tag{11.12a}$$

$$\widehat{B}^{(\infty)} := \lim_{\lambda \to \infty} \widehat{B}^{(\lambda)}. \tag{11.12b}$$

The following proposition shows the existence of the solution (11.12). That is, the limit in (11.12) is well-defined. This proposition is a generalization of Proposition 4.7 to any set $\Omega \subseteq [d] \times [n]$, and its proof is a straightforward generalization of the proof of Proposition 4.7 (Appendix 11.4).

Proposition 11.2 (Existence of the estimator at $\lambda = \infty$)**.** *For any $\Omega \subseteq [d] \times [n]$, the solution $(\widehat{x}^{(\infty)}, \widehat{B}^{(\infty)})$ defined in* (11.12) *exists. Moreover, we have*

$$[\widehat{x}^{(\infty)}]_i = \frac{1}{n_i} \sum_{j \in \Omega_i} y_{ij} \qquad \forall i \in [d]$$

$$\widehat{B}^{(\infty)} = 0.$$

The following lemma gives a relation between $\widehat{x}^{(\lambda)}$ and $\widehat{B}^{(\lambda)}$ for any $\lambda \in [0, \infty]$. This basic relation is used in proving multiple properties of the estimator to be presented subsequently in this section.

Lemma 11.3. *For any $\lambda \in [0, \infty]$, and any $\Omega \subseteq [d] \times [n]$, the solution $(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$ to the estimator* (11.9) *satisfies*

$$\widehat{x}_i^{(\lambda)} = \frac{1}{n_i} \sum_{j \in \Omega_i} \left(y_{ij} - \widehat{b}_{ij}^{(\lambda)} \right) \qquad \forall i \in [d]. \tag{11.13}$$

In particular, in the special case of $\Omega = [d] \times [n]$, we have

$$\widehat{x}_i^{(\lambda)} = \frac{1}{n} \sum_{j \in [n]} \left(y_{ij} - \widehat{b}_{ij}^{(\lambda)} \right) \qquad \forall i \in [d]. \tag{11.14}$$

The proof of this result is provided in Appendix 11.9.2 The following property gives expressions of the sum of the elements in $\widehat{x}$ and the sum of the elements in $\widehat{B}$.

Lemma 11.4. *For any $\lambda \in [0, \infty]$, any $\Omega \subseteq [d] \times [n]$, the solution $(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$ given any partial ordering $\mathcal{O}$ and any observations Y satisfies*

$$\sum_{(i,j) \in \Omega} \widehat{b}_{ij}^{(\lambda)} = 0 \tag{11.15a}$$

$$\sum_{i \in [d]} n_i \widehat{x}_i^{(\lambda)} = \sum_{(i,j) \in \Omega} y_{ij}. \tag{11.15b}$$

In particular, in the special case of $\Omega = [d] \times [n]$, we have

$$\sum_{i \in [d], j \in [d]} \widehat{b}_{ij}^{(\lambda)} = 0 \tag{11.16a}$$

$$n \sum_{i \in [d]} \widehat{x}_i^{(\lambda)} = \sum_{i \in [d], j \in [n]} y_{ij}. \tag{11.16b}$$

The proof of this result is provided in Appendix 11.9.3. The following property shows a shift-invariant property of our estimator. This property is used so that we assume $x^* = 0$ without loss of generality all the proofs.

Proposition 11.5 (Shift-invariance of the estimator)**.** *Consider any $\Omega \subseteq [d] \times [n]$, and any partial ordering $\mathcal{O}$. Fix any $\lambda \in [0, \infty]$. Let $(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$ be the solution of our estimator for any observations $Y \in \mathbb{R}^{d\times n}$ given $(\mathcal{O}, \lambda, \Omega)$. Consider any $\Delta x \in \mathbb{R}^d$. Then the solution of our estimator for the observations $Y + \Delta x \mathbf{1}^T$ given $(\mathcal{O}, \lambda, \Omega)$ is $(\widehat{x}^{(\lambda)} + \Delta x, \widehat{B}^{(\lambda)})$.*

The proof of this result is provided in Appendix 11.9.4. Note that the observation model (4.1) is shift-invariant by definition. That is, consider any fixed $B, Z \in \mathbb{R}^{d\times n}$, denote the observations with $x^* = 0$ as Y. Then the observations with $x^* = \Delta x$ is $(Y + \Delta x \mathbf{1}^T)$. Hence, Proposition 11.5 implies the following corollary.

Corollary 11.6. *Under the observation model* (4.1)*, consider any fixed bias $B \in \mathbb{R}^{d\times n}$ and noise $Z \in \mathbb{R}^{d\times n}$. Suppose the solution of our estimator under $x^* = 0$ is $(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$ given any $(\mathcal{O}, \lambda, \Omega)$. Then the solution under $x^* = \Delta x$ is $(\widehat{x}^{(\lambda)} + \Delta x, \widehat{B}^{(\lambda)})$.*

Based on the result of Corollary 11.6, it can be further verified that the cross-validation algorithm (Algorithm 2) that uses our estimator is shift-invariant. Therefore, *for all the proofs, we assume $x^* = 0$ without loss of generality*.

The following pair of lemmas (Lemma 11.7 and Lemma 11.8) converts between a bound on the difference of a pair of courses $|\widehat{x}_i - \widehat{x}_{i'}|$ and a bound on $\|\widehat{x}\|_2$. Lemma 11.7 is used in Theorem 4.9 and Theorem 4.10; Lemma 11.8 is used in Theorem 4.5. Recall the notation $\Lambda_\epsilon := \{\lambda \in [0, \infty] : \|\widehat{x}^{(\lambda)}\|_2 > \epsilon\}$.

Lemma 11.7. *Suppose $x^* = 0$. Consider random Ω^{t} obtained by Algorithm 2. Suppose the observations are generate from either:*

(a) The bias is marginally distributed as $\mathcal{N}(0, \sigma^2)$ following assumption (A2) and there is no noise, or

(b) The noise is generated from $\mathcal{N}(0, \eta^2)$ following assumption (A1), and there is no bias.

For any constant $\epsilon > 0$, our estimator $\widehat{x}^{(\lambda)}$ restricted to Ω^{t} satisfies

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i,i'\in[d]} \left(\widehat{x}^{(\lambda)}_i - \widehat{x}^{(\lambda)}_{i'}\right) > \frac{\epsilon}{\sqrt{d}}, \quad \forall \lambda \in \Lambda_\epsilon\right) = 1,$$

where the probability is taken over the randomness in the observations Y and the training set Ω^{t}.

The proof of this result is provided in Appendix 11.9.5.

Lemma 11.8. *Suppose $x^* = 0$. Suppose the observations follow part (a) of Lemma 11.7. Suppose the estimator is restricted to the set of either*

(a) $\Omega = [d] \times [n]$, or

(b) random Ω^{t} obtained by Algorithm 2.

Fix any $\lambda \in [0, \infty]$ and any $\epsilon > 0$. Suppose we have

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i,i'\in[n]} \left|\widehat{x}^{(\lambda)}_i - \widehat{x}^{(\lambda)}_{i'}\right| < \epsilon\right) = 1. \tag{11.17}$$

Then we have

$$\lim_{n\to\infty} \mathbb{P}\left(\|\widehat{x}^{(\lambda)}\|_2 < \epsilon\right) = 1,$$

where the probabilities are taken over the randomness in the observations Y and (for part (b)) in Ω^{t}.

The proof of this result is provided in Appendix 11.9.6. The following proposition gives a closed-form solution under $d = 2$ courses and $r = 2$ groups at $\lambda = 0$. This proposition is used for proving Theorem 4.5(b) and Proposition 4.13. Recall the definitions of $\overline{y}, \overline{y}_i$, $y_{ik,\min}$ and $y_{ik,\max}$ from (11.4), (11.5) and (11.6).

Proposition 11.9. *Consider $d = 2$ courses and any group ordering $\mathcal{O}$ with $r = 2$ groups. Let $\Omega = [d] \times [n]$. Suppose the bias B satisfies the partial ordering $\mathcal{O}$, and there is no noise. Then the solution of our estimator* (4.2) *at $\lambda = 0$ has the closed-form expression $\widehat{x}^{(0)} = \overline{y} + \begin{bmatrix}-1\\1\end{bmatrix} \cdot \frac{\gamma}{2}$, where*

$$\gamma = \begin{cases} y_{22,\min} - y_{11,\max} & \text{if } y_{22,\min} - y_{11,\max} < \overline{y}_2 - \overline{y}_1 \\ y_{21,\max} - y_{12,\min} & \text{if } y_{21,\max} - y_{12,\min} > \overline{y}_2 - \overline{y}_1 \\ \overline{y}_2 - \overline{y}_1 & \text{o.w.} \end{cases} \tag{11.18}$$

If some of $\{y_{11,\max}, y_{21,\max}, y_{12,\min}, y_{22,\min}\}$ do not exist (i.e., when a certain course doesn't have students of a certain group), then the corresponding case in (11.18) *is ignored.*

The proof of this result is provided in Appendix 11.9.7

11.2.2 Order statistics

This section presents a few standard properties of order statistics.

Consider n i.i.d. random variables $\{X_i\}_{i\in[n]}$ ordered as

$$X^{(1)} \leq \ldots \leq X^{(n)}.$$

Define the maximal spacing as

$$M_n := \max_{1\leq i\leq n-1} (X^{(i+1)} - X^{(i)}). \tag{11.19}$$

The following standard result from statistics states that the maximum difference between adjacent order statistics converges to 0 for the Gaussian distribution.

Lemma 11.10. *Let $n > 1$ be any integer. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(0, 1)$. Then for any $\epsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbb{P}(M_n < \epsilon) = 1.$$

For completeness, the proof of this result is provided in Appendix 11.9.8. Denote G^{-1} as the inverse c.d.f. of $\mathcal{N}(0, 1)$. The following standard result from statistics states that the order statistics converges to the inverse c.d.f.

Lemma 11.11. *Let $X_1, \ldots, X_n$ be $\mathcal{N}(0, 1)$. Fix constant $p \in (0, 1)$ and $c \in \mathbb{R}$. Let $\{k_n\}_{n=1}^{\infty}$ be a sequence such that $\frac{k_n}{n} = p + \frac{c}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$. We have*

$$X^{(k_n:n)} \xrightarrow{P} G^{-1}(p).$$

For completeness, the proof of this result is provided in Appendix 11.9.9.

The following standard result from statistics provides a simple bound on the maximum (and the minimum) of a set of i.i.d. Gaussian random variables.

Lemma 11.12. *Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(0, \sigma^2)$. Then we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i\in[n]} X_i < 2\sigma\sqrt{\log n}\right) = 1$$

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i\in[n]} X_i - \min_{i\in[n]} X_i < 4\sigma\sqrt{\log n}\right) = 1.$$

11.2.3 Additional preliminaries

In this section, we present several more additional preliminary results that are used in the subsequent proofs.

The following result considers the number of students under the all constant-fraction assumption given any training-validation split $(\Omega^{\mathrm{t}}, \Omega^{\mathrm{v}})$. Recall the definitions of $\ell_{ik}, \ell_k, \ell^{\mathrm{v}}_{ik}, \ell^{\mathrm{t}}_k$ and ℓ^{v}_k from (11.2) and (11.3).

Lemma 11.13. *Assume $\ell_{ik} \geq 4$ for each $i \in [d]$ and $k \in [r]$. Consider any training-validation split $(\Omega^{\mathrm{t}}, \Omega^{\mathrm{v}})$ obtained by Algorithm 2. Then we have the deterministic relations*

$$\frac{\ell_{ik}}{4} \leq \ell^{\mathrm{v}}_{ik} \leq \frac{3\ell_{ik}}{4} \qquad \forall i \in [d], k \in [r] \tag{11.20a}$$

$$\frac{\ell_{ik}}{4} \leq \ell^{\mathrm{t}}_{ik} \leq \frac{3\ell_{ik}}{4} \qquad \forall i \in [d], k \in [r] \tag{11.20b}$$

and

$$\frac{\ell_k}{4} \leq \ell^{\mathrm{v}}_k \leq \frac{3\ell_k}{4} \qquad \forall k \in [r] \tag{11.21a}$$

$$\frac{\ell_k}{4} \leq \ell^{\mathrm{t}}_k \leq \frac{3\ell_k}{4} \qquad \forall k \in [r]. \tag{11.21b}$$

The proof of this result is provided in Appendix 11.9.10. The following result considers any total ordering. It states that the ranks of the adjacent elements within Ω^{t}, or the ranks of the adjacent elements between Ω^{t} and Ω^{v} differ by at most a constant. Formally, for any $1 \leq k_1 < k_2 \leq dn$, the element of rank k_1 and the element of rank k_2 are said to be adjacent within Ω^{t}, if both elements are in Ω^{t}, and elements of ranks $k_1 + 1$ through $k_2 - 1$ are all in Ω^{v}. The two elements are said be be adjacent between Ω^{t} and Ω^{v}, if one of the following is true:

- The elements of ranks k_1 through $(k_2 - 1)$ are in Ω^{t}, and the element of rank k_2 is in Ω^{v};
- The elements of ranks k_1 through $(k_2 - 1)$ are in Ω^{v}, and the element of rank k_2 is in Ω^{t}.

Lemma 11.14. *For any partition $(\Omega^{\mathrm{t}}, \Omega^{\mathrm{v}})$ obtained by Algorithm 2, for any $1 \leq k_1 < k_2 \leq dn$, suppose that the element of rank k_1 and the element of rank k_2 are*

(a) adjacent within Ω^{t}, or

(b) adjacent between Ω^{t} and Ω^{v}.

Then we have

$$k_2 - k_1 \leq 2d + 1.$$

The proof of this result is provided in Appendix 11.9.11. The following lemma bounds the mean of the bias terms using standard concentration inequalities.

Lemma 11.15. *Consider any partial ordering $\mathcal{O}$ and any random Ω^{t} obtained by Algorithm 2. Suppose that the bias is marginally distributed as $\mathcal{N}(0,1)$ following assumption (A2). For any $\epsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n^{\mathrm{t}}}\sum_{j\in\Omega_i^{\mathrm{t}}} b_{ij} - \frac{1}{n}\sum_{j\in[n]} b_{ij}\right| < \epsilon\right) = 1 \qquad \forall i \in [d], \tag{11.22a}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{|\Omega^{\mathrm{t}}|}\sum_{(i,j)\in\Omega^{\mathrm{t}}} b_{ij}\right| < \epsilon\right) = 1, \tag{11.22b}$$

where the probabilities are over the randomness in B and in Ω^{t}.

The proof of this result is provided in Appendix 11.9.12.

11.3 Proof of Theorem 4.5

The proof follows notation in Appendix 11.1 and preliminaries in Appendix 11.2. By Corollary 11.6, we assume $x^* = 0$ throughout the proof without loss of generality. We also assume without loss of generality that the standard deviation of the Gaussian bias is $\sigma = 1$. Given $x^* = 0$ and the assumption that there is no noise, model (4.1) reduces to

$$Y = B. \tag{11.23}$$

Recall that ℓ_{ik} denotes the number of observations in course $i \in [d]$ of group $k \in [r]$, and ℓ_k denotes the number of observations of group k summed over all courses. For any positive constant $c > 0$, we define the set S_c as

$$S_c := \left\{(i,i') \in [d]^2 : \exists k \in [r] \text{ such that } \frac{\ell_{ik}}{\ell_k}, \frac{\ell_{i',k+1}}{\ell_{k+1}} \geq c\right\}. \tag{11.24}$$

In words, the definition (11.24) says that for any pair of courses $(i,i') \in S_c$, we have that course i takes at least c-fraction of observations in some group $k \in [r]$, and course i' takes at least c-fraction of observations in group $(k+1)$.

Before proving the three parts separately, we first state a few lemmas that are used for more than one part. The first lemma states that any $(i,i') \in S_c$ imposes a constraint on our estimator $\widehat{x}^{(0)}$ at $\lambda = 0$.

Lemma 11.16. *Assume $x^* = 0$. Consider bias marginally distributed as $\mathcal{N}(0,1)$ following assumption (A2) and no noise. Let $\widehat{x}^{(0)}$ be the solution of our estimator at $\lambda = 0$. Fix any $c > 0$. For any $(i,i') \in S_c$, we have that for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\Big(\widehat{x}_{i'}^{(0)} - \widehat{x}_i^{(0)} < \epsilon\Big) = 1. \tag{11.25}$$

The proof of this result is provided in Appendix 11.10.1. To state the next lemma, we first make the following definition of a "cycle" of courses.

Definition 11.17. *Let $L \geq 2$ be an integer. We say that $(i_1, i_2, \ldots, i_L) \in [d]^L$ is a "cycle" of courses with respect to S_c, if*

$$(i_m, i_{m+1}) \in S_c \qquad \forall m \in [L-1], \tag{11.26a}$$

$$\textit{and } \ (i_L, i_1) \in S_c. \tag{11.26b}$$

The following lemma states that if there exists a cycle of courses, then the difference of the estimated quality $\widehat{x}$ between any two courses in this cycle converges to 0 in probability.

Lemma 11.18. *Fix any $c > 0$. Suppose d is a fixed constant. Let $(i_1, i_2, \ldots, i_L) \in [d]^L$ for some $L \geq 2$ be a cycle with respect to S_c. Then for any $\epsilon > 0$ we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{m,m' \in [L]} \left|\widehat{x}_{i_{m'}} - \widehat{x}_{i_m}\right| < \epsilon\right) = 1.$$

The proof of this result is provided in Appendix 11.10.2. Now we prove the three parts of Theorem 4.5 respectively.

11.3.1 Proof of part (a)

For clarity of notation, we denote the constant in the all constant-fraction assumption as c_{f}. Consider any $i, i' \in [d]$ and any $k \in [r-1]$. We have

$$\frac{\ell_{ik}}{\ell_k} \overset{\text{(i)}}{\geq} \frac{c_{\mathrm{f}} n}{dn} = \frac{c_{\mathrm{f}}}{d},$$

where step (i) is true by the all c-fraction assumption from Definition 4.3. Hence, by the definition (11.24) of S_c, we have $(i, i') \in S_{\frac{c_{\mathrm{f}}}{d}}$ for every $i, i' \in [d]$. Hence, $(1, 2, \ldots, d)$ is a cycle with respect to $S_{\frac{c_{\mathrm{f}}}{d}}$ according to Definition 11.17. Applying Lemma 11.18 followed by Lemma 11.8(a) completes the proof.

11.3.2 Proof of part (b)

Without loss of generality we assume course 1 has more (or equal) students in group 1 than course 2, that is, we assume

$$\ell_{11} \geq \ell_{21}. \tag{11.27}$$

Since we assume there are only two courses and two groups, we have

$$\ell_{12} = n - \ell_{11} \leq n - \ell_{21} = \ell_{22}. \tag{11.28}$$

We fix any constant $\epsilon > 0$. We now bound the probability that $|\widehat{x}_2 - \widehat{x}_1| < \epsilon$. Specifically, we separately bound the probability of $\widehat{x}_2 - \widehat{x}_1 < \epsilon$, and the probability of $\widehat{x}_2 - \widehat{x}_1 > -\epsilon$. Finally, we invoke Lemma 11.8 to complete the proof.

Bounding the probability of $\widehat{x}_2 - \widehat{x}_1 < \epsilon$: By the definition (11.24) of S_c, it can be verified that given (11.27) and (11.28) we have $(1,2) \in S_{0.5}$ (taking $k = 1$). By Lemma 11.16, we have

$$\lim_{n\to\infty} \mathbb{P}(\widehat{x}_2 - \widehat{x}_1 < \epsilon) = 1. \tag{11.29}$$

Bounding the probability of $\widehat{x}_2 - \widehat{x}_1 > -\epsilon$: By the closed-form solution in Proposition 11.9, we have $\widehat{x}_2 - \widehat{x}_1 = \gamma$ where γ is defined in (11.18) as

$$\gamma = \begin{cases} y_{22,\min} - y_{11,\max} & \text{if } y_{22,\min} - y_{11,\max} < \overline{y}_2 - \overline{y}_1 \\ y_{21,\max} - y_{12,\min} & \text{if } y_{21,\max} - y_{12,\min} > \overline{y}_2 - \overline{y}_1 \\ \overline{y}_2 - \overline{y}_1 & \text{o.w.} \end{cases} \tag{11.30}$$

Recall from the model (11.23) that $Y = B$, and hence we have the deterministic relation $y_{22,\min} - y_{11,\max} = b_{22,\min} - b_{11,\max} \geq 0$ due to the assumption (A2) under the group ordering, and similarly we have the deterministic relation $y_{21,\max} - y_{12,\min} \leq 0$. Consider the case of $\overline{y}_2 - \overline{y}_1 \geq 0$. In this case, only the first and the third cases in (11.30) are possible, and therefore we have $0 \leq \gamma \leq \overline{y}_2 - \overline{y}_1$. Now consider the case of $\overline{y}_2 - \overline{y}_1 < 0$. In this case, only the second and the third cases in (11.30) are possible, and we have $\overline{y}_2 - \overline{y}_1 \leq \gamma \leq 0$. Combining the two cases, we have the relation

$$\widehat{x}_2 - \widehat{x}_1 = \gamma > -\epsilon \qquad \text{if } \overline{y}_2 - \overline{y}_1 > -\epsilon. \tag{11.31}$$

It suffices to bound the probability of $\overline{y}_2 - \overline{y}_1 > -\epsilon$.

In what follows we show that $\lim_{n\to\infty} \mathbb{P}(\overline{y}_2 - \overline{y}_1 > -\epsilon) = 1$. That is, we fix some small $\delta > 0$ and show that $\mathbb{P}(\overline{y}_2 - \overline{y}_1 > -\epsilon) \geq 1 - \delta$ for all sufficiently large d. The intuition is that course 2 has more students in group 2, which is the group of greater values of the bias. Since according to assumption (A2) the bias is assigned within each group uniformly at random, the set of observations in course 2 statistically dominates the set of observations in course 1. Therefore, $\overline{y}_2$ should not be less than $\overline{y}_1$ by a large amount.

We first condition on any fixed values of bias ranked as $b^{*(1)} \leq \ldots \leq b^{*(2n)}$ (since we assume the number of courses is $d = 2$). Denote the mean of bias of group 1 as $\overline{b}^*_{G_1} = \frac{1}{\ell_1}\sum_{k=1}^{\ell_1} b^{*(k)}$ and the mean of bias of group 2 as $\overline{b}^*_{G_2} = \frac{1}{\ell_2}\sum_{k=\ell_1+1}^{2n} b^{*(k)}$. Denote $\Delta_{B^*} := b^{*(2n)} - b^{*(1)}$ and denote $\Delta_B := b^{*(2n)} - b^{*(1)}$. By Hoeffding's inequality without replacement [86, Section 6] on group 1 of course 1, we have

$$\mathbb{P}\left[\left|\sum_{j\in G_{11}} b_{1j} - \ell_{11}\overline{b}^*_{G_1}\right| \geq \Delta_{B^*}\sqrt{\ell_{11}\log\left(\frac{1}{\delta}\right)} \;\middle|\; B^*\right] \leq 2\exp\left(-\frac{2\cdot\Delta^2_{B^*}\ell\log(\frac{1}{\delta})}{\ell\Delta^2_B}\right) = 2\delta^2 \overset{\text{(i)}}{\leq} \frac{\delta}{8},$$

where (i) holds for any $\delta \in (0, \frac{1}{16})$. We apply Hoeffding's inequaltiy without replacement for any $i \in \{1,2\}$ and any $k \in \{1,2\}$. Using the fact that $\ell_{ik} \leq n$ for any $i \in \{1,2\}$ and any $k \in \{1,2\}$, we have

$$\mathbb{P}\left[\left|\sum_{j\in G_{ik}} b_{ij} - \ell_{ik}\overline{b}^*_{G_k}\right| \geq \Delta_{B^*}\sqrt{n\log\left(\frac{1}{\delta}\right)} \;\middle|\; B^*\right] \leq \frac{\delta}{8}. \tag{11.32}$$

Taking a union bound of (11.32) over $i \in \{1,2\}$ and $k \in \{1,2\}$, we have that with probability at least $1 - \frac{\delta}{2}$,

$$\begin{aligned}
\overline{y}_2 - \overline{y}_1 &= \frac{1}{n}\left(\sum_{j \in G_{21}} b_{2j} + \sum_{j \in G_{22}} b_{2j} - \sum_{j \in G_{11}} b_{1j} - \sum_{j \in G_{12}} b_{1j}\right)\\
&\overset{\text{(i)}}{\geq} \frac{1}{n}\left(\ell_{21}\overline{b}^*_{G_1} + \ell_{22}\overline{b}^*_{G_2} - \ell_{11}\overline{b}^*_{G_1} - \ell_{12}\overline{b}^*_{G_2} - 4\Delta_{B^*}\sqrt{n \log\left(\frac{1}{\delta}\right)}\right)\\
&= \frac{1}{n}\left((\ell_{21} - \ell_{11})\overline{b}^*_{G_1} + (\ell_{22} - \ell_{12})\overline{b}^*_{G_2} - 4\Delta_{B^*}\sqrt{n \log\left(\frac{1}{\delta}\right)}\right)\\
&\overset{\text{(ii)}}{=} \frac{1}{n}\left((\ell_{21} - \ell_{11})(\overline{b}^*_{G_1} - \overline{b}^*_{G_2}) - 4\Delta_{B^*}\sqrt{n \log\left(\frac{1}{\delta}\right)}\right)\\
&\overset{\text{(iii)}}{\geq} -4\Delta_{B^*}\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}},
\end{aligned} \tag{11.33}$$

where inequality (i) is true by (11.32), step (ii) is true because $\ell_{11} + \ell_{12} = \ell_{21} + \ell_{22}$ and hence $\ell_{21} - \ell_{11} = -(\ell_{22} - \ell_{12})$, and finally step (iii) is true by $\overline{b}^*_{G_1} \leq \overline{b}^*_{G_2}$ due to the assumption (A2) of the bias and the group orderings.

Now we analyze the term Δ_B in (11.33). By Lemma 11.12, there exists integer n_0 such that for any $n \geq n_0$,

$$\mathbb{P}\left(\Delta_B \leq 4\sqrt{\log 2n}\right) \geq 1 - \frac{\delta}{2}. \tag{11.34}$$

Let n_1 be a sufficiently large such that $n_1 \geq n_0$ and $16\sqrt{\log 2n_1} \cdot \sqrt{\frac{\log(\frac{1}{\delta})}{n_1}} < \epsilon$. Then combining (11.34) with (11.33), we have that for any $n \geq n_0$,

$$\begin{aligned}
\mathbb{P}\left(\overline{y}_2 - \overline{y}_1 > -\epsilon\right) &= \int_{B \in \mathbb{R}^{2\times n}} \mathbb{P}\left(\overline{y}_2 - \overline{y}_1 > -\epsilon \mid B\right) \cdot \mathbb{P}(B)\,\mathrm{d}B\\
&\geq \int_{\substack{B \in \mathbb{R}^{2\times n}:\\ \Delta_B \leq 4\sqrt{\log n}}} \mathbb{P}(\overline{y}_2 - \overline{y}_1 > -\epsilon \mid B) \cdot \mathbb{P}(B)\,\mathrm{d}B\\
&\overset{\text{(i)}}{\geq} \left(1 - \frac{\delta}{2}\right) \cdot \mathbb{P}(\Delta_B \leq 4\sqrt{\log 2n})\\
&\overset{\text{(ii)}}{\geq} \left(1 - \frac{\delta}{2}\right)^2 \geq 1 - \delta,
\end{aligned} \tag{11.35}$$

where inequality (i) is true by (11.33) due to the choice of n_1, and inequality (ii) is true by (11.34). Combining (11.35) with (11.31), for any $n \geq n_1$, we have

$$\mathbb{P}(\widehat{x}_2 - \widehat{x}_1 = \gamma > -\epsilon) \geq \mathbb{P}(\overline{y}_2 - \overline{y}_1 > -\epsilon) \geq 1 - \delta.$$

That is,

$$\lim_{n\to\infty} \mathbb{P}(\widehat{x}_2 - \widehat{x}_1 > -\epsilon) = 1. \tag{11.36}$$

Finally, combining Step 1 and Step 2, we take a union bound of (11.29) and (11.36), we have

$$\lim_{n\to\infty} \mathbb{P}\Big(|\widehat{x}_2 - \widehat{x}_1| < \epsilon\Big) = 1. \tag{11.37}$$

Given (11.37), we invoke Lemma 11.8 and obtain

$$\lim_{n\to\infty} \mathbb{P}\Big(\|\widehat{x}\|_2 < \epsilon\Big) = 1,$$

completing the proof.

11.3.3 Proof of part (c)

For total orderings, each observation forms its own group of size 1 (that is, $\ell_k = 1$ for all $k \in [dn]$). A bias term belonging to group some $k \in [dn]$ is equivalent to the bias term being rank k. By the definition 11.24 of S_c, if course i contains rank k and course i' contains rank $k+1$ then we have $(i, i') \in S_1$, because $\frac{\ell_{ik}}{\ell_k} = \frac{\ell_{i',k+1}}{\ell_{k+1}} = 1$ due to the total ordering.

The proof consists of four steps:

- In Step 1, we find a partition of the courses, where each subset in this partition consists of courses i whose estimated qualities $\widehat{x}_i$ are close to each other.
- In Step 2, we use this partition to analyze $|\widehat{x}_i - \widehat{x}_{i'}|$.
- In Step 3, we upper-bound the probability that $|\widehat{x}_i - \widehat{x}_{i'}|$ is large. If $|\widehat{x}_i - \widehat{x}_{i'}|$ is large, then we construct an alternative solution according to the partition and derive a contradiction that $\widehat{x}$ cannot be the optimal compared to the alternative solution.
- In Step 4, we invoke Lemma 11.8 to convert the bound on $|\widehat{x}_i - \widehat{x}_{i'}|$ to a bound on $\|\widehat{x}\|_2$.

Step 1: Constructing the partition We describe the procedure to construct the partition of courses based on any given total ordering $\mathcal{O}$. Without loss of generality, we assume that the minimal rank in course i is strictly less than the minimal rank in course $(i+1)$ for every $i \in [d-1]$. That is, we have

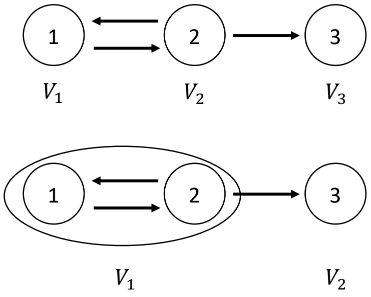
$$\min_{j\in[n]} t_{ij} < \min_{j\in[n]} t_{i+1,j} \qquad \forall i \in [d-1]. \tag{11.38}$$

The partition is constructed in steps. We first describe the initialization of the partition. After the partition is initialized, we specify a procedure to "merge" subsets in the partition. We continue merging the subsets until there are no more subsets to merge according to a specified condition, and arrive at the final partition.

Initialization We construct a directed graph of d nodes, where each node $i \in [d]$ represents course i. We put a directed edge from node i to node i' for every $(i, i') \in S_1$. Let $V_1, \ldots, V_d \subseteq [d]$ be a partition of the d nodes. We initialize the partition as $V_i = \{i\}$ for all $i \in [d]$. We also call each subset V_i as a "hypernode".

	students			
course 1	1	2	6	7
course 2	3	4	5	8
course 3	9	10	11	12

(a) The total ordering



(b) The procedure of constructing the partition

Figure 11.1: An example for constructing the partition of hypernodes.

Merging nodes We now merge the partition according to the following procedure. We find a cycle (of directed edges) in the constructed graph, such that the nodes (courses) in this cycle belong to at least two different hypernodes. If there are multiple such cycles, we arbitrarily choose one. We "merge" all the hypernodes involved in this cycle. Formally, we denote the hypernodes involved in this cycle as $V_{i_1}, V_{i_2}, \ldots, V_{i_L}$. To merge these hypernodes we construct a new hypernode $V = V_{i_1} \cup V_{i_2} \cup \ldots \cup V_{i_L}$. Then we remove the hypernodes $V_{i_1}, V_{i_2}, \ldots, V_{i_L}$ from the partition, and add the merged hypernode V to the partition.

We continue merging hypernodes, until there exist no such cycles that involve at least two different hypernodes. When we say we construct a partition we refer to this final partition after all possible merges are completed.

An example is provided in Fig. 11.1. In this example we consider $d = 3$ courses and $n = 4$ students per course. We consider the total ordering in Fig. 11.1(a), where each integer in the table represents the rank of the corresponding element with respect to this total ordering. The top graph of Fig. 11.1(b) shows the constructed graph and the initialized partition. At initialization there is a cycle between course 1 and course 2 (that belong to different hypernodes V_1 and V_2), so we merge the hypernodes V_1 and V_2 as shown in the bottom graph of Fig. 11.1(b). At this point, there are no more cycles that involve more than one hypernode, so the bottom graph is the final constructed partition.

In what follows we state two properties of the partition. We define the length of a cycle as the number of edges in this cycle. The first lemma states that within the same hypernode, any two courses included in a cycle whose length is upper-bounded.

Lemma 11.19. *Consider the partition constructed from any total ordering $\mathcal{O}$. Let V be any hypernode in this partition. Then for any $i, i' \in V$ with $i \neq i'$, there exists a cycle whose length is at most $2(d-1)$, such that the cycle includes both course i and course i'.*

The proof of this result is provided in Appendix 11.10.3. The following lemma provides further properties on the constructed partition. We say that there exists an edge from hypernode V to V', if and only if there exists an edge from some node $i \in V$ to some node $i' \in V'$. Denote s as the number of hypernodes in the partition. Denote the hypernodes as $V_1, \ldots, V_s$.

Lemma 11.20. *Consider the partition constructed from any total ordering $\mathcal{O}$. The hypernodes in this partition can be indexed in a way such that the only edges on the hypernodes are (V_m, V_{m+1}) for all $m \in [s-1]$. Under this indexing of hypernodes, the nodes within each hypernodes* *are consecutive, and increasing in the indexing of the hypernodes. That is, there exist integers* $0 = i_1 < i_2 < \ldots < i_{s+1} = d$*, such that* $V_m = \{i_m + 1, \ldots, i_{m+1}\}$ *for each* $m \in [s]$.

Moreover, for each $m \in [s]$*, the ranks of elements (with respect to the total ordering* $\mathcal{O}$*) contained in the nodes of hypernode* V_m *are consecutive and increasing in the indexing of the hypernodes. That is, there exists integers* $0 = t_1 < t_2 \ldots < t_{s+1} = dn$*, such that* $\cup_{i \in V_m} \cup_{j \in [n]} \{t_{ij}\} = \{t_m + 1, \ldots, t_{m+1}\}$.

The proof of this result is provided in Appendix 11.10.4. When we refer to a partition $(V_1, \ldots, V_s)$, we specifically refer to the indexing of the hypernodes that satisfies Lemma 11.20.

As an example, in Fig. 11.1 we have $V_1 = \{1, 2\}$ and $V_2 = \{3\}$. The ranks of elements in V_1 are $\{1, \ldots, 8\}$, and the ranks of elements in V_2 are $\{9, \ldots, 12\}$.

Step 2: Analyzing $|\widehat{x}_i - \widehat{x}_{i'}|$ using the partition Our goal in Step 2 and Step 3 is to prove the that for any $\epsilon > 0$, we have

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{i,i' \in [n]} |\widehat{x}_{i'} - \widehat{x}_i| < \epsilon \right) = 1.$$

Equivalently, denote the "bad" event as

$$E_{\text{bad}} := \left\{ \max_{i,i' \in [n]} |\widehat{x}_{i'} - \widehat{x}_i| > 4d^2\epsilon \right\}. \tag{11.39}$$

The goal is to prove $\lim_{n \to \infty} \mathbb{P}(E_{\text{bad}}) = 0$. In Step 2, we define some high-probability event (namely, $E_1 \cap E_2 \cap E_3$ to be presented), and show that it suffices to prove

$$\lim_{n \to \infty} \mathbb{P}(E_{\text{bad}}, E_1 \cap E_2 \cap E_3) = 0.$$

The event E_1 bounds $|\widehat{x}_{i'} - \widehat{x}_i|$ within each hypernode We first bound $|\widehat{x}_{i'} - \widehat{x}_i|$ for $i, i' \in [d]$ within each hypernode. By Lemam 11.19, there exists a cycle of length at most $2(n-1)$ between any two courses i, i' within the same hypernode. Given assumption (A3) that n is a constant, by Lemma 11.18 we have that for each hypernode V,

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{i,i' \in V} |\widehat{x}_i - \widehat{x}_{i'}| < \epsilon \right) = 1. \tag{11.40}$$

Since the number of hypernodes is at most d, taking a union bound of (11.40) across all hypernodes in the partition, we have

$$\lim_{n \to \infty} \mathbb{P}\Bigg(\underbrace{\max_{i,i' \in V} |\widehat{x}_i - \widehat{x}_{i'}| < \epsilon, \quad \forall V \text{ hypernode in the partition}}_{E_1} \Bigg) = 1. \tag{11.41}$$

We denote this event in (11.41) as E_1.

The event E_2 bounds $|\widehat{x}_{i'} - \widehat{x}_i|$ across hypernodes We then bound $|\widehat{x}_{i'} - \widehat{x}_i|$ across different hypernodes. We consider adjacent hypernodes V_m and V_{m+1} for any $m \in [s-1]$. By Lemma 11.20, there exists an edge from V_m to V_{m+1}. That is, there exists $i \in V_m$ and $i' \in V_{m+1}$ such that $(i, i') \in S_1$. By Lemma 11.16, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\widehat{x}_{i'} - \widehat{x}_i < \epsilon\right) = 1. \tag{11.42}$$

Since the number of hypernodes s is at most d, taking a union bound of (11.42) over all $m \in [s-1]$, we have

$$\lim_{n\to\infty} \mathbb{P}\Bigg(\underbrace{\min_{i\in V_m, i'\in V_{m+1}} \widehat{x}_{i'} - \widehat{x}_i < \epsilon, \quad \forall m \in [s-1]}_{E_2}\Bigg) = 1. \tag{11.43}$$

We denote this event in (11.43) as E_2.

Define E_3: Finally, we define E_3 as the event that B is not a constant matrix. That is,

$$E_3 = \{\exists i, i' \in [d], j, j' \in [n] : b_{ij} \neq b_{i'j'}\}.$$

Since by assumption (A2) (setting $\sigma = 1$) the bias terms $\{b_{ij}\}_{i\in[d],j\in[n]}$ are marginally distributed as $\mathcal{N}(0,1)$, it is straightforward to see that the event E_3 happens almost surely:

$$\mathbb{P}(E_3) = 1. \tag{11.44}$$

Decompose E_{bad}: We decompose the bad event E_{bad} as

$$\begin{aligned}\mathbb{P}(E_{\text{bad}}) &= \mathbb{P}(E_{\text{bad}}, E_1 \cap E_2 \cap E_3) + \mathbb{P}(E_{\text{bad}}, \overline{E_1 \cap E_2 \cap E_3}) \\ &\le \mathbb{P}(E_{\text{bad}}, E_1 \cap E_2 \cap E_3) + \mathbb{P}(\overline{E_1 \cap E_2 \cap E_3}).\end{aligned} \tag{11.45}$$

Combining (11.41), (11.43) and (11.44), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{E_1 \cap E_2 \cap E_3}\right) = \lim_{n\to\infty} \mathbb{P}(\overline{E_1} \cup \overline{E_2} \cup \overline{E_3}) \le \lim_{n\to\infty} \left[\mathbb{P}(\overline{E_1}) + \mathbb{P}(\overline{E_2}) + \mathbb{P}(\overline{E_3})\right] = 0. \tag{11.46}$$

Combining (11.45) and (11.46), in order to show $\lim_{n\to\infty} \mathbb{P}(E_{\text{bad}}) = 0$ it suffices to show $\lim_{n\to\infty} \mathbb{P}(E_{\text{bad}}, E_1 \cap E_2 \cap E_3) = 0$.

Step 3: Analyzing the event $E_{\text{bad}} \cap E_1 \cap E_2 \cap E_3$ In this step, we analyze the event $E_{\text{bad}} \cap E_1 \cap E_2 \cap E_3$, and identify a new partition (namely, $\{V_{\text{L}}, V_{\text{H}}\}$ to be defined) of the nodes. This new partition is used to drive a contradiction in Step 4.

First consider the case that the number of hypernodes is $s = 1$. In this case E_1 and E_{bad} gives a direct contradiction, and we have $E_{\text{bad}} \cap E_1 \cap E_2 \cap E_3 = \emptyset$. We now analyze the case when the number of hypernodes is $s \ge 2$. We arbitrarily find one course from each hypernode and denote them as $i_1 \in V_1, \ldots, i_s \in V_s$.

We condition on $E_{\text{bad}} \cap E_1 \cap E_2 \cap E_3$. Recall that by definition (11.39), the event E_{bad} requires that there exists $i, i' \in [d]$ such that

$$|\widehat{x}_{i'} - \widehat{x}_i| > 4d^2\epsilon. \tag{11.47}$$

By the definition (11.41) of E_1, we have that i and i' cannot be in the same hypernode. Hence, we assume $i \in V_m$ and $i' \in V_{m'}$, and assume $m < m'$ without loss of generality. We bound $\widehat{x}_{i'} - \widehat{x}_i$ as

$$\begin{aligned}\widehat{x}_{i'} - \widehat{x}_i &= (\widehat{x}_{i'} - \widehat{x}_{i_{m'}}) + (\widehat{x}_{i_{m'}} - \widehat{x}_{i_{m'-1}}) + \ldots + (\widehat{x}_{i_{m+1}} - \widehat{x}_{i_m}) + (\widehat{x}_{i_m} - \widehat{x}_{i'})\\ &\overset{\text{(i)}}{<} 2\epsilon + d\epsilon < 4d^2\epsilon, \end{aligned} \tag{11.48}$$

where (i) is true by events E_1 and E_2. Combining (11.47) and (11.48), we must have $\widehat{x}_{i'} - \widehat{x}_i < -4d^2\epsilon$, or equivalently

$$\widehat{x}_i - \widehat{x}_{i'} > 4d^2\epsilon. \tag{11.49}$$

We decompose $\widehat{x}_i - \widehat{x}_{i'}$ as

$$\begin{aligned}\widehat{x}_i - \widehat{x}_{i'} &= (\widehat{x}_i - \widehat{x}_{i_m}) + (\widehat{x}_{i_m} - \widehat{x}_{i_{m+1}}) + \ldots + (\widehat{x}_{i_{m'-1}} - \widehat{x}_{i_{m'}}) + (\widehat{x}_{i_{m'}} - \widehat{x}_{i'})\\ &\overset{\text{(i)}}{<} 2\epsilon + (\widehat{x}_{i_m} - \widehat{x}_{i_{m+1}}) + \ldots + (\widehat{x}_{i_{m'-1}} - \widehat{x}_{i_{m'}}), \end{aligned} \tag{11.50}$$

where (i) is due to event E_1. Combining (11.49) and (11.50), we have

$$\begin{aligned}2\epsilon + (\widehat{x}_{i_m} - \widehat{x}_{i_{m+1}}) + \ldots + (\widehat{x}_{i_{m'-1}} - \widehat{x}_{i_{m'}}) &> \widehat{x}_i - \widehat{x}_{i'} > 4d^2\epsilon\\ (\widehat{x}_{i_m} - \widehat{x}_{i_{m+1}}) + \ldots + (\widehat{x}_{i_{m'-1}} - \widehat{x}_{i_{m'}}) &> (4d^2 - 2)\epsilon > 3d^2\epsilon. \end{aligned}$$

Hence, we have

$$\begin{aligned}d \cdot \max\{(\widehat{x}_{i_m} - \widehat{x}_{i_{m+1}}), \ldots, (\widehat{x}_{i_{m'-1}} - \widehat{x}_{i_{m'}})\} &> 3d^2\epsilon\\ \max\{(\widehat{x}_{i_m} - \widehat{x}_{i_{m+1}}), \ldots, (\widehat{x}_{i_{m'-1}} - \widehat{x}_{i_{m'}})\} &> 3d\epsilon. \end{aligned} \tag{11.51}$$

Without loss of generality, we assume that in (11.51) we have integer m^* with $m \leq m^* < m'$ such that

$$\widehat{x}_{i_{m^*}} - \widehat{x}_{i_{m^*+1}} > 3d\epsilon. \tag{11.52}$$

Now consider any $m, m' \in [s]$ such that $m \leq m^* < m'$, and for any $i \in V_m$ and $i' \in V_{m'}$, we have

$$\begin{aligned}\widehat{x}_i - \widehat{x}_{i'} &= (\widehat{x}_i - \widehat{x}_{i_m}) + (\widehat{x}_{i_m} - \widehat{x}_{i_{m+1}}) + \ldots + (\widehat{x}_{i^*_m} - \widehat{x}_{i_{m^*+1}}) + \ldots + (\widehat{x}_{i_{m'-1}} - \widehat{x}_{i_{m'}}) + (\widehat{x}_{i_{m'}} - \widehat{x}_{i'})\\ &\overset{\text{(i)}}{>} -2\epsilon + 3d\epsilon - d\epsilon > \epsilon, \end{aligned}$$

where (i) is by events E_1 and E_2 combined with (11.52). Equivalently, denote $V_{\text{L}} := V_1 \cup \ldots \cup V_{m^*}$ and $V_{\text{H}} := V_{m^*+1} \cup \ldots \cup V_s$, we have

$$\widehat{x}_i - \widehat{x}_{i'} > \epsilon \qquad \forall i \in V_{\text{L}},\ i' \in V_{\text{H}}. \tag{11.53}$$

Step 4: Showing $\mathbb{P}(E_{\text{bad}}, E_1 \cap E_2 \cap E_3) = 0$ by deriving a contradiction We consider any solution $(\widehat{x}, \widehat{B})$ of our estimator at $\lambda = 0$ conditional on $E_{\text{bad}} \cap E_1 \cap E_2 \cap E_3$, and derive a contradiction. Hence, we have $\mathbb{P}(E_{\text{bad}}, E_1 \cap E_2 \cap E_3) = 0$.

Analyzing properties of $\widehat{B}$ By Lemma 11.20, any bias term $\widehat{b}_{ij}$ for $i \in V_{\text{L}}$ has a smaller rank than any bias term $\widehat{b}_{ij}$ for $i \in V_{\text{H}}$. Therefore, the mean of $\widehat{B}$ over elements in V_{L} is less than or equal to the mean of $\widehat{B}$ over V_{H}. That is, with the definition of $\widehat{b}_{\text{L}}$ and $\widehat{b}_{\text{H}}$ as

$$\widehat{b}_{\text{L}} := \frac{1}{|V_{\text{L}}| \cdot n} \sum_{i \in V_{\text{L}}} \sum_{j \in [n]} \widehat{b}_{ij} \tag{11.54a}$$

$$\widehat{b}_{\text{H}} := \frac{1}{|V_{\text{H}}| \cdot n} \sum_{i \in V_{\text{H}}} \sum_{j \in [n]} \widehat{b}_{ij}, \tag{11.54b}$$

We have the deterministic relation $\widehat{b}_{\text{L}} \leq \widehat{b}_{\text{H}}$.

First consider the case of $\widehat{b}_{\text{L}} = \widehat{b}_{\text{H}}$. Since $\widehat{B}$ obeys the total ordering $\mathcal{O}$, we have $\widehat{B} = c$ for some constant c. Conditional on E_3, it can be verified that for any $c \in \mathbb{R}$, the objective (4.2) attained at $(\widehat{x}, \widehat{B})$ is strictly positive. Recall from the model (11.23) that $Y = B$. Hence, an objective (4.2) of 0 can be attained by the solution $(0, B)$. Contradiction to the assumption that $(\widehat{x}, \widehat{B})$ is the minimizer of the objective.

Now we consider the case of $\widehat{b}_{\text{L}} < \widehat{b}_{\text{H}}$. We have that either $\widehat{b}_{\text{L}} < 0$ or $\widehat{b}_{\text{H}} > 0$ (or both). Without loss of generality we assume $\widehat{b}_{\text{H}} > 0$.

Constructing an alternative solution We now construct an alternative solution by increasing $\widehat{x}_i$ for every course $i \in V_{\text{H}}$ by a tiny amount, and prove for contradiction that this alternative solution is preferred by the tie-breaking rule of minimizing $\|B\|_F^2$. We construct the alternative solution $(\widehat{x}', \widehat{B}')$ as

$$\begin{aligned} \widehat{x}'_i &= \begin{cases} \widehat{x}_i & \text{if } i \in V_{\text{L}} \\ \widehat{x}_i + \Delta & \text{if } i \in V_{\text{H}} \end{cases} \\ \widehat{B}' &= Y - \widehat{x}' \mathbf{1}^T, \end{aligned} \tag{11.55}$$

for some sufficiently small $\Delta > 0$ whose value is specified later. Since $(\widehat{x}, \widehat{B})$ is a solution, as discussed previously it has to attain an objective of 0. By the construction (11.55), it can be verified that $(\widehat{x}', \widehat{B}')$ also attains an objective of 0. In what remains for this step, we first show that the alternative solution $(\widehat{x}', \widehat{B}')$ satisfies all ordering constraints by the total ordering $\mathcal{O}$. Then we show that $\|\widehat{B}'\|_F^2 < \|\widehat{B}\|_F^2$, and therefore $(\widehat{x}', \widehat{B}')$ is preferred by the tie-breaking rule over $(\widehat{x}, \widehat{B})$, giving a contradiction.

The alternative solution $(\widehat{x}', \widehat{B}')$ satisfies all ordering constraints in $\mathcal{O}$ Since both $(\widehat{x}, \widehat{B})$ and $(\widehat{x}', \widehat{B}')$ attain an objective of 0, we have the deterministic relation

$$y_{ij} = \widehat{x}_i + \widehat{b}_{ij} = \widehat{x}'_i + \widehat{b}'_{ij} \qquad \forall i \in [d], j \in [n]. \tag{11.56}$$

Consider any constraint $((i,j),(i',j')) \in \mathcal{O}$. If $i, i' \in V_{\mathrm{L}}$, then we have

$$\begin{aligned}\widehat{b}'_{ij} - \widehat{b}'_{i'j'} &= y_{ij} - \widehat{x}'_i - (y_{i'j'} - \widehat{x}'_{i'})\\ &= y_{ij} - \widehat{x}_i - (y_{i'j'} - \widehat{x}_{i'})\\ &= \widehat{b}_{ij} - \widehat{b}_{i'j'} \overset{\text{(i)}}{<} 0,\end{aligned}$$

where (i) is true because by assumption $(\widehat{x}, \widehat{B})$ is the optimal solution, and hence $\widehat{B}$ satisfies the ordering constraint of $\widehat{b}_{ij} \leq \widehat{b}_{i'j'}$. Similarly if $i, i' \in V_{\mathrm{H}}$, then $(\widehat{x}', \widehat{B}')$ also satisfies this ordering constraint. Finally, consider the case where one of $\{i, i'\}$ is in V_{L} and the other is in V_{H}. Due to Lemma 11.20 regarding the ranks combined with the definition of $(V_{\mathrm{L}}, V_{\mathrm{H}})$, it can only be the case that $i \in V_{\mathrm{L}}$ and $i' \in V_{\mathrm{H}}$. For any $\Delta \in (0, \epsilon)$, we have that conditional on $E_{\text{bad}} \cap E_1 \cap E_2 \cap E_3$,

$$\begin{aligned}\widehat{b}'_{ij} - \widehat{b}'_{i'j'} &= (y_{ij} - \widehat{x}'_i) - (y_{i'j'} - \widehat{x}'_{i'})\\ &= (b_{ij} - \widehat{x}_i) - (b_{i'j'} - \widehat{x}_{i'} - \Delta)\\ &= (b_{ij} - b_{i'j'}) + (\widehat{x}_{i'} + \Delta - \widehat{x}_i) \overset{\text{(i)}}{<} 0,\end{aligned}$$

where (i) is true because the ordering constraint $((i,j),(i',j'))$ gives $b_{ij} \leq b_{i'j'}$. Moreover, we have $\widehat{x}_{i'} - \widehat{x}_i < -\epsilon$ due to (11.53). Hence, all ordering constraints are satisfied by the alternative solution $(\widehat{x}', \widehat{B}')$.

The alternative solution $(\widehat{x}', \widehat{B}')$ satisfies $\|\widehat{B}'\|_F < \|\widehat{B}\|_F$, thus preferred by tie-breaking Plugging in the construction (11.55), we compute $\|\widehat{B}'\|_F^2$ as

$$\begin{aligned}\|\widehat{B}'\|_F^2 &= \sum_{i \in V_{\mathrm{L}}} \sum_{j \in [n]} (y_{ij} - \widehat{x}_i)^2 + \sum_{i \in V_{\mathrm{H}}} \sum_{j \in [n]} (y_{ij} - \widehat{x}_i - \Delta)^2\\ &\overset{\text{(i)}}{=} \sum_{i \in V_{\mathrm{L}}} \sum_{j \in [n]} (\widehat{b}_{ij})^2 + \sum_{i \in V_{\mathrm{H}}} \sum_{j \in [n]} (\widehat{b}_{ij} - \Delta)^2, \qquad (11.57)\end{aligned}$$

where (i) is true by (11.56). Taking the partial derivative of (11.57) with respect to Δ, we have

$$\frac{\partial \|\widehat{B}'\|_F^2}{\partial \Delta} = 2\left(|V_{\mathrm{H}}| \cdot n\Delta - \sum_{i \in V_{\mathrm{H}}} \sum_{j \in [n]} \widehat{b}_{ij}\right) = 2|V_{\mathrm{H}}| \cdot n(\Delta - \widehat{b}_{\mathrm{H}}). \qquad (11.58)$$

By the assumption of $\widehat{b}_{\mathrm{H}} > 0$, the partial derivative (11.58) is strictly negative for any $\Delta \in \left[0, \widehat{b}_{\mathrm{H}}\right)$. Contradiction to the fact that $\widehat{B}$ (corresponding to $\Delta = 0$) is the solution with the minimal Frobenius norm $\|\widehat{B}\|_F^2$. Hence, $(\widehat{x}, \widehat{B})$ cannot be a solution, and we have

$$\mathbb{P}(E_{\text{bad}}, E_1 \cap E_2 \cap E_3) = 0.$$

Step 4: Invoking Lemma 11.8 Recall from Step 2 that $\lim_{n\to\infty} \mathbb{P}(E_{\text{bad}}, E_1 \cap E_2 \cap E_3) = 0$ implies $\lim_{n\to\infty} \mathbb{P}(E_{\text{bad}}) = 0$. Equivalently, for any $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i,i' \in [d]} |\widehat{x}_{i'} - \widehat{x}_i| < \epsilon\right) = 1.$$

Invoking Lemma 11.8 completes the proof.

11.4 Proof of Proposition 4.7

We denote $(\widehat{x}^{(\infty)}, B^{(\infty)})$ as the values given by expression (4.3). We prove that

$$(\widehat{x}^{(\infty)}, B^{(\infty)}) = \lim_{\lambda \to \infty} (\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)}).$$

Denote the minimal value of the first term in the objective (4.2) as

$$V^* := \min_{\substack{x \in \mathbb{R}^d, B \in \mathbb{R}^{d \times n} \\ B \text{ satisfies } \mathcal{O}}} \left\| Y - x\mathbf{1}^T - B \right\|_F^2.$$

Denote V as the value of the first term attained at $(\widehat{x}^{(\infty)}, \widehat{B}^{(\infty)})$. By the definition of V^* as the minimal value over the domain, we have $V \geq V^*$. We discuss the following two cases depending on the value of V.

Case of $V = V^*$: We have that $(\widehat{x}^{(\infty)}, \widehat{B}^{(\infty)})$ is the solution for any $\lambda \in (0, \infty)$, because it attains the minimal value separately for the two terms in the objective (4.2). By Proposition 11.1, a unique solution exists for any $\lambda \in (0, \infty)$. Hence, the limit $\lim_{\lambda \to \infty} (\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$ exists and we have $(\widehat{x}^{(\infty)}, \widehat{B}^{(\infty)}) = \lim_{\lambda \to \infty} (\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$.

Case of $V > V^*$: We first show that $\lim_{\lambda \to \infty} \widehat{B}^{(\lambda)} = 0$. That is, we show that for any $\epsilon > 0$, there exists some $\lambda_0 > 0$, such that $\|\widehat{B}^{(\lambda)}\|_F^2 < \epsilon$ for all $\lambda \in (\lambda_0, \infty)$.

Take $\lambda_0 = \frac{V - V^*}{\epsilon}$, and assume for contradiction that there exists some $\lambda^* > \lambda_0$ such that $\|\widehat{B}^{(\lambda^*)}\|_F^2 > \epsilon$. The objective (4.2) (setting $\lambda = \lambda^*$) attained by $(\widehat{x}^{(\lambda^*)}, \widehat{B}^{(\lambda^*)})$ is lower-bounded by

$$\|Y - \widehat{x}^{(\lambda^*)} - \widehat{B}^{(\lambda^*)}\|_2^2 + \lambda^* \|\widehat{B}^{(\lambda^*)}\|_F^2 > V^* + \lambda_0 \epsilon > V^* + (V - V^*) = V.$$

On the other hand, the objective attained by $(\widehat{x}^{(\infty)}, \widehat{B}^{(\infty)})$ is V. Hence, $(\widehat{x}^{(\infty)}, \widehat{B}^{(\infty)})$ attains a strictly smaller value of the objective than $(\widehat{x}^{(\lambda^*)}, \widehat{B}^{(\lambda^*)})$ at $\lambda = \lambda^*$. Contradiction to the assumption that $(\widehat{x}^{(\lambda^*)}, \widehat{B}^{(\lambda^*)})$ is the solution at $\lambda = \lambda^*$. Hence, we have $\lim_{\lambda \to \infty} \widehat{B}^{(\lambda)} = 0$.

Combining the fact that $\lim_{\lambda \to \infty} \widehat{B}^{(\lambda)} = 0$ with the relation (11.14) in Lemma 11.3 (at any $\lambda \in [0, \infty)$), we have that for each $i \in [d]$,

$$\widehat{x}_i^{(\lambda)} = \frac{1}{n} \sum_{j \in [n]} \left(y_{ij} - \widehat{b}_{ij}^{(\lambda)} \right) \to \frac{1}{n} \sum_{j \in [n]} y_{ij} \qquad \text{as } \lambda \to \infty,$$

completing the proof.

11.5 Proof of Theorem 4.9

The proof follows notation in Appendix 11.1 and preliminaries in Appendix 11.2. By Corollary 11.6, we assume $x^* = 0$ without loss of generality. We also assume without loss of generality that the standard deviation of the Gaussian bias distribution is $\sigma = 1$. Given $x^* = 0$ and the assumption that there is no noise, model (4.1) reduces to:

$$Y = B. \tag{11.59}$$

Both part (a) and part (b) consist of 3 similar steps. We start with the first step, and proceed separately for the two remaining steps for the two parts.

Step 1: Showing the consistency of our estimator at $\lambda = 0$ restricted to the training set Ω^{t}.

In the first step, we show that our estimator is consistent under group orderings satisfying part (a) and part (b), on any fixed training set $\Omega^{\text{t}} \subseteq [d] \times [n]$ obtained by Algorithm 2. Note that Theorem 4.5(a) and Theorem 4.5(c) give the desired consistency result when the data is full observations $\Omega = [d] \times [n]$. It remains to extend the proof of Theorem 4.5(a) and Theorem 4.5(c) to any Ω^{t} given by Algorithm 2. The following theorem states that part (a) and part (c) of Theorem 4.5 still hold for the estimator (11.9) restricted to Ω^{t}. We use $(\widehat{x}^{(0)}, \widehat{B}^{(0)})$ to denote the solution to (11.9) restricted to Ω^{t} for the remaining of the proof of Theorem 4.9.

Theorem 11.21 (Generalization of Theorem 4.5 to any Ω^{t})**.** *Consider any fixed $\Omega^{\text{t}} \subseteq [d] \times [n]$ obtained by Algorithm 2. Suppose the partial ordering is one of*

(a) any group ordering satisfying the all c-fraction assumption, or

(b) any total ordering.

Then for any $\epsilon > 0$ and $\delta > 0$, there exists an integer n_0 (dependent on ϵ, δ, c, d), such that for every $n \geq n_0$ and every partial ordering satisfying one of the conditions (a) or (b), the estimator $\widehat{x}^{(0)}$ (as the solution to (11.9) *restricted to Ω^{t}) satisfies*

$$\mathbb{P}\Big(\|\widehat{x}^{(0)} - x^*\|_2 < \epsilon\Big) \geq 1 - \delta. \tag{11.60}$$

Equivalently, for any $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\Big(\|\widehat{x}^{(0)} - x^*\|_2 < \epsilon\Big) = 1. \tag{11.61}$$

The proof of this theorem is in Appendix 11.11.1. Now we consider the consistency of the bias term $\widehat{B}$. Given the model (11.59), the objective (11.9) at $\lambda = 0$ equals 0 at the values of $(\widehat{x}, \widehat{B}) = (0, B)$. Hence, objective (11.9) attains a value of 0 at the solution $(\widehat{x}^{(0)}, \widehat{B}^{(0)})$. Therefore, we have the deterministic relation $Y_{\Omega^{\text{t}}} = [\widehat{x}^{(0)}\mathbf{1}^T + \widehat{B}^{(0)}]_{\Omega^{\text{t}}}$. For any $(i, j) \in \Omega^{\text{t}}$, we have

$$\widehat{b}_{ij}^{(0)} = Y_{ij} - \widehat{x}_i^{(0)} \overset{\text{(i)}}{=} b_{ij} - \widehat{x}_i^{(0)}, \tag{11.62}$$

where equality (i) is true because of the model (11.59). Combining (11.62) with (11.61), we have that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|\widehat{b}_{ij}^{(0)} - b_{ij}\Big| < \epsilon, \quad \forall (i, j) \in \Omega^{\text{t}}\Big) = 1. \tag{11.63}$$

This completes Step 1 of the proof. The remaining two steps are presented separately for the two parts.

11.5.1 Proof of part (a)

We fix some constant $\epsilon_1 > 0$ whose value is determined later. For clarity of notation, we denote the constant in the all constant-fraction assumption as c_{f}.

Step 2: Computing the validation error at $\lambda = 0$

We first analyze the interpolated bias $\widetilde{B}^{(0)}$. Recall that G_k^{t} and G_k^{v} denote the set of elements of group $k \in [r]$ in the training set Ω^{t} and the validation set Ω^{v}, respectively. By symmetry of the interpolation expression in Line 15 of Algorithm 2 and Definition 4.1 of the group ordering, it can be verified that the interpolated bias $\widetilde{b}_{ij}$ is identical for all elements within any group $k \in [r]$. That is, for each $k \in [r]$, we have

$$\widetilde{b}_{ij} = \widetilde{b}_{i'j'}, \text{ for any } (i,j),(i',j') \in G_k^{\mathrm{v}}. \tag{11.64}$$

Denote $\widetilde{b}_k := \widetilde{b}_{ij}$ for any $(i,j) \in G_k^{\mathrm{t}}$. By (11.64), we have that $\widetilde{b}_k$ is well-defined. Denote the random variables b_k^{t} and b_k^{v} as the mean of the (random) bias B in group $k \in [r]$, over G_k^{t} and G_k^{v}, respectively. Denote the random variable b_{ik}^{v} as the mean of the (random) B of group $k \in [r]$ in course $i \in [d]$ over Ω^{v}. That is, we define

$$b_k^{\mathrm{t}} := \frac{1}{|G_k^{\mathrm{t}}|} \sum_{(i,j)\in G_k^{\mathrm{t}}} b_{ij} \tag{11.65}$$

$$b_k^{\mathrm{v}} := \frac{1}{|G_k^{\mathrm{v}}|} \sum_{(i,j)\in G_k^{\mathrm{v}}} b_{ij} \tag{11.66}$$

$$b_{ik}^{\mathrm{v}} := \frac{1}{|G_{ik}^{\mathrm{v}}|} \sum_{j\in G_{ik}^{\mathrm{v}}} b_{ij}. \tag{11.67}$$

Denote $\widehat{b}_k^{\mathrm{t}}$ likewise as the mean of the estimated bias $\widehat{B}$ over G_k^{t}. Given $Y = B$ from model (11.59), the validation error at $\lambda = 0$ is computed as:

$$\begin{aligned} e^{(0)} &= \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in\Omega^{\mathrm{v}}} \left(y_{ij} - \widehat{x}_i^{(0)} - \widetilde{b}_{ij}\right)^2 \\ &= \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{i\in[d],k\in[r]} \sum_{j\in G_{ik}^{\mathrm{v}}} \left(b_{ij} - \widehat{x}_i^{(0)} - \widetilde{b}_k\right)^2. \end{aligned} \tag{11.68}$$

We first analyze the term $\widetilde{b}_k$ in (11.68). The following lemma shows that the interpolation procedure in Algorithm 2 ensures that $\widetilde{b}_k$ is close to $\widehat{b}_k^{\mathrm{t}}$, the mean of the estimated bias over G_k^{t}.

Lemma 11.22. *Consider any group ordering $\mathcal{O}$ that satisfies the all c_{f}-fraction assumption, and any $\Omega^{\mathrm{t}} \subseteq [d] \times [n]$ obtained by Algorithm 2. Then for any $\lambda \in [0, \infty]$ we have the deterministic relation:*

$$\left|\widetilde{b}_k - \widehat{b}_k^{\mathrm{t}}\right| \leq \frac{12}{c_{\mathrm{f}} dn} \cdot \max_{(i,j)\in\Omega^{\mathrm{t}}} \left|\widehat{b}_{ij}\right| \qquad \forall k \in [r].$$

The proof of this result is provided in Appendix 11.11.2. Combining Lemma 11.22 with the consistency (11.63) of $\widehat{B}^{(0)}$ from Step 1 and a bound on $\max_{(i,j)\in\Omega^{\mathrm{t}}}|b_{ij}|$ from Lemma 11.12, we have the following lemma.

Lemma 11.23. *Under the same condition as Lemma 11.22, the interpolated bias at* $\lambda = 0$ *satisfies*

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|\widetilde{b}_k - b^{\mathrm{t}}_k\Big| < \epsilon, \quad \forall k \in [r]\Big) = 1.$$

The proof of this result is provided in Appendix 11.11.3. Recall that $\overline{b}_{G_k}$ denotes the the mean of the bias of any group $k \in [r]$. The following lemma gives concentration inequality results that the quantities b^{v}_{ik} and b^{t}_k are close to b_k. Note that this lemma is on the bias B and does not involve any estimator.

Lemma 11.24. *Consider any group ordering* $\mathcal{O}$ *that satisfies the all* c_{f}*-fraction assumption. Consider any fixed training-validation split* $(\Omega^{\mathrm{t}}, \Omega^{\mathrm{v}})$ *obtained by Algorithm 2. For any* $\epsilon > 0$*, we have*

$$\lim_{n\to\infty} \mathbb{P}\Big(\big|b^{\mathrm{v}}_{ik} - \overline{b}_{G_k}\big| < \epsilon, \quad \forall i \in [d], k \in [r]\Big) = 1 \tag{11.69a}$$

$$\lim_{n\to\infty} \mathbb{P}\Big(\big|b^{\mathrm{t}}_{k} - \overline{b}_{G_k}\big| < \epsilon, \quad \forall k \in [r]\Big) = 1. \tag{11.69b}$$

The proof of this result is provided in Appendix 11.11.4. Combining Lemma 11.23 and (11.69) from Lemma 11.24 with a union bound, we have the following corollary.

Corollary 11.25. *Consider any group ordering* $\mathcal{O}$ *that satisfies the all* c_{f}*-fraction assumption. Consider any fixed* $\Omega^{\mathrm{t}} \subseteq [d] \times [n]$ *obtained by Algorithm 2. For any* $\epsilon > 0$*, the interpolated bias at* $\lambda = 0$ *satisfies*

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|b^{\mathrm{v}}_{ik} - \widetilde{b}_k\Big| < \epsilon, \quad \forall i \in [d], k \in [r]\Big) = 1.$$

Consider each $i \in [d]$ and $k \in [r]$. The terms in the validation error (11.68) involving course i and group k are:

$$\begin{aligned} e^{(0)}_{ik} := \frac{1}{|\Omega^{\mathrm{v}}|}\sum_{j\in G^{\mathrm{v}}_{ik}} \Big(b_{ij} - \widehat{x}^{(0)}_i - \widetilde{b}_k\Big)^2 &= \frac{1}{|\Omega^{\mathrm{v}}|}\left[\sum_{j\in G^{\mathrm{v}}_{ik}} \Big(b_{ij} - \widetilde{b}_k\Big)^2 + |G^{\mathrm{v}}_{ik}| \cdot \widehat{x}^2_i - 2\sum_{j\in G^{\mathrm{v}}_{ik}} \Big(b_{ij} - \widetilde{b}_k\Big)\widehat{x}_i\right] \\ &\overset{\text{(i)}}{=} \underbrace{\frac{1}{|\Omega^{\mathrm{v}}|}\sum_{j\in G^{\mathrm{v}}_{ik}} \Big(b_{ij} - \widetilde{b}_k\Big)^2}_{T_1} + \underbrace{\frac{|G^{\mathrm{v}}_{ik}|}{|\Omega^{\mathrm{v}}|}\widehat{x}^2_i}_{T_2} - \underbrace{\frac{2|G^{\mathrm{v}}_{ik}|}{|\Omega^{\mathrm{v}}|} \cdot (b^{\mathrm{v}}_{ik} - \widetilde{b}_k)\widehat{x}_i}_{T_3}, \end{aligned}$$

where (i) is true by the definition (11.67) of b^{v}_{ik}. We now consider the three terms T_1, T_2 and T_3 (dependent on i and k), respectively.

Term T_2: By the convergence (11.61) of $\widehat{x}^{(0)}$ in Theorem 11.21(a), we have

$$\lim_{n\to\infty} \mathbb{P}\left(T_2 \le \frac{|G^{\mathrm{v}}_{ik}|}{|\Omega^{\mathrm{v}}|}\epsilon^2_1, \quad \forall i \in [d], k \in [r]\right) = 1. \tag{11.70}$$

Term T_3: We have

$$T_3 \leq 2\frac{|G^{\mathrm{v}}_{ik}|}{|\Omega^{\mathrm{v}}|} \cdot \left| b^{\mathrm{v}}_{ik} - \widetilde{b}_k \right| \cdot |\widehat{x}_i| \leq 2 \left| b^{\mathrm{v}}_{ik} - \widetilde{b}_k \right| \cdot |\widehat{x}_i|.$$

By combining the convergence (11.61) of $\widehat{x}^{(0)}$ in Theorem 11.21(a) and Corollary 11.25 with a union bound, we have

$$\lim_{n\to\infty} \mathbb{P}\Big(T_3 \leq \frac{2|G^{\mathrm{v}}_{ik}|}{|\Omega^{\mathrm{v}}|}\epsilon_1^2, \quad \forall i \in [d], k \in [r] \Big) = 1. \tag{11.71}$$

Term T_1: We have

$$\begin{aligned} T_1 = \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{j \in G^{\mathrm{v}}_{ik}} \left(b_{ij} - \widetilde{b}_k \right)^2 &= \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{j \in G^{\mathrm{v}}_{ik}} \left(b_{ij} - b^{\mathrm{v}}_{ik} + b^{\mathrm{v}}_{ik} - \widetilde{b}_k \right)^2 \\ &= \frac{1}{|\Omega^{\mathrm{v}}|} \left[\sum_{j \in G^{\mathrm{v}}_{ik}} (b_{ij} - b^{\mathrm{v}}_{ik})^2 + |G^{\mathrm{v}}_{ik}| \cdot (b^{\mathrm{v}}_{ik} - \widetilde{b}_k)^2 + 2 \sum_{j \in G^{\mathrm{v}}_{ik}} (b_{ij} - b^{\mathrm{v}}_{ik})(b^{\mathrm{v}}_{ik} - \widetilde{b}_k) \right] \\ &\overset{\text{(i)}}{=} \frac{1}{|\Omega^{\mathrm{v}}|} \left[\sum_{j \in G^{\mathrm{v}}_{ik}} (b_{ij} - b^{\mathrm{v}}_{ik})^2 + |G^{\mathrm{v}}_{ik}| \cdot (b^{\mathrm{v}}_{ik} - \widetilde{b}_k)^2 \right] \end{aligned}$$

where inequality (i) holds because $\sum_{j \in G^{\mathrm{v}}_{ik}} (b_{ij} - b^{\mathrm{v}}_{ik}) = 0$ by the definition (11.67) of b^{v}_{ik}. By Corollary 11.25, we have

$$\lim_{n\to\infty} \left(T_1 < \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{j \in G^{\mathrm{v}}_{ik}} (b_{ij} - b^{\mathrm{v}}_{ik})^2 + \frac{|G^{\mathrm{v}}_{ik}|}{|\Omega^{\mathrm{v}}|}\epsilon_1^2, \quad \forall i \in [d], k \in [r] \right) = 1. \tag{11.72}$$

Combining the three terms from (11.70), (11.71) and (11.72), we bound $e^{(0)}_{ik}$ as

$$\lim_{n\to\infty} \left(e^{(0)}_{ik} = T_1 + T_2 + T_3 < \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{j \in G^{\mathrm{v}}_{ik}} (b_{ij} - b^{\mathrm{v}}_{ik})^2 + \frac{4|G^{\mathrm{v}}_{ik}|}{|\Omega^{\mathrm{v}}|}\epsilon_1^2, \quad \forall i \in [d], k \in [r] \right) = 1. \tag{11.73}$$

By the all c_{f}-fraction assumption, the number of groups is upper-bounded by a constant as $r \leq \frac{1}{c_{\mathrm{f}}}$. Taking a union bound of (11.73) over $i \in [d]$ and $k \in [r]$, we have

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P} \left(e^{(0)} = \sum_{i \in [d], k \in [r]} e^{(0)}_{ik} < \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{i \in [d], k \in [r]} \left[\sum_{j \in G^{\mathrm{v}}_{ik}} (b_{ij} - b^{\mathrm{v}}_{ik})^2 + 4|G^{\mathrm{v}}_{ik}| \cdot \epsilon_1^2 \right] \right) &= 1 \\ \lim_{n\to\infty} \mathbb{P} \left(e^{(0)} < \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{i \in [d], k \in [r]} \sum_{j \in G^{\mathrm{v}}_{ik}} (b_{ij} - b^{\mathrm{v}}_{ik})^2 + 4\epsilon_1^2 \right) &= 1. \end{aligned} \tag{11.74}$$

This completes Step 2 of bounding the validation error at $\lambda = 0$.

Step 3: Computing the validation error at general $\lambda \in \Lambda_\epsilon$, and showing that it is greater than the validation error at $\lambda = 0$

Recall from (11.10) the definition of the random set $\Lambda_\epsilon := \{\lambda \in [0, \infty] : \|\widehat{x}^{(\lambda)}\|_2 > \epsilon\}$. In this step, we show that

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} > e^{(0)}, \quad \forall \lambda \in \Lambda_\epsilon\right) = 1. \tag{11.75}$$

From (11.75), we have that the estimated quality $\widehat{x}^{(\lambda_{\text{cv}})}$ by cross-validation satisfies

$$\lim_{n\to\infty} (\lambda_{\text{cv}} \notin \Lambda_\epsilon) = 1$$

and consequently by the definition of Λ_ϵ

$$\lim_{n\to\infty} \mathbb{P}\left(\|\widehat{x}^{(\lambda_{\text{cv}})}\|_2 < \epsilon\right) = 1.$$

It remains to prove (11.75).

Proof of (11.75) For any $i \in [d]$ and $k \in [r]$, the terms in the validation error at any $\lambda \in [0, \infty]$ involving course i and group k are computed as:

$$\begin{aligned} e_{ik}^{(\lambda)} = \frac{1}{|\Omega^{\text{v}}|} \sum_{j \in G_{ik}^{\text{v}}} \left(b_{ij} - \widehat{x}_i^{(\lambda)} - \widetilde{b}_k^{(\lambda)}\right)^2 &= \frac{1}{|\Omega^{\text{v}}|} \sum_{j \in G_{ik}^{\text{v}}} \left(b_{ij} - b_{ik}^{\text{v}} + b_{ik}^{\text{v}} - \widehat{x}_i - \widetilde{b}_k\right)^2 \\ &\overset{\text{(i)}}{=} \frac{1}{|\Omega^{\text{v}}|} \sum_{j \in G_{ik}^{\text{v}}} (b_{ij} - b_{ik}^{\text{v}})^2 + \underbrace{\frac{|G_{ik}^{\text{v}}|}{|\Omega^{\text{v}}|} \left(b_{ik}^{\text{v}} - \widehat{x}_i - \widetilde{b}_k\right)^2}_{T_{ik}}, \end{aligned} \tag{11.76}$$

where (i) is true because $\sum_{j \in G_{ik}^{\text{v}}} (b_{ij} - b_{ik}^{\text{v}}) = 0$ by the definition (11.67) of b_{ik}^{v}. Note that the first term in (11.76) is identical to the first term in (11.73) from Step 2. We now analyze the second term T_{ik} in (11.76). On the one hand, by Lemma 11.7(a), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i,i' \in [d]} \widehat{x}_i - \widehat{x}_{i'} > \frac{\epsilon}{\sqrt{d}}, \quad \forall \lambda \in \Lambda_\epsilon\right) = 1. \tag{11.77}$$

On the other hand, taking a union bound of (11.69a) in Lemma 11.24 over $i, i' \in [d]$, we have

$$\lim_{n\to\infty} \mathbb{P}\Big(|b_{ik}^{\text{v}} - b_{i'k}^{\text{v}}| < \frac{\epsilon}{2\sqrt{d}}, \quad \forall i, i' \in [d], k \in [r]\Big) = 1. \tag{11.78}$$

Conditional on (11.77) and (11.78), for every $\lambda \in \Lambda_\epsilon$ and for every $k \in [r]$,

$$\begin{aligned} \max_{i,i' \in [d]} \left| \left(b_{ik}^{\text{v}} - \widehat{x}_i - \widetilde{b}_k\right) - \left(b_{i'k}^{\text{v}} - \widehat{x}_{i'} - \widetilde{b}_k\right) \right| &= \max_{i,i' \in [d]} |(b_{ik}^{\text{v}} - b_{i'k}^{\text{v}}) - (\widehat{x}_i - \widehat{x}_{i'})| \\ &\geq \max_{i,i' \in [d]} (\widehat{x}_i - \widehat{x}_{i'}) - \max_{i,i' \in [d]} |b_{ik}^{\text{v}} - b_{i'k}^{\text{v}}| \\ &> \frac{\epsilon}{\sqrt{d}} - \frac{\epsilon}{2\sqrt{d}} = \frac{\epsilon}{2\sqrt{d}}. \end{aligned}$$

Hence, conditional on (11.77) and (11.78),

$$\max_{i,i'\in[d]}\left\{(b^{\mathrm{v}}_{ik}-\widehat{x}_i-\widetilde{b}_k)^2,(b^{\mathrm{v}}_{i'k}-\widehat{x}_{i'}-\widetilde{b}_k)^2\right\}\geq\frac{\epsilon^2}{16d}\qquad\forall k\in[r],\forall\lambda\in\Lambda_\epsilon.\tag{11.79}$$

Now consider the terms T_{ik}. By (11.20a) from Lemma 11.13 combined with the all c_{f}-fraction assumption, we have

$$\frac{|G^{\mathrm{v}}_{ik}|}{|\Omega^{\mathrm{v}}|}\geq\frac{1}{|\Omega^{\mathrm{v}}|}\cdot\frac{|G_{ik}|}{4}\geq\frac{c_{\mathrm{f}}n}{4|\Omega^{\mathrm{v}}|}=\frac{c_{\mathrm{f}}}{2d}.\tag{11.80}$$

Conditional on (11.77) and (11.78), for every $\lambda\in\Lambda_\epsilon$ and $i\in[d]$,

$$\begin{aligned}\max_{i,i'\in[d]}(T_{ik}+T_{i'k})&\overset{\text{(i)}}{\geq}\frac{c_{\mathrm{f}}}{2d}\left[\left(b^{\mathrm{v}}_{ik}-\widehat{x}_i-\widehat{b}^{\mathrm{t}}_k\right)^2+\left(b^{\mathrm{v}}_{i'k}-\widehat{x}_{i'}-\widehat{b}^{\mathrm{t}}_k\right)^2\right]\\&\overset{\text{(ii)}}{\geq}\frac{c_{\mathrm{f}}}{2d}\frac{\epsilon^2}{16d}=\frac{c_{\mathrm{f}}\epsilon^2}{32d^2},\end{aligned}$$

where inequality (i) is true by (11.80), and inequality (ii) is true by (11.79). Now consider the validation error $e^{(\lambda)}$. Conditional on (11.77) and (11.78), for every $\lambda\in\Lambda_\epsilon$,

$$\begin{aligned}e^{(\lambda)}=\sum_{i\in[d],k\in[r]}e^{(\lambda)}_{ik}&\overset{\text{(i)}}{\geq}\frac{1}{|\Omega^{\mathrm{v}}|}\sum_{i\in[d],k\in[r]}\sum_{j\in G^{\mathrm{v}}_{ik}}(b_{ij}-b^{\mathrm{v}}_{ik})^2+\sum_{i\in[d],k\in[r]}(T_{ik}+T_{i'k})\\&>\frac{1}{|\Omega^{\mathrm{v}}|}\sum_{i\in[d],k\in[r]}\sum_{j\in G^{\mathrm{v}}_{ik}}(b_{ij}-b^{\mathrm{v}}_{ik})^2+\frac{c_{\mathrm{f}}\epsilon^2}{32d^2},\end{aligned}$$

where inequality (i) is true by plugging in (11.76). Hence,

$$\lim_{n\to\infty}\left(e^{(\lambda)}>\frac{1}{|\Omega^{\mathrm{v}}|}\sum_{i\in[d],k\in[r]}\sum_{j\in G^{\mathrm{v}}_{ik}}(b_{ij}-b^{\mathrm{v}}_{ik})^2+\frac{c_{\mathrm{f}}\epsilon^2}{32d^2},\quad\forall\lambda\in\Lambda_\epsilon\right)=1.\tag{11.81}$$

We set ϵ_1 to be sufficient small such that $4\epsilon_1^2<\frac{c_{\mathrm{f}}\epsilon^2}{32d^2}$. Taking a union bound of (11.81) with (11.74) from Step 2, we have

$$\lim_{n\to\infty}\mathbb{P}\left(e^{(\lambda)}>e^{(0)},\quad\forall\lambda\in\Lambda_\epsilon\right)=1,$$

completing the proof of (11.75).

11.5.2 Proof of part (b)

We fix some constant $\epsilon_1>0$ whose value is determined later. Since the partial ordering $\mathcal{O}$ is assumed to be a total ordering, we also denote it as π.

Step 2: Computing the validation error at $\lambda=0$

For any element $(i,j)\in\Omega^{\mathrm{v}}$, recall that $\mathrm{NN}(i,j;\pi)\subseteq[d]\times[n]$ denotes the set (of size 1 or 2) of its nearest neighbors in the training set Ω^{t} with respect to the total ordering π. We use

$\mathrm{NN}(i,j)$ as the shorthand notation for $\mathrm{NN}(i,j;\pi)$. For any $\lambda \in [0,\infty]$, we define the mean of the estimated bias over the nearest-neighbor set

$$\widehat{b}^{(\lambda)}_{\mathrm{NN}(i,j)} := \frac{1}{|\mathrm{NN}(i,j)|} \sum_{(i',j')\in \mathrm{NN}(i,j)} \widehat{b}^{(\lambda)}_{i'j'}$$

Similarly, we define

$$b_{\mathrm{NN}(i,j)} := \frac{1}{|\mathrm{NN}(i,j)|} \sum_{(i',j')\in \mathrm{NN}(i,j)} b_{i'j'}.$$

Since $\mathcal{O}$ is a total ordering, the set of total orderings consistent with $\mathcal{O} = \pi$ is trivially itself, that is, $\mathcal{T} = \{\pi\}$. Then in Line 15 of Algorithm 2, the interpolated bias for any element $(i,j) \in \Omega^{\mathrm{v}}$ is $\widetilde{b}^{(\lambda)}_{ij} = \widehat{b}^{(\lambda)}_{\mathrm{NN}(i,j)}$.

Recall from the model (11.59) that $Y = B$. The validation error at $\lambda = 0$ is computed as:

$$\begin{aligned} e^{(0)} &= \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in\Omega^{\mathrm{v}}} \left(b_{ij} - \widehat{b}^{(0)}_{\mathrm{NN}(i,j)} - \widehat{x}^{(0)}_i\right)^2 \\ &\leq \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in\Omega^{\mathrm{v}}} \left(\left|b_{ij} - b_{\mathrm{NN}(i,j)}\right| + \left|b_{\mathrm{NN}(i,j)} - \widehat{b}^{(0)}_{\mathrm{NN}(i,j)}\right| + \left|\widehat{x}^{(0)}_i\right|\right)^2. \end{aligned} \tag{11.82}$$

We consider the three terms inside the summation in (11.82) separately. For the first term $\left|b_{ij} - b_{\mathrm{NN}(i,j)}\right|$, combining Lemma 11.14(b) with Lemma 11.10, we have

$$\lim_{n\to\infty} \mathbb{P}\Big(\left|b_{ij} - b_{\mathrm{NN}(i,j)}\right| < \epsilon_1, \quad \forall (i,j) \in \Omega^{\mathrm{v}}\Big) = 1 \tag{11.83}$$

For the second term $|b_{\mathrm{NN}(i,j)} - \widehat{b}^{(0)}_{\mathrm{NN}(i,j)}|$, we have $|b_{\mathrm{NN}(i,j)} - \widehat{b}^{(0)}_{\mathrm{NN}(i,j)}| \leq \max_{i\in[d],j\in[n]} |b_{ij} - \widehat{b}^{(0)}_{ij}|$. By the consistency (11.63) of $\widehat{B}^{(0)}$ from Step 1, we have

$$\lim_{n\to\infty} \mathbb{P}\left(|b_{\mathrm{NN}(i,j)} - \widehat{b}^{(0)}_{\mathrm{NN}(i,j)}| < \epsilon_1, \quad \forall (i,j) \in \Omega^{\mathrm{v}}\right) = 1. \tag{11.84}$$

For the third term $\widehat{x}^{(0)}_i$, by (11.61) in Theorem 11.21(b), we have

$$\lim_{n\to\infty} \mathbb{P}\Big(|\widehat{x}_i| < \epsilon_1, \quad \forall i \in [d]\Big) = 1. \tag{11.85}$$

Taking a union bound over the three terms (11.83), (11.84) and (11.85) and plugging them back to (11.82), the validation error at $\lambda = 0$ satisfies

$$\lim_{n\to\infty} \mathbb{P}\Big(e^{(0)} \leq 9\epsilon_1^2\Big) = 1. \tag{11.86}$$

Step 3: Computing the validation error at general $\lambda \in \Lambda_\epsilon$, and showing that it is greater than the validation error at $\lambda = 0$

Recall the definition $\Lambda_\epsilon := \{\lambda \in [0,\infty] : \|\widehat{x}^{(\lambda)}\|_2 > \epsilon\}$. In this step, we establish

$$\lim_{n\to\infty} (\lambda_{\text{cv}} \notin \Lambda_\epsilon) = 1.$$

By Lemma 11.7(a) combined with the assumption that $d = 2$, we have

$$\lim_{n\to\infty} \mathbb{P}\underbrace{\Big(\Big|\widehat{x}_1^{(\lambda)} - \widehat{x}_2^{(\lambda)}\Big| > \frac{\epsilon}{\sqrt{2}}, \quad \forall \lambda \in \Lambda_\epsilon \Big)}_{E} = 1. \tag{11.87}$$

We denote the the event in (11.87) as E. We define

$$\Lambda_{2>1} := \left\{\lambda \in [0,\infty] : \widehat{x}_2^{(\lambda)} - \widehat{x}_1^{(\lambda)} > \frac{\epsilon}{\sqrt{2}}\right\} \tag{11.88a}$$

$$\Lambda_{1>2} := \left\{\lambda \in [0,\infty] : \widehat{x}_1^{(\lambda)} - \widehat{x}_2^{(\lambda)} > \frac{\epsilon}{\sqrt{2}}\right\}. \tag{11.88b}$$

Then we have

$$\Lambda_\epsilon \subseteq \Lambda_{2>1} \cup \Lambda_{1>2} \;\mid\; E. \tag{11.89}$$

We first analyze $\Lambda_{2>1}$. We discuss the following two cases, depending on the comparison of the mean of the bias for the two courses.

Case 1: $\sum_{j\in[n]} b_{1j} \geq \sum_{j\in[n]} b_{2j}$

We denote the event that Case 1 happens as $E_1 := \{\sum_{j\in[n]} b_{1j} \geq \sum_{j\in[n]} b_{2j}\}$. In this case, our goal is to show

$$\lim_{n\to\infty} \mathbb{P}\Big(\lambda_{\text{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E_1\Big) = \lim_{n\to\infty} (E_1). \tag{11.90}$$

To show (11.90) it suffices to prove

$$\lim_{n\to\infty} \mathbb{P}\Big(\Lambda_\epsilon \cap \Lambda_{2>1} = \emptyset, E_1\Big) = \lim_{n\to\infty} \mathbb{P}(E_1).$$

We separately discuss the cases of $\lambda = \infty$ and $\lambda \notin \infty$.

Showing $\infty \notin \Lambda_\epsilon \cap \Lambda_{2>1}$: Denote the mean of the bias in each course in the training set Ω^{t} as $b_i^{\text{t}} := \frac{1}{n^{\text{t}}} \sum_{j\in\Omega_i^{\text{t}}} b_{ij}$ for $i \in \{1,2\}$. By (11.22a) in Lemma 11.15, we have

$$\lim_{n\to\infty} \mathbb{P}\left(b_1^{\text{t}} - \frac{1}{n}\sum_{j\in[n]} b_{1j} < -\frac{\epsilon}{8}\right) = 0 \tag{11.91a}$$

$$\lim_{n\to\infty} \mathbb{P}\left(b_2^{\text{t}} - \frac{1}{n}\sum_{j\in[n]} b_{2j} > \frac{\epsilon}{8}\right) = 0 \tag{11.91b}$$

Taking a union bound of (11.91), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\underbrace{b_1^{\mathrm{t}} - b_2^{\mathrm{t}} > \frac{1}{n}\sum_{j\in[n]}(b_{1j} - b_{2j}) - \frac{\epsilon}{4}}_{E'}\right) = 1. \tag{11.92}$$

Denote this event in (11.92) as E'. Hence, we have

$$b_1^{\mathrm{t}} - b_2^{\mathrm{t}} > -\frac{\epsilon}{4} \;\Big|\; (E', E_1) \tag{11.93}$$

Recall from Proposition 11.2 that we have our estimator at $\lambda = \infty$ equals to the sample mean per course. That is, $\widehat{x}^{(\infty)} = \begin{bmatrix} b_1^{\mathrm{t}} \\ b_2^{\mathrm{t}} \end{bmatrix}$. Hence, we have

$$\widehat{x}_2^{(\infty)} - \widehat{x}_1^{(\infty)} < \frac{\epsilon}{4} \;\Big|\; (E', E_1).$$

By the definition of $\Lambda_{2>1}$, we have

$$\infty \notin \Lambda_\epsilon \cap \Lambda_{2>1} \mid (E', E_1). \tag{11.94}$$

Showing $\lambda \notin \Lambda_\epsilon \cap \Lambda_{2>1}$ for general $\lambda \in [0, \infty)$: As an overview, we assume there exists some $\lambda \in \Lambda_\epsilon \cap \Lambda_{2>1} \setminus \{\infty\}$ and derive a contradiction.

Denote the mean of the bias in the training set Ω^{t} as $b^{\mathrm{t}} := \frac{1}{|\Omega^{\mathrm{v}}|}\sum_{(i,j)\in\Omega^{\mathrm{v}}} b_{ij} = \frac{b_1^{\mathrm{t}} + b_2^{\mathrm{t}}}{2}$. Since $\lambda \in \Lambda_{2>1}$, we have $\widehat{x}_2^{(\lambda)} - \widehat{x}_1^{(\lambda)} > \frac{\epsilon}{\sqrt{2}}$. By (11.15b) in Lemma 11.4, we have

$$\widehat{x}^{(\lambda_1)} + \widehat{x}^{(\lambda_2)} = 2b^{\mathrm{t}},$$

and hence $\widehat{x}^{(\lambda)}$ can be reparameterized as

$$\widehat{x}^{(\lambda)} = b^{\mathrm{t}} + \Delta \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \text{ for some } \Delta > \frac{\epsilon}{2\sqrt{2}}. \tag{11.95}$$

The following lemma gives a closed-form formula for ℓ_2-regularized isotonic regression. Recall that $\mathcal{M}$ denotes the monotonic cone, and the isotonic projection for any $y \in \mathbb{R}^d$ is defined in (11.8) as $\Pi_{\mathcal{M}}(y) = \operatorname{argmin}_{u\in\mathcal{M}}\|y - u\|_2^2$.

Lemma 11.26. *Consider any $y \in \mathbb{R}^d$ and any $\lambda \in [0, \infty)$. Then we have*

$$\min_{u\in\mathcal{M}}\left(\|y - u\|_2^2 + \lambda\|u\|_2^2\right) = \frac{1}{1+\lambda}\|y - \Pi_{\mathcal{M}}(y)\|_2^2 + \frac{\lambda}{1+\lambda}\|y\|_2^2. \tag{11.96}$$

The proof of this result is provided in Appendix 11.11.5. We denote the objective (11.9) under any fixed $x \in \mathbb{R}^d$ as

$$\begin{aligned} L(x) &:= \min_{B \text{ obeys } \pi} \left\|Y - x\mathbf{1}^T - B\right\|_{\Omega^{\mathrm{t}}}^2 + \lambda\|B\|_{\Omega^{\mathrm{t}}}^2 \\ &\overset{\text{(i)}}{=} \frac{1}{1+\lambda}\underbrace{\left\|(Y - x\mathbf{1}^T) - \Pi_\pi(Y - x\mathbf{1}^T)\right\|_{\Omega^{\mathrm{t}}}^2}_{L_1(x)} + \frac{\lambda}{1+\lambda}\underbrace{\left\|Y - x\mathbf{1}^T\right\|_{\Omega^{\mathrm{t}}}^2}_{L_2(x)}, \end{aligned} \tag{11.97}$$

where equality (i) is true by (11.96) in Lemma 11.26. We now construct an alternative estimate $\widehat{x}' = b^{\mathrm{t}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, and show that

$$L(\widehat{x}) > L(\widehat{x}') \qquad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1} \setminus \{\infty\}.$$

We consider the two terms $L_1(x)$ and $L_2(x)$ in (11.97) separately.

Term L_1: Recall from the model (11.59) that $Y = B$. Hence, Y satisfies the total ordering π, and hence $Y - \widehat{x}'\mathbf{1}^T = Y - b^{\mathrm{t}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mathbf{1}_n^T$ satisfies the total ordering π. That is,

$$\Pi_\pi(Y - \widehat{x}'\mathbf{1}^T) = Y - \widehat{x}'\mathbf{1}^T.$$

Hence,

$$0 = L_1(\widehat{x}') \le L_1(\widehat{x}^{(\lambda)}) \qquad \forall \lambda \in [0, \infty]. \tag{11.98}$$

Term L_2: We have

$$\begin{aligned}
L_2(\widehat{x}) - L_2(\widehat{x}') &= \|Y - \widehat{x}^{(\lambda)}\mathbf{1}^T\|_{\Omega^{\mathrm{t}}}^2 - \|Y - \widehat{x}'\mathbf{1}^T\|_{\Omega^{\mathrm{t}}}^2 \\
&= \sum_{j \in \Omega_1^{\mathrm{t}}} (b_{1j} - \widehat{x}_1^{(\lambda)})^2 + \sum_{j \in \Omega_2^{\mathrm{t}}} (b_{2j} - \widehat{x}_2^{(\lambda)})^2 - \left[\sum_{j \in \Omega_1^{\mathrm{t}}} (b_{1j} - \widehat{x}_1')^2 + \sum_{j \in \Omega_2^{\mathrm{t}}} (b_{2j} - \widehat{x}_2')^2 \right] \\
&= n^{\mathrm{t}} \left[2b_1^{\mathrm{t}}(\widehat{x}_1' - \widehat{x}_1^{(\lambda)}) + 2b_2^{\mathrm{t}}(\widehat{x}_2' - \widehat{x}_2^{(\lambda)}) + ((\widehat{x}_1^{(\lambda)})^2 - (\widehat{x}_1')^2) + ((\widehat{x}_2^{(\lambda)})^2 - (\widehat{x}_2')^2) \right] \\
&= n^{\mathrm{t}}[2\Delta(b_1^{\mathrm{t}} - b_2^{\mathrm{t}}) + 2\Delta^2] \\
&= 2n^{\mathrm{t}}\Delta(b_1^{\mathrm{t}} - b_2^{\mathrm{t}} + \Delta) \overset{\text{(i)}}{>} 0 \;\mid\; (E', E_1),
\end{aligned}$$

where inequality (i) is true by combining (11.93) with (11.95). Hence, we have

$$L_2(\widehat{x}) > L_2(\widehat{x}'), \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1} \setminus \{\infty\} \mid (E', E_1). \tag{11.99}$$

Combining the term L_1 from (11.98) and the term L_2 from (11.99), we have

$$L(\widehat{x}^{(\lambda)}) > L(\widehat{x}'), \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1} \setminus \{\infty\} \;\Big|\; (E', E_1).$$

Contradiction to the assumption that $\widehat{x}^{(\lambda)}$ is optimal. Hence, we have

$$\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1} \setminus \{\infty\} \mid (E', E_1). \tag{11.100}$$

Combining the cases of $\lambda = \infty$ from (11.94) and $\lambda \neq \infty$ from (11.100), we have

$$\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1} \mid (E', E_1).$$

Hence,

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\text{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E_1\right) &\geq \mathbb{P}(E', E_1) \\
&= \mathbb{P}(E_1) - \mathbb{P}(E_1 \cap \overline{E'}) \\
&\geq \mathbb{P}(E_1) - \mathbb{P}(\overline{E'}) \qquad (11.101)
\end{aligned}$$

Taking the limit of (11.101), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\text{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E_1\right) \overset{\text{(i)}}{=} \lim_{n\to\infty} \mathbb{P}(E_1), \qquad (11.102)$$

where (i) is true by (11.92).

Case 2: $\sum_{j\in[n]} b_{1j} < \sum_{j\in[n]} b_{2j}$

Denote the event that Case 2 happens as $E_2 := \left\{\sum_{j\in[n]} b_{1j} < \sum_{j\in[n]} b_{2j}\right\}$. Our goal is to find a set of elements on which the validation error is large. For any constant $c > 0$, we define the set:

$$S_c := \{(j, j') \in [n]^2 : 0 < b_{2j'} - b_{1j} < c\}. \qquad (11.103)$$

Let $c' > 0$ be a constant. Denote $E^{\text{v}}_{c',c}$ as the event that there exists distinct values $(j_1, \ldots, j_{c'n})$ and distinct values $(j'_1, \ldots, j'_{c'n})$, such that $(j_k, j'_k) \in S_c \cap \Omega^{\text{v}}$ for all $k \in [c'n]$. That is, the set $S_c \cap \Omega^{\text{v}}$ contains a subset of size at least $c'n$ of pairs (j, j'), such that each element b_{1j} and $b_{2j'}$ appears at most once in this subset. We denote this subset as S'.

The following lemma bounds the probability that $E^{\text{v}}_{c',c}$ happens under case E_2.

Lemma 11.27. *Suppose $d = 2$. Assume the bias is distributed according to assumption (A2) with $\sigma = 1$. For any $c > 0$, there exists a constant $c' > 0$ such that*

$$\lim_{n\to\infty} \mathbb{P}\left(E^{\text{v}}_{c',c} \cap E_2\right) = \lim_{n\to\infty} \mathbb{P}(E_2).$$

The proof of this result is provided in Appendix 11.11.6. Now consider the the validation error contributed by the pairs in the set S'. We have

$$e^{(\lambda)} \leq \frac{1}{|\Omega^{\text{v}}|} \sum_{(j,j')\in S'} \left[\left(b_{1j} - \widehat{b}^{(\lambda)}_{\text{NN}(1,j)} - \widehat{x}^{(\lambda)}_1\right)^2 + \left(b_{2j'} - \widehat{b}^{(\lambda)}_{\text{NN}(2,j')} - \widehat{x}^{(\lambda)}_2\right)^2\right]. \qquad (11.104)$$

We consider each individual term $(j, j') \in S'$. On the one hand, we have $b_{1j} < b_{2j'}$ by the definition (11.103) of S_c. Therefore, the element $(1, j)$ is ranked lower than $(2, j')$ in the total ordering $\mathcal{T}$. According to Algorithm 2, it can be verified that their interpolated bias satisfies

$$\widetilde{b}^{(\lambda)}_{\text{NN}(1,j)} \leq \widetilde{b}^{(\lambda)}_{\text{NN}(2,j')} \qquad \forall \lambda \in [0, \infty]. \qquad (11.105)$$

On the other hand, we have

$$b_{1j} - \widehat{x}_1 - (b_{2j'} - \widehat{x}_2) = (b_{1j} - b_{2j'}) + (\widehat{x}_2 - \widehat{x}_1) \overset{\text{(i)}}{>} -\frac{\epsilon}{2} + \frac{\epsilon}{\sqrt{2}} = \frac{\epsilon}{5}, \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1} \,\Big|\, (E^{\text{v}}_{c',\frac{\epsilon}{2}}, E), \qquad (11.106)$$

where (i) is true by the definition of S_c in (11.103) (setting $c = \frac{\epsilon}{2}$), and the definition 11.88 of $\Lambda_{2>1}$. Combining (11.105) and (11.106), we have that for all $(j, j') \in S'$:

$$\begin{aligned}\left(b_{1j} - \widetilde{b}^{(\lambda)}_{\mathrm{NN}(1,j)} - \widehat{x}^{(\lambda)}_1\right)^2 + \left(b_{2j'} - \widetilde{b}^{(\lambda)}_{\mathrm{NN}(2,j')} - \widehat{x}^{(\lambda)}_2\right)^2 &\geq \min_{\substack{u_1,u_2\in\mathbb{R}\\ u_1\leq u_2}} \min_{\substack{v_1,v_2\in\mathbb{R}\\ v_1-v_2>\frac{\epsilon}{5}}} (v_1-u_1)^2 + (v_2-u_2)^2\\ &> \frac{\epsilon^2}{50}, \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1} \;\Big|\; (E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}, E).\end{aligned} \tag{11.107}$$

Conditional on $E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}$, there are at least $c'n$ such non-overlapping pairs. Plugging (11.107) to (11.104), the validation error is lower-bounded as

$$e^{(\lambda)} \geq \frac{1}{|\Omega^{\mathrm{v}}|} c'n \cdot \frac{\epsilon^2}{50} \geq \frac{2}{dn} c'n \cdot \frac{\epsilon^2}{50} = \frac{c'\epsilon^2}{25d}, \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1} \;\Big|\; (E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}, E). \tag{11.108}$$

Setting the constant ϵ_1 to be a sufficiently small constant such that $9\epsilon_1^2 < \frac{c'\epsilon^2}{25d}$, we have

$$\begin{aligned}\mathbb{P}\left(e^{(\lambda)} \geq e^{(0)}, \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1}, E_2\right) &\geq \mathbb{P}\left(e^{(\lambda)} > \frac{c'\epsilon^2}{25d} > 9\epsilon_1^2 > e^{(0)}, \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1}, E_2\right)\\ &\geq \mathbb{P}\left(e^{(\lambda)} > \frac{c'\epsilon^2}{25d}, E_2\right) - \mathbb{P}\left(e^{(0)} > 9\epsilon_1^2, E_2\right)\\ &\overset{\text{(i)}}{\geq} \mathbb{P}\left(E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}, E, E_2\right) - \mathbb{P}\left(e^{(0)} > 9\epsilon_1^2\right) \qquad (11.109)\\ &= \mathbb{P}\left(E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}, E\right) - \mathbb{P}\left(E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}, E, \overline{E_2}\right) - \mathbb{P}\left(e^{(0)} > 9\epsilon_1^2\right),\end{aligned} \tag{11.110}$$

where (i) is true by (11.108). Taking the limit of $n \to \infty$ in (11.110), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} \geq e^{(0)}, \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1}, E_2\right) = \lim_{n\to\infty} \mathbb{P}(E_2).$$

and (ii) is true by combining Lemma 11.27, (11.87) and (11.86) from Step 2. Equivalently,

$$\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E_2\right) = 1. \tag{11.111}$$

Finally, combining the two cases from (11.102) and (11.111), we have

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}\right) &= \lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E_1\right) + \lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E_2\right)\\ &= \lim_{n\to\infty} \mathbb{P}(E_1) + \lim_{n\to\infty} \mathbb{P}(E_2) = 1.\end{aligned} \tag{11.112a}$$

By a symmetric argument on the set $\Lambda_{1>2}$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{1>2}\right) = 1. \tag{11.112b}$$

Hence, we have

$$
\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\text{cv}} \notin \Lambda_\epsilon\right) &\geq \lim_{n\to\infty} \mathbb{P}\left(\lambda_{\text{cv}} \notin \Lambda_\epsilon, E\right) \\
&\overset{\text{(i)}}{\geq} \lim_{n\to\infty} \mathbb{P}\left(\lambda_{\text{cv}} \notin \Lambda_\epsilon \cap \Lambda_{1>2}, E\right) + \lim_{n\to\infty} \mathbb{P}\left(\lambda_{\text{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E\right) \\
&\geq \lim_{n\to\infty} \mathbb{P}\left(\lambda_{\text{cv}} \notin \Lambda_\epsilon \cap \Lambda_{1>2}\right) + \mathbb{P}\left(\lambda_{\text{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}\right) - 2 \lim_{n\to\infty} \mathbb{P}(\overline{E}) \overset{\text{(ii)}}{=} 1,
\end{aligned}
$$

where inequality (i) is true by (11.89), and equality (ii) is true by combining (11.112) with (11.87). This completes the proof.

11.6 Proof of Theorem 4.10

The proof follows notation in Appendix 11.1 and preliminaries in Appendix 11.2. Similar to the proof of Theorem 4.9, without loss of generality we assume $x^* = 0$ and the standard deviation of the Gaussian noise is $\eta = 1$. Under this setting, the model (4.1) reduces to:

$$Y = Z. \tag{11.113}$$

The proof consists of 3 steps that are similar to the steps in Theorem 4.9. Both part (a) and part (b) share the same first two steps as follows. We fix some constants $\epsilon_1, \epsilon_2 > 0$, whose values are determined later.

Step 1: Showing the consistency of our estimator at $\lambda = \infty$ restricted to the training set Ω^{t}

By Proposition 11.2, our estimator $\widehat{x}^{(\infty)}$ at $\lambda = \infty$ is identical to taking the sample mean of each course. By the model (11.113), conditional on any training-validation split $(\Omega^{\text{t}}, \Omega^{\text{v}})$ given by Algorithm 2, each observation is i.i.d. noise of $\mathcal{N}(0, 1)$. Recall from (11.1) that the number of observations in each course restricted to the training set Ω^{t} is $n^{\text{t}} = \frac{n}{2}$. Given the assumption (A3) that the number of courses d is a constant, sample mean on the training set Ω^{t} is consistent. That is,

$$\lim_{n\to\infty} \mathbb{P}\Big(\|\widehat{x}^{(\infty)}\|_\infty < \epsilon_1\Big) = 1. \tag{11.114}$$

By Proposition 11.2, we have $\widehat{B}^{(\infty)} = 0$.

Step 2: Computing the validation error at $\lambda = \infty$

Recall from Algorithm 2 that the interpolated bias $\widetilde{b}_{ij}$ for any element $(i, j) \in \Omega^{\text{v}}$ is computed as the mean of the estimated bias $\widehat{B}$ from its nearest neighbor set in the training set Ω^{t}. Since the estimated bias is $\widehat{B}^{(\infty)} = 0$, the interpolated bias is $\widetilde{B}^{(\infty)} = 0$. Recall the model (11.113) of $Y = Z$. The validation error at $\lambda = \infty$ is computed as

$$
\begin{aligned}
e^{(\infty)} = \frac{1}{|\Omega^{\text{v}}|} \sum_{(i,j)\in\Omega^{\text{v}}} \left(y_{ij} - \widehat{x}_i^{(\infty)} - \widetilde{b}_{ij}^{(\infty)}\right)^2 &= \frac{1}{|\Omega^{\text{v}}|} \sum_{(i,j)\in\Omega^{\text{v}}} \left(z_{ij} - \widehat{x}_i^{(\infty)}\right)^2 \\
&= \frac{1}{|\Omega^{\text{v}}|} \Bigg[\underbrace{\sum_{(i,j)\in\Omega^{\text{v}}} {z_{ij}}^2}_{T_1} - 2 \underbrace{\sum_{(i,j)\in\Omega^{\text{v}}} z_{ij}\widehat{x}_i^{(\infty)}}_{T_2} + \underbrace{\sum_{(i,j)\in\Omega^{\text{v}}} (\widehat{x}_i^{(\infty)})^2}_{T_3} \Bigg].
\end{aligned}
\tag{11.115}
$$

We consider the three terms T_1, T_2 and T_3 in (11.115) separately. For the term T_1, we have $\mathbb{E}[z_{ij}^2] = \eta^2 = 1$. The number of samples is $|\Omega^{\text{v}}| = dn^{\text{v}} = d\frac{n}{2}$. By Hoeffding's inequality we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\text{v}}|}\sum_{(i,j)\in\Omega^{\text{v}}} z_{ij}^2 < 1 + \epsilon_1\right) = 1. \tag{11.116}$$

For the term T_2, we have $\mathbb{E}[z_{ij}] = 0$. By Hoeffding's inequality and a union bound over $i \in [d]$ we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\text{v}}|}\left|\sum_{j\in\Omega_i^{\text{v}}} z_{ij}\right| < \epsilon_1, \quad \forall i \in [d]\right) = 1. \tag{11.117}$$

Combining (11.117) with the consistency result (11.114) on $\widehat{x}^{(\infty)}$ from Step 1, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\text{v}}|}|T_2| < d\epsilon_1^2\right) = 1. \tag{11.118}$$

For the term T_3, we have

$$\frac{1}{|\Omega^{\text{v}}|}T_3 \leq \max_{i\in[d]}|\widehat{x}_i|^2. \tag{11.119}$$

Combining (11.119) with the consistency result (11.114) on $\widehat{x}^{(\infty)}$ from Step 1, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\text{v}}|}T_3 < \epsilon_1^2\right) = 1. \tag{11.120}$$

Taking a union bound of the terms T_1, T_2 and T_3 from (11.116), (11.118) and (11.120) and plugging them back to (11.115), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\infty)} \leq (1+\epsilon_1) + d\epsilon_1^2 + \epsilon_1^2 = 1 + \epsilon_1 + (d+1)\epsilon_1^2\right) = 1. \tag{11.121}$$

Step 3 (preliminaries): Computing the validation error at general $\lambda \in \Lambda_\epsilon$, and showing that it is greater than the validation error at $\lambda = \infty$

We set up some preliminaries for this step that are shared between part (a) and part (b). Then we discuss the two parts separately.

Recall from (11.10) the definition of $\Lambda_\epsilon := \{\lambda \in [0,\infty] : \|\widehat{x}^{(\lambda)}\|_2 > \epsilon\}$. In this step, we show that

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} > e^{(\infty)}, \quad \forall \lambda \in \Lambda_\epsilon\right) = 1. \tag{11.122}$$

Then from (11.122) we have

$$\lim_{n\to\infty} (\lambda_{\text{cv}} \notin \Lambda_\epsilon) = 1,$$

yielding the result of Theorem 4.10. It is sufficient to establish (11.122).

We now give some additional preliminary results for this step. By Lemma 11.7, we have

$$\lim_{n\to\infty} \underbrace{\mathbb{P}\left(\max_{i,i'\in[d]} \widehat{x}_i - \widehat{x}_{i'} > \frac{\epsilon}{\sqrt{d}}, \quad \forall \lambda \in \Lambda_\epsilon\right)}_{E} = 1. \tag{11.123}$$

We denote this event in (11.123) as E.

Both parts also use the following lemma that bounds the magnitude of the estimated bias $\widehat{B}$ given some value of $\widehat{x}$.

Lemma 11.28. *Let $\Omega \subseteq [d] \times [n]$ be any non-empty set. For any $\lambda \in [0, \infty]$, the solution $(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$ restricted to the set Ω satisfies the deterministic relation*

$$\max_{(i,j)\in\Omega} \left|\widehat{b}_{ij}^{(\lambda)}\right| \leq \max_{(i,j)\in\Omega} |y_{ij}| + \|\widehat{x}^{(\lambda)}\|_\infty. \tag{11.124}$$

The proof of this result is provided in Appendix 11.12.1. Now we proceed differently for Step 3 for part (a) and part (b).

11.6.1 Proof of part (a)

Step 3 (continued): For clarity of notation, we denote the constant in the single constant-fraction as c_{f}.

We analyze the validation error at any $\lambda \in \Lambda_\epsilon$ similar to Step 2. The difference is that Step 2 (at $\lambda = \infty$) uses the consistency of $\widehat{x}^{(\infty)}$ from Step 1 on to bound the validation error. However, $\widehat{x}^{(\lambda)}$ may not be consistent for any general $\lambda \in \Lambda_\epsilon$. Hence, we consider the following two subsets of Λ_ϵ depending on the value of $\widehat{x}$.

Similar to the proof of Theorem 4.9(a), by Algorithm 2 the interpolated bias for elements in each group $k \in [r]$ is identical for all $(i, j) \in G_k^{\mathrm{v}}$. That is,

$$\widetilde{b}_{ij} = \widetilde{b}_{i'j'} \qquad \forall (i, j), (i', j') \in G_k^{\mathrm{v}}. \tag{11.125}$$

We denote the interpolated bias for group k as $\widetilde{b}_k := \widetilde{b}_{ij}$ for $(i, j) \in G_k^{\mathrm{v}}$.

Case 1: $\Lambda_1 := \left\{\lambda \in [0, \infty] : \max_{i,i'\in[d]} \widehat{x}_i - \widehat{x}_{i'} > 8\sqrt{\frac{d}{c_{\mathrm{f}}}}\right\}$.

Let $k_{\mathrm{f}} \in [r]$ be a group that satisfies the single c_{f}-fraction assumption. By the definition of Λ_1 we have $\max_{i,i'\in[d]} \left[(\widehat{x}_i + \widetilde{b}_{k_{\mathrm{f}}}) - (\widehat{x}_{i'} + \widetilde{b}_{k_{\mathrm{f}}})\right] > 8\sqrt{\frac{d}{c_{\mathrm{f}}}}$ for any $\lambda \in \Lambda_1$, which implies that

$$\max_{i\in[d]} \left|\widehat{x}_i + \widetilde{b}_{k_{\mathrm{f}}}\right| > 4\sqrt{\frac{d}{c_{\mathrm{f}}}} \qquad \forall \lambda \in \Lambda_1. \tag{11.126}$$

Combining (11.20a) from Lemma 11.13 with the single c_{f}-fraction assumption, one can see

$$\ell_{ik_{\mathrm{f}}}^{\mathrm{v}} \geq \frac{\ell_{ik_{\mathrm{f}}}}{4} > \frac{c_{\mathrm{f}} n}{4}. \tag{11.127}$$

Given (11.127), by Hoeffding's inequality we have

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{j\in G^{\mathrm{v}}_{ik_{\mathrm{f}}}} \mathbb{1}\{z_{ij} > 0\} \geq \frac{c_{\mathrm{f}} n}{12}\right) = 1 \tag{11.128a}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{j\in G^{\mathrm{v}}_{ik_{\mathrm{f}}}} \mathbb{1}\{z_{ij} < 0\} \geq \frac{c_{\mathrm{f}} n}{12}\right) = 1. \tag{11.128b}$$

We denote the event

$$E_1 := \left\{\sum_{j\in G^{\mathrm{v}}_{ik_{\mathrm{f}}}} \mathbb{1}\{z_{ij} > 0\} \geq \frac{c_{\mathrm{f}} n}{12}, \quad \forall i \in [d]\right\} \cap \left\{\sum_{j\in G^{\mathrm{v}}_{ik_{\mathrm{f}}}} \mathbb{1}\{z_{ij} < 0\} \geq \frac{c_{\mathrm{f}} n}{12}, \quad \forall i \in [d]\right\}. \tag{11.129}$$

Given that d is a constant by the assumption (A3), taking (11.128) with a union bound over $i \in [d]$, we have

$$\lim_{n\to\infty} \mathbb{P}(E_1) = 1. \tag{11.130}$$

Let i^* be a random variable (as a function of λ) defined as $i^* := \mathrm{argmax}_{i\in[d]}\left|\widehat{x}_i + \widetilde{b}_{k_{\mathrm{f}}}\right|$ where the tie is broken arbitrarily. Conditional on E_1, for any $\lambda \in \Lambda_1$ we have the deterministic relation

$$\begin{aligned}
e^{(\lambda)} = \frac{1}{|\Omega^{\mathrm{v}}|}\sum_{k\in[r]}\sum_{(i,j)\in G^{\mathrm{v}}_k}\left(z_{ij} - \widehat{x}^{(\lambda)}_i - \widetilde{b}^{(\lambda)}_k\right)^2 &\geq \frac{1}{|\Omega^{\mathrm{v}}|}\sum_{(i,j)\in G^{\mathrm{v}}_{k_{\mathrm{f}}}} (z_{ij} - \widehat{x}_i - \widetilde{b}_{k_{\mathrm{f}}})^2 \\
&\geq \frac{1}{|\Omega^{\mathrm{v}}|}\sum_{j\in G^{\mathrm{v}}_{i^*k_{\mathrm{f}}}} (z_{i^*j} - \widehat{x}_{i^*} - \widetilde{b}_{k_{\mathrm{f}}})^2 \\
&\overset{\text{(i)}}{\geq} \frac{1}{|\Omega^{\mathrm{v}}|}\frac{c_{\mathrm{f}} n}{12}\left(4\sqrt{\frac{d}{c_{\mathrm{f}}}}\right)^2 \\
&= \frac{2}{dn}\cdot\frac{c_{\mathrm{f}} n}{12}\frac{16d}{c_{\mathrm{f}}} = \frac{8}{3}, \quad \forall\lambda \in \Lambda_1 \;\Big|\; E_1.
\end{aligned} \tag{11.131}$$

where (i) is true by (11.126) and the definition (11.129) of E_1. Combining (11.131) with (11.130), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} \geq \frac{4}{3}, \quad \forall\lambda \in \Lambda_1\right) \geq \mathbb{P}\left(E_1\right) = 1. \tag{11.132}$$

Case 2: $\Lambda_2 = \Lambda_\epsilon \cap \left\{\lambda \in [0,\infty] : \max_{i,i'\in[d]} \widehat{x}_i - \widehat{x}_{i'} \leq 8\sqrt{\frac{d}{c_{\mathrm{f}}}}\right\}$.

Note that we have $\Lambda_\epsilon \subseteq \Lambda_1 \cup \Lambda_2$ by the definition of Λ_1 and Λ_2. We decompose the validation error as:

$$\begin{aligned}
e^{(\lambda)} &= \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{k\in[r]} \sum_{(i,j)\in G_k^{\mathrm{v}}} \left(z_{ij} - \widehat{x}_i^{(\lambda)} - \widetilde{b}_k^{(\lambda)} \right)^2 \\
&= \frac{1}{|\Omega^{\mathrm{v}}|} \left[\sum_{(i,j)\in\Omega^{\mathrm{v}}} z_{ij}^2 - 2 \sum_{k\in[r]} \sum_{(i,j)\in G_k^{\mathrm{v}}} z_{ij} \left(\widehat{x}_i^{(\lambda)} + \widetilde{b}_k^{(\lambda)} \right) + \sum_{k\in[r]} \sum_{(i,j)\in G_k^{\mathrm{v}}} \left(\widehat{x}_i^{(\lambda)} + \widetilde{b}_k^{(\lambda)} \right)^2 \right] \\
&= \frac{1}{|\Omega^{\mathrm{v}}|} \left[\underbrace{\sum_{(i,j)\in\Omega^{\mathrm{v}}} z_{ij}^2}_{T_1} \underbrace{-2 \sum_{(i,j)\in\Omega^{\mathrm{v}}} z_{ij} \widehat{x}_i^{(\lambda)}}_{T_2} \underbrace{+2 \sum_{k\in[r]} \sum_{(i,j)\in G_k^{\mathrm{v}}} z_{ij} \widetilde{b}_k^{(\lambda)}}_{T_3} \underbrace{+ \sum_{k\in[r]} \sum_{(i,j)\in G_k^{\mathrm{v}}} \left(\widehat{x}_i^{(\lambda)} + \widetilde{b}_k^{(\lambda)} \right)^2}_{T_4} \right].
\end{aligned} \tag{11.133}$$

We analyze the four terms T_1, T_2, T_3 and T_4 in (11.133) separately.

Term T_1: Similar to (11.116) from Step 2, by Hoeffding's inequality we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in\Omega^{\mathrm{v}}} z_{ij}^2 > 1 - \epsilon_2 \right) = 1. \tag{11.134}$$

Term T_2: Recall that d is a constant by the assumption (A3). Similar to (11.117) from Step 2, by Hoeffding with a union bound over $i \in [d]$, we have

$$\lim_{n\to\infty} \underbrace{\mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|} \left| \sum_{j\in\Omega_i^{\mathrm{v}}} z_{ij} \right| < \epsilon, \quad \forall i \in [d] \right)}_{E_2} = 1. \tag{11.135}$$

Denote this event in (11.135) as E_2.

We now bound $\|\widehat{x}\|_\infty$. By Hoeffding's inequality, on the training Ω^{t} we have:

$$\lim_{n\to\infty} \underbrace{\mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{t}}|} \left| \sum_{(i,j)\in\Omega^{\mathrm{t}}} z_{ij} \right| < \sqrt{\frac{1}{dc_{\mathrm{f}}}} \right)}_{E_2'} = 1. \tag{11.136}$$

Plugging (11.15b) in Lemma 11.4 to (11.136), we have

$$\left| \sum_{i\in[d]} \widehat{x}_i^{(\lambda)} \right| = \frac{1}{n^{\mathrm{t}}} \left| \sum_{(i,j)\in\Omega^{\mathrm{t}}} z_{ij} \right| < \sqrt{\frac{d}{c_{\mathrm{f}}}} \qquad \forall \lambda \in \Lambda_2, \quad \text{conditional on } E_2'. \tag{11.137}$$

Combining (11.137) with the definition of Λ_2, we have

$$\|\widehat{x}\|_\infty \leq 8\sqrt{\frac{d}{c_\mathrm{f}}} \qquad \forall \lambda \in \Lambda_2 \,\Bigg|\, E_2'. \tag{11.138}$$

To see (11.138), assume for contradiction that (11.138) does not hold. Consider the case of $\widehat{x}_{i^*} > 8\sqrt{\frac{d}{c_\mathrm{f}}}$ for some $i^* \in [d]$. Then by the definition of Λ_2, we have $\widehat{x}_i > 0$ for all $i \in [d]$. Then we have $\left|\sum_{i\in[d]} \widehat{x}_i\right| > 8\sqrt{\frac{d}{c_\mathrm{f}}}$. Contradiction to (11.137). A similar argument applies if $\widehat{x}_{i^*} < -8\sqrt{\frac{d}{c_\mathrm{f}}}$. Hence, (11.138) holds.

Finally, combining (11.138) with (11.135), we have:

$$\frac{1}{|\Omega^\mathrm{v}|}|T_2| = \frac{1}{|\Omega^\mathrm{v}|}\left|\sum_{(i,j)\in\Omega^\mathrm{v}} z_{ij}\widehat{x}_i\right| \tag{11.139}$$

$$\leq \frac{d}{|\Omega^\mathrm{v}|}\max_{i\in[d]}\left|\sum_{(i,j)\in\Omega^\mathrm{v}} z_{ij}\right| \cdot \|\widehat{x}\|_\infty < 8d\sqrt{\frac{d}{c_\mathrm{f}}}\epsilon_2 \qquad \forall \lambda \in \Lambda_2, \quad \text{conditional on } (E_2, E_2'). \tag{11.140}$$

Hence, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^\mathrm{v}|}|T_2| < 8d\sqrt{\frac{d}{c_\mathrm{f}}}\epsilon, \quad \forall \lambda \in \Lambda_2\right) \geq \lim_{n\to\infty} \mathbb{P}\left(E_2 \cap E_2'\right) \overset{\text{(i)}}{=} 1,$$

where (i) is true by (11.135) and (11.136).

Term T_3: We use the following standard result derived from statistics.

Lemma 11.29. *Consider any fixed $d \geq 1$. Let $Z \sim \mathcal{N}(0, I_d)$. Then we have*

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{\substack{\|\theta\|_2=1\\ \theta_1\leq\cdots\leq\theta_d}} \theta^T Z \leq d^{\frac{1}{4}}\right) = 1.$$

For completeness, the proof of this lemma is in Appendix 11.12.2. We now explain how to apply Lemma 11.29 on $\widetilde{B}_{\Omega^\mathrm{t}}$.

The ordering of $\widetilde{B}$: Take any arbitrary total ordering $\pi \in \mathcal{T}$ that is consistent with the partial ordering $\mathcal{O}$. Recall from (11.125) that the interpolated bias within each group $k \in [r]$ is identical, so $\widetilde{B}$ satisfies the total ordering π.

Bounding $\|\widetilde{B}\|_{\Omega^{\mathrm{t}}}$: We bound each $\widetilde{b}_k$. Recall that each $\widetilde{b}_k$ is a mean of $\widehat{B}$ on its nearest-neighbor set. Hence, we have

$$\max_{k\in[r]}|\widetilde{b}_k| \leq \max_{(i,j)\in\Omega^{\mathrm{t}}}\left|\widehat{b}_{ij}^{(\lambda)}\right| \overset{\text{(i)}}{\leq} \max_{(i,j)\in\Omega^{\mathrm{t}}}|y_{ij}| + \|\widehat{x}^{(\lambda)}\|_\infty \qquad \forall\lambda\in[0,\infty], \tag{11.141}$$

where (i) is true by (11.124) in Lemma 11.28. We consider the term $\max_{(i,j)\in\Omega^{\mathrm{v}}}|y_{ij}|$ on the RHS of (11.141). Recall from the model (11.113) that $Y = Z$. Hence, we have

$$\lim_{n\to\infty}\underbrace{\mathbb{P}\left(\max_{(i,j)\in\Omega^{\mathrm{v}}}|y_{ij}| < 2\sqrt{\log dn}\right)}_{E_2''} \overset{\text{(i)}}{=} 1, \tag{11.142}$$

where (i) is true by Lemma 11.12. Plugging (11.142) and the bound on $\|\widehat{x}\|_\infty$ from (11.138) to (11.141), we have that conditional on E_2'' and E_2',

$$\begin{aligned}\max_{k\in[r]}|\widetilde{b}_k| &\leq \max_{(i,j)\in\Omega^{\mathrm{t}}}|y_{ij}| + \|\widehat{x}^{(\lambda)}\|_\infty \\ &\leq 2\sqrt{\log dn} + 8\sqrt{\frac{d}{c_{\mathrm{f}}}} \qquad \forall\lambda\in\Lambda_2 \;\Bigg|\; (E_2', E_2'').\end{aligned}$$

Hence, we have

$$\|\widetilde{B}\|_{\Omega^{\mathrm{t}}} \leq \sqrt{|\Omega^{\mathrm{t}}|}\cdot\max_{k\in[r]}\left|\widetilde{b}_k\right| \leq \sqrt{dn^{\mathrm{v}}}\left(2\sqrt{\log dn} + 8\sqrt{\frac{d}{c_{\mathrm{f}}}}\right) \qquad \forall\lambda\in\Lambda_2 \;\Bigg|\; (E_2', E_2'').$$

and therefore

$$\lim_{n\to\infty}\mathbb{P}\left(\|\widetilde{B}\|_{\Omega^{\mathrm{t}}} \leq \sqrt{dn^{\mathrm{v}}}\left(2\sqrt{\log dn} + 8\sqrt{\frac{d}{c_{\mathrm{f}}}}\right), \quad \forall\lambda\in\Lambda_2\right) \geq \lim_{n\to\infty}\mathbb{P}(E_2'\cap E_2'') = 1. \tag{11.143}$$

Applying Lemma 11.29: For the term T_3, for any constant $C > 0$, we have

$$\mathbb{P}\left(|T_3| < C(dn^{\mathrm{t}})^{\frac{1}{4}}, \quad \forall\lambda\in\Lambda_2\right) \geq \mathbb{P}\Bigg(\underbrace{\left\{\left|\frac{T_3}{C}\right| < (dn^{\mathrm{t}})^{\frac{1}{4}}, \quad \forall\lambda\in\Lambda_2\right\}}_{E_3} \cap \underbrace{\left\{\left\|\frac{\widetilde{B}}{C}\right\|_{\Omega^{\mathrm{t}}} \leq 1, \quad \forall\lambda\in\Lambda_2\right\}}_{E_4}\Bigg) \tag{11.144}$$

We have

$$\mathbb{P}(\overline{E_3\cap E_4}) = \mathbb{P}(\overline{E_4}) + \mathbb{P}(\overline{E_3}\cap E_4) \tag{11.145}$$

Setting $C = \sqrt{dn^{\mathrm{v}}}\left(2\sqrt{\log dn} + 8\sqrt{\frac{d}{c_{\mathrm{f}}}}\right)$, by (11.143) we have

$$\mathbb{P}(\overline{E_4}) = 0. \tag{11.146}$$

Applying Lemma 11.29 on $\frac{\widetilde{B}_{\Omega^{\mathrm{t}}}}{C}$, we have

$$\lim_{n\to\infty} \mathbb{P}(\overline{E_3} \cap E_4) = 0. \tag{11.147}$$

Plugging (11.146) and (11.147) to (11.145), we have

$$\lim_{n\to\infty} \mathbb{P}(\overline{E_3 \cap E_4}) = 0. \tag{11.148}$$

Combining (11.148) with (11.144), we have

$$\lim_{n\to\infty} \mathbb{P}\left(|T_3| < C(dn^{\mathrm{t}})^{\frac{1}{4}} = (dn^{\mathrm{t}})^{\frac{3}{4}} \left(2\sqrt{\log dn} + 8\sqrt{\frac{d}{c_{\mathrm{f}}}} \right), \quad \forall \lambda \in \Lambda_2 \right) = 1.$$

Hence, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|} |T_3| < \epsilon_2 \right) = 1. \tag{11.149}$$

Term T_4: Recall that k_{f} denotes a group k_{f} that satisfies the single c_{f}-fraction assumption. By the definition of E from (11.123), we have

$$\max_{i,i' \in [d]} (\widehat{x}_i + \widetilde{b}_{k_{\mathrm{f}}}) - (\widehat{x}_{i'} + \widetilde{b}_{k_{\mathrm{f}}}) > \frac{\epsilon}{\sqrt{d}} \qquad \forall \lambda \in \Lambda_2, \Bigg| \, E. \tag{11.150}$$

Therefore, we have

$$\max_{i,i' \in [d]} \left[(\widehat{x}_i + \widetilde{b}_{k_{\mathrm{f}}})^2 + (\widehat{x}_{i'} + \widetilde{b}_{k_{\mathrm{f}}})^2 \right] > \frac{\epsilon^2}{4d} \qquad \forall \lambda \in \Lambda_2 \Bigg| \, E. \tag{11.151}$$

We bound the term T_4 as

$$\frac{1}{|\Omega^{\mathrm{v}}|} T_4 \geq \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j) \in G^{\mathrm{v}}_{k_{\mathrm{f}}}} (\widehat{x}_i + \widetilde{b}_{k_{\mathrm{f}}})^2 \overset{\text{(i)}}{\geq} \frac{2}{dn} \cdot \frac{c_{\mathrm{f}} n}{4} \cdot \frac{\epsilon^2}{4d} = \frac{c_{\mathrm{f}} \epsilon^2}{8d^2} \qquad \forall \lambda \in \Lambda_2 \Bigg| \, E,$$

where (i) is true by combining (11.127) and (11.151). Hence,

$$\mathbb{P}\left(T_4 \geq \frac{c_{\mathrm{f}} \epsilon^2}{8d^2} \quad \forall \lambda \in \Lambda_2 \right) \geq \mathbb{P}(E) = 1. \tag{11.152}$$

Putting things together: Plugging the four terms from (11.134), (11.135), (11.149) and (11.152) respectively back to (11.133), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} > (1 - \epsilon_2) + 8d\sqrt{\frac{d}{c_{\mathrm{f}}}} \epsilon_2 + \epsilon_2 + \frac{c_{\mathrm{f}} \epsilon^2}{8d^2}, \quad \forall \lambda \in \Lambda_2 \right) = 1. \tag{11.153}$$

Finally, combining the two cases from (11.132) and (11.153), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} \geq \frac{8}{3} \wedge \left(1 + 16d\sqrt{\frac{d}{c_{\mathrm{f}}}}\epsilon_2 + \frac{c_{\mathrm{f}}\epsilon^2}{8d^2}\right), \quad \forall \lambda \in \Lambda_\epsilon\right) = 1. \tag{11.154}$$

Recall from (11.121) that the validation error at $\lambda = \infty$ is bounded as

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\infty)} \leq 1 + \epsilon_1 + (d+1)\epsilon_1^2\right) = 1. \tag{11.155}$$

Combining (11.154) and (11.155) with choices of (ϵ_1, ϵ_2) (dependent on $\epsilon, d, c_{\mathrm{f}}$) such that $\frac{8}{3} \wedge \left(1 + 16d\sqrt{\frac{d}{c_{\mathrm{f}}}}\epsilon_2 + \frac{c_{\mathrm{f}}\epsilon^2}{8d^2}\right) > 1 + \epsilon_1 + (d+1)\epsilon_1^2$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\infty)} > e^{(0)}, \quad \forall \lambda \in \Lambda_\epsilon\right) = 1,$$

completing the proof.

11.6.2 Proof of part (b)

For clarity of notation, we denote the constant in the constant-fraction interleaving assumption as c_{f}. Since $\mathcal{O}$ is a total ordering, we also denote it as π.

Step 3 (continued): Combining (11.15b) with Hoeffding's inequality, we have

$$\lim_{n\to\infty} \mathbb{P}\Bigg(\underbrace{|\widehat{x}_1 + \widehat{x}_2| = \frac{1}{n^{\mathrm{t}}}\left|\sum_{(i,j)\in\Omega^{\mathrm{t}}} z_{ij}\right| < \epsilon \wedge \frac{16}{\sqrt{c_{\mathrm{f}}}}, \quad \forall \lambda \in \Lambda_\epsilon}_{E_1}\Bigg) = 1. \tag{11.156}$$

We denote this event in (11.156) as E_1.

Analyzing the number of interleaving points Let $S \subseteq [2n-1]$ denotes the interleaving points. Recall that (i_t, j_t) denotes element of rank t, and t_{ij} denotes the rank of the element (i, j). We slightly abuse the notation to say $(i, j) \in S$ if $t_{ij} \in S$, and also for other definitions of subsets of interleaving points later in the proof. Denote $S_i \subseteq S$ as the set of interleaving points in course $i \in \{1, 2\}$:

$$S_i = S \cap \{t \in [2n-1] : i_t = i\}.$$

Denote S_i^{v} as the set of interleaving points in S_i that are in the validation set:

$$S_i^{\mathrm{v}} = S_i \cap \Omega^{\mathrm{v}}.$$

We define S_{pairs} as a set of pairs of interleaving points as:

$$S_{\text{pairs}} := \{(t, t') \in [2n-1]^2 : t \in S_1^{\mathrm{v}},\ t' \in S_2^{\mathrm{v}},\ t < t'\}.$$

Define E_c as the event that there exists distinct values $(t_1, t'_1, \dots, t_{cn}, t'_{cn})$ such that $(t_k, t'_k) \in S_{\text{pairs}}$ for all $k \in [cn]$. That is, S_{pairs} includes cn distinct pairs where each interleaving point appears at most once. We define S'_{pairs} likewise as

$$S'_{\text{pairs}} := \{(t, t') \in [2n-1]^2 : t \in S_2^{\text{v}},\ t' \in S_1^{\text{v}},\ t < t'\}.$$

and define E'_c likewise.

The following lemma bounds the probability of the event $E_{\frac{1}{36}}$ and $E'_{\frac{1}{36}}$.

Lemma 11.30. *Suppose $d = 2$. Then we have*

$$\lim_{n \to \infty} \mathbb{P}\left(E_{\frac{1}{36}} \cap E'_{\frac{1}{36}}\right) = 1.$$

The proof of this result is provided in Appendix 11.12.3. Denote S^+ as the set of the half of the highest interleaving points and S^- as the set of the half of the lowest interleaving points. That is, we define

$$\begin{aligned} S^+ &:= S \cap \{t \in [2n-1] : t > \text{median}(S)\} \\ S^- &:= S \cap \{t \in [2n-1] : t < \text{median}(S)\}. \end{aligned}$$

Furthermore, for $i \in \{1, 2\}$, we define

$$\begin{aligned} S_i^{\text{v}+} &:= S^+ \cap S_i \cap \Omega^{\text{v}} \\ S_i^{\text{v}-} &:= S^- \cap S_i \cap \Omega^{\text{v}}. \end{aligned}$$

The following lemma lower-bounds the size of $S_i^{\text{v}+}$ and $S_i^{\text{v}-}$.

Lemma 11.31. *We have*

$$\lim_{n \to \infty} \mathbb{P}\underbrace{\left(|T| \geq \frac{c_{\text{f}} n}{36}, \quad \forall T \in \{S_1^{\text{v}+}, S_1^{\text{v}-}, S_2^{\text{v}+}, S_2^{\text{v}-}\}\right)}_{E_2} = 1.$$

The proof of this result is provided in Appendix 11.12.4. We denote this event in Lemma 11.31 as E_2.

Bounding the validation error Similar to part (a), we discuss the following two cases depending on the value of $\widehat{x}$.

Case 1: $\Lambda_1 = \Lambda_\epsilon \cap \left\{\lambda \in [0, \infty] : \widehat{x}_1^{(\lambda)} < -\frac{32}{\sqrt{c_{\text{f}}}}\right\}$ It can be verified that due to (11.156), we have

$$\widehat{x}_1^{(\lambda)} < -\frac{32}{\sqrt{c_{\text{f}}}} < \frac{16}{\sqrt{c_{\text{f}}}} < \widehat{x}_2^{(\lambda)} \quad \forall \lambda \in \Lambda_1 \,\Big|\, E. \tag{11.157}$$

By Hoeffding's inequality combined with Lemma 11.31, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{(i,j)\in S_1^{\mathrm{v}-}} \mathbb{1}\{z_{ij} > 0\} > \frac{c_{\mathrm{f}} n}{96}\right) = 1 \tag{11.158a}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{(i,j)\in S_2^{\mathrm{v}+}} \mathbb{1}\{z_{ij} < 0\} > \frac{c_{\mathrm{f}} n}{96}\right) = 1. \tag{11.158b}$$

Denote the event

$$E_3 := \left\{\sum_{(i,j)\in S_1^{\mathrm{v}-}} \mathbb{1}\{z_{ij} > 0\} > \frac{c_{\mathrm{f}} n}{96}\right\} \cap \left\{\sum_{(i,j)\in S_2^{\mathrm{v}+}} \mathbb{1}\{z_{ij} < 0\} > \frac{c_{\mathrm{f}} n}{96}\right\}.$$

Taking a union bound of (11.158), we have

$$\lim_{n\to\infty} \mathbb{P}(E_3) = 1. \tag{11.159}$$

We slightly abuse the notation and denote $\widetilde{b}_t$ as the value of the interpolated bias on the element of rank t. That is, we define $\widetilde{b}_t := \widetilde{b}_{i_t j_t}$. It can be verified that $\widetilde{b}_t$ is non-decreasing in t due to the nearest-neighbor interpolation in Algorithm 2. Hence, $\widetilde{b}_t \le 0$ for all $t \in S^-$ or $\widetilde{b}_t \ge 0$ for all $t \in S^+$.

First consider the case $\widetilde{b}_t \le 0$ for all $t \in S^-$. We bound the validation error at $\lambda \in \Lambda_1$ as:

$$e^{(\lambda)} \ge \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in S_1^{\mathrm{v}-}} \left(z_{ij} - \widehat{x}_1^{(\lambda)} - \widetilde{b}_{ij}^{(\lambda)}\right)^2 \tag{11.160}$$

$$\overset{\text{(i)}}{\ge} \frac{1}{|\Omega^{\mathrm{v}}|} \cdot \left|S_1^{\mathrm{v}-}\right| \cdot \left(0 + \frac{16}{\sqrt{c_{\mathrm{f}}}} + 0\right)^2 \overset{\text{(i)}}{\ge} \frac{1}{n} \frac{c_{\mathrm{f}} n}{96} \frac{256}{c_{\mathrm{f}}} = \frac{8}{3}, \quad \forall \lambda \in \Lambda_1 \;\Bigg|\; (E_1, E_2, E_3), \tag{11.161}$$

where (i) is true by (11.157) and the definition of E_3, and (ii) is true by the definition of E_2. Hence, we have

$$\lim_{n\to\infty} \left(e^{(\lambda)} \ge \frac{8}{3} \quad \forall \lambda \in \Lambda_1, \{\widetilde{b}_t \le 0 \text{ for all } t \in S^-\}\right) \overset{\text{(i)}}{\ge} \mathbb{P}\left(\widetilde{b}_t \le 0 \text{ for all } t \in S^-\right), \tag{11.162a}$$

where (i) is true by (11.156), Lemma 11.31 and (11.159). By a similar argument, we have

$$\lim_{n\to\infty} \left(e^{(\lambda)} \ge \frac{8}{3} \quad \forall \lambda \in \Lambda_1, \{\widetilde{b}_t \ge 0 \text{ for all } t \in S^+\}\right) \ge \mathbb{P}\left(\widetilde{b}_t \ge 0 \text{ for all } t \in S^+\right), \tag{11.162b}$$

Summing over (11.162), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} \ge \frac{8}{3}, \quad \forall \lambda \in \Lambda_1\right) = 1. \tag{11.163}$$

Case 2: $\Lambda_2 = \Lambda_\epsilon \cap \left\{\lambda \in [0,\infty] : \widehat{x}_1^{(\lambda)} > -\frac{32}{\sqrt{c_\mathrm{f}}}\right\}$ It can be verified that due to (11.156), we have

$$-\frac{32}{\sqrt{c_\mathrm{f}}} < \{\widehat{x}_1, \widehat{x}_2\} < \frac{48}{\sqrt{c_\mathrm{f}}}. \tag{11.164}$$

Similar to Case 2 in part (a), we decompose the validation error at $\lambda \in \Lambda_2$ as

$$\begin{aligned} e^{(\lambda)} &= \frac{1}{|\Omega^\mathrm{v}|} \sum_{(i,j)\in\Omega^\mathrm{v}} \left(z_{ij} - \widehat{x}_i^{(\lambda)} - \widetilde{b}_{ij}^{(\lambda)}\right)^2 \\ &= \frac{1}{|\Omega^\mathrm{v}|}\left[\underbrace{\sum_{(i,j)\in\Omega^\mathrm{v}} z_{ij}^2}_{T_1} \underbrace{-2\sum_{(i,j)\in\Omega^\mathrm{v}} z_{ij}\widehat{x}_i^{(\lambda)}}_{T_2} \underbrace{-2\sum_{(i,j)} z_{ij}\widetilde{b}_{ij}^{(\lambda)}}_{T_3} + \underbrace{\sum_{(i,j)}\left(\widehat{x}_i^{(\lambda)} + \widetilde{b}_{ij}^{(\lambda)}\right)^2}_{T_4}\right]. \end{aligned}$$

Given that $\|\widehat{x}\|_\infty$ is bounded by a constant by (11.164), the analysis of the terms T_1, T_2 and T_3 follows the proof in part (a). We have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^\mathrm{v}|}T_1 > 1 - \epsilon_2\right) = 1. \tag{11.165a}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^\mathrm{v}|}\sum_{(i,j)\in\Omega^\mathrm{v}} |T_2| < \frac{96}{\sqrt{c_\mathrm{f}}}\epsilon_2\right) = 1. \tag{11.165b}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^\mathrm{v}|}|T_3| < \epsilon_2\right) = 1. \tag{11.165c}$$

Now we consider the last term T_4. Recall from (11.123) that

$$|\widehat{x}_2 - \widehat{x}_1| > \frac{\epsilon}{\sqrt{2}} \quad \forall\lambda \in \Lambda_2 \,\Big|\, E.$$

First consider the case of $\Lambda_{2>1} := \left\{\lambda \in [0,\infty] : \widehat{x}_2^{(\lambda)} - \widehat{x}_1^{(\lambda)} > \frac{\epsilon}{\sqrt{2}}\right\}$. Consider any $(t, t') \in S_\mathrm{pairs}$. By the definition of S_pairs we have $t < t'$. Hence, we have $\widetilde{b}_t \leq \widetilde{b}_{t'}$ due to the nearest-neighbor interpolation in Algorithm 2. Hence, we have $\widehat{x}_2 + \widetilde{b}_{t'} - (\widehat{x}_1 + \widetilde{b}_t) > \frac{\epsilon}{\sqrt{2}}$ and consequently

$$(\widehat{x}_1 + \widetilde{b}_t)^2 + (\widehat{x}_2 + \widetilde{b}_{t'})^2 > \frac{\epsilon^2}{8} \quad \forall\lambda \in \Lambda_2 \cap \Lambda_{2>1} \,\Big|\, E.$$

We bound the term T_4 as:

$$\begin{aligned} \frac{1}{|\Omega^\mathrm{v}|}T_4 &\geq \frac{1}{|\Omega^\mathrm{v}|}\sum_{(t,t')\in S_\mathrm{pairs}} \left[(\widehat{x}_1 + \widetilde{b}_t)^2 + (\widehat{x}_2 + \widetilde{b}_{t'})^2\right] \\ &\overset{\text{(i)}}{\geq} \frac{1}{2n}\cdot\frac{c_\mathrm{f} n}{36}\cdot\frac{\epsilon^2}{8} = \frac{c_\mathrm{f}\epsilon^2}{576} \qquad \forall\lambda \in \Lambda_2 \cap \Lambda_{2>1} \,\Big|\, (E_{\frac{1}{36}}, E), \end{aligned} \tag{11.166a}$$

where inequality (i) is true by the definition of $E_{\frac{1}{36}}$. Define $\Lambda_{1>2} := \left\{\lambda \in [0,\infty] : \widehat{x}_1^{(\lambda)} - \widehat{x}_2^{(\lambda)} > \frac{\epsilon}{\sqrt{2}}\right\}$. With a similar argument, we have

$$\frac{1}{|\Omega^{\mathrm{v}}|}T_4 \geq \frac{c_{\mathrm{f}}\epsilon^2}{576}, \quad \forall \lambda \in \Lambda_2 \cap \Lambda_{1>2} \,\Big|\, (E'_{\frac{1}{36}}, E). \tag{11.166b}$$

Combining (11.166), we have

$$\frac{1}{|\Omega^{\mathrm{v}}|}T_4 \geq \frac{c_{\mathrm{f}}\epsilon^2}{576}, \quad \forall \lambda \in \Lambda_2 \,\Big|\, (E_{\frac{1}{36}}, E'_{\frac{1}{36}}, E).$$

By Lemma 11.30 and (11.123), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|}T_4 \geq \frac{c_{\mathrm{f}}\epsilon^2}{576}, \quad \forall \lambda \in \Lambda_2\right) \geq \lim_{n\to\infty}\mathbb{P}\left(E_{\frac{1}{36}}, E'_{\frac{1}{36}}, E\right) = 1. \tag{11.167}$$

Putting things together: Combining the four terms from (11.165) and (11.167), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} > 1 - \epsilon_2 - \frac{128}{\sqrt{c_{\mathrm{f}}}}\epsilon_2 - 2\epsilon_2 + \frac{c_{\mathrm{f}}\epsilon^2}{576} = 1 - \left(3 + \frac{128}{\sqrt{c_{\mathrm{f}}}}\right)\epsilon_2 + \frac{c_{\mathrm{f}}\epsilon^2}{576}, \quad \forall \lambda \in \Lambda_2\right) = 1. \tag{11.168}$$

Combining the two cases from (11.163) and (11.168), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} > \frac{8}{3} \wedge \left[1 - \left(3 + \frac{128}{\sqrt{c_{\mathrm{f}}}}\right)\epsilon_2 + \frac{c_{\mathrm{f}}\epsilon^2}{576}\right], \quad \forall \lambda \in \Lambda_2\right) = 1. \tag{11.169}$$

Recall from (11.121) that the validation error at $\lambda = \infty$ is bounded as (taking $d = 2$):

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\infty)} \leq 1 + \epsilon_1 + 3\epsilon_1^2, \quad \forall \lambda \in \Lambda_\epsilon\right) = 1. \tag{11.170}$$

Combining (11.169) and (11.170) with choices of (ϵ_1, ϵ_2) (dependent on ϵ, c_{f}) such that $\frac{8}{3} \wedge \left[1 - \left(3 + \frac{128}{\sqrt{c_{\mathrm{f}}}}\right)\epsilon_2 + \frac{c_{\mathrm{f}}\epsilon^2}{576}\right] > 1 + \epsilon_1 + 3\epsilon_1^2$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\infty)} > e^{(0)}, \quad \forall \lambda \in \Lambda_\epsilon\right) = 1,$$

completing the proof.

11.7 Proof of Proposition 4.11

To prove the claimed result, we construct partial orderings that satisfy each of the conditions (a), (b), and (c) separately, and show that the mean estimator fails under each construction. Intuitively, the mean estimator does not account for any bias, so we construct partial orderings where the mean of the bias differs significantly across courses, and show that the mean estimator fails on these construction. Without loss of generality we assume that the standard deviation parameter for the Gaussian distribution of the bias is $\sigma = 1$.

11.7.1 Proof of part (a)

We first construct a partial ordering that satisfies the condition (a), and then bound the mean of each course to derive the claimed result. For clarity of notation, we denote the constant in the all constant-fraction assumption as c_{f}.

Constructing the partial ordering: Recall from Definition 4.3 that the all c_{f}-fraction assumption requires that each course $i \in [d]$ has at least $\ell_{ik} \geq c_{\text{f}} n$ students in each group $k \in [r]$. Let $c_0 = 1 - c_{\text{f}} r$. Due to the assumption that $c_{\text{f}} \in (0, \frac{1}{r})$, we have that $c_0 > 0$ is a constant. We construct the following group ordering $\mathcal{O}$, where the number of students in each course from each group is specified as

- **Course 1:** The course has $(c_{\text{f}} + c_0)n$ students from group 1, and $c_{\text{f}} n$ students from each remaining group $k \in \{2, \ldots, r\}$. That is,

$$\ell_{1k} = \begin{cases} (c_{\text{f}} + c_0)n & \text{if } k = 1 \\ c_{\text{f}} n & \text{if } 2 \leq k \leq r. \end{cases} \tag{11.171a}$$

- **Course 2:** The course has $(c_{\text{f}} + c_0)n$ students from group r, and $c_{\text{f}} n$ students from each remaining group $k \in [r-1]$. That is,

$$\ell_{2k} = \begin{cases} (c_{\text{f}} + c_0)n & \text{if } 1 \leq k \leq r-1 \\ c_{\text{f}} n. & \text{if } k = r. \end{cases} \tag{11.171b}$$

- **Course** $i \geq 3$: The course has an equal number of students from each group $k \in [r]$. That is, for every $3 \leq i \leq d$,

$$\ell_{ik} = \frac{n}{r} \qquad \forall k \in [r].$$

It can be seen that this construction of the group ordering $\mathcal{O}$ is valid, satisfying the equality $\sum_{k \in [r]} \ell_{ik} = n$ for each $i \in [d]$. Moreover, the group ordering $\mathcal{O}$ satisfies the all c_{f}-fraction assumption. Intuitively, course 1 contains more students associated with negative bias (from group 1), and course 2 contains more students associated with positive bias (from group k). The mean estimator underestimates the quality of course 1, and overestimates the quality of course 2. We construct some true qualities x^* with $x_1^* > x_2^*$, whose values are specified later in the proof.

Bounding the mean of each course: Denote the mean of the bias in any course $i \in \{1, 2\}$ of group $k \in [r]$ as $b_{ik} := \frac{1}{\ell_{ik}} \sum_{j \in G_{ik}} b_{ij}$. Similar to the proof of Lemma 11.24 (see Appendix 11.3.1 for its statement and Appendix 11.11.4 for its proof), due to assumptions (A2) and (A3) we establish the following lemma.

Lemma 11.32. *Consider any group ordering $\mathcal{O}$ that satisfies the all c_{f}-fraction assumption. For any $\epsilon > 0$, we have*

$$\lim_{n \to \infty} \mathbb{P}\Big(\underbrace{\left| b_{ik} - \bar{b}_{G_k} \right| < \epsilon, \quad \forall i \in [d], k \in [r]}_{E_1} \Big) = 1.$$

Denote this event in Lemma 11.32 as E_1. Recall that ℓ_k denotes the number of students in each group $k \in [r]$. From the construction of the group ordering $\mathcal{O}$, we have $\ell_0 := \ell_1 = \ell_r = (2c_\mathrm{f} + c_0 + \frac{d-2}{r})n$. Recall that $b^{(k)}$ denotes the k^{th} order statistics of $\{b_{ij}\}_{i\in[d],j\in[n]}$. By the assumption (A2) of the bias and the construction of the partial ordering $\mathcal{O}$, the group 1 contains the ℓ_1 lowest bias terms, $\{b^{(1)}, \ldots, b^{(\ell_0)}\}$, and the group r contains the ℓ_r highest bias terms, $\{b^{(dn-\ell_0+1)}, \ldots, b^{(dn)}\}$. Hence, we have

$$\begin{aligned}
\bar{b}_{G_1} &< \frac{b^{(\frac{\ell_0}{2})} + b^{(\ell_0)}}{2} \\
\bar{b}_{G_r} &> \frac{b^{(dn-\ell_0)} + b^{(dn-\frac{\ell_0}{2})}}{2}.
\end{aligned}$$

By the convergence of the order statistics from Lemma 11.11, it can be shown that there exists some constant $c > 0$ (dependent on d, r and c_f), such that

$$\lim_{n\to\infty} \mathbb{P}\Big(\underbrace{\bar{b}_{G_r} - \bar{b}_{G_1} > c}_{E_2}\Big) = 1. \tag{11.172}$$

Denote this event in (11.172) as E_2. The mean estimator is computed as

$$[\widehat{x}_\text{mean}]_1 = x_1^* + \frac{1}{n}\sum_{k\in[r]} \ell_{1k} b_{1k} \tag{11.173a}$$

$$[\widehat{x}_\text{mean}]_2 = x_2^* + \frac{1}{n}\sum_{k\in[r]} \ell_{2k} b_{2k} \tag{11.173b}$$

Taking the difference on (11.172), conditional on E_1 and E_2,

$$\begin{aligned}
[\widehat{x}_\text{mean}]_2 - [\widehat{x}_\text{mean}]_1 &= (x_2^* - x_1^*) + \frac{1}{n}\sum_{k\in[r]}(\ell_{2k}b_{2k} - \ell_{1k}b_{1k}) \\
&\overset{\text{(i)}}{>} (x_2^* - x_1^*) + \frac{1}{n}\sum_{k\in[r]}(\ell_{2k}\bar{b}_{G_k} - \ell_{1k}\bar{b}_{G_k}) - 2\epsilon \\
&\overset{\text{(i)}}{=} (x_2^* - x_1^*) + c_0(b_r - \bar{b}_{G_1}) - 2\epsilon \\
&\overset{\text{(iii)}}{>} (x_2^* - x_1^*) + c_0 c - 2\epsilon.
\end{aligned} \tag{11.174}$$

where inequality (i) is true by the event E_1, and equality (i) is true by plugging in the construction of the group ordering from (11.171), and inequality (iii) is true by the definition (11.172) of E_2. We set $\epsilon = \frac{c_0 c}{4}$, and set $x_1^* = \frac{c_0 c}{2}$ and $x_2^* = 0$. Then by (11.174) we have

$$\mathbb{P}([\widehat{x}_\text{mean}]_2 - [\widehat{x}_\text{mean}]_1 > 0) = 1. \tag{11.175}$$

Combining (11.175) with the fact that $x_2^* - x_1^* < 0$, completing the proof of part (a).

11.7.2 Proof of part (b)

To construct the partial ordering, we set $r = 2$ and $d = 2$ in construction we used for part (a). This completes the proof of part (b).

11.7.3 Proof of part (c)

We construct a total ordering where the bias obeys the following order (same as the "non-interleaving" total ordering described in Section 4.5.1):

$$b_{11} \leq \ldots \leq b_{1n} \leq b_{21} \leq \ldots \leq b_{2n} \leq \ldots \leq b_{d1} \leq \ldots \leq b_{dn}.$$

In this construction, course 1 contains the n students with the lowest bias, and course d contains the n students with the highest bias. Recall that $\overline{b}_i$ denotes the mean of the bias in course $i \in [d]$. We have

$$\begin{aligned}\overline{b}_1 &= \frac{1}{n}\sum_{j\in[n]} b_{1j} < \frac{b^{(\frac{n}{2})} + b^{(n)}}{2}\\ \overline{b}_r &= \frac{1}{n}\sum_{j\in[n]} b_{2j} > \frac{b^{(dn-\frac{n}{2})} + b^{(dn)}}{2}.\end{aligned}$$

Similar to part (a), by Lemma 11.11, there exists a positive constant $c > 0$ (dependent on d), such that

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{b}_r - \overline{b}_1 > c\right) = 1.$$

Let $x_1^* = c$ and $x_2^* = 0$. We have

$$\lim_{n\to\infty} \mathbb{P}([\widehat{x}_{\text{mean}}]_r - [\widehat{x}_{\text{mean}}]_1 = x_2^* - x_1^* + \overline{b}_2 - \overline{b}_1 > 0) = 1. \tag{11.176}$$

Combining (11.176) with the fact that $x_1^* > x_r^*$ completes the proof of part (c).

11.8 Proof of Proposition 4.13

By Corollary 11.6, we assume $x^* = 0$ without loss of generality. Denote the bias of course 1 as $\{U_j\}_{j\in[rn]}$ in group 1, and $\{V_j\}_{j\in[(1-r)n]}$ in group 2. Denote the bias of course 2 as $\{U_j'\}_{j\in[(1-r)n]}$ in group 1 and $\{V_j'\}_{j\in[rn]}$ in group 2. We have $U_j, U_j' \sim \text{Unif}[-1,0]$ and $V_j, V_j' \sim \text{Unif}[0,1]$. Denote the mean of $\{U_j\}, \{V_j\}, \{U_j'\}$ and $\{V_j'\}$ as $\overline{U}, \overline{V}, \overline{U}'$ and $\overline{V}'$ respectively. We prove the claimed result respectively for the reweighted mean estimator (Appendix 11.8.1) and for our estimator at $\lambda = 0$ (Appendix 11.8.2). Both parts use the following standard result regarding the uniform distribution.

Lemma 11.33. *Let $X_1, \ldots, X_n$ be i.i.d. Unif$[0,1]$, we have*

$$\mathbb{E}\left(\frac{\sum_{i=1}^n X_i}{n}\right)^2 = \frac{1}{4} + \frac{1}{12n}.$$

11.8.1 The reweighted mean estimator

We follow the definition of the reweighted mean estimator defined in Appendix 4.6.2. In the reweighting step, by (4.4) we have

$$\widehat{x}_{\text{rw}} = \frac{1}{2}\begin{bmatrix} \overline{U} + \overline{V} \\ \overline{U}' + \overline{V}' \end{bmatrix}. \tag{11.177}$$

In the recentering step, by (4.6) we have

$$\begin{aligned}
\widehat{x}_{\text{rw}} &\leftarrow \widehat{x}_{\text{rw}} + \left(-\frac{1}{2} \sum_{i\in\{1,2\}} [\widehat{x}_{\text{rw}}]_i + \frac{1}{2n} \sum_{i\in\{1,2\}, j\in[n]} y_{ij} \right) \mathbf{1} \\
&= \widehat{x}_{\text{rw}} + \left(-\frac{[\widehat{x}_{\text{rw}}]_1 + [\widehat{x}_{\text{rw}}]_2}{2} + \frac{rn\overline{U} + (1-r)n\overline{V} + (1-r)n\overline{U}' + rn\overline{V}'}{2n} \right) \mathbf{1} \\
&= \frac{[\widehat{x}_{\text{rw}}]_1 - [\widehat{x}_{\text{rw}}]_2}{2} \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \left(\frac{r\overline{U} + (1-r)\overline{V} + (1-r)\overline{U}' + r\overline{V}'}{2} \right) \mathbf{1} \\
&\overset{\text{(i)}}{=} \frac{\overline{U} + \overline{V} - \overline{U}' - \overline{V}'}{4} \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \left(\frac{r\overline{U} + (1-r)\overline{V} + (1-r)\overline{U}' + r\overline{V}'}{2} \right) \mathbf{1},
\end{aligned} \tag{11.178}$$

where equality (i) is true by plugging in (11.177) from the reweighting step. By symmetry, we have $\mathbb{E}[\widehat{x}_{\text{rw}}]_1^2 = \mathbb{E}[\widehat{x}_{\text{rw}}]_2^2$, so we only consider course 1. By (11.178), we have

$$\begin{aligned}
\mathbb{E}[\widehat{x}_{\text{rw}}]_1^2 &\overset{\text{(i)}}{=} \mathbb{E}\left(\frac{\overline{U} + \overline{V} - \overline{U}' - \overline{V}'}{4} \right)^2 + \mathbb{E}\left(\frac{r\overline{U} + (1-r)\overline{V} + (1-r)\overline{U}' + r\overline{V}'}{2} \right)^2 \\
&= \frac{1}{16}\mathbb{E}\left[\overline{U}'^2 + \overline{V}'^2 + \overline{U}^2 + \overline{V}^2 - 4 \cdot \frac{1}{2}\frac{1}{2} \right] \\
&\qquad + \frac{1}{4}\mathbb{E}\left[(1-r)^2\overline{U}'^2 + r^2\overline{V}'^2 + r^2\overline{U}^2 + (1-r)^2\overline{V}^2 - 2\left(\frac{r^2}{4} + \frac{(1-r)^2}{4} \right) \right] \\
&= \frac{1}{8}\mathbb{E}\left[\overline{U}^2 + \overline{V}^2 - \frac{1}{2} \right] + \frac{1}{2}\mathbb{E}\left[r^2\overline{U}^2 + (1-r)^2\overline{V}^2 - \frac{r^2 + (1-r)^2}{4} \right] \\
&\overset{\text{(ii)}}{=} \frac{1}{8}\left[\frac{1}{4} + \frac{1}{12rn} + \frac{1}{4} + \frac{1}{12(1-r)n} - \frac{1}{2} \right] + \frac{1}{2}\mathbb{E}\left[\frac{r^2}{4} + \frac{r^2}{12rn} + \frac{(1-r)^2}{4} + \frac{(1-r)^2}{12(1-r)n} - \frac{r^2 + (1-r)}{4} \right. \\
&= \frac{1}{96n}\left(\frac{1}{r} + \frac{1}{1-r} \right) + \frac{1}{24n} \\
&= \frac{1}{24n} + \frac{1}{96r(1-r)n}.
\end{aligned}$$

where (i) is true because it can be verified by algebra that $\mathbb{E}\left[\left(\frac{\overline{U}+\overline{V}-\overline{U}'-\overline{V}'}{4} \right) \left(\frac{r\overline{U}+(1-r)\overline{V}+(1-r)\overline{U}'+r\overline{V}'}{2} \right) \right] = 0$, and (ii) is true by Lemma 11.33. Finally, we have

$$\frac{1}{2}\mathbb{E}\|\widehat{x}_{\text{rw}}\|]_2^2 = \frac{1}{2}\left(\mathbb{E}[\widehat{x}_{\text{rw}}]_1^2] + \mathbb{E}[\widehat{x}_{\text{rw}}]_2^2 \right) = \mathbb{E}[\widehat{x}_{\text{rw}}]_1^2 = \frac{1}{24n} + \frac{1}{96r(1-r)n} \geq \frac{1}{24n} + \frac{1}{24n} = \frac{1}{12n},$$

where the inequality holds because $r(1-r) \leq \frac{1}{4}$ for every $r \in (0,1)$.

11.8.2 Our estimator at $\lambda = 0$

Recall from Proposition 11.9 that for $d = 2$ courses and $r = 2$ groups, our estimator at $\lambda = 0$ has the closed-form expression $\widehat{x}^{(0)} = \overline{y} + \begin{bmatrix} -1 \\ 1 \end{bmatrix} \cdot \frac{\gamma}{2}$, where

$$\gamma = \begin{cases} y_{22,\min} - y_{11,\max} & \text{if } y_{22,\min} - y_{11,\max} < \overline{y}_2 - \overline{y}_1 \\ y_{21,\max} - y_{12,\min} & \text{if } y_{21,\max} - y_{12,\min} > \overline{y}_2 - \overline{y}_1 \\ \overline{y}_2 - \overline{y}_1 & \text{o.w.} \end{cases} \tag{11.179}$$

By (11.179), we have

$$\frac{1}{2}\mathbb{E}\|\widehat{x}^{(0)}\|_2^2 = \frac{1}{2}\mathbb{E}\left[\left(\overline{y} - \frac{\gamma}{2}\right)^2 + \left(\overline{y} + \frac{\gamma}{2}\right)^2\right] = \mathbb{E}[\overline{y}^2] + \frac{1}{4}\mathbb{E}[\gamma^2]. \tag{11.180}$$

We analyze the two terms in (11.180) separately.

Term of $\mathbb{E}[\overline{y}^2]$: For ease of notation, we denote the random variables

$$\begin{aligned} \{\widetilde{U}_j\}_{j\in[n]} &:= \{U_j\}_{j\in[rn]} \cup \{U_j'\}_{j\in[(1-r)n]} \\ \{\widetilde{V}_j\}_{j\in[n]} &:= \{V_j\}_{j\in[(1-r)n]} \cup \{V_j'\}_{j\in[rn]} \end{aligned}$$

Then $\{\widetilde{U}_j\}_{j\in[n]}$ is i.i.d. Unif$[-1, 0]$ and $\{\widetilde{V}_j\}_{j\in[n]}$ is i.i.d. Unif$[0, 1]$. We have

$$\begin{aligned} \mathbb{E}[\overline{y}^2] &= \mathbb{E}\left(\frac{\sum_{i\in[n]}\widetilde{U}_i + \sum_{i\in[n]}\widetilde{V}_i}{2n}\right)^2 \\ &= \frac{1}{4n^2}\mathbb{E}\left[\sum_{i\in[n]}\widetilde{U}_i^2 + \sum_{i\in[n]}\widetilde{V}_i^2 + 2\sum_{i\in[n],j\in[n]}\widetilde{U}_i\widetilde{V}_j + \sum_{i\in[n]}\sum_{j\neq i}\widetilde{U}_i\widetilde{U}_j + \sum_{i\in[n]}\sum_{j\neq i}\widetilde{V}_i\widetilde{V}_j\right] \\ &= \frac{1}{4n^2}\left[\frac{n}{3} + \frac{n}{3} + 2n^2\left(-\frac{1}{4}\right) + n(n-1)\frac{1}{4} + n(n-1)\frac{1}{4}\right] \\ &= \frac{1}{24n}. \end{aligned} \tag{11.181}$$

Term of $\mathbb{E}[\gamma^2]$: To analyze the term $\mathbb{E}[\gamma^2]$, we use the following standard result from statistics.

Lemma 11.34. *Let $X_1, \ldots, X_n \sim$ Unif$[0, 1]$. Let $X_{\min} = \min_{i\in[n]} X_i$. We have*

$$\begin{aligned} \mathbb{E}[X_{\min}] &= \frac{1}{n+1} \\ \mathbb{E}[X_{\min}^2] &= \frac{2}{(n+1)(n+2)}. \end{aligned}$$

We define

$$U_{\max} := \max_{j\in[rn]} U_j$$
$$V_{\min} := \min_{j\in[(1-r)n]} V_j,$$

and define $U'_{\max}$ and $V'_{\min}$ likewise. By (11.179) it can be verified that we have the deterministic relation

$$\begin{aligned}|\gamma| &\leq (y_{22,\min} - y_{11,\max}) \vee (y_{12,\min} - y_{21,\max})\\ &\overset{\text{(i)}}{=} (V'_{\min} - U_{\max}) \vee (V_{\min} - U'_{\max})\\ &\leq V'_{\min} - U_{\max} + V_{\min} - U'_{\max},\end{aligned}$$

where equality (i) is true by the assumption that there is no noise and the assumption of $x^* = 0$. Therefore,

$$\begin{aligned}\mathbb{E}[\gamma^2] &\leq \mathbb{E}\left[(V'_{\min} - U_{\max}) + (V_{\min} - U'_{\max})\right]^2\\ &= \underbrace{\mathbb{E}(V'_{\min} - U_{\max})^2}_{T_1} + \underbrace{\mathbb{E}(V_{\min} - U'_{\max})^2}_{T_2} + 2\underbrace{\mathbb{E}(V'_{\min} - U_{\max})(V_{\min} - U'_{\max})}_{T_3}.\end{aligned} \tag{11.182}$$

We consider the three terms T_1, T_2 and T_3 separately. For the term T_1, by Lemma 11.34 we have

$$\begin{aligned}T_1 &= \mathbb{E}[V'_{\min}]^2 + \mathbb{E}[U^2_{\max}] - 2\mathbb{E}[V'_{\min}U_{\max}]\\ &= 2\cdot\frac{2}{(rn+1)(rn+2)} + 2\cdot\frac{1}{(rn+1)^2} \leq \frac{6}{r^2n^2}.\end{aligned}$$

Likewise, for the term T_2 we have

$$T_2 \leq \frac{6}{(1-r)^2n^2}.$$

For the term T_3, by Lemma 11.34 we have

$$T_3 = \frac{2}{rn+1}\cdot\frac{2}{(1-r)n+1} \leq \frac{4}{r(1-r)n^2}.$$

Plugging the three terms back to (11.182), we have

$$\mathbb{E}[\gamma^2] \leq \frac{6}{r^2n^2} + \frac{6}{(1-r)^2n^2} + \frac{8}{r(1-r)n^2} = \frac{c}{n^2}, \tag{11.183}$$

for some constant $c > 0$.

Finally, plugging (11.181) and (11.183) back to (11.180), we have

$$\frac{1}{2}\mathbb{E}\|\widehat{x}^{(0)}\|_2 \leq \frac{1}{24n} + \frac{c}{4n^2},$$

completing the proof.

11.9 Proof of preliminaries

In this section, we present the proofs of the preliminary results presented in Appendix 11.2.

11.9.1 Proof of Proposition 11.1

To avoid clutter of notation, we first prove the case for $\Omega = [d] \times [n]$, and then comment on the general case of $\Omega \subseteq [d] \times [n]$.

Now consider $\Omega = [d] \times [n]$, where our estimator (11.9) reduces to (4.2). We separately consider the cases of $\lambda = 0$ and $\lambda \in (0, \infty)$.

Case of $\lambda = 0$ The objective (4.2) becomes

$$\min_{x \in \mathbb{R}^d} \min_{\substack{B \in \mathbb{R}^{d \times n} \\ B \text{ satisfies } \mathcal{O}}} \left\| Y - x\mathbf{1}^T - B \right\|_F^2 = \min_{\substack{W \in \mathbb{R}^{d \times n} \\ W \in \{x\mathbf{1}^T + B \mid x \in \mathbb{R}^d,\, B \in \mathbb{R}^{d \times n},\, B \text{ satisfies } \mathcal{O}\}}} \|Y - W\|_F^2. \tag{11.184}$$

It can be verified that the set $\{x\mathbf{1}^T + B \mid x \in \mathbb{R}^d,\, B \in \mathbb{R}^{d \times n},\, B \text{ satisfies } \mathcal{O}\}$ is a closed convex set. By the Projection Theorem [21, Proposition 1.1.9], a unique minimizer W_0 to the RHS of (11.184) exists. Therefore, the set of minimizers to the LHS of (11.184) can be written as $\{(x, W_0 - x\mathbf{1}^T) \mid x \in \mathbb{R}^d\}$. The tie-breaking rule minimizes the Frobenius norm $\|B\|_F^2$. That is, we solve

$$\min_{x \in \mathbb{R}^d} \left\| W_0 - x\mathbf{1}^T \right\|_F^2. \tag{11.185}$$

It can be verified that a unique solution to (11.185) exists, because the objective is quadratic in x. Hence, the tie-breaking rule defines a unique solution (x, B).

Case of $\lambda \in (0, \infty)$ It can be verified that the objective (4.2) is strictly convex in (x, B). Therefore, there exists at most one minimizer [21, Proposition 3.1.1].

It remains to prove that there exists a minimizer. It is straightforward to see that the objective is continuous in (x, B). We now prove that the objective is coercive on $\{(x, B) : x \in \mathbb{R}^d, B \in \mathbb{R}^{d \times n}, B \text{ satisfies } \mathcal{O}\}$. That is, for any constant $M > 0$, there exists a constant $R_M > 0$, such that the objective at (x, B) is greater than M for all (x, B) in the domain $\{(x, B) : x \in \mathbb{R}^d, B \in \mathbb{R}^{d \times n}, B \text{ satisfies } \mathcal{O}\}$ with

$$\|x\|_2^2 + \|B\|_F^2 > R_M \tag{11.186}$$

Given coercivity, invoking Weierstrass' Theorem [21, Proposition 3.2.1] completes the proof.

We set

$$R_M = d\left[\left(1 + \frac{1}{\sqrt{\lambda}}\right)\sqrt{M} + \max_{i \in [d], j \in [n]} Y\right]^2 + \frac{1}{\lambda}M. \tag{11.187}$$

We discuss the following two cases depending on the value of $\|B\|_F^2$.

Case of $\|B\|_F^2 \geq \frac{M}{\lambda}$ The second term of the objective (11.9) is lower-bounded as $\lambda\|B\|_F^2 \geq M$. Hence, the objective (4.2) is at least M.

Case of $\|B\|_F^2 < \frac{M}{\lambda}$: Combining (11.186) and (11.187), we have

$$\|x\|_2^2 > R_M - \|B\|_F^2 > d\left[(1+\frac{1}{\sqrt{\lambda}})\sqrt{M} + \max_{i\in[d],j\in[n]} y_{ij}\right]^2.$$

Hence, there exists some $i^* \in [d]$ such that

$$|x_{i^*}| > (1+\frac{1}{\sqrt{\lambda}})\sqrt{M} + \max_{i\in[d],j\in[n]} y_{ij}. \tag{11.188}$$

Consider the (i^*, j) entry in the matrix $(Y - x\mathbf{1}^T - B)$ for any $j \in [n]$. We have

$$\begin{aligned}\left|(Y - x\mathbf{1}^T - B)_{i^*j}\right| &\geq |x_{i^*}| - |y_{i^*j}| - |b_{i^*j}| \\ &\geq |x_{i^*}| - \max_{i\in[d],j\in[n]} y_{ij} - \|B\|_F \\ &\overset{\text{(i)}}{>} \left(1+\frac{1}{\sqrt{\lambda}}\right)\sqrt{M} - \sqrt{\frac{M}{\lambda}} = \sqrt{M},\end{aligned}$$

where (i) is true by (11.188) and the assumption of the case that $\|B\|_F^2 < \frac{1}{\lambda}M$. Hence, the second term in the objective (4.2) is lower-bounded by

$$\left\|Y - x\mathbf{1}^T - B\right\|_F^2 \geq \left|(Y - x\mathbf{1}^T - B)_{i^*j}\right|^2 > M,$$

and therefore the objective (4.2) is greater than M.

Combining the two cases depending on $\|B\|_F^2$ completes the proof of the coercivity of the objective (4.2) in terms of (x, B). Invoking the Weierstrass' Theorem [21, Proposition 3.2.1] completes the proof of $\Omega = [d] \times [n]$.

Extending the proof to general $\Omega \subseteq [d]\times[n]$: For general $\Omega \subseteq [d]\times[n]$, by a similar argument the solution $(\widehat{x}, \{\widehat{b}_{ij}\}_{(i,j)\in\Omega})$ exists and is unique. Note that the objective (11.9) is independent from $\{b_{ij}\}_{(i,j)\notin\Omega}$,so we have $\widehat{b}_{ij} = 0$ for each $(i,j) \notin \Omega$. Hence, a unique solution $(\widehat{x}, \widehat{B})$ to (11.9) exists for general Ω.

11.9.2 Proof of Lemma 11.3

It is sufficient to prove the general version (11.13). First consider $\lambda = \infty$. It can be verified that the closed-form expression (4.3) for the solution at $\lambda = \infty$ satisfies the claimed relation (11.13).

It remains to consider the case of $\lambda \in [0, \infty)$. Given the value of the solution $\widehat{B}^{(\lambda)}$, we solve for $\widehat{x}^{(\lambda)}$ by minimizing the first term of the objective (4.2) as

$$\min_{x\in\mathbb{R}^d} \|Y - x\mathbf{1}^T - \widehat{B}^{(\lambda)}\|_F^2. \tag{11.189}$$

Writing out all the terms in (11.189) and completing the square yields the claimed relation (11.13).

11.9.3 Proof of Lemma 11.4

It is sufficient to prove the general version (11.15). First consider the case of $\lambda = \infty$. It can be verified that the closed-form expression expressions (4.3) for the solution at $\lambda = \infty$ satisfies the claimed relations (11.16).

It remains to consider the case of $\lambda \in [0, \infty)$. First we prove (11.15a). Assume for contradiction that $\sum_{(i,j)\in\Omega} \widehat{b}_{ij} \neq 0$. Consider the set of alternative solutions $(\widehat{x}_\gamma, \widehat{B}_\gamma)$ parameterized by some $\gamma \in \mathbb{R}$ as

$$\widehat{x}_\gamma = \widehat{x} + \gamma \mathbf{1}_d \tag{11.190a}$$

$$\widehat{B}_\gamma = \widehat{B} - \gamma \mathbf{1}_d \mathbf{1}_n^T. \tag{11.190b}$$

Note that the original solution $(\widehat{x}, \widehat{B})$ corresponds to $\gamma = 0$.

Since $\widehat{B}_\gamma$ in (11.190) is obtained by subtracting all entries in the matrix by a constant γ, the bias term $\widehat{b}_\gamma$ satisfies the partial ordering $\mathcal{O}$ for any $\gamma \in \mathbb{R}$. Moreover, since by construction (11.190) the value of $(\widehat{x}_\gamma \mathbf{1}_d + \widehat{b}_\gamma)$ is the same for all $\gamma \in \mathbb{R}$, the first term in the objective (4.2) is equal for all $\gamma \in \mathbb{R}^d$. Now consider the second term $\|\widehat{B}_\gamma\|_\Omega^2$. Writing out the terms in $\|\widehat{B}_\gamma\|_\Omega^2$ and completing the square, we have $\|\widehat{b}_\gamma\|_\Omega^2$ is minimized at $\gamma = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \widehat{b}_{ij} \neq 0$. Contradiction to the assumption that the solution at $\gamma = 0$ minimizes the objective, completing the proof of (11.15a).

Now we prove (11.15b). By (11.13) from Lemma 11.3 and summing over $i \in [d]$, we have

$$\sum_{i\in[d]} n_i \widehat{x}_i = \sum_{i\in[d]} \sum_{j\Omega_i} (y_{ij} - \widehat{b}_{ij}) = \sum_{(i,j)\in\Omega} (y_{ij} - \widehat{b}_{ij}) \overset{\text{(i)}}{=} \sum_{(i,j)\in\Omega} y_{ij},$$

where equality (i) is true by (11.15a), completing the proof of (11.15b).

11.9.4 Proof of Proposition 11.5

First consider the case of $\lambda = \infty$, the claimed result can be verified using the closed-form expressions (4.3) at $\lambda = \infty$. It remains to consider the case of any $\lambda \in [0, \infty)$. Assume for contradiction that the solution at $Y + \Delta x \mathbf{1}^T$ is not $(\widehat{x} + \Delta x, \widehat{B})$, but instead $(\widehat{x} + \Delta x + u, \widehat{B}')$ for some non-zero $u \in \mathbb{R}^d$. By the optimality of $(\widehat{x} + \Delta x + u, \widehat{B}')$, we have

$$\|(Y + \Delta x \mathbf{1}^T) - (\widehat{x} + \Delta x + u)\mathbf{1}^T - \widehat{B}'\|_\Omega^2 + \lambda \|\widehat{B}'\|_\Omega^2 \leq \|(Y + \Delta x \mathbf{1}^T) - (\widehat{x} + \Delta x)\mathbf{1}^T - \widehat{B}\|_\Omega^2 + \lambda \|\widehat{B}\|_\Omega^2 \tag{11.191}$$

$$\|Y - (\widehat{x} + u)\mathbf{1}^T - \widehat{B}'\|_\Omega^2 + \lambda \|\widehat{B}'\|_\Omega^2 \leq \|Y - \widehat{x}\mathbf{1}^T - \widehat{B}\|_\Omega^2 + \lambda \|\widehat{B}\|_\Omega^2. \tag{11.192}$$

If strict inequality in (11.192) holds, then $(\widehat{x} + u, \widehat{B}')$ attains a strictly smaller objective on observations Y given $(\mathcal{O}, \lambda, \Omega)$ than $(\widehat{x}, \widehat{B})$. Contradiction to the assumption that $(\widehat{x}, \widehat{B})$ is optimal on the observations Y. Otherwise, equality holds in (11.192) and hence in (11.191). By the tie-breaking rule of the equality (11.191) on the observations $(Y + \Delta x \mathbf{1}^T)$, we have

$$\|\widehat{B}'\|_\Omega^2 < \|\widehat{B}\|_\Omega^2, \tag{11.193}$$

Combining (11.193) with the equality of (11.192) yields a contradiction to the assumption that $(\widehat{x}, \widehat{B})$ is optimal on the observations Y, and hence is chosen by the tie-breaking rule over the alternative solution $(\widehat{x} + u, \widehat{B}')$.

11.9.5 Proof of Lemma 11.7

The proof relies on (11.15b) from Lemma 11.4. Assume without loss of generality that $x^* = 0$. We first show that on the RHS of (11.15b), we have that $\sum_{(i,j)\in\Omega^{\mathrm{t}}} y_{ij}$ converges to 0 for random Ω^{t} obtained by Algorithm 2.

Fix some constant $\epsilon_1 > 0$ whose value is determined later.

Part (b): For any fixed Ω^{t}, by Hoeffding's inequality, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \frac{1}{|\Omega^{\mathrm{t}}|} \sum_{(i,j)\in\Omega^{\mathrm{t}}} y_{ij} \right| < \epsilon_1 \right) = 1. \tag{11.194a}$$

Part (a): Given the assumption that $x^* = 0$ and the assumption that there is no noise, we have $Y = B$. By (11.22b) from Lemma 11.15, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \frac{1}{|\Omega^{\mathrm{t}}|} \sum_{(i,j)\in\Omega^{\mathrm{t}}} y_{ij} \right| < \epsilon_1 \right) = 1. \tag{11.194b}$$

The rest of the proof is the same for both parts. Denote the event in (11.194) as E. We now condition on E and consider the LHS of (11.15b). By (11.1), the number of students in each course $i \in [d]$ is $n^{\mathrm{t}} = \frac{1}{2}n$. Consider any $\lambda \in [0, \infty] \in \Lambda_\epsilon$. By the definition of Λ_ϵ we have $\|\widehat{x}^{(\lambda)}\|_2 \geq \epsilon$. There exists some i^* such that $|\widehat{x}_{i^*}| \geq \frac{\epsilon}{\sqrt{d}}$. Assume without loss of generality that $\widehat{x}_{i^*} > \frac{\epsilon}{\sqrt{d}}$. We now show that there exists some i' such that $\widehat{x}_{i'} \leq 0$. Assume for contradiction that $\widehat{x}_i > 0$ for all $i \in [d]$. Then by (11.15b), we have

$$\sum_{(i,j)\in\Omega^{\mathrm{t}}} y_{ij} = n^{\mathrm{t}} \sum_{i\in[d]} \widehat{x}_i \geq n^{\mathrm{t}} \widehat{x}_{i^*} > \frac{n}{2}\frac{\epsilon}{\sqrt{d}}.$$

Therefore,

$$\frac{1}{|\Omega^{\mathrm{t}}|} \sum_{(i,j)\in\Omega} y_{ij} = \frac{2}{dn}\frac{n}{3}\frac{\epsilon}{\sqrt{d}} = \frac{2\epsilon}{3d^{\frac{3}{2}}}.$$

Setting ϵ_1 to be sufficiently small such that $\epsilon_1 < \frac{2\epsilon}{3d^{\frac{3}{2}}}$ yields a contradiction with E. Hence, conditional on E, there exists some i_2^* such that $\widehat{x}_{i_2^*} \leq 0$. Therefore, $\max_{i,i'\in[d]}(\widehat{x}_i - \widehat{x}_{i'}) \geq \widehat{x}_{i^*} - \widehat{x}_{i_2^*} > \frac{\epsilon}{\sqrt{d}}$. A similar argument applies to the case of $\widehat{x}_{i^*} < -\frac{\epsilon}{\sqrt{d}}$. Hence, we have

$$\max_{i,i'\in[d]} (\widehat{x}_i - \widehat{x}_{i'}) > \frac{\epsilon}{\sqrt{d}}, \quad \forall \lambda \in \Lambda_\epsilon \;\Big|\; E. \tag{11.195}$$

Combining (11.195) with (11.194), we have

$$\lim_{n\to\infty}\left(\max_{i,i'\in[d]}(\widehat{x}_i-\widehat{x}_{i'}),\quad \forall\lambda\in\Lambda_\epsilon\right)\geq\mathbb{P}(E)=1,$$

completing the proof.

11.9.6 Proof of Lemma 11.8

We follow the proof of Lemma 11.7, we assume $x^*=0$ without loss of generality. Then fix some constant $\epsilon_1>0$, and estalish concentration inequalities on the RHS of (11.15b).

Part (b): Same as (11.194b) from Lemma 11.7, we have

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{1}{|\Omega^{\mathrm{t}}|}\sum_{(i,j)\in\Omega^{\mathrm{t}}}y_{ij}\right|<\epsilon_1\right)=1. \tag{11.196a}$$

Part (a): By Hoeffding's inequality, we have

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{1}{dn}\left|\sum_{i\in[d],j\in[n]}y_{ij}\right|<\epsilon_1\right)=1. \tag{11.196b}$$

The rest of the proof is the same for both parts. Combining (11.196) with (11.15b), we have

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{1}{d}\sum_{i\in[d]}\widehat{x}_i\right|<\epsilon_1\right)=1. \tag{11.197}$$

Fix any value $\epsilon>0$. Denote E as the event that the events in both (11.17) and (11.197) hold. By a union bound of (11.17) and (11.197), we have

$$\lim_{n\to\infty}(E)=1. \tag{11.198}$$

Condition on E and consider the value of $\widehat{x}_1^{(\lambda)}$. First consider the case of $\widehat{x}_1>\epsilon$, then by (11.17) we have $\widehat{x}_i>0$ for each $i\in[d]$. Then

$$\frac{1}{d}\left|\sum_{i\in[d]}\widehat{x}_i\right|=\frac{1}{d}\sum_{i\in[d]}\widehat{x}_i>\frac{\epsilon}{d}\quad\left|\ \widehat{x}_1>\epsilon,E\right.$$

A similar argument applies to the case of e $\widehat{x}_1<-\epsilon$, and we have

$$\frac{1}{d}\left|\sum_{i\in[d]}\widehat{x}_i\right|>\frac{\epsilon}{d}\quad\left|\ |\widehat{x}_1|>\epsilon,E\right.$$

The same argument applies to each $i \in [d]$. We have

$$\frac{1}{d}\left|\sum_{i\in[d]} \widehat{x}_i\right| > \frac{\epsilon}{d} \quad \middle| \quad \|\widehat{x}\|_\infty > \epsilon, E$$

Taking a sufficiently small ϵ_1 such that $\epsilon_1 < \frac{\epsilon}{d}$ in (11.197) yields a contradiction. Hence, we have

$$\lim_{n\to\infty} \mathbb{P}(\|\widehat{x}\|_\infty > \epsilon, E) = 0. \tag{11.199}$$

Hence,

$$\lim_{n\to\infty} \mathbb{P}\left(\|\widehat{x}\|_2 > \sqrt{d}\epsilon\right) \leq \lim_{n\to\infty} \mathbb{P}\left(\|\widehat{x}\|_\infty > \epsilon\right) \overset{\text{(i)}}{=} \lim_{n\to\infty} \left(\|\widehat{x}\|_\infty > \epsilon, \overline{E}\right) \leq \lim_{n\to\infty} \mathbb{P}(\overline{E}) \overset{\text{(ii)}}{=} 0,$$

where inequality (i) is true by (11.199) and (ii) is true by (11.198), completing the proof.

11.9.7 Proof of Proposition 11.9

Without loss of generality we assume $x^* = 0$. By (11.16b) from Lemma 11.4 with the assumption that $d = 2$, we have $\frac{1}{2}(\widehat{x}_1 + \widehat{x}_2) = \overline{y}$, and hence without loss of generality we parameterize $\widehat{x}$ with some $\gamma \in \mathbb{R}$ as

$$\widehat{x}_\gamma = \overline{y} + \begin{bmatrix} -1 \\ 1 \end{bmatrix} \cdot \frac{\gamma}{2} \tag{11.200}$$

It remains to determine the value of γ.

Given $x^* = 0$ and the assumption that there is no noise, we have $Y = B$. By the assumption (A2) on the bias, we have B obeys the ordering constraints $\mathcal{O}$. Hence, setting $(\widehat{x}, \widehat{B}) = (0, B)$ gives an objective of 0 in (4.2). Hence, at the optimal solution $(\widehat{x}_\gamma, \widehat{B}_\gamma)$, the objective (4.2) equals 0. At the optimal solution, we have

$$\widehat{B}_\gamma = Y - \widehat{x}_\gamma \mathbf{1}^T. \tag{11.201}$$

The rest of the proof consists of two steps in determining the value of γ. First, we find the set of γ such that $\widehat{B}_\gamma$ satisfies the ordering constraint $\mathcal{O}$. Then we find the optimal γ from this set that is chosen by tie-breaking, minimizing $\|\widehat{B}_\gamma\|_F^2$.

Step 1: Finding the set of γ that satisfies the ordering constraint Given $Y = B$, for any $\gamma\mathbb{R}$ we have that $\widehat{B}_\gamma$ satisfies all ordering constraints in $\mathcal{O}$ that are within the same course, that is, the ordering constraints in the form of $((i,j),(i,j')) \in \mathcal{O}$ with $i \in \{1,2\}$. Hence, we only need to consider ordering constraints involving both courses, that is, the ordering constraints in the form of $((i,j),(i',j'))$ with $\{i,i'\} = \{1,2\}$. It can be verified that these constraints involving both courses are satisfied if and only if

$$\begin{cases} y_{11,\max} - \widehat{x}_1 \leq y_{22,\min} - \widehat{x}_2 \\ y_{21,\max} - \widehat{x}_2 \leq y_{12,\min} - \widehat{x}_1. \end{cases} \tag{11.202}$$

Plugging the parameterization (11.200) of $\widehat{x}_\gamma$ into (11.202), we have

$$y_{21,\max} - y_{12,\min} \le \gamma \le y_{22,\min} - y_{11,\max}. \tag{11.203}$$

Note that the range in (11.203) is always non-empty, because given $Y = B$, we have $y_{11,\max} \le y_{12,\min}$ and $y_{21,\max} \le y_{22,\min}$ and hence $y_{21,\max} - y_{12,\min} \le y_{22,\min} - y_{11,\max}$.

Step 2: Finding the optimal γ from the range (11.203) **minimizing** $\|\widehat{B}_\gamma\|_F^2$ Using the parameterizations (11.200) and (11.201), we write $\|\widehat{B}_\gamma\|_F^2$ as

$$\begin{aligned}\|\widehat{B}_\gamma\|_F^2 &= \|Y - \widehat{x}_\gamma \mathbf{1}^T\|_F^2 \\ &\overset{\text{(i)}}{=} \sum_{j\in[n]} \left(y_{1j} - \overline{y} + \frac{\gamma}{2}\right)^2 + \sum_{j\in[n]} \left(y_{2j} - \overline{y} - \frac{\gamma}{2}\right)^2.\end{aligned} \tag{11.204}$$

Writing out the terms in (11.204) and completing the square, we have that minimizing $\|\widehat{b}_\gamma\|_F^2$ is equivalent to minimizing the term:

$$\frac{n}{2}\left(\gamma - (\overline{y}_2 - \overline{y}_1)\right)^2 \tag{11.205}$$

Combining (11.203) and (11.205) gives the yields expression (11.18) for the optimal γ.

11.9.8 Proof of Lemma 11.10

The lemma is a direct consequence of the following result (given that almost-sure convergence implying convergence in probability).

Lemma 11.35 (Theorem 2 in [52]). *Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(0,1)$. We have*

$$\limsup_{n\to\infty} \frac{\sqrt{2\log n}}{\log\log n} M_n = 1 \quad \textit{almost surely},$$

where $\log$ *is the logarithm of base* 2.

11.9.9 Proof of Lemma 11.11

Let g be the p.d.f. of $\mathcal{N}(0,1)$. Let G_n be the empirical c.d.f. and the empirical inverse c.d.f. of n i.i.d. samples from $\mathcal{N}(0,1)$ and let G_n^{-1} be the inverse of G_n.

The claim is a straightforward combination of the following two lemmas. The first lemma states that the empirical inverse c.d.f. converges to the true inverse c.d.f. The second lemma states that order statistics converges to the empirical inverse c.d.f.

Lemma 11.36 (Example 3.9.21 of [182], or Corollary 21.5 of [181]). *Consider any fixed $p \in (0,1)$. Assume that G is differentiable at $G^{-1}(p)$ and $g(G^{-1}(p)) > 0$. Then we have*

$$\sqrt{n}\left[G_n^{-1}(p) - G^{-1}(p)\right] \xrightarrow{d} N\left(0, \frac{p(1-p)}{g^2(G^{-1}(p))}\right).$$

Lemma 11.37 (Lemma 21.7 in [181]). *Fix constant $p \in (0, 1)$. Let $\{k_n\}_{n=1}^{\infty}$ be a sequence of integers such that $\frac{k_n}{n} = p + \frac{c}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$ for some constant c. Then*

$$\sqrt{n}\left[X^{(k_n:n)} - G_n^{-1}(p)\right] \xrightarrow{P} \frac{c}{g(G^{-1}(p))}$$

11.9.10 Proof of Lemma 11.13

We consider any fixed $i \in [d], k \in [r]$, and any fixed total ordering π_0 generated by Line 2 of Algorithm 2. Note that the ℓ_{ik} elements in G_{ik} are consecutive with respect to the sub-ordering of π_0 restricted to course i in Line 4 of Algorithm 2. Then it can be verified from Line 5-7 of Algorithm 2 that

$$\frac{\ell_{ik}}{2} - 1 \leq \ell_{ik}^{\mathrm{v}} \leq \frac{\ell_{ik}}{2} + 1, \tag{11.206}$$

It can be verified that (11.206) along with the assumption that $\ell_{ik} \geq 4$ yields (11.20a). Summing (11.20a) over $i \in [d]$ yields (11.21a). Finally, replacing the validation set Ω^{v} by the training set Ω^{t} in the proof of (11.20a) and (11.21a) yields (11.20b) and (11.21b), respectively.

11.9.11 Proof of Lemma 11.14

We prove part (a) and part (b) together. Note that if the element of rank k_1 and the element of rank k_2 are adjacent within Ω^{t}, or adjacent between Ω^{t} and Ω^{v}, the $(k_2 - k_1 - 1)$ elements of ranks from $k_1 + 1$ through $k_2 - 1$ are within the same set (i.e., Ω^{t} or Ω^{v}). Assume for contradiction that $k_2 - k_1 \geq 2d + 2$. Then the number of elements from rank $k_1 + 1$ through $k_2 - 1$ is at least $k_2 - k_1 - 1 \geq 2d + 1$. Consider these elements. There exists a course i^* such that the number of such elements within this course is at least 3. Given that these elements have consecutive ranks, they are consecutive within course i^*. Hence, two of these elements in course i^* appear as the same pair of elements in Line 7 of Algorithm 2. According to Line 7 of Algorithm 2, one element in this pair is assigned to Ω^{t} and the other element is assigned to Ω^{v}. Contradiction to the assumption that all of these elements are from the same set.

11.9.12 Proof of Lemma 11.15

Proof of (11.22a)**:** We consider any course $i \in [d]$. We first fix any value of $B = B^*$. Fix any π_0 of the dn elements (in Line 2 of Algorithm 2). Recall from Line 4 of Algorithm 2 that the sub-ordering of the n elements in course i according to π_0 is denoted as $(i, j^{(1)}), \ldots, (i, j^{(n)})$.

Consider each pair $(i, j^{(2t-1)})$ and $(i, j^{(2t)})$ for $t \in \left[\frac{n}{2}\right]$. Algorithm 2 randomly assigns one of the two elements to the training set Ω^{t} uniformly at random. Denote U_t as the the value from this pair that is assigned to training set. Then we have

$$U_t = \begin{cases} b^*_{i,j^{(2t-1)}} & \text{with probability } 0.5 \\ b^*_{i,j^{(2t)}} & \text{with probability } 0.5. \end{cases}$$

Denote $\Delta_B := \max_{j\in[n]} b_{ij} - \min_{j\in[n]} b_{ij}$ and denote $\Delta_{B^*} = \max_{j\in[n]} b^*_{ij} - \min_{j\in[n]} b^*_{ij}$. Recall from (11.1) that $n^{\mathrm{t}} = \frac{n}{2}$. Fix any $\delta > 0$. By Hoeffding's inequality, there exists n_1 such that for all $n \geq n_1$,

$$\mathbb{P}\left(\left|\frac{1}{n^{\mathrm{t}}}\sum_{t\in\left[\frac{n}{2}\right]} U_t - \frac{1}{n^{\mathrm{t}}}\mathbb{E}[U_t]\right| < \Delta_{B^*}\sqrt{\frac{\log n}{n}} \;\middle|\; B = B^*\right) \geq 1 - \frac{\delta}{2}.$$

Equivalently, for all $n \geq n_1$,

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{1}{n^{\mathrm{t}}}\sum_{j\in\Omega_i^{\mathrm{t}}} b_{ij} - \frac{1}{n}\sum_{j\in[n]} b_{ij}\right| < \Delta_{B^*}\sqrt{\frac{\log n}{n}} \;\middle|\; B = B^*\right) \geq 1 - \frac{\delta}{2}. \tag{11.207}$$

Now we analyze the term Δ_B. By Lemma 11.12, we have that there exists n_2 such that for all $n \geq n_2$,

$$\mathbb{P}\Big(\Delta_B \leq 4\sqrt{\log n}\Big) \geq 1 - \frac{\delta}{2}. \tag{11.208}$$

Fix any $\epsilon > 0$. Take n_0 to be sufficiently large such that $n_0 \geq \max\{n_1, n_2\}$ and $\frac{4\log n_0}{\sqrt{n_0}} < \epsilon$. We have that for all $n \geq n_0$,

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{n^{\mathrm{t}}}\sum_{j\in\Omega_i^{\mathrm{t}}} b_{ij} - \frac{1}{n}\sum_{j\in[n]} b_{ij}\right| < \epsilon\right) &= \int_{B^*\in\mathbb{R}^{d\times n}} \mathbb{P}\left(\left|\frac{1}{n^{\mathrm{t}}}\sum_{j\in\Omega_i^{\mathrm{t}}} b_{ij} - \frac{1}{n}\sum_{j\in[n]} b_{ij}\right| < \epsilon \;\middle|\; B^*\right)\cdot\mathbb{P}(B^*)\,\mathrm{d}B^*\\
&\geq \int_{\substack{B^*\in\mathbb{R}^{d\times n}:\\ \Delta_{B^*}\leq 4\sqrt{\log n}}} \mathbb{P}\left(\left|\frac{1}{n^{\mathrm{t}}}\sum_{j:(i,j)\in\Omega^{\mathrm{t}}} b_{ij} - \frac{1}{n}\sum_{j\in[n]} b_{ij}\right| < \epsilon \;\middle|\; B\right)\cdot\mathbb{P}(B^*)\,\mathrm{d}B^*\\
&\overset{\text{(i)}}{\geq} \left(1 - \frac{\delta}{2}\right)\cdot\mathbb{P}\left(\Delta_B \leq 4\sqrt{\log n}\right)\\
&\overset{\text{(ii)}}{\geq} \left(1 - \frac{\delta}{2}\right)^2 \geq 1 - \delta,
\end{aligned}$$

where inequality (i) is true by (11.207) and inequality (ii) is true by (11.208), completing the proof.

Proof of (11.22b)**:** By Hoeffding's inequality, we have that for any $\epsilon > 0$,

$$\lim_{n\to\infty}\mathbb{P}\left(\frac{1}{dn}\left|\sum_{i\in[d],j\in[n]} b_{ij}\right| < \epsilon\right) = 1. \tag{11.209}$$

Recall from assumption (A3) that d is assumed to be a constant. Taking a union bound of (11.22a) over $i \in [d]$ and (11.209), folloed by using the triangle inequality yields the claimed result.

11.10 Proof of auxiliary results for Theorem 4.5

In this section, we present the proofs of the auxiliary results for Theorem 4.5.

11.10.1 Proof of Lemma 11.16

Fix any $c > 0$ and fix any $(i, i') \in S_c$. Suppose $k \in [r]$ satisfies the definition (11.24) corresponding to (i, i'). We prove that for any $\epsilon > 0$ and $\delta > 0$, there exists some n_0 such that for all $n \geq n_0$,

$$\mathbb{P}\Big(\widehat{x}_{i'}^{(0)} - \widehat{x}_i^{(0)} < \epsilon\Big) \geq 1 - \delta.$$

The proof consists of two steps. In the first step, we consider the rank of the maximum bias in course i of group k (that is, $\max_{(i,j)\in G_{ik}} t_{ij}$), and the rank of the minimum bias in course i' of group $(k+1)$ (that is, $\min_{(i,j)\in G_{i'k+1}} t_{ij}$). We bound the difference between these two ranks, and then bound the difference between the values of these two terms. In the second step, we show that the ordering constraint imposed by this pair of bias terms leads to the claimed bound (11.25) on $\widehat{x}_{i'}^{(0)} - \widehat{x}_i^{(0)}$.

Step 1: Bounding the difference of a pair of bias terms Recall from (11.7) that $b_{k,\max}$ denotes the largest bias of group k, and $b_{k+1,\min}$ denotes the smallest bias of group $k+1$. We denote the rank of $b_{k,\max}$ as t. By the definition of group ordering, the value of t is deterministic and we have $t = \sum_{k'=1}^{k} \ell_{k'}$. Then the rank of $b_{k+1,\min}$ is $(t+1)$.

Recall that $b_{ik,\max}$ denotes the largest bias in course i of group k, and $b_{ik,\min}$ denotes the smallest bias in course i of group k. Let T_k be a random variable denoting the difference between the ranks of $b_{k,\max}$ and $b_{ik,\max}$, and let T_{k+1} be a random variable denoting the difference between the ranks of $b_{k+1,\min}$ and $b_{i,k+1,\min}$. Equivalently, the ranks of $b_{ik,\max}$ and $b_{i+1,k+1,\min}$ are $(t - T_k)$ and $(t + 1 + T_{k+1})$, respectively, and we have $T_k, T_{k+1} \geq 0$.

Recall that the biases within a group are ordered uniformly at random among all courses. For any constant integer $t_0 > 0$, if we have $T_k \geq t_0$, then the bias terms corresponding to ranks of $(t - t_0 + 1), \ldots, t$ are not assigned to course i. Recall that $\ell_{-i,k} = \ell_k - \ell_{ik}$ denotes the number of observations in group k that are not in course i. We bound the random variable T_k as

$$\mathbb{P}(T_k \geq t_0) = \prod_{m=0}^{t_0-1} \frac{\ell_{-i,k} - m}{\ell_k - m} < \left(\frac{\ell_{-i,k}}{\ell_k}\right)^{t_0} \overset{\text{(i)}}{\leq} (1-c)^{t_0}, \tag{11.210}$$

where step (i) is true by the definition (11.24) of S_c. Similarly we have

$$\mathbb{P}(T_{k+1} \geq t_0) \leq (1-c)^{t_0}. \tag{11.211}$$

Taking $t_0 = \frac{\log(\frac{4}{\delta})}{\log(1-c)}$ and taking a union bound of (11.210) and (11.211), we have

$$\mathbb{P}\Big(T_k + T_{k+1} < 2t_0\Big) \geq \mathbb{P}\Big(T_k < t_0, T_{k+1} < t_0\Big) \geq 1 - 2(1-c)^{t_0} = 1 - \frac{\delta}{2}. \tag{11.212}$$

By Lemma 11.10, there exists n_0 such that for all $n \geq n_0$, we have

$$\mathbb{P}\left(M < \frac{\epsilon}{2t_0 + 1}\right) > 1 - \frac{\delta}{2}, \tag{11.213}$$

where M is the maximum difference between a pair of bias terms of adjacent ranks, defined as $M := \max_{i \in [dn-1]} b^{(i+1)} - b^{(i)}$. Taking a union bound of (11.213) with (11.212), we have that for all $n \geq n_0$

$$\begin{aligned} b_{i',k+1,\min} - b_{ik,\max} &< [(t + 1 + T_{k+1}) - (t - T_k) + 1] \cdot M \\ &\leq (2t_0 + 1)M < \epsilon, \qquad \text{with probability at least } 1 - \delta. \end{aligned} \tag{11.214}$$

Due to the assumption of no noise and the assumption of $x^* = 0$, the observation model (4.1) reduces to $Y = B$. In particular, we have $y_{ik,\max} = b_{ik,\max}$ and $y_{i',k+1,\min} = b_{i',k+1,\min}$. Moreover, the solution $(\widehat{x}, \widehat{B}) = (0, B)$ gives an objective (4.2) of 0 at $\lambda = 0$ due to $Y = B$. Therefore the solution $(\widehat{x}^{(0)}, \widehat{B}^{(0)})$ by our estimator gives an objective of 0, satisfying the deterministic relation $y_{ij} = \widehat{x}_i^{(0)} + \widehat{b}_{ij}^{(0)}$. By definition of the group ordering, the group ordering includes the constraint requiring $\widehat{b}_{ik,\max}^{(0)} \leq \widehat{b}_{i',k+1,\min}^{(0)}$. Therefore, this ordering constraint requires the solution $(\widehat{x}^{(0)}, \widehat{B}^{(0)})$ to satisfy

$$\begin{aligned} \widehat{b}_{i',k+1,\min}^{(0)} - \widehat{b}_{ik,\max}^{(0)} &= (y_{i',k+1,\min} - \widehat{x}_{i'}^{(0)}) - (y_{ik,\max} - \widehat{x}_i^{(0)}) \\ &= (b_{i',k+1,\min} - \widehat{x}_{i'}^{(0)}) - (b_{ik,\max} - \widehat{x}_i^{(0)}) \geq 0 \end{aligned} \tag{11.215}$$

Rearranging (11.215)and combining it with (11.214), we have that for all $n \geq n_0$,

$$\mathbb{P}\Big(\widehat{x}_{i'}^{(0)} - \widehat{x}_i^{(0)} \leq b_{i',k+1,\min} - b_{ik,\max} < \epsilon\Big) \geq 1 - \delta,$$

completing the proof.

11.10.2 Proof of Lemma 11.18

First of all, we assume that $L \leq d$ without loss of generality. This is because if $L > d$, then there exists a course i that appears twice in this cycle. We write the cycle as $(i_1, \ldots, i, \ldots, i', \ldots, i, \ldots, i_L)$, where $i' \in [d]$ denotes some course appearing in between the two occurrences of i. We obtain a shortened cycle by replacing the segment $(i, \ldots, i', \ldots i)$ with a single i. By shortening the cycle the set of courses that appear in this cycle remain the same. We keep shortening the cycle until $L \leq d$.

Fix any $\epsilon > 0$ and $\delta > 0$. Recall from assumption (A3) that d is assumed to be a constant. By applying Lemma 11.16 on the L pairs in (11.26) of S_c, and taking a union bound over these L pairs, we have that there exists n_0 such that for all $n \geq n_0$, with probability at least $1 - \delta$ we simultaneously have

$$\begin{aligned}
\widehat{x}_{m_2} - \widehat{x}_{m_1} &< \frac{\epsilon}{d}, \\
\widehat{x}_{m_3} - \widehat{x}_{m_2} &< \frac{\epsilon}{d}, \\
&\vdots \\
\widehat{x}_{m_L} - \widehat{x}_{m_{L-1}} &< \frac{\epsilon}{d}, \\
\widehat{x}_{m_1} - \widehat{x}_{m_L} &< \frac{\epsilon}{d}.
\end{aligned} \tag{11.216}$$

Consider any $m < m'$ with $m, m' \in [L]$. Conditional on (11.216) we have

$$\widehat{x}_{i_{m'}} - \widehat{x}_{i_m} = (\widehat{x}_{i_{m'}} - \widehat{x}_{i_{m'-1}}) + \ldots + (\widehat{x}_{i_{m+1}} - \widehat{x}_{i_m}) < \epsilon. \tag{11.217}$$

On the other hand, conditional on (11.216) we also have

$$\widehat{x}_{i_m} - \widehat{x}_{i_{m'}} = (\widehat{x}_{i_m} - \widehat{x}_{i_{m-1}}) + \ldots + (\widehat{x}_{i_2} - \widehat{x}_{i_1}) + (\widehat{x}_{i_1} - \widehat{x}_{i_L}) + \ldots + (\widehat{x}_{i_{m'+1}} - \widehat{x}_{i_{m'}}) < \epsilon \tag{11.218}$$

Combining (11.217) and (11.218), we have that for all $n \geq n_0$,

$$\mathbb{P}\Big(\big|\widehat{x}_{i_{m'}} - \widehat{x}_{i_m}\big| < \epsilon, \quad \forall m, m' \in [L]\Big) \geq 1 - \delta.$$

Equivalently,

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{m,m'\in[L]} |\widehat{x}_{i'} - \widehat{x}_i| < \epsilon\right) = 1,$$

completing the proof.

11.10.3 Proof of Lemma 11.19

The proof consists of two steps. We first show that if there exists a cycle including the nodes $i, i' \in V$, then this cycle can be modified to construct a cycle of length at most $2(d-1)$ including i and i'. In the second step, we prove the existence of a cycle.

Constructing a cycle of length at most $2(d-1)$ given a cycle of arbitrary length Fix any hypernode V and any $i, i' \in V$. We assume that there exists a cycle including the nodes i and i'. By the definition of a cycle, this cycle includes a directed path $i \to i'$ and a directed path $i' \to i$. If the directed path $i \to i'$ has length greater than $(d-1)$, then there exists some course $i'' \in [d]$ (which may or may not equal to i or i') that appears at least twice in this cycle. Then we decompose the path into three sub-paths of $i \to i''$, $i'' \to i''$, and $i'' \to i'$. We remove the sub-path $i'' \to i''$, and concatenate the subpaths $i \to i''$ and $i'' \to i'$, giving a new path $i \to i'$ of strictly smaller length than the original path. We continue shortening the path until each course appears at most once in the path, and hence the path is of length at most $(d-1)$. Likewise we shorten the path $i' \to i$ to have length at most $(d-1)$. Finally, combining these two paths $i \to i'$ and $i' \to i$ gives a cycle of length at most $2(d-1)$, including nodes i and i'.

Existence of a cycle of arbitrary length We prove the existence of a cycle including i and i' by induction on the procedure that constructs the partition. At initialization, each hypernode contains a single course. The claim is trivially satisfied because for any hypernode V there do not exist $i, i' \in V$ with $i \neq i'$. Now consider any merge step that merges hypernodes $V_1, \ldots, V_L$ for some $L \geq 2$ during the construction of the partition. By definition, the merge occurs because there is a cycle that includes at least one course from each of the hypernodes $V_1, \ldots, V_L$. We denote the course from V_m that is included the cycle as $i_m \in V_m$ for each $m \in [L]$. If there exist multiple courses from V_m included in the cycle, we arbitrarily choose one as i_m). Denote the merged hypernode as $V = V_1 \cup \ldots \cup V_L$. Now consider any two courses i and i' from the same hypernode.

First consider the case of i and i' are from a hypernode that is not V, then by the induction hypothesis there is a cycle including both i and i'.

Now consider the case of $i, i' \in V$. We have that $i \in V_m$ and $i' \in V_{m'}$ for some $m, m' \in [L]$. If $m = m'$, then by the induction hypothesis there is a cycle that includes both m and m'. If $m \neq m'$, then by the induction hypothesis, there is a directed path $i \to i_m$ within V_m (trivially if $i = i_m$), and a directed path $i_{m'} \to i'$ within $V_{m'}$ (trivially if $i' = i_{m'}$). Moreover, by the definition of i_m and $i_{m'}$, we have that i_m and $i_{m'}$ are included in a cycle. Hence, there exists a directed path $i_m \to i_{m'}$. Concatenating the paths $i \to i_m$, $i_m \to i_{m'}$ and $i_{m'} \to i'$ gives a path $i \to i'$. Likewise there exists a path $i' \to i$. Hence, for any $i, i' \in V$, there exists a cycle that includes both i and i'.

11.10.4 Proof of Lemma 11.20

The proof consists of four steps. The first step gives a preliminary property on the graph, to be used in the later steps. The second step shows that each hypernode contains courses that are consecutive. The third step shows that the ranks of elements in each hypernode are consecutive. The fourth step shows that the edges only exist between hypernodes that are adjacent in their indexing.

Step 1: There exists a path from any course i to any course i' with $i < i'$ Denote the minimal rank in course i and in course i' as t and t', respectivly. By the assumption (11.38), we have $t < t'$. We consider the courses corresponding to the elements of ranks t through t', denoted as $(i_t, \ldots, i_{t'})$. For any integer $k \in \{t, \ldots, t'-1\}$ if $i_k \neq i_{k+1}$, then by the definition of S_c from (11.24) we have $(i_k, i_{k+1}) \in S_1$ because these two elements have consecutive ranks. Hence, there is an edge $i_k \to i_{k+1}$ by the construction of the graph. Concatenating all such edges $\{i_k \to i_{k+1}\}_{k \in \{t, \ldots, t'-1\} : i_k \neq i_{k+1}\}}$ gives a path $i \to i'$.

Step 2: Each hypernode contains consecutive nodes We prove that the nodes within each hypernode are consecutive. That is, for each hypernode V, there exist courses $i, i' \in [d]$ with $i < i'$ such that $V = \{i, i+1, \ldots, i'\}$. It suffices to consider any course i'' such that $i < i'' < i'$ and show that $i'' \in V$. Assume for contradiction that $i'' \notin V$. By Step 1, there exists a path $i \to i''$ and also a path $i'' \to i'$. Since $i, i' \in V$, by Lemma 11.19 there exists a path $i' \to i$. Hence, by concatenating these three paths $i \to i'', i'' \to i'$ and $i' \to i$, we have a cycle that includes courses i, i'' and i' that are involved in two different hypernodes. Contradiction to the definition of the partition that there are no cycles including nodes from more than one hypernode in the final partition, completing the proof that each hypernode contains consecutive nodes. Hence, we order the hypernodes as $V_1, \dots V_s$, such that the indexing of the nodes increases with respect to the indexing of the hypernodes.

Step 3: The ranks in each hypernode are consecutive We show that the ranks of the elements within each hypernode are consecutive, and also in the increasing order of the indexing of the hypernodes. Assume for contradiction that there exists some element of rank t' in $V_{m'}$, and some element of rank t in V_m with $m < m'$ and $t > t'$. Denote the corresponding courses as $i \in V_m$ and $i' \in V_{m'}$. On the one hand, by Step 2 we have $i < i'$ due to $m < m'$. Then by Step 1, we have a path $i \to i'$. On the other hand, we consider the elements of ranks $\{t', \dots, t\}$ and construct a path $i' \to i$ similar to the construction of the path in Step 1. Concatenating the paths $i \to i'$ and $i' \to i$ gives a cycle that include courses $i \in V_m$ and $i' \in V_{m'}$ that from two different hypernodes. Contradiction to the definition of the partition that there does not exist cycles including more than one hypernode.

Step 4: The only edges on the hypernodes are (V_m, V_{m+1}) for all $m \in [s-1]$ For total orderings, the edges exist between elements of adjacent ranks. That is, consider the elements of ranks t and $t+1$ for any $t \in [dn-1]$. If their corresponding courses i_t and i_{t+1} are different, then there exists an edge $i_t \to i_{t+1}$. Then Step 4 is a direct consequence of Step 3.

11.11 Proof of auxiliary results for Theorem 4.9

In this section, we present the proofs of the auxiliary results for Theorem 4.9.

11.11.1 Proof of Theorem 11.21

The proof closely follows part (a) and part (c) of Theorem 4.5 (see Appendix 11.3). Therefore, we outline the modifications to the proof of Theorem 4.5, in order to extend to any $\Omega^{\mathrm{t}} \subseteq [d] \times [n]$ obtained by Algorithm 2.

Proof Theorem 11.21(a) The proof closely follows the proof of Theorem 4.5(a) (see Appendix 11.3.1) with the modifications discussed in what follows.

Extending S_c to S_c^{t} Recall from (11.3) that ℓ_{ik}^{t} denotes the number of students in course $i \in [d]$ of group $k \in [r]$ restricted to the training set Ω^{t}, and ℓ_k^{t} denotes the number of students in group k restricted to the training set Ω^{t}. We extend the definition (11.24) of S_c and define

$$S_c^{\mathrm{t}} := \left\{ (i, i') \in [d]^2 : \exists k \in [r] \text{ such that } \frac{\ell_{ik}^{\mathrm{t}}}{\ell_k^{\mathrm{t}}}, \frac{\ell_{i'k+1}^{\mathrm{t}}}{\ell_{k+1}^{\mathrm{t}}} \geq c \right\}.$$

Extending Lemma 11.16 to S^{t}_c restricted to the training set Ω^{t} We show that Lemma 11.16 holds for any $(i, i') \in S^{\mathrm{t}}_c$, and the estimator (11.9) $\widehat{x}^{(0)}$ restricted to Ω^{t}.

Denote $b^{\mathrm{t}}_{ik,\max}$ as the largest bias in course i of group k restricted to the training set Ω^{t}, and denote $b^{\mathrm{t}}_{k,\max}$ as the largest bias of group k restricted to the training set Ω^{t}. We extend (11.210) to show that the difference between the ranks of $b^{\mathrm{t}}_{ik,\max}$ and $b^{\mathrm{t}}_{k,\max}$ is bounded by some constant with high probability.

Moreover, it can be verified that the difference between the ranks of $b^{\mathrm{t}}_{k,\max}$ and $b_{k,\max}$ is bounded by a constant with high probability. Combining these two bounds, the difference between the ranks of $b^{\mathrm{t}}_{ik,\max}$ and $b_{k,\max}$ is bounded by a constant with high probability. We define $b^{\mathrm{t}}_{i'k+1,\min}$ and $b_{k+1,\min}$ likewise, and extend (11.211) to show that the difference between the ranks of $b^{\mathrm{t}}_{i'k+1,\min}$ and $b_{k+1,\min}$ is bounded by a constant with high probability. Therefore, we extend 11.214 to:

$$b^{\mathrm{t}}_{i'k+1,\min} - b^{\mathrm{t}}_{ik,\max} < \epsilon, \quad \text{with probability at least } 1-\delta.$$

Following the rest of the original arguments for Lemma 11.16 (see Appendix 11.10) completes the extension of Lemma 11.16 to being restricted to Ω^{t}.

Extending Lemma 11.18 to S^{t}_c restricted to Ω^{t} We replace the set S_c in Lemma 11.18 by the set S^{t}_c. It can be verified that Lemma 11.18 holds under this extension following its original proof (see Appendix 11.10).

Extending the rest of the arguments For any $i \in [d], k \in [r]$, by (11.20b) and (11.21b) from Lemma 11.13 we have

$$\frac{\ell^{\mathrm{t}}_{ik}}{\ell^{\mathrm{t}}_k} \geq \frac{\frac{\ell_{ik}}{4}}{\frac{3\ell_k}{4}} = \frac{\ell_{ik}}{3\ell_k}.$$

Hence, any $(i, i') \in S_{\frac{c_{\mathrm{f}}}{d}}$, we have $(i, i') \in S^{\mathrm{t}}_{\frac{c_{\mathrm{f}}}{3d}}$. The rest of the arguments follow from the original proof of Theorem 4.5(a) (see Appendix 11.3.1).

Proof of Theorem 11.21(b) The proof closely follows the proof of Theorem 4.5(c) (see Appendix 11.3.3) with the modifications discussed in what follows.

Extending S_c to $S^{\mathrm{t}'}_c$ Recall that for total orderings, we have $(i, i') \in S_1$ if and only if there exists some $k \in [dn-1]$ such that course i contains the element of rank k, and course i' contains the element of rank $(k+1)$. We define the following set $S^{\mathrm{t}'}$, where we consider the rank with respect to the total ordering restricted to the elements in Ω^{t}. That is, we extend the definition (11.24) of S_c and define

$$S^{\mathrm{t}'} := \left\{ \begin{array}{lll} (i, i') \in [d]^2 : \exists 1 \leq k < k' \leq |\Omega^{\mathrm{t}}| & & \\ & \text{such that} & \text{the element of rank } k \text{ is in } \Omega^{\mathrm{t}}_i, \\ & & \text{the element of rank } k' \text{ is in } \Omega^{\mathrm{t}}_{i+1}, \\ & & \text{the elements of ranks } (k+1) \text{ through } (k'-1) \text{ are in } \Omega^{\mathrm{v}} \end{array} \right\}. \tag{11.219}$$

Extending Lemma 11.16 By Lemma 11.14(a) we have that for any $(i, i') \in S^{\mathrm{t}'}$, the corresponding values of k and k' in (11.219) satisfy $k' - k \leq 2d + 1$. We define M' as the maximal difference between elements that are adjacent within Ω^{t}. Then by Lemma 11.10 we extend the bound of M in (11.213) to M' as

$$\mathbb{P}\left(M' < \epsilon\right) > 1 - \frac{\delta}{2}.$$

Following the rest of the arguments in Appendix 11.10.1, we have that Lemma 11.16 holds restricted to the training set Ω^{t}.

Extending Lemma 11.18 to S_c^{t} restricted to Ω^{t} We replace the set S_c in Lemma 11.18 by the set $S^{\mathrm{t}'}$. It can be verified that Lemma 11.18 holds under this extension following its original proof (see Appendix 11.10).

Extending the rest of the arguments The rest of the arguments follow from the original proof of Theorem 4.5(c) (see Appendix 11.3.3). Specifically, we replace the set S_1 by $S^{\mathrm{t}'}$. We consider the total ordering restricted to the training set Ω^{t}. We extend the definition (11.54) of $(\widehat{b}_{\mathrm{L}}, \widehat{b}_{\mathrm{H}})$ to $(\widehat{b}'_{\mathrm{L}}, \widehat{b}'_{\mathrm{H}})$ defined as:

$$\widehat{b}'_{\mathrm{L}} := \frac{1}{\sum_{i \in V_{\mathrm{L}}} |\Omega_i^{\mathrm{t}}|} \sum_{i \in V_{\mathrm{L}}} \sum_{j \in \Omega_i^{\mathrm{t}}} \widehat{b}_{ij}$$

$$\widehat{b}'_{\mathrm{H}} := \frac{1}{\sum_{i \in V_{\mathrm{H}}} |\Omega_i^{\mathrm{t}}|} \sum_{i \in V_{\mathrm{H}}} \sum_{j \in \Omega_i^{\mathrm{t}}} \widehat{b}_{ij}.$$

11.11.2 Proof of Lemma 11.22

We fix any partial ordering $\mathcal{O}$ that satisfies the all c_{f}-fraction assumption, and fix any training-validation split $(\Omega^{\mathrm{t}}, \Omega^{\mathrm{v}})$ obtained by Algorithm 2. Recall that $\mathcal{T}$ denotes the set of all total orderings that are consistent with the partial ordering $\mathcal{O}$. Recall from Line 15 of Algorithm 2 that the interpolated bias is computed as:

$$\widetilde{B}^{(\lambda)} = \frac{1}{|\mathcal{T}|} \sum_{\pi \in \mathcal{T}} \widetilde{B}_{\pi}^{(\lambda)}, \tag{11.220}$$

where recall from Line 13 of Algorithm 2 that $[\widetilde{B}_{\pi}^{(\lambda)}]_{ij}$ for any $(i, j) \in \Omega^{\mathrm{v}}$ is computed as the mean value of $\widehat{B}$ on the nearest-neighbor(s) of (i, j) with respect to the total ordering π. Recall that $\mathrm{NN}(i, j; \pi)$ denotes the set (of size 1 or 2) of the nearest neighbor(s) of (i, j). We have

$$[\widetilde{B}_{\pi}^{(\lambda)}]_{ij} = \frac{1}{|\mathrm{NN}(i, j; \pi)|} \sum_{(i^{\pi}, j^{\pi}) \in \mathrm{NN}} \widehat{B}_{i^{\pi} j^{\pi}}^{(\lambda)}. \tag{11.221}$$

Plugging (11.221) to (11.220), we have

$$\widetilde{B}_{ij}^{(\lambda)} = \frac{1}{|\mathcal{T}|} \sum_{\pi \in \mathcal{T}} \frac{1}{|\mathrm{NN}(i, j; \pi)|} \sum_{(i^{\pi}, j^{\pi}) \in \mathrm{NN}} \widehat{B}_{i^{\pi} j^{\pi}}^{(\lambda)}.$$

The remaining of the proof is outlined as follows. We decompose the summation over $\pi \in \mathcal{T}$ on the RHS of (11.220) into two parts: total orderings $\pi \in \mathcal{T}$ where the set of nearest-neighbors $\mathrm{NN}(i, j; \pi)$ is within group k, and total orderings $\pi \in \mathcal{T}$ where at least one nearest-neighbor in NN is outside group k. We show $\widetilde{b}_k = \widehat{b}^{\mathrm{t}}_k$ in the first case, and then show that the second case happens with low probability.

We consider any group $k \in [r]$, and any element in the validation set of group k, that is, $(i, j) \in G^{\mathrm{v}}_k$. Let $\mathcal{T}_{\mathrm{in}} \subseteq \mathcal{T}$ denote the subset of total orderings where the nearest-neighbor set $\mathrm{NN}(i, j; \pi)$ is contained within group k:

$$\mathcal{T}_{\mathrm{in}} := \{\pi \in \mathcal{T} : \mathrm{NN}(i, j; \pi) \subseteq G^{\mathrm{t}}_k\}.$$

Let $\mathcal{T}_{\mathrm{out}} := \mathcal{T} \setminus \mathcal{T}_{\mathrm{in}}$ denote the subset of total orderings where at least one nearest-neighbor from $\mathrm{NN}(i, j; \pi)$ is from outside group k. It can be verified by symmetry that the value of $\widetilde{B}^{(\lambda)}_{ij}$ is identical for all $(i, j) \in G^{\mathrm{v}}_k$. Recall that we denote this value as $\widetilde{b}_k := \widetilde{B}^{(\lambda)}_{ij}$ for $(i, j) \in G^{\mathrm{v}}_k$.

Case of $\pi \in \mathcal{T}_{\mathrm{in}}$**:** By the definition of $\mathcal{T}_{\mathrm{in}}$, we have $\mathrm{NN}(i, j; \pi) \subseteq G^{\mathrm{t}}_k$. By symmetry, it can be verified that the mean of the nearest-neighbor set of the element (i, j) over $\mathcal{T}_{\mathrm{in}}$ is simply the mean of all training elements in G^{t}_k. That is,

$$\frac{1}{|\mathcal{T}_{\mathrm{in}}|} \sum_{\pi \in \mathcal{T}_{\mathrm{in}}} [\widetilde{B}^{(\lambda)}_{\pi}]_{ij} = \frac{1}{|G^{\mathrm{t}}_k|} \sum_{(i', j') \in G^{\mathrm{t}}_k} \widehat{b}^{(\lambda)}_{i'j'} \overset{\text{(i)}}{=} \widehat{b}^{\mathrm{t}}_k, \tag{11.222}$$

where step (i) is true by the definition of $\widehat{b}^{\mathrm{t}}_k$.

Case of $\pi \in \mathcal{T}_{\mathrm{out}}$**:** We bound the size of $\mathcal{T}_{\mathrm{out}}$. If a nearest-neighbor of the element (i, j) is outside group k, then this nearest-neighbor can only come from group $(k-1)$ or $(k+1)$. First consider the case where a nearest-neighbor is from group $(k-1)$. Assume that the element (i, j) is ranked $t \in [\ell_k]$ within the set G_k of all elements from group k with respect to π. A nearest-neighbor is from group $(k-1)$, only if all elements ranked 1 through $t-1$ are all in the validation set (otherwise there is some training element whose rank is between 1 and $(t-1)$ within group k, and this element is closer to (i, j) than any element from group $(k-1)$, giving a contradiction). Out of the total orderings in $\mathcal{T}$ where (i, j) is ranked t within group k, the fraction of total orderings that the elements ranked 1 through $(t-1)$ within group k are all in the validation set Ω^{v} is:

$$\prod_{i=1}^{t-1} \frac{\ell^{\mathrm{v}}_k - i}{\ell_k - i} \leq \left(\frac{\ell^{\mathrm{v}}_k}{\ell_k}\right)^{t-1} \overset{\text{(i)}}{<} \left(\frac{3}{4}\right)^t,$$

where (i) is true due to (11.21a) from Lemma 11.13. By symmetry, the fraction of $\pi \in \mathcal{T}$ such that (i, j) is placed in each position $t \in [\ell_k]$ is $\frac{1}{\ell_k}$. Therefore, the fraction of total orderings that a nearest-neighbor is from group $(k-1)$ is upper-bounded by:

$$\frac{1}{\ell_k} \sum_{t=1}^{\ell_k} \left(\frac{3}{4}\right)^t \leq \frac{3}{\ell_k} \overset{\text{(i)}}{<} \frac{3}{d c_{\mathrm{f}} n},$$

where inequality (i) holds because $\ell_k = \sum_{i\in[d]} \ell_{ik} > dc_{\mathrm{f}}n$ due to the all c_{f}-fraction assumption. By the same argument, the fraction of total orderings that at least one nearest-neighbor is from group $(k+1)$ is also upper-bounded by $\frac{3}{dc_{\mathrm{f}}n}$. Hence, we have

$$\frac{|\mathcal{T}_{\mathrm{out}}|}{|\mathcal{T}|} < \frac{6}{dc_{\mathrm{f}}n}. \tag{11.223}$$

For any $(i,j) \in G_k^{\mathrm{v}}$, we have

$$\widetilde{b}_k = \frac{1}{|\mathcal{T}|}\left(\sum_{\pi\in\mathcal{T}_{\mathrm{in}}} [\widetilde{B}_\pi^{(\lambda)}]_{ij} + \sum_{\pi\in\mathcal{T}_{\mathrm{out}}} [\widetilde{B}_\pi^{(\lambda)}]_{ij}\right) \overset{\text{(i)}}{=} \frac{1}{|\mathcal{T}|}\left(|\mathcal{T}_{\mathrm{in}}| \cdot \widehat{b}_k^{\mathrm{t}} + \sum_{\pi\in\mathcal{T}_{\mathrm{out}}} [\widetilde{B}_\pi^{(\lambda)}]_{ij}\right),$$

where equality (i) is true by plugging in (11.222). Hence, we have

$$\begin{aligned}
\left|\widetilde{b}_k - \widehat{b}_k^{\mathrm{t}}\right| &= \frac{1}{|\mathcal{T}|}\left|\sum_{\pi\in\mathcal{T}_{\mathrm{out}}} [\widetilde{B}_\pi^{(\lambda)}]_{ij} - \widehat{b}_k^{\mathrm{t}}\right| \\
&\le \frac{1}{|\mathcal{T}|}\sum_{\pi\in\mathcal{T}_{\mathrm{out}}}\left(\left|[\widetilde{B}_\pi^{(\lambda)}]_{ij}\right| + \left|\widehat{b}_k^{\mathrm{t}}\right|\right) \\
&\overset{\text{(i)}}{\le} \frac{2|\mathcal{T}_{\mathrm{out}}|}{|\mathcal{T}|} \max_{i\in[d],j\in[n]}\left|\widehat{b}_{ij}\right| \overset{\text{(ii)}}{\le} \frac{12}{c_{\mathrm{f}}dn} \cdot \max_{i\in[d],j\in[n]}\left|\widehat{b}_{ij}\right|,
\end{aligned}$$

where inequality (i) is true because $[\widetilde{B}_\pi^{(\lambda)}]_{ij}$ and $\widehat{b}_k^{\mathrm{t}}$ are both the mean of $\widehat{B}$ on a subset of its elements, so we have $\left|[\widetilde{B}_\pi^{(\lambda)}]_{ij}\right| \le \max_{i\in[d],j\in[n]}\left|\widehat{b}_{ij}\right|$ and $\left|\widehat{b}_k^{\mathrm{t}}\right| \le \max_{i\in[d],j\in[n]}\left|\widehat{b}_{ij}\right|$. Then step (ii) is true by plugging in (11.223). This completes the proof.

11.11.3 Proof of Corollary 11.23

Fix any $\epsilon > 0$. By the consistency of $\widehat{B}^{(0)}$ from (11.63), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{B}_{ij}^{(0)} - B_{ij}\right| < \frac{\epsilon}{2}, \quad \forall (i,j) \in \Omega^{\mathrm{t}}\right) = 1. \tag{11.224}$$

Since $\widehat{b}_k^{\mathrm{t}}$ and b_k^{t} are simply the mean of $\widehat{B}$ and B over $G_k^{\mathrm{t}} \subseteq \Omega^{\mathrm{t}}$. We have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{b}_k^{\mathrm{t}} - b_k^{\mathrm{t}}\right| < \frac{\epsilon}{2}, \quad \forall k \in [r]\right) = 1. \tag{11.225}$$

For each $k \in [r]$, we have

$$\begin{aligned}
\left|\widetilde{b}_k - b_k^{\mathrm{t}}\right| &\le \left|\widetilde{b}_k - \widehat{b}_k^{\mathrm{t}}\right| + \left|\widehat{b}_k^{\mathrm{t}} - b_k^{\mathrm{t}}\right| \\
&\overset{\text{(i)}}{\le} \frac{12}{c_{\mathrm{f}}dn} \cdot \max_{i\in[d],j\in[n]}\left|\widehat{b}_{ij}\right| + \left|\widehat{b}_k^{\mathrm{t}} - b_k^{\mathrm{t}}\right| \\
&\le \frac{12}{c_{\mathrm{f}}dn}\left(\max_{i\in[d],j\in[n]}|b_{ij}| + \max_{i\in[d],j\in[n]}\left|b_{ij} - \widehat{b}_{ij}\right|\right) + \left|\widehat{b}_k^{\mathrm{t}} - b_k^{\mathrm{t}}\right|,
\end{aligned} \tag{11.226}$$

where (i) is true by combining Lemma 11.22. In (11.226), we bound the term $\max_{i\in[d],j\in[n]}|b_{ij}|$ by Lemma 11.12 as

$$\lim_{n\to\infty}\mathbb{P}\left(\max_{i\in[d],j\in[n]}|b_{ij}|<2\sqrt{\log dn}\right)=1. \tag{11.227}$$

We bound the term $\max_{i\in[d],j\in[n]}\left|b_{ij}-\widehat{b}_{ij}\right|$ by (11.224), and the term $\left|\widehat{b}^{\mathrm{t}}_k-b^{\mathrm{t}}_k\right|$ by (11.225). Hence, plugging (11.227), (11.224) and (11.225) into (11.226), we have

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\widetilde{b}_k-b^{\mathrm{t}}_k\right|\le\frac{12}{c_{\mathrm{f}}dn}\left(2\sqrt{\log dn}+\frac{\epsilon}{2}\right)+\frac{\epsilon}{2},\quad\forall k\in[r]\right)=1.$$

Equivalently,

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\widetilde{b}_k-b^{\mathrm{t}}_k\right|\le\epsilon,\quad\forall k\in[r]\right)=1,$$

completing the proof.

11.11.4 Proof of Lemma 11.24

We fix any training-validation split $(\Omega^{\mathrm{t}},\Omega^{\mathrm{v}})$ and fix any $\epsilon>0$ and $\delta>0$. We first condition on any value of the bias as $B=B^*$. Then the bias terms in G^{v}_{ik} (whose mean is b^{v}_{ik}) can be considered as randomly sampling ℓ^{v}_{ik} values from the ℓ_k terms in G_k (whose mean is b_k). Denote $\Delta_{B^*}:=\max_{i\in[d],j\in[n]}b^*_{ij}-\min_{i\in[d],j\in[n]}b^*_{ij}$, and denote $\Delta_B:=\max_{i\in[d],j\in[n]}b_{ij}-\min_{i\in[d],j\in[n]}b_{ij}$. By Hoeffding's inequality without replacement [86, Section 6], we have

$$\mathbb{P}\left(|b^{\mathrm{v}}_{ik}-b^*_k|>\Delta_{B^*}\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\ell^{\mathrm{v}}_{ik}}}\;\middle|\;B=B^*\right)\le2\exp\left(-\frac{2\ell^{\mathrm{v}}_{ik}\Delta^2_{B^*}\log\left(\frac{1}{\delta}\right)}{\ell^{\mathrm{v}}_{ik}\Delta^2_{B^*}}\right)=2\delta^2\overset{\text{(i)}}{<}\frac{\delta}{2}, \tag{11.228}$$

where inequality (i) is true for any $\delta\in(0,\frac{1}{4})$. Invoking (11.20a) from Lemma 11.13 and using the all c_{f}-fraction assumption, we have

$$\ell^{\mathrm{v}}_{ik}\ge\frac{\ell_{ik}}{4}>\frac{c_{\mathrm{f}}n}{4}. \tag{11.229}$$

Combining (11.228) with (11.229), we have that for any $\delta\in(0,\frac{1}{4})$,

$$\mathbb{P}\left(|b^{\mathrm{v}}_{ik}-b^*_k|>2\Delta_{B^*}\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{c_{\mathrm{f}}n}}\;\middle|\;B=B^*\right)<\frac{\delta}{2}. \tag{11.230}$$

Now we analyze the term Δ_B in (11.230). By Lemma 11.12, there exists integer n_0 such that for any $n\ge n_0$,

$$\mathbb{P}\left(\Delta_B\le4\sqrt{\log dn}\right)\ge1-\frac{\delta}{2}. \tag{11.231}$$

Let n_1 be a sufficiently large constant such that $n_1 \geq n_0$ and $8\sqrt{\log dn} \cdot \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{c_{\text{f}} n}} < \epsilon$. Then combining (11.231) with (11.230), for any $n \geq n_1$,

$$\begin{aligned}
\mathbb{P}\Big(|b^{\text{v}}_{ik} - b_k| < \epsilon\Big) &= \int_{B^* \in \mathbb{R}^{d\times n}} \mathbb{P}\Big(|b^{\text{v}}_{ik} - b_k| < \epsilon \;\Big|\; B = B^*\Big) \cdot \mathbb{P}(B^*)\, \mathrm{d}B^* \\
&\geq \int_{\substack{B^* \in \mathbb{R}^{d\times n} \\ \Delta_{B^*} \leq 4\sqrt{\log dn}}} \mathbb{P}\Big(|b^{\text{v}}_{ik} - b_k| < \epsilon \;\Big|\; B\Big) \cdot \mathbb{P}(B)\, \mathrm{d}B^* \\
&\overset{\text{(i)}}{\geq} \left(1 - \frac{\delta}{2}\right) \cdot \mathbb{P}\Big(\Delta_B \leq \sqrt{4 \log dn}\Big) \\
&\overset{\text{(ii)}}{\geq} \left(1 - \frac{\delta}{2}\right)^2 \geq 1 - \delta,
\end{aligned}$$

where inequality (i) is true by (11.230) and inequality (ii) is true by (11.231). Equivalently, we have

$$\lim_{n\to\infty} \mathbb{P}\Big(|b^{\text{v}}_{ik} - b_k| < \epsilon\Big) = 1. \tag{11.232}$$

Due to the all c-fraction assumption, the number of groups is upper-bounded as $r \leq \frac{1}{c_{\text{f}}}$. Taking a union bound of (11.232) over $i \in [d], k \in [r]$, we have

$$\lim_{n\to\infty} \mathbb{P}\Big(|b^{\text{v}}_{ik} - b_k| < \epsilon, \quad \forall i \in [d], k \in [r]\Big) = 1,$$

completing the proof of (11.69a). A similar argument yields (11.69b), where in (11.229) we invoke (11.21b) from Lemma 11.13 instead of (11.20a).

11.11.5 Proof of Lemma 11.26

In the proof, we use the following lemma.

Lemma 11.38. *Let $d \geq 1$ be an integer. For any $y \in \mathbb{R}^d$, we have*

$$\operatorname*{argmin}_{u\in\mathcal{M}} \|y - u\|_2^2 + \lambda\|u\|_2^2 = \operatorname*{argmin}_{u\in\mathcal{M}} \|\Pi_{\mathcal{M}}(y) - u\|_2^2 + \lambda\|u\|_2^2 \tag{11.233}$$

The proof of Lemma 11.38 is presented at the end of this section. We now derive a the closed-form solution to (11.233). Consider the optimization problem on the RHS of (11.233). We take the derivative of the objective with respect to u, and solve for u by setting the derivative to 0. It can be verified that the unconstrained solution u^*_{un} to the RHS of (11.233) is:

$$u^*_{\text{un}} = \frac{1}{1+\lambda} \Pi_{\mathcal{M}}(y). \tag{11.234}$$

Note that this unconstrained solution u^*_{un} satisfies $u^*_{\text{un}} \in \mathcal{M}$, so u^*_{un} is also the (constrained) solution to (11.233). Plugging (11.234) to the objective on the LHS of (11.233) and rearranging the terms complete the proof.

Algorithm 6: The Pool-Adjacent-Violators algorithm (PAVA). Input: $y \in \mathbb{R}^d$.

```
1 Initialize $u = y$
2 Initialize the partition $P = \{S_1, \ldots, S_d\}$, where $S_i = \{i\}$ for every $i \in [d]$.
3 while $u \notin \mathcal{M}$ do
4     Find any $i \in [d]$ such that $u_i > u_{i+1}$.
5     Find $S, S' \in P$ such that $i \in S$ and $i + 1 \in S'$.
6     Update $u_r \leftarrow \frac{1}{|S|+|S'|}(\sum_{i \in S} u_i + \sum_{i \in S'} u_i)$ for each $r \in S \cup S'$.
7     Update the partition as $P \leftarrow P \setminus \{S, S'\} + \{S \cup S'\}$.
8 end
9 return $u$
```

Proof of Lemma 11.38 We apply induction on the Pool-Adjacent-Violators algorithm (PAVA) [12, Section 1.2]. For completeness, the Pool-Adjacent-Violators algorithm is shown in Algorithm 6. For any integer $d \geq 1$ and any input $y \in \mathbb{R}^d$, PAVA returns $\operatorname{argmin}_{u \in \mathcal{M}} \|y - u\|_2^2$.

Assume that the while loop in Algorithm 6 is executed T times. Let $u^{(0)} \to u^{(1)} \to \ldots \to u^{(T)}$ be any sequence of the value of x obtained in Algorithm 6. We have $u^{(0)} = y$ and $u^{(T)} = \Pi_{\mathcal{M}} y$. In what follows, we show that for any $0 \leq t \leq T - 1$,

$$\operatorname*{argmin}_{u \in \mathcal{M}} \|u^{(t)} - u\|_2^2 + \lambda \|u\|_2^2 = \operatorname*{argmin}_{u \in \mathcal{M}} \|u^{(t+1)} - u\|_2^2 + \lambda \|u\|_2^2. \tag{11.235}$$

By induction on (11.235), we have

$$\operatorname*{argmin}_{u \in \mathcal{M}} \|u^{(0)} - u\|_2^2 + \lambda \|u\|_2^2 = \operatorname*{argmin}_{u \in \mathcal{M}} \|u^{(T)} - u\|_2^2 + \lambda \|u\|_2^2. \tag{11.236}$$

Combining (11.236) with the fact that $u^{(0)} = y$ and $u^{(T)} = \Pi_{\mathcal{M}} y$ completes the proof.

Proof of (11.235)**:** Consider any t such that $0 \leq t \leq T - 1$. We consider Line 4-6 of PAVA in Algorithm 6. For clarity of notation, we denote the partition corresponding to $u^{(t)}$ as $P^{(t)}$ and the partition corresponding to $u^{(t+1)}$ as $P^{(t+1)}$. Then we have $S, S' \in P^{(t)}$ and $S \cup S' \in P^{(t+1)}$.

First, by PAVA it is straightforward to verify that S and S' both contain consecutive indices. That is, there exists integers m_1, m_2 such that $1 \leq m_1 \leq i < m_2 \leq d$, such that

$$S = \{m_1, \ldots, i\}$$
$$S' = \{i + 1, \ldots, m_2\}.$$

Furthermore, by PAVA it can be verified that

$$a := u_i^{(t)} = u_{i'}^{(t)} \qquad \forall i, i' \in S \tag{11.237a}$$

$$b := u_i^{(t)} = u_{i'}^{(t)} \qquad \forall i, i' \in S' \tag{11.237b}$$

$$z := u_i^{(t+1)} = u_{i'}^{(t+1)} \qquad \forall i, i' \in S \cup S'. \tag{11.237c}$$

Denote these values in (11.237) as a, b and z, respectively. By the update of u in Line 6 of Algorithm 6, we have the relation

$$z = \frac{1}{|S| + |S'|} \left(|S| \cdot a + |S'| \cdot b \right). \tag{11.238}$$

Denote $u^{*(t)}$ and $u^{*(t+1)}$ as the minimizer to the LHS and RHS of (11.235), respectively. Using (11.237), it can be verified that

$$\begin{aligned} a^* :=& u_i^{*(t)} = u_{i'}^{*(t)} \qquad \forall i, i' \in S && (11.239a)\\ b^* :=& u_i^{*(t)} = u_{i'}^{*(t)} \qquad \forall i, i' \in S' && (11.239b)\\ & u_i^{*(t+1)} = u_{i'}^{*(t+1)} \qquad \forall i, i' \in S \cup S'. && (11.239c) \end{aligned}$$

Denote the values in (11.239a) and (11.239b) as a^* and b^*, respectively.

We now show that $a^* = b^*$. Assume for contradiction that $a^* \neq b^*$. Since the solution $u^{*(t)} \in \mathcal{M}$, we have $a^* \leq b^*$. Hence, we have $a^* < b^*$. By Line 4 of Algorithm 6, we have $a > b$. We construct the alternative solution

$$v_i^{*(t)} = \begin{cases} u_i^{*(t)} & i \notin S \cup S \\ \frac{1}{|S|+|S'|}(|S| \cdot a^* + |S| \cdot b^*) & i \in S \cup S'. \end{cases}$$

It can be verified that $v^{*(t)}$ attains a strict strictly smaller objective than $u^{*(t)}$ for the objective on the LHS of (11.235). Contradiction to the assumption that $u^{*(t)}$ is the minimizer to the LHS of (11.235). Hence, we have $a^* = b^*$, implying

$$u_i^{*(t)} = u_{i'}^{*(t)} \qquad \forall i, i' \in S \cup S'.$$

The LHS of (11.235) is equivalent to

$$\begin{aligned} &\underset{\substack{u \in \mathcal{M}, t \in \mathbb{R} \\ t = u_i,\, \forall i, i' \in S \cup S'}}{\operatorname{argmin}} \sum_{i \notin S \cup S'} (u_i^{(t)} - x_i)^2 + \sum_{i \in S \cup S'} (u_i^{(t)} - x_i)^2 + \lambda \|u\|_2^2 \\ &\underset{\substack{u \in \mathcal{M} \\ t = u_i,\, \forall i, i' \in S \cup S'}}{\operatorname{argmin}} \sum_{i \notin S \cup S'} (u_i^{(t)} - x_i)^2 + \underbrace{|S| \cdot (a-t)^2 + |S'| \cdot (b-t)^2}_{T} + \lambda \|u\|_2^2. \end{aligned} \tag{11.240}$$

We write the term T as

$$\begin{aligned} T &= |S| \cdot a^2 + |S'| \cdot b^2 - 2 \left(|S| \cdot a + |S'| \cdot b \right) \cdot t + \left(|S| + |S'| \right) \cdot t^2 \\ &= \left(|S| + |S'| \right) \cdot \left(\frac{|S| \cdot a + |S'| b}{|S| + |S'|} - t \right)^2 + \text{term}(a, b, S, S') \\ &\overset{\text{(i)}}{=} \left(|S| + |S'| \right) \cdot (z - t)^2 + \text{term}(a, b, S, S'), \end{aligned} \tag{11.241}$$

where equality (i) is true by (11.238).

Using the relation $u_i^{(t)} = u_i^{(t+1)}$ for every $i \notin S \cup S'$, the RHS of (11.235) is equivalent to

$$\begin{aligned}
&\underset{\substack{u\in\mathcal{M}, t\in\mathbb{R}\\ t=u_i,\, \forall i\in S\cup S'}}{\operatorname{argmin}} \sum_{i\notin S\cup S'} (u_i^{(t+1)} - x_i)^2 + \sum_{i\in S\cup S'} (u_i^{(t+1)} - x_i)^2 + \lambda\|u\|_2^2 \\
&\underset{\substack{u\in\mathcal{M}, t\in\mathbb{R}\\ t=u_i,\, \forall i\in S\cup S'}}{\operatorname{argmin}} \sum_{i\notin S\cup S'} (u_i^{(t)} - x_i)^2 + (|S| + |S'|)\cdot(z-t)^2 + \lambda\|u\|_2^2. \qquad (11.242)
\end{aligned}$$

The equivalence of the LHS and RHS of (11.235) can be verified by combining (11.240), (11.241), and (11.242).

11.11.6 Proof of Lemma 11.27

Let $c' > 0$ be a constant. Denote $E_{c',c}$ as the event that the number of non-overlapping pairs in S_c (instead of $S_c \cap \Omega^{\text{v}}$ defined for the event $E^{\text{v}}_{c',c}$) is at least $c'n$. We delegate the main part of this proof to the following lemma.

Lemma 11.39. *Suppose $d = 2$. Assume the bias is distributed according to assumption (A2) with $\sigma = 1$. For any $c > 0$, there exists a constant $c' > 0$ such that*

$$\lim_{n\to\infty} \mathbb{P}\left(E_{c',c} \cap E_2\right) = \lim_{n\to\infty} \mathbb{P}(E_2).$$

The proof this result is provided at the end of this section. We first explain how to complete the proof of Lemma 11.27 given Lemma 11.39. The proof of Lemma 11.39 is presented at the end of this section.

Conditional on $E_{c',c}$, consider the $c'n$ non-overlapping pairs in S_c. We denote this subset of non-overlapping pairs as S''. For each $t \in [\frac{n}{2}]$ in Lines 5-7 in Algorithm 2, consider the elements $(1, j^{(2t-1)})$ and $(1, j^{(2t)})$ in Line 6 of Algorithm 2. If both $(1, j^{(2t-1)})$ and $(1, j^{(2t)})$ are involved in some pairs in S'', then we arbitrarily remove one of the pairs involving either $(1, j^{(2t-1)})$ or $(1, j^{(2t)})$ from S''. After the removal, the size of the remaining S'' is at least $\frac{c'n}{2}$. We repeat the same procedure to consider the elements $(2, j^{(2t-1)})$ and $(2, j^{(2t)})$ and remove elements. After this second removal, the size of the remaining S'' is at least $\frac{c'n}{4}$. We now denote this set of non-overlapping pairs after the two removals as S''. Now consider any remaining pair $(j, j') \in S''$. The probability of $(1, j) \in \Omega^{\text{v}}$ is $\frac{1}{2}$ and the probability of $(2, j') \in \Omega^{\text{v}}$ is $\frac{1}{2}$. Hence, the probability of $(j, j') \in S'' \cap \Omega^{\text{v}}$ is $\frac{1}{4}$. Due to the removal, all of the elements involved in S'' appear in different pairs during the training-validation split in Lines 5-7 in Algorithm 2. Hence, the probability of $(j, j') \in \Omega^{\text{v}}$ is independent for each pair $(j, j') \in S''$. By Hoeffding's inequality, we have

$$\lim_{n\to\infty} \mathbb{P}\left(|S'' \cap \Omega^{\text{v}}| \geq \frac{c'n}{32} \,\middle|\, E_{c',c} \right) = 1.$$

That is,

$$\lim_{n\to\infty} \mathbb{P}\left(E^{\text{v}}_{\frac{c'}{32},c} \,\middle|\, E_{c',c} \right) = 1. \qquad (11.243)$$

Hence, we have

$$\begin{aligned}\mathbb{P}(E^{\text{v}}_{\frac{c'}{32},c} \cap E_2) &\geq \mathbb{P}(E^{\text{v}}_{\frac{c'}{32},c} \cap E_{c',c} \cap E_2)\\ &= \mathbb{P}(E_{c',c} \cap E_2) - \mathbb{P}(\overline{E^{\text{v}}_{\frac{c'}{32},c}} \cap E_{c',c} \cap E_2)\\ &\geq \mathbb{P}(E_{c',c} \cap E_2) - \mathbb{P}(\overline{E^{\text{v}}_{\frac{c'}{32},c}} \cap E_{c',c}). \end{aligned} \tag{11.244}$$

Taking the limit of $n \to \infty$ in (11.244), we have

$$\lim_{n\to\infty} \mathbb{P}(E^{\text{v}}_{\frac{c'}{32},c} \cap E_2) \overset{\text{(i)}}{\geq} \lim_{n\to\infty} \mathbb{P}(E_2),$$

where inequality (i) is true by combining Lemma 11.39 and (11.243), completing the proof of Lemma 11.27. It remains to prove Lemma 11.39.

Proof of Lemma 11.39 Recall the definition (11.103) of $S_c = \{(j,j') \in [n]^2 : 0 < b_{2j'} - b_{1j} < c\}$. We first convert the constraint $0 < b_{2j'} - b_{1j} < c$ to a constraint on the ranks of the elements $(1,j)$ and $(2,j')$.

Recall that g denotes the p.d.f. of $\mathcal{N}(0,1)$. Recall that $t(ij)$ is the rank of the element (i,j) (in the total ordering of all $2n$ elements since we assume $d=2$). For any constant $\gamma \in (0, 1/2)$, we define the following set of pairs:

$$R_{\gamma,c} = \left\{ \begin{array}{ll} (j,j') \in [n]^2 : & \gamma n < t_{1j} < t_{2j'} < (2-\gamma)n, \\ & t_{2j'} - t_{1j} \leq cg(\frac{\gamma}{2})n \end{array} \right\}.$$

The following lemma shows that $R_{\gamma,c}$ is a subset of S_c for each $\gamma > 0$ with high probability, and therefore we only need to lower-bound the number of non-overlapping pairs in $R_{\gamma,c}$.

Lemma 11.40. *For each $c > 0$, for any $\gamma \in \left(0, \frac{1}{2}\right)$, we have*

$$\lim_{n\to\infty} \mathbb{P}\Big(R_{\gamma,c} \subseteq S_{2c}\Big) = 1.$$

The proof of this result is provided in Appendix 11.11.7. Denote $E_{\gamma,c',c}$ as the event that the set $R_{\gamma,c}$ contains at least $c'n$ non-overlapping pairs. We have that $E_{\gamma,c',c}$ is deterministic (depending on γ, c', c and the total ordering π). Then Lemma 11.40 implies that for any $\gamma \in \left(0, \frac{1}{2}\right)$ and any $c' \in (0,1)$,

$$\lim_{n\to\infty} \mathbb{P}\Big(E_{\gamma,c',c} \cap \overline{E_{c',2c}}\Big) = 0. \tag{11.245}$$

In what follows, we establish that there exists $\gamma > 0$ and $c' > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\Big(\overline{E_{\gamma,c',c}} \cap E_2\Big) = 0, \tag{11.246}$$

where the choices of γ and c' are specified later.

Proof of (11.246)**:** Assume there exists maximally t such non-overlapping pairs in $R_{\gamma,c}$ (that is, $R_{\gamma,c}$ does not have any subset of non-overlapping pairs of size greater than t). Assume for contradiction that

$$t < \min\left\{\frac{cg(\frac{\gamma}{2})}{2}, \gamma\right\} \cdot n. \tag{11.247}$$

We "remove" these t pairs from the total ordering of $2n$ elements, and then there are $2(n-t)$ remaining elements after the removal. In what follows, we derive a contradiction by using the fact that theses elements are not in $R_{\gamma,c}$.

Denote the ranks corresponding to the remaining elements from course 2 with rank between $(\gamma n, (2-\gamma)n]$ as $j_1 < \ldots < j_T$. Since t elements are removed from each course, we have

$$T \leq n - t. \tag{11.248}$$

Since there are $(n-t)$ remaining elements in course 2, and the number of elements whose rank is outside the range $(\gamma n, (2-\gamma)n]$ is $2\gamma n$, we also have $T \geq n - t - 2\gamma n > 0$. Denote the difference of the ranks between adjacent remaining elements in course 2 as

$$\ell_i = \begin{cases} j_1 - \gamma n - 1 & \text{if } i = 0 \\ j_{i+1} - j_i - 1 & \text{if } 1 \leq i \leq T-1 \\ (2-\gamma)n - j_i & \text{if } i = T. \end{cases} \tag{11.249}$$

The definition (11.249) of ℓ is also visualized in Fig. 11.2.

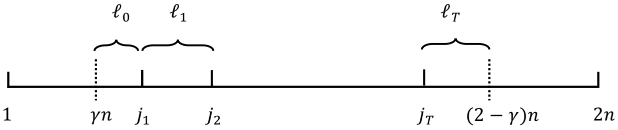


Figure 11.2: The definition (11.249) of ℓ.

By in the definition of (11.249), we have

$$\sum_{i=0}^{T} \ell_i = (2-2\gamma)n - T \overset{\text{(i)}}{\geq} (1-2\gamma)n + t,$$

where inequality (i) is true by (11.248).

There are also $(n-t)$ remaining elements in course 1. We consider the ranks where these elements can be placed. Again, the number of positions outside the range $(\gamma n, (2-\gamma)n]$ is $2\gamma n$. Therefore, at least $(1-2\gamma)n - t$ elements form course 1 need to placed within the range of $(\gamma n, (2-\gamma)n]$. Inside this range, the $cg\left(\frac{\gamma}{2}\right)n$ ranks before each element in course 2 cannot be placed, because otherwise this element from course 1 and the corresponding element from course 2 form a pair in $R_{\gamma,c}$. Contradiction to the assumption that a maximal subset of non-overlapping pairs has been removed. Hence, inside the range, the number of ranks where elements from course 1 can be placed is

$$\sum_{i=0}^{T-1} \max\left\{\ell_i - cg\left(\frac{\gamma}{2}\right)n, 0\right\} + \ell_T.$$

Since we need to place at least $(1-2\gamma)n-t$ elements from course 1 to these ranks, we have

$$\sum_{i=0}^{T-1} \max\left\{\ell_i - cg\left(\frac{\gamma}{2}\right)n, 0\right\} + \ell_T \geq (1-2\gamma)n - t. \tag{11.250}$$

Now we separately discuss the following two cases.

Case 1: $\ell_i \geq cg\left(\frac{\gamma}{2}\right)n$ for some $0 \leq i \leq T-1$. Then consider the interval $[j_i - cg(\frac{\gamma}{2})n, j_i)$. On the one hand, there cannot be elements from course 2 in this interval, because we define ℓ_i as the difference of ranks between elements j_{i+1} and j_i that are already adjacent among elements in course 2. On the other hand, there cannot be elements j from course 1 in this interval, because otherwise we have $(j, i_i) \in R_{\gamma,c}$. Contradiction to the assumption that the removed subset of non-overlapping pairs is maximal. Hence, all of the $cg\left(\frac{\gamma}{2}\right)n$ elements from this interval $[j_i - cg(\frac{\gamma}{2})n, j_i)$ have been removed, and we have $t \geq \frac{cg(\frac{\gamma}{2})n}{2}$. Contradiction to the assumption (11.247).

Case 2: $\ell_i < cg\left(\frac{\gamma}{2}\right)n$ for all $0 \leq i \leq T-1$. Then inequality (11.250) reduces to

$$\ell_T \geq (1-2\gamma)n - t \overset{\text{(i)}}{\geq} (1-3\gamma)n, \tag{11.251}$$

where inequality (i) is true by the assumption (11.247) that $t < \gamma n$.

In what follows, we consider the construction of ranks of all elements (either removed or not) that maximizes $\sum_{j\in[n]}(b_{2j} - b_{1j})$. Then we show that under the assumption (11.247), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{j\in[n]}(b_{2j} - b_{1j}) < 0\right) = 1.$$

Construction of the ranks: To maximize $\sum_j(b_{2j} - b_{1j})$, we want to assign elements in course 2 to higher ranks, and elements in course 1 to lower ranks. We consider the course assigned to the following ranges of the rank.

- **Ranks** $((2-\gamma)n, 2n]$: The size of this range is $2\gamma n$. We assign elements from the course 2 to these ranks, since these are the highest possible ranks.
- **Ranks** $((1+2\gamma)n, (2-\gamma)n]$**:** The size of this range is $(1-3\gamma)n$. Note that the rank j_T is

  $$\begin{aligned} j_T &\overset{\text{(i)}}{=} (2-\gamma)n - \ell_T \\ &\overset{\text{(ii)}}{\leq} (2-\gamma)n - (1-3\gamma)n = (1+2\gamma)n, \end{aligned}$$

  where equality (i) is true by the definition (11.249), and inequality (ii) is true by (11.251). We consider the number of elements from course 2 in this range, remaining or removed. By the definition of j_T from (11.249) there cannot exist remaining elements from course 2 in this range. The number of removed elements from course 2 is $t \leq \gamma n$ by assumption (11.247). Hence, the number of elements from course 2 in this range is at most γn. The other elements in this range are from course 1. Hence, the number of elements from course 1 in this range is at least $(1-4\gamma)n$. We assign the elements in course 2 to higher ranks than the elements in course 1.

- **Ranks** $[1, (1-2\gamma)n]$ There are $4\gamma n$ elements from course 1, and $(1-2\gamma)n$ elements from course 2 that have not been assigned to ranks. We simply assign the $(1-2\gamma)n$ elements from course 2 to be higher ranks than the $4\gamma n$ elements from course 1.

This construction of ranks is also shown in Fig. 11.3. We denote $S_{1L}, S_{2L}, S_{1H}, S_{2H}$ respectively as the sums of the subset of elements as shown in Fig. 11.3.

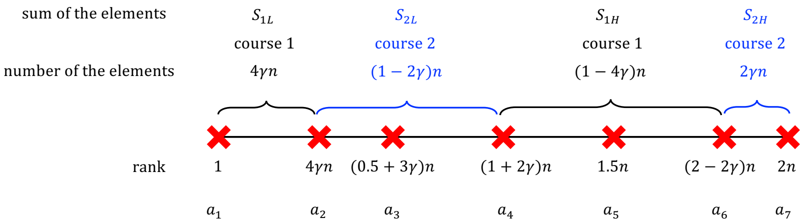


Figure 11.3: Assignment of biases to the 2 courses.

The following lemma now bounds the difference between the sums of the bias in the two courses, under this construction.

Lemma 11.41. *Consider $2n$ i.i.d. samples from $\mathcal{N}(0,1)$, ordered as $X^{(1)} \leq \ldots \leq X^{(2n)}$. Let*

$$\begin{aligned} I_{1L} &:= \{1, \ldots, 4\gamma n\} \\ I_{2L} &:= \{4\gamma n + 1, \ldots, (1+2\gamma)n\} \\ I_{1H} &:= \{(2-2\gamma)n, \ldots, 2n\} \\ I_{2H} &:= \{(2-2\gamma)n, \ldots, 2n\}, \end{aligned}$$

and let

$$\begin{aligned} I_1 &:= I_{1L} \cup I_{1H}, \\ I_2 &:= I_{2L} \cup I_{2H}. \end{aligned}$$

Then there exists some constant $\gamma > 0$, such that

$$\lim_{n\to\infty} \left(\sum_{i \in I_2} X^{(i)} - \sum_{i \in I_1} X^{(i)} < 0 \right) = 1.$$

The proof of this result is provided in Appendix 11.11.8. Denote the constant γ in Lemma 11.41 as γ_0. By Lemma 11.41, we have that under the assumption (11.247) of $t < \min\left\{ \frac{cg\left(\frac{\gamma_0}{2}\right)}{2}, \gamma_0 \right\} n$, then

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{j\in[n]} (b_{2j} - b_{1j}) < 0 \right) = 1.$$

Equivalently, let $c_0' = \min\left\{ \frac{cg\left(\frac{\gamma_0}{2}\right)}{\gamma_0} \right\}$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{E_{\gamma_0, c_0', c}} \cap E_2 \right) = 0,$$

completing the proof of (11.246).

Combining (11.245) **and** (11.246)**:** We have

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left(E_{c_0',c} \cap E_2\right) &= \mathbb{P}(E_2) - \mathbb{P}(E_2 \cap \overline{E_{c',c}})\\ &= \mathbb{P}(E_2) - \mathbb{P}(E_2 \cap \overline{E_{c',c}})\\ &= \mathbb{P}(E_2) - \mathbb{P}(E_2 \cap \overline{E_{c',c}} \cap E_{\gamma_0,c_0',c}) - \mathbb{P}(E_2 \cap \overline{E_{c_0',c}} \cap \overline{E_{\gamma_0,c_0',c}}). \quad (11.252)\end{aligned}$$

Taking the limit of $n \to \infty$ in (11.252), we have

$$\mathbb{P}\left(E_{c_0',c} \cap E_2\right) \overset{\text{(i)}}{=} \lim_{n\to\infty} \mathbb{P}(E_2),$$

where equality (i) is true by combining (11.245) and (11.246). This completes the proof of Lemma 11.39.

11.11.7 Proof of Lemma 11.40

We show that for any $(j, j') \in R_{\gamma,c}$ we have $(j, j') \in S_{2c}$ due to the assumption ((A2)). First, by the definition of $R_{\gamma,c}$ we have $t_{1j} < t_{2j'}$, and hence $b_{2j'} > b_{1j}$. It remains to show that $b_{2j'} - b_{1j} < c$. We denote $(t_0, \ldots, t_T) := (\gamma, \gamma + cg(\frac{\gamma}{2}), \ldots, (2-\gamma))$, where $T = \frac{2-2\gamma}{cg(\frac{\gamma}{2})}$ which is a constant. Recall that $b^{(k\,:\,2n)}$ denotes the k^{th} order statistics among the $2n$ random variables. Recall that G^{-1} denotes the inverse c.d.f. of $\mathcal{N}(0,1)$. By Lemma 11.11 we have

$$b^{(t_i n\,:\,2n)} \xrightarrow{P} G^{-1}\left(\frac{t_i}{2}\right) \qquad \forall 0 \le i \le T. \tag{11.253}$$

Taking a union bound of (11.253) over $0 \le i \le T$, we have

$$\lim_{n\to\infty}\Bigg(\underbrace{\left|b^{(t_i n\,:\,2n)} - G^{-1}\left(\frac{t_i}{2}\right)\right| < \frac{c}{2} \quad \forall 0 \le i \le T}_{E}\Bigg) = 1. \tag{11.254}$$

Denote this event in (11.254) as E. By the definition of $R_{\gamma,c}$, for any $(j, j') \in R_{\gamma,c}$ we have $\gamma n < t_{1j} < t_{2j'} < (2-\gamma)n$ and $t_{2j'} - t_{1j} < cg(\frac{\gamma}{2})n$. Hence, there exists some integer $0 \le i \le T - 2$ such that $t_i n \le t_{1j} < t_{2j'} \le t_{i+2} n$. Conditional on the event E from (11.254), for any $(j, j') \in R_{\gamma,c}$,

$$\begin{aligned} b_{2j'} - b_{1j} \le b^{(t_{i+2}n\,:\,2n)} - b_{(t_i n\,:\,2n)} &< G^{-1}\left(\frac{t_{i+2}}{2}\right) - G^{-1}\left(\frac{t_i}{2}\right) + c\\ &< \frac{(t_{i+2} - t_i)}{2} \cdot \max_{x \in (\frac{\gamma}{2}, 1-\frac{\gamma}{2})} (G^{-1})'(x) + c\\ &\overset{\text{(i)}}{=} cg\left(\frac{\gamma}{2}\right) \cdot \max_{x \in \left(\frac{\gamma}{2}, 1-\frac{\gamma}{2}\right)} \frac{1}{g(x)} + c\\ &= cg\left(\frac{\gamma}{2}\right) \cdot \frac{1}{g\left(\frac{\gamma}{2}\right)} + c = 2c \,\Bigg|\, E. \end{aligned}$$

where (i) holds due to the equality $(G^{-1})'(x) = \frac{1}{G'(x)} = \frac{1}{g(x)}$ for all $x \in (0,1)$. Hence, $R_{\gamma,c} \subseteq S_{2c}$ conditional on E, and we have

$$\lim_{n\to\infty} \mathbb{P}(R_{\gamma,c} \subseteq S_{2c}) \geq \lim_{n\to\infty} \mathbb{P}(E) \overset{\text{(i)}}{=} 1,$$

where equality (i) is true by (11.254), completing the proof.

11.11.8 Proof of Lemma 11.41

We denote the random variables S_{1L}, S_{2L}, S_{1H} and S_{2H} as the sums over I_{1L}, I_{2L}, I_{1H} and I_{2H}, respectively. To bound these sums, we consider the values of $X^{(i)}$ at the following 7 ranks:

$$i \in \{1, 4\gamma n, (0.5+3\gamma)n, (1+2\gamma)n, 1.5n, (2-2\gamma)n, 2n\},$$

as shown by the cross marks in Fig. 11.3. Let $a \in \mathbb{R}^7$. In what follows we condition on the event that

$$\left[X^{(1)}, X^{(4\gamma n)}, X^{((0.5+3\gamma)n)}, X^{((1+2\gamma)n)}, X^{(1.5n)}, X^{((2-2\gamma)n)}, X^{(2n)}\right]^T = a.$$

Denote the expected means of S_{1L}, S_{2L}, S_{1H} and S_{2H} conditional on a as $\mu_{1L|a}, \mu_{2L|a}, \mu_{1H|a}$ and $\mu_{2H|a}$, respectively.

Bounding the sums S_{1L}, S_{2L}, S_{1H} and S_{2H} conditional on a: We first consider the sum S_{2H}. By Hoeffding's inequality, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|S_{1L} - 4\gamma n\mu_{1L|a}\right| < (a_7 - a_1)\sqrt{n\log n} \;\middle|\; a\right) = 1 \tag{11.255a}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\left|S_{2L} - (1-2\gamma) n\mu_{2L|a}\right| < (a_7 - a_1)\sqrt{n\log n} \;\middle|\; a\right) = 1 \tag{11.255b}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\left|S_{1H} - (1-4\gamma) n\mu_{1H|a}\right| < (a_7 - a_1)\sqrt{n\log n} \;\middle|\; a\right) = 1 \tag{11.255c}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\left|S_{2H} - 2\gamma n\mu_{2H|a}\right| < (a_7 - a_1)\sqrt{n\log n} \;\middle|\; a\right) = 1. \tag{11.255d}$$

Taking a union bound of (11.255) and using the equality $\sum_{i\in I_2} X^{(i)} - \sum_{i \in I_1} X^{(i)} = S_{2L} + S_{2H} - S_{1L} - S_{1H}$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{i\in I_2} X^{(i)} - \sum_{i\in I_1} X^{(i)} \right.$$
$$\left. \leq n \left[\underbrace{(1-2\gamma)\mu_{2L|a} - (1-4\gamma)\mu_{1H|a} + 2\gamma\mu_{2H|a} - 4\gamma\mu_{1L|a} + 4(a_7 - a_1)\sqrt{\frac{\log n}{n}}}_{T} \;\middle|\; a \right] \right) = 1.$$

We rearrange the terms in T as

$$T = (1-4\gamma)(\mu_{2L|a} - \mu_{1H|a}) + 4\gamma(\mu_{2H|a} - \mu_{1L|a}) + 2\gamma(\mu_{2L|a} - \mu_{2H|a}) + 4(a_7 - a_1)\sqrt{\frac{\log n}{n}}. \tag{11.256}$$

In what follows, we define a range A on the values of a, show that $\lim_{n\to\infty} \mathbb{P}(a \in A) = 1$ and show that $T < 0$ conditional on any $a \in A$.

Defining the range A and showing $\lim_{n\to\infty}\mathbb{P}(a\in A)=1$: We define the range $A\subseteq\mathbb{R}^7$ as

$$A := \left\{\begin{array}{l} a_1 < G^{-1}(1.5\gamma) \\ a_2 > G^{-1}(1.99\gamma) \\ a_3 < G^{-1}(0.25+1.5\gamma)+0.01 \\ a_5 > G^{-1}(0.75)-0.01 \\ a_6 < G^{-1}(1-0.99\gamma) \\ a_7 > G^{-1}(1-0.5\gamma) \end{array}\right\} \cap \left\{\begin{array}{l} a_1 > -2\sqrt{\log 2n} \\ a_7 < 2\sqrt{\log 2n} \end{array}\right\}. \tag{11.257}$$

By Lemma 11.11, we have

$$a_2 \xrightarrow{P} G^{-1}(2\gamma) \tag{11.258a}$$

$$a_3 \xrightarrow{P} G^{-1}(0.25+1.5\gamma) \tag{11.258b}$$

$$a_5 \xrightarrow{P} G^{-1}(0.75) \tag{11.258c}$$

$$a_6 \xrightarrow{P} G^{-1}(1-\gamma). \tag{11.258d}$$

Moreover, for the extremal values a_1 and a_7, we have that for any $c\in\mathbb{R}$,

$$\lim_{n\to\infty}\mathbb{P}(a_1 < c) = 1 \tag{11.259a}$$

$$\lim_{n\to\infty}\mathbb{P}(a_7 > c) = 1. \tag{11.259b}$$

Combining (11.258), (11.259) and Lemma 11.12, we have that for any $\gamma > 0$,

$$\lim_{n\to\infty}\mathbb{P}(E) = 1.$$

Analyzing the expected means $\mu_{1L|a}, \mu_{2L|a}, \mu_{1H|a}, \mu_{2H|a}$: We analyze the terms on the RHS of (11.256).

Term $(\mu_{2L|a}-\mu_{1H|a})$: We have $\mu_{2L}\leq\frac{a_3+a_4}{2}$ and $\mu_{1H}\geq\frac{a_4+a_5}{2}$. Therefore, conditional on any $a\in A$, for any $\gamma < 0.1$,

$$\mu_{2L|a}-\mu_{1H|a} \leq \frac{a_3-a_5}{2} \overset{\text{(i)}}{\leq} -0.5, \tag{11.260}$$

where inequality (i) is true by the definition (11.257) of A.

Term $(\mu_{2H}-\mu_{1L})$: Let X denote a random variable of $\mathcal{N}(0,1)$. Conditional on any $a\in A$,

$$\begin{aligned} \mu_{2H|a} &= \frac{1}{\sqrt{2\pi}}\frac{1}{\mathbb{P}\left(a_6 < X < a_7\right)}\int_{a_6}^{a_7} xe^{-\frac{x^2}{2}}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{\mathbb{P}\left(a_6 < X < a_7\right)}\left[-e^{-\frac{x^2}{2}}\right]_{x=a_6}^{a_7} \\ &\leq \frac{1}{\sqrt{2\pi}}\frac{1}{\mathbb{P}\left(a_6 < X < a_7\right)}e^{-\frac{a_6^2}{2}} \\ &\overset{\text{(i)}}{\leq} \frac{1}{\sqrt{2\pi}}\frac{1}{0.49\gamma}e^{-\frac{\left[G^{-1}(1-0.99\gamma)\right]^2}{2}}, \end{aligned} \tag{11.261a}$$

where (i) is true by the definition (11.257) of A. Similarly, conditional on the event E and on any a,

$$\mu_{1L|a} > -\frac{1}{\sqrt{2\pi}}\frac{1}{0.49\gamma}e^{-\frac{\left[G^{-1}(1.99\gamma)\right]^2}{2}}. \tag{11.261b}$$

Term: $(\mu_{2L|a} - \mu_{2H|a})$**:** For any $a \in \mathbb{R}^7$, we have

$$(\mu_{2L|a} - \mu_{2H|a}) < 0. \tag{11.262}$$

Showing $T < 0$**:** Plugging the three terms from (11.260), (11.261) and (11.262) back to (11.256), conditional on any $a \in A$,

$$T < -0.5(1-4\gamma) + 4\cdot\frac{1}{\sqrt{2\pi}}\frac{1}{0.49}\left(e^{-\frac{[G^{-1}(1-0.99\gamma)]^2}{2}} + e^{-\frac{[G^{-1}(1.99\gamma)]^2}{2}}\right) + 8\sqrt{\log n}\sqrt{\frac{\log 2n}{n}}.$$

As $\gamma \to 0$, we have $G^{-1}(1.99\gamma) \to -\infty$ and $G^{-1}(1-0.99\gamma) \to \infty$. It can be verified that there exists some sufficiently small $\gamma_0 > 0$, such that

$$\lim_{n\to\infty} T < 0 \,\Big|\, a \in A.$$

Hence, we have

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}\left(\sum_{i\in I_2}X^{(i)} - \sum_{i\in I_1}X^{(i)} \leq 0\right) &\geq \lim_{n\to\infty}\int_{a\in\mathbb{R}^7}\mathbb{P}\left(T<0 \mid a\right)\mathbb{P}(a)\\ &\geq \lim_{n\to\infty}\mathbb{P}(a\in A) = 1,\end{aligned}$$

completing the proof.

11.12 Proof of auxiliary results for Theorem 4.10

In this section, we present the proofs of the auxiliary results for Theorem 4.10.

11.12.1 Proof of Lemma 11.28

First, at $\lambda = \infty$ we have $\widehat{B}^{(\infty)} = 0$ by Proposition 4.7, and hence the claimed result is trivially true.

Now consider any $\lambda \in [0,\infty)$. We fix any value of $Y \in \mathbb{R}^{d\times n}$ and any value of $x \in \mathbb{R}^d$. Denote $U := Y - x\mathbf{1}^T$. By triangle's inequality, we have $\max_{(i,j)\in\Omega}|u_{ij}| \leq \max_{(i,j)\in\Omega}|y_{ij}| + \|x\|_\infty$. It then suffices to establish the inequality

$$\max_{(i,j)\in\Omega}|b_{ij}^{(\lambda)}| \leq \max_{(i,j)\in\Omega}|u_{ij}|,$$

where $B^{(\lambda)}$ is the solution to the optimization

$$\underset{B \text{ satisfies } \mathcal{O}}{\operatorname{argmin}} \|U - B\|_\Omega^2 + \lambda \|B\|_\Omega^2, \tag{11.263}$$

with ties broken by minimizing $\|B\|_\Omega^2$. Assume for contradiction that we have

$$\max_{(i,j)\in\Omega} |b_{ij}^{(\lambda)}| > \max_{(i,j)\in\Omega} |u_{ij}|. \tag{11.264}$$

Denote $u_{\max} := \max_{(i,j)\in\Omega} u_{ij}$ and $u_{\min} := \min_{(i,j)\in\Omega} u_{ij}$. Then we consider an alternative solution B' constructed from $B^{(\lambda)}$ as:

$$b'_{ij} = \begin{cases} \max_{(i,j)\in\Omega} u_{ij} & \text{if } b_{ij}^{(\lambda)} \in (u_{\max}, \infty) \\ b_{ij}^{(\lambda)} & b_{ij}^{(\lambda)} \in [u_{\min}, u_{\max}] \\ \min_{(i,j)\in\Omega} u_{ij} & \text{if } b_{ij}^{(\lambda)} \in (-\infty, u_{\min}). \end{cases}$$

By the assumption (11.264), there exists some $(i, j) \in \Omega$ such that $b_{ij}^{(\lambda)} \notin [u_{\min}, u_{\max}]$. Hence, we have $B' \neq B^{(\lambda)}$. It can be verified that B' satisfies the partial ordering $\mathcal{O}$ because $B^{(\lambda)}$ satisfies $\mathcal{O}$. Furthermore, it can be verified that

$$\|U - B'\|_\Omega^2 < \|U - B^{(\lambda)}\|_\Omega^2$$

and also

$$\|B'\|_\Omega^2 < \|B^{(\lambda)}\|_\Omega^2$$

Hence, B' attains a strictly smaller objective of (11.263) than $B^{(\lambda)}$. Contradiction to the assumption that $\widehat{B}^{(\lambda)}$ is the optimal solution of (11.263).

11.12.2 Proof of Lemma 11.29

Recall that the monotone cone is denoted as $M := \{\theta \in \mathbb{R}^d : \theta_1 \leq \ldots \leq \theta_d\}$, and Π_M denotes the projection (11.8) onto M.

From known results on the monotone cone (see [6, Section 3.5]), we have $\mathbb{E}[\Pi_M Z] \leq c\sqrt{\log d}$ for some fixed constant $c > 0$. Using the Moreau decomposition, we have (see [190, Eq. 20]):

$$\mathbb{E}\left[\sup_{\substack{\|\theta\|_2=1\\ \theta\in M}} \theta^T Z\right] = \mathbb{E}\|\Pi_M Z\|_2 \leq c\sqrt{\log d}.$$

Note that we have the deterministic equality $\sup_{\theta\in M, \|\theta\|_2=1} \theta^T Z \geq 0$ by taking $\theta = 0$. By Markov's inequality, we have

$$\mathbb{P}\left(\sup_{\substack{\|\theta\|_2=1\\ \theta\in M}} \theta^T Z > d^{\frac{1}{4}}\right) \leq \frac{\mathbb{E}\left[\sup_{\theta\in M, \|\theta\|_2=1} \theta^T Z\right]}{d^{\frac{1}{4}}} \leq \frac{c\sqrt{\log d}}{d^{\frac{1}{4}}},$$

completing the proof.

11.12.3 Proof of Lemma 11.30

In the proof, we first bound the event $E_{\frac{1}{36}}$, and then combine the events $E_{\frac{1}{36}}$ and $E'_{\frac{1}{36}}$.

Bounding $E_{\frac{1}{36}}$ We denote the interleaving points in S_{pairs} as $t^{(1)} < \ldots < t^{(|S_{\text{pairs}}|)}$. It can be verified that for any $k \in [|S_{\text{pairs}}| - 1]$, if $t^{(k)} \in S_1$ then then we have $t^{(k+1)} \in S_2$, and vice versa. Hence, we have

$$-1 \leq |S_1| - |S_2| \leq 1. \tag{11.265}$$

By Definition 4.4 of the c_{f}-fraction interleaving assumption, we have

$$|S_1| + |S_2| = |S| \geq c_{\text{f}} n. \tag{11.266}$$

Combining (11.265) and (11.266), we have

$$|S_1|, |S_2| > \frac{c_{\text{f}} n}{3}.$$

Suppose the smallest interleaving point in S_1 is $t_1 := \min S_1$. We now denote the interleaving points in the increasing order of their rank as:

$$\ldots < t_1 < t'_1 < \ldots < t_{\frac{c_{\text{f}} n}{3}} < t'_{\frac{c_{\text{f}} n}{3}} < \ldots .$$

Then we have $t_k \in S_1$ and $t'_k \in S_2$ for all $k \in \left[\frac{c_{\text{f}} n}{3}\right]$.

we construct the set of distinct pairs as:

$$S^{\text{v}} := \left\{ (t_{2k-1}, t'_{2k}) : k \in \left[\frac{c_{\text{f}} n}{6}\right] \right\} \cap (\Omega^{\text{v}} \times \Omega^{\text{v}}).$$

Now we lower-bound the size of S^{v}. For each $k \in \left[\frac{c_{\text{f}} n}{6}\right]$, consider the probability that the pair (t_{2k-1}, t'_{2k}) is in Ω^{v}. It can be verified that the elements of ranks $\{t_{2k-1}\}_{k \in \left[\frac{c_{\text{f}} n}{6}\right]}$ are not adjacent in the sub-ordering of π restricted to course 1, and hence appear in distinct pairs in Line 5-7 of Algorithm 2 when generating the training-validation split of $(\Omega^{\text{t}}, \Omega^{\text{v}})$. Hence, the probability that each element $\{t_{2k-1}\}_{k \in \left[\frac{c_{\text{f}} n}{6}\right]}$ is assigned to Ω^{v} is independently $\frac{1}{2}$. Similarly, the probability that each element $\{t'_{2k}\}_{k \in \left[\frac{c_{\text{f}} n}{6}\right]}$ is assigned to Ω^{v} is $\frac{1}{2}$. Hence, the probability of each pair (t_{2k-1}, t'_{2k}) is assigned to Ω^{v} is $\frac{1}{4}$. By Hoeffding's inequality, we have

$$\lim_{n \to \infty} \mathbb{P}\left(|S^{\text{v}}| > \frac{c_{\text{f}} n}{36} \right) = 1.$$

That is, $\lim_{n \to \infty} \mathbb{P}\left(E_{\frac{1}{36}}\right) = 1$.

Combining $E_{\frac{1}{36}}$ and $E'_{\frac{1}{36}}$ By a similar argument, we have $\lim_{n \to \infty} \mathbb{P}\left(E'_{\frac{1}{36}}\right) = 1$. Taking a union bound of $E_{\frac{1}{36}}$ and $E'_{\frac{1}{36}}$ completes the proof.

11.12.4 Proof of Lemma 11.31

Consider any $T' \in \{S^+ \cap S_1, S^- \cap S_1, S^+ \cap S_2, S^- \cap S_2\}$. Similar to the proof of Lemma 11.30, using the fact that the interleaving points alternate between S_1 and S_2, we have

$$|T'| > \frac{c_\mathrm{f} n}{6}.$$

We write the elements in T' in the increasing order as $k_1 < \ldots < k_{\frac{c_\mathrm{f} n}{6}} < \ldots < k_{|T'|}$. It can be verified that the elements in $\{t_{2k}\}_{k \in \left[\frac{c_\mathrm{f} n}{12}\right]}$ appear in different pairs when generating the training-validation split $(\Omega^\mathrm{t}, \Omega^\mathrm{v})$ in Line 5-7 of Algorithm 2. Hence, each element in $\{t_{2k}\}_{k \in \left[\frac{c_\mathrm{f} n}{12}\right]}$ is assigned to Ω^v independently with probability $\frac{1}{2}$. Using Hoeffding's inequality, we lower-bound the size of $T' \cap \Omega^\mathrm{v}$ as:

$$\lim_{n \to \infty} \mathbb{P}\left(|T' \cap \Omega^\mathrm{v}| > \frac{c_\mathrm{f} n}{36}\right) = 1. \tag{11.267}$$

Taking a union bound of (11.267) over $T' \in \{S^+ \cap S_1, S^- \cap S_1, S^+ \cap S_2, S^- \cap S_2\}$ completes the proof.

Chapter 12

Proofs of Chapter 5

In this section, we present all proofs for results in Section 5.

12.1 Proof of Theorem 5.4

In this chapter, we present the proof of Theorem 5.4. We first introduce notation and preliminaries in Section 12.1.1, to be used subsequently in proving both parts of Theorem 5.4. The proof of Theorem 5.4(b) is presented in Section 12.1.2. The proof of Theorem 5.4(a) is presented in Section 12.1.3. We first present the proof of Theorem 5.4(b) followed by Theorem 5.4(a), because the proof of Theorem 5.4(a) depends on the proof of Theorem 5.4(b).

In the proof of Theorem 5.4(a), the constants are allowed to depend only on the constant B. In the proof of Theorem 5.4(b), the constants are allowed to depend only on the constants A and B. The proofs for all the lemmas are presented in Section 12.1.4.

12.1.1 Notation and preliminaries

In this section, we introduce notation and preliminaries that are used subsequently in the proofs of both Theorem 5.4(b) and Theorem 5.4(a).

(i) Notation

Recall that d denotes the number of items, and k denotes the number of comparisons per pair of items. The d items are associated to a true parameter vector $\theta^* = [\theta_1^*, \ldots, \theta_d^*]$. We have the set $\Theta_B = \{\theta \in \mathbb{R}^d \mid \|\theta\|_\infty \leq B, \sum_{i=1}^d \theta_i = 0\}$ and the set $\Theta_A = \{\theta \in \mathbb{R}^d \mid \|\theta\|_\infty \leq A, \sum_{i=1}^d \theta_i = 0\}$, where A and B are finite constants such that $A > B > 0$. The true parameter vector satisfies $\theta^* \in \Theta_B$.

Denote μ_{ij}^* as the probability that item $i \in [d]$ beats item $j \in [d]$. Under the BTL model, we have

$$\mu_{ij}^* = \frac{1}{1 + e^{-(\theta_i^* - \theta_j^*)}}. \tag{12.1}$$

For every $r \in [k]$, denote the outcome of the r^{th} comparison between item $i \in [d]$ and item $j \in [d]$ as

$$X_{ij}^{(r)} := \mathbb{1}\{\text{item } i \text{ beats item } j \text{ in their } r^{th} \text{ comparison}\}.$$

We have $X_{ij}^{(r)} \sim \text{Bernoulli}(\mu_{ij}^*)$, independent across all $r \in [k]$ and all $i < j$. Recall that W_{ij} denotes the number of times that item i beats j. We have $W_{ij} = \sum_{r=1}^{k} X_{ij}^{(r)}$ and therefore $W_{ij} \sim \text{Binom}(k, \mu_{ij}^*)$. Denote μ_{ij} as the fraction of times that item i beats item j. That is,

$$\mu_{ij} := \frac{1}{k} W_{ij} = \frac{1}{k} \sum_{r=1}^{k} X_{ij}^{(r)}. \tag{12.2}$$

We have $\mu_{ij} \sim \frac{1}{k}\text{Binom}(k, \mu_{ij}^*)$, independent across all $i < j$.

Finally, we use c, c', c_1, c_2, etc. to denote finite constants whose values may change from line to line. We write $f(n) \lesssim g(n)$ if there exists a constant c such that $f(n) \leq c \cdot g(n)$ for all $n \geq 1$. The notation $f(n) \gtrsim g(n)$ is defined analogously.

(ii) Notion of conditioning

Let E be any event. The conditional bias of any estimator $\widehat{\theta}$ conditioned on the event E is defined as:

$$\beta(\widehat{\theta} \mid E) := \sup_{\theta^* \in \Theta_B} \|\mathbb{E}[\widehat{\theta} \mid E] - \theta^*\|_\infty.$$

We use "w.h.p.($\frac{1}{dk}$)" to denote that an event E happens with probability at least

$$\mathbb{P}(E) > 1 - \frac{c}{dk},$$

for all $d \geq d_0$ and $k \geq k_0$, where d_0, k_0 and c are positive constants.

Similarly, we use "w.h.p.($\frac{1}{dk} \mid E$)" to denote that conditioned on some event E, some other event E' happens with probability at least

$$\mathbb{P}(E' \mid E) \geq 1 - \frac{c}{dk},$$

for all $d \geq d_0$ and $k \geq k_0$, where d_0, k_0 and c are positive constants.

(iii) The negative log-likelihood function and its derivative

Recall that ℓ denotes the negative log-likelihood function. Under the BTL model, we have

$$\begin{aligned}\ell(\theta) := \ell(\{W_{ij}\}; \theta) &= -\sum_{1 \leq i < j \leq d} \left[W_{ij} \log\left(\frac{1}{1 + e^{-(\theta_i - \theta_j)}}\right) + W_{ji} \log\left(\frac{1}{1 + e^{-(\theta_j - \theta_i)}}\right)\right] \\ &= -k \sum_{1 \leq i < j \leq d} \left[\mu_{ij} \log\left(\frac{1}{1 + e^{-(\theta_i - \theta_j)}}\right) + \mu_{ji} \log\left(\frac{1}{1 + e^{-(\theta_j - \theta_i)}}\right)\right] \\ &= k \sum_{1 \leq i < j \leq d} \left[\log(e^{\theta_i} + e^{\theta_j}) - \mu_{ij}\theta_i - \mu_{ji}\theta_j\right]. \end{aligned} \tag{12.3}$$

Since $\{\mu_{ij}\}$ is simply a normalized version of $\{W_{ij}\}$, we equivalently denote the negative log-likelihood function as $\ell(\{\mu_{ij}\};\theta)$.

From the expression of ℓ in (12.3), we compute the gradient $\frac{\partial \ell}{\partial \theta_m}$ for every $m \in [d]$ as

$$\frac{\partial \ell}{\partial \theta_m} = k \sum_{i \neq m} \left(\frac{1}{1+e^{-(\theta_m - \theta_i)}} - \mu_{mi} \right). \tag{12.4}$$

Finally, the following lemma from [89] shows the strict convexity of the negative log-likelihood function ℓ.

Lemma 12.1 (Lemma 2(a) from [89])**.** *The negative log-likelihood function $\ell(\theta)$ is strictly convex in $\theta \in \mathbb{R}^d$.*

(iv) The sigmoid function and its derivatives

Denote the function $f : (-\infty, \infty) \to (0,1)$ as the sigmoid function $f(x) = \frac{1}{1+e^{-x}}$. It is straightforward to verify that the function f has the following two properties.

- The first derivative f' is positive on $(-\infty, \infty)$. Moreover, on any bounded interval, the first derivative f' is bounded above and below. That is, for any constants $c_1 < c_2$, there exist constants $c_3, c_4 > 0$ such that

$$0 < c_3 < f'(x) < c_4, \qquad \text{for all } x \in (c_1, c_2). \tag{12.5a}$$

- The second derivative f'' is bounded on any bounded interval. That is, for any constants $c_1 < c_2$, there exists a constant c_5 such that

$$|f''(x)| < c_5, \qquad \text{for all } x \in (c_1, c_2). \tag{12.5b}$$

(v) Existence and uniqueness of MLE

Recall that the MLE (5.3), the unconstrained MLE (5.4), and the stretched-MLE (5.5) are respectively defined as:

$$\widehat{\theta}^{(B)}(\{\mu_{ij}\}) = \operatorname*{argmin}_{\theta \in \Theta_B} \ell(\{\mu_{ij}\};\theta), \tag{12.6}$$

$$\widehat{\theta}^{(\infty)}(\{\mu_{ij}\}) = \operatorname*{argmin}_{\theta \in \Theta_\infty} \ell(\{\mu_{ij}\};\theta), \tag{12.7}$$

$$\widehat{\theta}^{(A)}(\{\mu_{ij}\}) = \operatorname*{argmin}_{\theta \in \Theta_A} \ell(\{\mu_{ij}\};\theta). \tag{12.8}$$

The following lemma shows the existence and uniqueness of the stretched-MLE $\widehat{\theta}^{(A)}$ (12.8) for any constant $A > 0$, which incorporates the standard MLE $\widehat{\theta}^{(B)}$ by setting $A = B$.

Lemma 12.2. *For any finite constant $A > 0$, there always exists a unique solution $\widehat{\theta}^{(A)}$ to the stretched-MLE* (12.8).

See Section 12.1.4 for the proof of Lemma 12.2.

For the unconstrained MLE, due to the removal of the box constraint in (12.7), a finite solution $\widehat{\theta}^{(\infty)}$ may not exist. However, the following lemma shows that a unique finite solution exists with high probability.

Lemma 12.3. *There exists a unique finite solution $\widehat{\theta}^{(\infty)}$ to the unconstrained MLE* (12.7) *w.h.p.*$(\frac{1}{dk})$.

See Section 12.1.4 for the proof of Lemma 12.3.

In the subsequent proofs of Theorem 5.4(b) and Theorem 5.4(a), we heavily use the unconstrained MLE as an intermediate quantity to analyze the MLE and the stretched-MLE.

12.1.2 Proof of Theorem 5.4(b)

In this section, we present the proof of Theorem 5.4(b). To describe the main steps involved, we first present a proof sketch of a simple case of $d = 2$ items (Section 12.1.2), followed by the complete proof of the general case (Section 12.1.2). The reader may pass to the complete proof in Section 12.1.2 without loss of continuity.

Simple case: 2 items

We first present an informal proof sketch for a simple case where there are $d = 2$ items. The proof for the general case in Section 12.1.2 follows the same outline. In the case of $d = 2$ items, due to the centering constraint on the true parameter vector θ^*, we have $\theta_2^* = -\theta_1^*$. Similarly, we have $\widehat{\theta}_2 = -\widehat{\theta}_1$ for any estimator that satisfies the centering constraint (in particular, for the stretched-MLE $\widehat{\theta}^{(A)}$ and the unconstrained MLE $\widehat{\theta}^{(\infty)}$). Therefore, it suffices to focus only on item 1. Since there are only two items, for ease of notation, we denote $\mu = \mu_{12}$ and $\mu^* = \mu_{12}^*$. We now present the main steps of the proof sketch.

Proof sketch of the 2-item case (informal):

In the proof sketch, we fix any $\theta^* \in \Theta_B$, and any finite constants A and B such that $A > B > 0$.

Step 1: Establish concentration of μ

By Hoeffding's inequality, we have

$$|\mu - \mu^*| \lesssim \sqrt{\frac{\log k}{k}}, \qquad \text{w.h.p.} \tag{12.9}$$

Since $|\theta^*| \leq B$, we have that μ^* is bounded away from 0 and 1 by a constant. Hence, for sufficiently large k, there exist constants c_L, c_U where $0 < c_L < c_U < 1$, such that

$$\mu, \mu^* \in (c_L, c_U). \tag{12.10}$$

Step 2: Write the first-order optimality condition for $\widehat{\theta}^{(\infty)}$

The unconstrained MLE $\widehat{\theta}^{(\infty)}$ minimizes the negative log-likelihood ℓ. If a finite unconstrained MLE $\widehat{\theta}^{(\infty)}$ exists[1], we have $\nabla_{\theta=\widehat{\theta}^{(\infty)}} \ell(\theta) = 0$. Setting $m = 1$ in the gradient

[1] For the proof sketch, we ignore the high-probability nature of Lemma 12.3, and assume that a finite $\widehat{\theta}^{(\infty)}$ always exists. It is made precise in the complete proof in Section 12.1.2.

expression (12.4) and plugging in $\widehat{\theta}^{(\infty)}$, we have

$$\begin{aligned} \frac{\partial \ell}{\partial \theta_1}\bigg|_{\theta=\widehat{\theta}^{(\infty)}} &= k\left(\frac{1}{1+e^{-(\widehat{\theta}_1^{(\infty)}-\widehat{\theta}_2^{(\infty)})}} - \mu_{12}\right) \\ &= k\left(\frac{1}{1+e^{-2\widehat{\theta}_1^{(\infty)}}} - \mu\right). \end{aligned} \tag{12.11}$$

Setting the derivative (12.11) to 0, we have

$$\widehat{\theta}_1^{(\infty)} = -\frac{1}{2}\log\left(\frac{1}{\mu}-1\right). \tag{12.12}$$

By the definition of $\{\mu^*_{ij}\}$ in (12.1), we have $\mu^* = \frac{1}{1+e^{-(\theta_1^*-\theta_2^*)}} = \frac{1}{1+e^{-2\theta_1^*}}$, which can be written as

$$\theta_1^* = -\frac{1}{2}\log\left(\frac{1}{\mu^*}-1\right). \tag{12.13}$$

Define a function $h : [0,1] \to \mathbb{R} \cup \{\pm\infty\}$ as

$$h(t) = -\frac{1}{2}\log\left(\frac{1}{t}-1\right). \tag{12.14}$$

Subtracting (12.13) from (12.12) and using the definition of h from (12.14), we have

$$\widehat{\theta}_1^{(\infty)} - \theta_1^* = h(\mu) - h(\mu^*). \tag{12.15}$$

Step 3: Bound the difference between $\widehat{\theta}^{(\infty)}$ and θ^*, by the first-order mean value theorem

It can be verified that h has positive first-order derivative on $(0,1)$. Moreover, there exists some constant c_1 such that $0 < h'(t) < c_1$ for all $t \in (c_L, c_U)$. Applying the first-order mean value theorem on (12.15), we have the deterministic relation

$$\widehat{\theta}_1^{(\infty)} - \theta_1^* = h'(\lambda)\cdot(\mu-\mu^*), \tag{12.16}$$

where λ is a random variable that depends on μ and μ^*, and takes values between μ and μ^*. By (12.10), we have $\lambda \in (c_L, c_U)$. From (12.16) we have

$$|\widehat{\theta}_1^{(\infty)} - \theta_1^*| \leq c_1|\mu-\mu^*|. \tag{12.17}$$

Combining (12.17) with (12.9), we have

$$|\widehat{\theta}_1^{(\infty)} - \theta_1^*| \lesssim \sqrt{\frac{\log k}{k}}, \qquad \text{w.h.p.} \tag{12.18}$$

Step 4: Bound the *expected* difference between $\widehat{\theta}^{(\infty)}$ and θ^*, by the second-order mean value theorem

By the second-order mean value theorem on (12.15), we have the deterministic relation

$$\widehat{\theta}_1^{(\infty)} - \theta_1^* = h(\mu) - h(\mu^*) = h'(\mu^*) \cdot (\mu - \mu^*) + h''(\widetilde{\lambda}) \cdot (\mu - \mu^*)^2, \tag{12.19}$$

where $\widetilde{\lambda}$ is a random variable that depends on μ and μ^*, and takes values between μ and μ^*. By (12.10), we have $\widetilde{\lambda} \in (c_L, c_U)$.

It can be verified that h has bounded second-order derivative. That is, $|h''(t)| < c_2$ for all $t \in (c_L, c_U)$. Taking an expectation over (12.19), we have

$$\begin{aligned}\mathbb{E}[\widehat{\theta}_1^{(\infty)}] - \theta_1^* &= h'(\mu^*) \cdot (\mathbb{E}[\mu] - \mu^*) + \mathbb{E}[h''(\widetilde{\lambda}) \cdot (\mu - \mu^*)^2] \qquad (12.20)\\ &\overset{\text{(i)}}{\leq} c_2 \mathbb{E}[(\mu - \mu^*)^2] \\ &\overset{\text{(ii)}}{\lesssim} \frac{\log k}{k}, \qquad (12.21)\end{aligned}$$

where (i) is true because $\mathbb{E}[\mu] = \mu^*$ combined with the fact that $|h''| < c_2$ on (c_L, c_U), and (ii) is true[2] by (12.9).

Step 5: Connect $\widehat{\theta}^{(\infty)}$ back to $\widehat{\theta}^{(A)}$

From (12.18), we have $|\widehat{\theta}_1^{(\infty)} - \theta_1^*| \leq A - B$ w.h.p. for sufficiently large k. Hence,

$$|\widehat{\theta}_1^{(\infty)}| \leq |\theta_1^*| + |\widehat{\theta}_1^{(\infty)} - \theta_1^*| \leq B + (A - B) = A, \qquad \text{w.h.p.}$$

Moreover, we have $\left|\widehat{\theta}_2^{(\infty)}\right| = \left|\widehat{\theta}_1^{(\infty)}\right| \leq A$. Therefore, with high probability, the unconstrained MLE $\widehat{\theta}^{(\infty)}$ does not violate the box constraint at A, and therefore $\widehat{\theta}^{(\infty)}$ is identical to the stretched-MLE $\widehat{\theta}^{(A)}$. Hence, the bound (12.21) holds[3] for the stretched-MLE, completing the proof sketch.

Complete Proof

In this section, we present the proof of Theorem 5.4(b), by formally extending the 5 steps outlined for the simple case in Section 12.1.2. In the general case, one notable challenge is that one can no longer write a closed-form solution of the MLE as we did in (12.12) of Step 2. The first-order optimality condition now becomes a system of equations that describe an implicit relation between θ and μ, requiring more involved analysis.

In the proof, we fix any $\theta^* \in \Theta_B$, and fix any finite constants A and B such that $A > B > 0$.

Step 1: Establish concentration of $\{\mu_{ij}\}$

We first use standard concentration inequalities to establish the following lemma, to be used in the subsequent steps of the proof.

[2] For the proof sketch, we ignore the high-probability nature of (12.9) and treat it as a deterministic relation. It is made precise in the complete proof in Section 12.1.2.

[3] For the proof sketch, we ignore the high-probability nature of the fact that $\widehat{\theta}^{(\infty)} = \widehat{\theta}^{(A)}$, and treat it as a deterministic relation. It is made precise in the complete proof in Section 12.1.2.

Lemma 12.4. *There exists a constant $c > 0$, such that*

$$\left|\sum_{i\neq m}\mu_{mi} - \sum_{i\neq m}\mu^*_{mi}\right| \leq c\sqrt{\frac{d(\log d + \log k)}{k}},$$

simultaneously for all $m \in [d]$ w.h.p.($\frac{1}{dk}$).

See Section 12.1.4 for the proof of Lemma 12.4.

Recall that Lemma 12.3 states that a finite unconstrained MLE $\widehat{\theta}^{(\infty)}$ exists w.h.p.($\frac{1}{dk}$). We denote E_0 as the event that Lemma 12.3 and Lemma 12.4 both hold. For the rest of the proof, we condition on E_0. Since both Lemma 12.3 and Lemma 12.4 hold w.h.p.($\frac{1}{dk}$), taking a union bound, we have that E_0 holds w.h.p.($\frac{1}{dk}$). That is,

$$\mathbb{P}(E_0) \geq 1 - \frac{c}{dk}, \qquad \text{for some constant } c > 0. \tag{12.22}$$

Step 2: Write the first-order optimality condition for the unconstrained MLE $\widehat{\theta}^{(\infty)}$

Recall from Lemma 12.1 that the negative log-likelihood function ℓ is convex in θ. In this step, we first justify that the whenever a finite unconstrained MLE $\widehat{\theta}^{(\infty)}$ exists, it satisfies the first-order optimality condition $\nabla_{\theta=\widehat{\theta}^{(\infty)}}\ell(\theta) = 0$. (Note that for any optimization problem with constraints, it is in general not true that the derivative of the convex objective equals 0 at the optimal solution.) Then we derive a specific form of the first-order optimality condition, to be used in subsequent steps of the proof.

Given that we have conditioned on E_0 (and therefore on Lemma 12.3), a finite solution $\widehat{\theta}^{(\infty)}$ to the unconstrained MLE exists. To show that $\widehat{\theta}^{(\infty)}$ satisfies the first-order optimality condition, we show that $\widehat{\theta}^{(\infty)}$ is also a solution to the following MLE without any constraint at all (that is, we remove the centering constraint too):

$$\operatorname*{argmin}_{\theta\in\mathbb{R}^d} \ell(\theta). \tag{12.23}$$

If the unconstrained MLE $\widehat{\theta}^{(\infty)}$ is a solution to (12.23), then it satisfies the first-order condition $\nabla_\theta\ell(\widehat{\theta}^{(\infty)}) = 0$. Now we prove that $\widehat{\theta}^{(\infty)}$ is a solution to (12.23). Note that the solutions to (12.23) are shift-invariant. That is, if θ is a solution to (12.23), then $\theta + c\mathbf{1}$ is also a solution, where $\mathbf{1}$ is the d-dimensional all-one vector, and c is any constant. Now suppose by contradiction that $\widehat{\theta}^{(\infty)}$ is not a solution to (12.23). Then there exists some finite $\theta \in \mathbb{R}^d$ such that $\ell(\theta) < \ell(\widehat{\theta}^{(\infty)})$. Now consider $\theta' := \theta - (\frac{1}{d}\sum_{i=1}^d \theta_i)\mathbf{1}$. We have $\theta' \in \Theta_\infty$ because it satisfies the centering constraint, and we have $\ell(\theta') = \ell(\theta) < \ell(\widehat{\theta}^{(\infty)})$ because the solutions to (12.23) are shift-invariant. The construction of θ' thus contradicts the assumption that $\widehat{\theta}^{(\infty)}$ is optimal for the unconstrained MLE. Hence, $\widehat{\theta}^{(\infty)}$ is a solution to (12.23), and $\widehat{\theta}^{(\infty)}$ satisfies the first-order optimality condition.

Now we derive a specific form of the first-order optimality condition. Plugging $\widehat{\theta}^{(\infty)}$ into the gradient expression (12.4) and setting the gradient to 0, we have the deterministic equality

$$\sum_{i\neq m}\frac{1}{1 + e^{-(\widehat{\theta}^{(\infty)}_m - \widehat{\theta}^{(\infty)}_i)}} = \sum_{i\neq m}\mu_{mi}, \qquad \text{for every } m \in [d]. \tag{12.24}$$

In words, the first-order optimality condition (12.24) means that for any item $m \in [d]$, the probability that item m wins (among all comparisons in which item m is involved) as predicted by the unconstrained MLE $\widehat{\theta}^{(\infty)}$ equals the fraction of wins by item m from the observed comparisons. We now subtract (12.1) from both sides of (12.24):

$$\sum_{i \neq m} \left(\frac{1}{1+e^{-(\widehat{\theta}_m^{(\infty)} - \widehat{\theta}_i^{(\infty)})}} - \frac{1}{1+e^{-(\theta_m^* - \theta_i^*)}} \right) = \sum_{i \neq m} (\mu_{mi} - \mu_{mi}^*)$$

$$\sum_{i=1}^{d} \left(\frac{1}{1+e^{-(\widehat{\theta}_m^{(\infty)} - \widehat{\theta}_i^{(\infty)})}} - \frac{1}{1+e^{-(\theta_m^* - \theta_i^*)}} \right) = \sum_{i \neq m} (\mu_{mi} - \mu_{mi}^*). \tag{12.25}$$

For ease of notation, we denote the random vector $\delta := \widehat{\theta}^{(\infty)} - \theta^*$. Equivalently, we have $\widehat{\theta}^{(\infty)} = \theta^* + \delta$. Using the definition of δ, we rewrite (12.25) as:

$$\sum_{i=1}^{d} \left(\frac{1}{1+e^{-(\theta_m^* - \theta_i^* + \delta_m - \delta_i)}} - \frac{1}{1+e^{-(\theta_m^* - \theta_i^*)}} \right) = \sum_{i \neq m} (\mu_{mi} - \mu_{mi}^*). \tag{12.26}$$

Using the definition of the sigmoid function $f(x) = \frac{1}{1+e^{-x}}$, we rewrite (12.26) as:

$$\sum_{i=1}^{d} \left[f(\theta_m^* - \theta_i^* + \delta_m - \delta_i) - f(\theta_m^* - \theta_i^*) \right] = \sum_{i \neq m} (\mu_{mi} - \mu_{mi}^*). \tag{12.27}$$

In the rest of the proof, we primarily work with the first-order optimality condition in the form of (12.27).

Step 3: Bound the difference between the unconstrained MLE $\widehat{\theta}^{(\infty)}$ and the true parameter vector θ^*

The first-order optimality condition (12.27) can be thought of as a system of equations that describes some implicit relation between the unconstrained MLE $\widehat{\theta}^{(\infty)}$ and the observations $\{\mu_{mi}\}$. Intuitively, the concentration of $\{\mu_{mi}\}$ on the RHS of (12.27) (by Lemma 12.4) should imply the concentration of the unconstrained MLE $\widehat{\theta}^{(\infty)}$ on the LHS. The following lemma formalizes this intuition about the concentration of $\widehat{\theta}^{(\infty)}$.

Lemma 12.5. *Conditioned on E_0, we have the deterministic relation*

$$|\delta_m| = |\widehat{\theta}_m^{(\infty)} - \theta_m^*| \lesssim \sqrt{\frac{\log d + \log k}{dk}}, \qquad \textit{for every } m \in [d],$$

for all $d \geq d_0$ and $k \geq k_0$, where d_0 and k_0 are constants.

See Section 12.1.4 for the proof of Lemma 12.5.

This lemma provides a deterministic bound on the difference between $\widehat{\theta}^{(\infty)}$ and θ^*. Now we move to analyze the difference between $\widehat{\theta}^{(\infty)}$ and θ^* in expectation.

Step 4: Bound the *expected* difference between the unconstrained MLE $\widehat{\theta}^{(\infty)}$ and the true parameter vector θ^*, using the second-order mean value theorem

In Step 1 we bound the difference between $\{\mu_{mi}\}$ and $\{\mu^*_{mi}\}$ with high-probability. However, if we consider the difference in expectation, we have $\mathbb{E}[\mu_{mi}] = \mu^*_{mi}$. The *expected* difference between $\{\mu_{mi}\}$ and $\{\mu^*_{mi}\}$ is 0, significantly smaller than the high-probability bound in Step 1. Intuitively, we may also expect that the *expected* difference between $\widehat{\theta}^{(\infty)}$ and θ^* is smaller than the deterministic bound in Lemma 12.5. In this step, we formalize this intuition.

By the second-order mean value theorem on the LHS of the first-order optimality condition (12.27), we have the deterministic relation that for every $m \in [d]$,

$$\sum_{i=1}^{d} \left[f'(\theta^*_m - \theta^*_i) \cdot (\delta_m - \delta_i) + \frac{1}{2} f''(\lambda_{mi}) \cdot (\delta_m - \delta_i)^2 \right] = \sum_{i \neq m} (\mu_{mi} - \mu^*_{mi})$$

$$\sum_{i=1}^{d} f'(\theta^*_m - \theta^*_i) \cdot (\delta_m - \delta_i) = \sum_{i \neq m} (\mu_{mi} - \mu^*_{mi}) - \frac{1}{2} \sum_{i=1}^{d} f''(\lambda_{mi}) \cdot (\delta_m - \delta_i)^2, \tag{12.28}$$

where each λ_{mi} is a random variable that takes values between $\theta^*_m - \theta^*_i$ and $\theta^*_m - \theta^*_i + (\delta_m - \delta_i)$. Taking an expectation over (12.28) conditional on E_0, we have that for every $m \in [d]$:

$$\sum_{i=1}^{d} f'(\theta^*_m - \theta^*_i) \cdot \mathbb{E}\left[\delta_m - \delta_i \mid E_0\right] = \sum_{i \neq m} (\mathbb{E}[\mu_{mi} \mid E_0] - \mu^*_{mi}) - \frac{1}{2} \sum_{i=1}^{d} \mathbb{E}[f''(\lambda_{mi})(\delta_m - \delta_i)^2 \mid E_0]. \tag{12.29}$$

Denote the vector $\Delta := \mathbb{E}[\delta \mid E_0] = \mathbb{E}[\widehat{\theta}^{(\infty)} \mid E_0] - \theta^*$. Plugging this definition of Δ into (12.29) yields

$$\sum_{i=1}^{d} f'(\theta^*_m - \theta^*_i) \cdot (\Delta_m - \Delta_i) = \sum_{i \neq m} (\mathbb{E}[\mu_{mi} \mid E_0] - \mu^*_{mi}) - \frac{1}{2} \sum_{i=1}^{d} \mathbb{E}[f''(\lambda_{mi})(\delta_m - \delta_i)^2 \mid E_0]. \tag{12.30}$$

We first bound the RHS of (12.30), and then derive a bound regarding Δ_i on the LHS accordingly.

To bound the RHS of (12.30), we first consider the term $\mathbb{E}[\mu_{mi} \mid E_0] - \mu^*_{mi}$. In what follows, we state a lemma that is slightly more general than what is needed here. The more general version is used in the subsequent proof of Theorem 5.4(a). To state the lemma, recall the definition that an event E' happens w.h.p.($\frac{1}{dk} \mid E$), if the conditional probability $\mathbb{P}(E' \mid E) \geq 1 - \frac{c}{dk}$, for some constant $c > 0$.

Lemma 12.6. *Let E be any event, and let E' be any event that happens w.h.p.($\frac{1}{dk} \mid E$). Then for any $m \neq i$, we have*

$$|\mathbb{E}[\mu_{mi} \mid E', E] - \mathbb{E}[\mu_{mi} \mid E]| \lesssim \frac{1}{dk}. \tag{12.31}$$

See Section 12.1.4 for the proof of Lemma 12.6.

To apply Lemma 12.6, we set E to be the (trivial) event of the entire probability space, and set E' to be E_0 in (12.31). We have

$$|\mathbb{E}[\mu_{mi} \mid E_0] - \mathbb{E}[\mu_{mi}]| = |\mathbb{E}[\mu_{mi} \mid E_0] - \mu^*_{mi}| \lesssim \frac{1}{dk}. \tag{12.32}$$

The remaining terms in (12.30) are handled in the following lemma. This lemma bounds the expected difference between $\widehat{\theta}^{(\infty)}$ and θ^* conditioned on E_0, that is, the quantity $|\Delta_m| = |\mathbb{E}[\widehat{\theta}^{(\infty)}_m \mid E_0] - \theta^*_m|$.

Lemma 12.7. *Conditioned on E_0, we have*

$$|\Delta_m| \lesssim \frac{\log d + \log k}{dk}, \qquad \textit{for every } m \in [d],$$

for all $d \geq d_0$ and all $k \geq k_0$, where d_0 and k_0 are constants. Equivalently,

$$\beta(\widehat{\theta}^{(\infty)} \mid E_0) = \|\mathbb{E}[\widehat{\theta}^{(\infty)} \mid E_0] - \theta^*\|_\infty = \|\Delta\|_\infty \lesssim \frac{\log d + \log k}{dk}, \tag{12.33}$$

for all $d \geq d_0$ and all $k \geq k_0$, where d_0 and k_0 are constants.

See Section 12.1.4 for the proof of Lemma 12.7.

Note that (12.33) yields the desired rate on the quantity $\beta(\widehat{\theta}^{(\infty)} \mid E_0)$. It remains to show that $\beta(\widehat{\theta}^{(\infty)} \mid E_0)$ is sufficiently close to $\beta(\widehat{\theta}^{(A)})$.

Step 5: Show that the box constraint at A is vacuous for the unconstrained MLE $\widehat{\theta}^{(\infty)}$ and hence $\widehat{\theta}^{(\infty)}$ is the same as the stretched-MLE $\widehat{\theta}^{(A)}$ with high probability, using the deterministic bound in Step 3

To show that $\beta(\widehat{\theta}^{(\infty)} \mid E_0)$ is sufficiently close to $\beta(\widehat{\theta}^{(A)})$, we divide the argument into two parts. First, we show that $\beta(\widehat{\theta}^{(\infty)} \mid E_0) = \beta(\widehat{\theta}^{(A)} \mid E_0)$. Second, we show that $\beta(\widehat{\theta}^{(A)} \mid E_0)$ is close to $\beta(\widehat{\theta}^{(A)})$.

We first show that $\beta(\widehat{\theta}^{(\infty)} \mid E_0) = \beta(\widehat{\theta}^{(A)} \mid E_0)$. Recall that A and B are constants such that $A > B$. Recall from Lemma 12.5 that $\|\widehat{\theta}^{(\infty)} - \theta^*\|_\infty \lesssim \frac{\log d + \log k}{dk}$ conditioned on E_0. Hence, there exist constants d_0 and k_0, such that for any $d \geq d_0$ and $k \geq k_0$, we have $\|\widehat{\theta}^{(\infty)} - \theta^*\|_\infty < A - B$ conditioned on E_0. In this case, we have

$$\|\widehat{\theta}^{(\infty)}\|_\infty \leq \|\theta^*\|_\infty + \|\widehat{\theta}^{(\infty)} - \theta^*\|_\infty < B + (A - B) = A, \qquad \text{conditioned on } E_0.$$

Conditioned on E_0, the unconstrained MLE $\widehat{\theta}^{(\infty)}$ obeys the box constraint $\|\widehat{\theta}^{(\infty)}\|_\infty \leq A$. Therefore, $\widehat{\theta}^{(\infty)}$ is also a solution to the stretched-MLE $\widehat{\theta}^{(A)}$. By the uniqueness of $\widehat{\theta}^{(A)}$ from Lemma 12.2, we have

$$\widehat{\theta}^{(A)} = \widehat{\theta}^{(\infty)}, \qquad \text{conditioned on } E_0.$$

Hence, we have the relation

$$\beta(\widehat{\theta}^{(\infty)} \mid E_0) = \beta(\widehat{\theta}^{(A)} \mid E_0), \tag{12.34}$$

completing the first part of the argument.

It remains to show that $\beta(\widehat{\theta}^{(A)} \mid E_0)$ is sufficiently close to $\beta(\widehat{\theta}^{(A)})$. We have

$$\begin{aligned}\beta(\widehat{\theta}^{(A)}) &= \|\mathbb{E}[\widehat{\theta}^{(A)}] - \theta^*\|_\infty \\ &\overset{\text{(i)}}{=} \|\mathbb{E}[\widehat{\theta}^{(A)} \mid E_0] \cdot \mathbb{P}(E_0) + \mathbb{E}[\widehat{\theta}^{(A)} \mid \overline{E_0}] \cdot \mathbb{P}(\overline{E_0}) - \theta^*\|_\infty \\ &\overset{\text{(ii)}}{\leq} \|\mathbb{E}[\widehat{\theta}^{(A)} \mid E_0] - \theta^*\|_\infty \cdot \mathbb{P}(E_0) + \|\mathbb{E}[\widehat{\theta}^{(A)} \mid \overline{E_0}] - \theta^*\|_\infty \cdot \mathbb{P}(\overline{E_0}) \\ &= \underbrace{\beta(\widehat{\theta}^{(A)} \mid E_0) \cdot \mathbb{P}(E_0)}_{R_1} + \underbrace{\|\mathbb{E}[\widehat{\theta}^{(A)} \mid \overline{E_0}] - \theta^*\|_\infty \cdot \mathbb{P}(\overline{E_0})}_{R_2}. \end{aligned} \tag{12.35}$$

where step (i) is true by the law of iterated expectation, and step (ii) is true by the triangle inequality.

Consider the two terms in (12.35). For R_1, combining (12.33) and (12.34) yields

$$\beta(\widehat{\theta}^{(A)} \mid E_0) = \beta(\widehat{\theta}^{(\infty)} \mid E_0) \lesssim \frac{\log d + \log k}{dk}.$$

Therefore,

$$R_1 \lesssim \frac{\log d + \log k}{dk}. \tag{12.36}$$

Now consider R_2. By the box constraint $\|\widehat{\theta}^{(A)}\|_\infty \leq A$, we have

$$\|\mathbb{E}[\widehat{\theta}^{(A)} \mid \overline{E_0}] - \theta^*\|_\infty \overset{\text{(i)}}{\leq} \|\mathbb{E}[\widehat{\theta}^{(A)} \mid \overline{E_0}]\|_\infty + \|\theta^*\|_\infty \leq A + B, \tag{12.37}$$

where step (i) is true by the triangle inequality. Recall from (12.22), the event E_0 happens w.h.p.($\frac{1}{dk}$). Therefore,

$$\mathbb{P}(\overline{E_0}) \lesssim \frac{1}{dk}. \tag{12.38}$$

Combining (12.37) and (12.38) yields

$$R_2 \lesssim \frac{1}{dk}. \tag{12.39}$$

Plugging the term R_1 from (12.36) and the term R_2 from (12.39) back into (12.35), we have

$$\beta(\widehat{\theta}^{(A)}) \lesssim \frac{\log d + \log k}{dk},$$

completing the proof of Theorem 5.4(b).

12.1.3 Proof of Theorem 5.4(a)

Similar to the proof of Theorem 5.4(b), we first present a proof of the simple case of $d = 2$ items. It is important to note that although we present proofs of the 2-item case for both Theorem 5.4(b) and Theorem 5.4(a), their purposes are different. In Theorem 5.4(b) presented in Section 12.1.2, the proof sketch of the 2-item case is informal. It serves as a guideline for the general case. Then the main work involved in the general case is to generalize the arguments in the 2-item case step-by-step. On the other hand, in Theorem 5.4(a), the proof of the 2-item case to be presented is formal. It serves as a core sub-problem of the general case. Then the main work involved in the general case is to reduce the problem to the 2-item case, and then the results from the 2-item case directly.

Simple case: 2 items

As in Section 12.1.2, we first consider the simple case where there are $d = 2$ items. Again, due to the centering constraint, we have $\theta_2^* = -\theta_1^*$ for the true parameter vector θ^*, and we have $\widehat{\theta}_2 = -\widehat{\theta}_1$ for any estimator $\widehat{\theta}$ that satisfies the centering constraint (in particular, for the standard MLE $\widehat{\theta}^{(B)}$ and the unconstrained MLE $\widehat{\theta}^{(\infty)}$). Therefore, it suffices to focus only on item 1. Since there are only two items, for ease of notation, we denote $\mu = \mu_{12}$ and $\mu^* = \mu_{12}^*$.

We consider the true parameter vector $\theta^* = [B, -B]$. By the definition of $\{\mu_{ij}^*\}$ in (12.1), we have

$$\mu^* = \frac{1}{1 + e^{-(\theta_1^* - \theta_2^*)}} = \frac{1}{1 + e^{-2B}}.$$

The following proposition now lower bounds the bias of the standard MLE $\widehat{\theta}^{(B)}$.

Proposition 12.8. *Under $\theta^* = [B, -B]$, the bias of the MLE $\widehat{\theta}^{(B)}$ is bounded as*

$$\beta(\widehat{\theta}^{(B)}) = \|\mathbb{E}[\widehat{\theta}^{(B)}] - \theta^*\|_\infty = |\mathbb{E}[\widehat{\theta}_1^{(B)}] - B| \gtrsim \frac{1}{\sqrt{k}}.$$

Specifically, the bias is negative, that is,

$$\mathbb{E}[\widehat{\theta}_1^{(B)}] - B \leq -\frac{c}{\sqrt{k}}, \tag{12.40}$$

for some constant $c > 0$.

The rest of this section is devoted to proving (12.40) in Proposition 12.8.

For ease of notation, denote $\mu_+ = \mu^* = \frac{1}{1+e^{-2B}}$, and $\mu_- = 1 - \mu^* = \frac{1}{1+e^{2B}}$. In the proof sketch of Theorem 5.4(b) of the case of $d = 2$ items (Section 12.1.2), we derived the following expression (12.12) for the unconstrained MLE:

$$\widehat{\theta}_1^{(\infty)}(\mu) = -\frac{1}{2}\log\left(\frac{1}{\mu} - 1\right).$$

Now consider the standard MLE $\widehat{\theta}^{(B)}$. By straightforward analysis, one can derive the following closed-form expression for the standard MLE:

$$\widehat{\theta}_1^{(B)}(\mu) = \begin{cases} -B & \text{if } \mu \in [0, \mu_-] \\ -\frac{1}{2}\log\left(\frac{1}{\mu} - 1\right) & \text{if } \mu \in (\mu_-, \mu_+) \\ B & \text{if } \mu \in [\mu_+, 1]. \end{cases} \tag{12.41}$$

For ease of notation, we denote a function $h : [0, 1] \to [-B, B]$ as

$$h(t) = \begin{cases} -B & \text{if } t \in [0, \mu_-] \\ -\frac{1}{2}\log\left(\frac{1}{t} - 1\right) & \text{if } t \in (\mu_-, \mu_+) \\ B & \text{if } t \in [\mu_+, 1], \end{cases} \tag{12.42}$$

where $h(t) = \widehat{\theta}_1^{(B)}(\mu = t)$ for any $t \in [0, 1]$. Then the standard MLE (12.41) can be equivalently written as $h(\mu)$. To make the computation of the bias incurred by $\widehat{\theta}^{(B)}$ more tractable, we also define the following auxiliary function $h^+ : [0, 1] \to [-B, B]$ as:

$$h^+(t) := \begin{cases} \frac{2B}{\mu_+}(t - \mu_+) + B & \text{if } t \in [0, \mu_+) \\ B & \text{if } t \in [\mu_+, 1]. \end{cases} \tag{12.43}$$

In words, the function h^+ is piecewise linear. On the interval $[0, \mu_+]$, it is a line passing through the points $(0, -B)$ and (μ_+, B). On the interval $[\mu_+, 1]$, its value equals the constant B. The following lemma now states a relation between $h^+(\mu)$ and $h(\mu)$ in expectation with respect to μ.

Lemma 12.9. *Under $\theta^* = [B, -B]$, we have*

$$\mathbb{E}[h(\mu)] \leq \mathbb{E}[h^+(\mu)]. \tag{12.44}$$

See Section 12.1.4 for the proof of Lemma 12.9.

Now subtracting B from both sides of (12.44), we have

$$\mathbb{E}[\widehat{\theta}_1^{(B)}] - \theta_1^* = \mathbb{E}[h(\mu)] - B \leq \mathbb{E}[h^+(\mu)] - B. \tag{12.45}$$

The following lemma states that the bias introduced by $h^+(\mu)$ satisfies the desired rate from Proposition 12.8.

Lemma 12.10. *Under $\theta^* = [B, -B]$, we have*

$$\mathbb{E}[h^+(\mu)] - B \leq -\frac{c}{\sqrt{k}}, \tag{12.46}$$

for some constant $c > 0$.

See Section 12.1.4 for the proof of Lemma 12.10.

Combining (12.45) and (12.46), we have

$$\mathbb{E}[\widehat{\theta}_1^{(B)}] - \theta_1^* \leq -\frac{c}{\sqrt{k}},$$

completing the proof of (12.40) in Proposition 12.8.

Complete Proof

In this section, we present the proof of Theorem 5.4(a). The proof reduces the general case to the 2-item case presented in Section 12.1.3. In the reduction, we construct an "oracle" MLE, such that the oracle MLE yields identical estimates for item 2 through item d. Specifically, we consider an unconstrained oracle denoted by $\widetilde{\theta}^{(\infty)}$ (without the box constraint), and a constrained oracle denoted by $\widetilde{\theta}^{(B)}$ (with the box constraint at B), to be defined precisely in the proof shortly. Then we derive the closed-form expressions for $\widetilde{\theta}^{(\infty)}$ and $\widetilde{\theta}^{(B)}$, which bear resemblance to the expressions of the the unconstrained MLE and the standard MLE in the 2-item case. Using the proof of the 2-item case, we prove that the constrained oracle $\widetilde{\theta}^{(B)}$ incurs a negative bias of $\Omega(\frac{1}{\sqrt{dk}})$. Given this result, it remains to show that $\widetilde{\theta}^{(B)}$ and $\widehat{\theta}^{(B)}$ differ by $o(\frac{1}{\sqrt{dk}})$ in terms of bias. We decompose the difference between $\widetilde{\theta}^{(B)}$ and $\widehat{\theta}^{(B)}$ into three terms: from $\widetilde{\theta}^{(B)}$ to $\widetilde{\theta}^{(\infty)}$, from $\widetilde{\theta}^{(\infty)}$ to $\widehat{\theta}^{(\infty)}$, and from $\widehat{\theta}^{(\infty)}$ to $\widehat{\theta}^{(B)}$, The second term is bounded by $\widetilde{\mathcal{O}}(\frac{1}{dk})$ by modifying the upper-bound proof of Theorem 5.4(b). The first and the third terms are bounded by carefully analyzing the effect of the box constraint on the oracle MLE and the standard MLE, respectively.

In the proof, we fix any constant $B > 0$, and consider the true parameter vector:

$$\theta^* = \left[B, -\frac{B}{d-1}, -\frac{B}{d-1}, \ldots, -\frac{B}{d-1}\right]. \tag{12.47}$$

It can be verified that θ^* satisfies both the box constraint at B and the centering constraint, so we have $\theta^* \in \Theta_B$. We prove that the bias on item 1 is negative, and its magnitude is $\Omega(\frac{1}{\sqrt{dk}})$. That is, we prove that

$$\mathbb{E}[\widehat{\theta}_1^{(B)}] - \theta_1^* = \mathbb{E}[\widehat{\theta}_1^{(B)}] - B \leq -\frac{c}{\sqrt{dk}},$$

for some constant $c > 0$. The proof consists of the following 5 steps.

Step 1: Construct oracle estimators $\widetilde{\theta}^{(\infty)}$ (unconstrained) and $\widetilde{\theta}^{(B)}$ (constrained)

Recall that $\mu_{ij} \sim \frac{1}{k}\text{Binom}(k, \mu_{ij}^*)$ is a random variable representing the fraction of times that item i beats item j. We define μ_1 as fraction of wins by item 1, among all comparisons in which item 1 is involved:

$$\mu_1 := \frac{1}{d-1}\sum_{m=2}^{d} \mu_{1m}. \tag{12.48}$$

We similarly define the true probability $\mu_1^* = \frac{1}{d-1}\sum_{m=2}^{d} \mu_{1m}^*$. With the construction (12.47) of θ^*, we have $\mu_1^* = \frac{1}{1+e^{-\frac{d}{d-1}B}}$. Now we construct the following random quantities $\{\widetilde{\mu}_{ij}\}_{i\neq j}$ as a function of $\{\mu_{ij}\}_{i\neq j}$:

$$\widetilde{\mu}_{ij} = \begin{cases} \mu_1 & \text{if } i = 1,\ j \in \{2, \ldots, d\} \\ 1 - \mu_1 & \text{if } j = 1,\ i \in \{2, \ldots, d\} \\ \frac{1}{2} & \text{otherwise.} \end{cases} \tag{12.49}$$

Recall that $\widehat{\theta}^{(\infty)}(\{\mu_{ij}\})$ denotes the unconstrained MLE (12.7). Now define an "unconstrained oracle" MLE $\widetilde{\theta}^{(\infty)}$ as:

$$\begin{aligned}\widetilde{\theta}^{(\infty)}(\{\mu_{ij}\}) &:= \widehat{\theta}^{(\infty)}(\{\widetilde{\mu}_{ij}\})\\ &= \operatorname*{argmin}_{\theta\in\Theta_\infty} \ell(\{\widetilde{\mu}_{ij}\};\theta).\end{aligned} \tag{12.50a}$$

Similarly, define a "constrained oracle" MLE $\widetilde{\theta}^{(B)}$ as:

$$\begin{aligned}\widetilde{\theta}^{(B)}(\{\mu_{ij}\}) &:= \widehat{\theta}^{(B)}(\{\widetilde{\mu}_{ij}\})\\ &= \operatorname*{argmin}_{\theta\in\Theta_B} \ell(\{\widetilde{\mu}_{ij}\};\theta).\end{aligned} \tag{12.50b}$$

In the subsequent steps, these oracle estimators are used to reduce the general case to the 2-item case.

Step 2: Formalize the oracle information contained in the unconstrained oracle $\widetilde{\theta}^{(\infty)}$ and the constrained oracle $\widetilde{\theta}^{(B)}$

Note that the construction of $\{\widetilde{\mu}_{ij}\}$ in (12.49) is symmetric with respect to item 2 through item d, that is, for any two items i and i' where $i, i' \in \{2,\ldots,d\}$, we have $\widetilde{\mu}_{ij} = \widetilde{\mu}_{i'j}$ and $\widetilde{\mu}_{ji} = \widetilde{\mu}_{ji'}$ for every $i \in [d]\backslash\{j, j'\}$. Therefore, the construction of $\{\widetilde{\mu}_{ij}\}$ intuitively encodes the "oracle" that item 2 through item d have identical parameters. Formally, define the set $\Theta_{\text{oracle}} := \{\theta \in \mathbb{R}^d \mid \theta_2 = \cdots = \theta_d\}$. The following lemma states that the unconstrained oracle and the constrained oracle incorporate the set Θ_{oracle} into the domain of optimization without altering their solutions.

Lemma 12.11. *The unconstrained oracle $\widetilde{\theta}^{(\infty)}$ can be equivalently written as*

$$\widetilde{\theta}^{(\infty)} = \operatorname*{argmin}_{\Theta_\infty\cap\Theta_{\text{oracle}}} \ell(\{\widetilde{\mu}_{ij}\};\theta). \tag{12.51a}$$

That is, a solution to (12.50a) *exists if and only if a solution to* (12.51a) *exists. Moreover, when the solutions to* (12.50a) *and* (12.51a) *exist, they are identical.*

Similarly, the constrained oracle $\widetilde{\theta}^{(B)}$ can be equivalently written as

$$\widetilde{\theta}^{(B)} = \operatorname*{argmin}_{\theta\in\Theta_B\cap\Theta_{\text{oracle}}} \ell(\{\widetilde{\mu}_{ij}\};\theta). \tag{12.51b}$$

See Section 12.1.4 for the proof of Lemma 12.11.

Given Lemma 12.11 combined with the centering constraint, we parameterize the unconstrained oracle $\widetilde{\theta}^{(\infty)}$ and the constrained oracle $\widetilde{\theta}^{(B)}$ as:

$$\widetilde{\theta}^{(\infty)} = \left[\widetilde{\theta}_1^{(\infty)}, -\frac{1}{d-1}\widetilde{\theta}_1^{(\infty)}, \ldots, -\frac{1}{d-1}\widetilde{\theta}_1^{(\infty)}\right], \tag{12.52a}$$

$$\widetilde{\theta}^{(B)} = \left[\widetilde{\theta}_1^{(B)}, -\frac{1}{d-1}\widetilde{\theta}_1^{(B)}, \ldots, -\frac{1}{d-1}\widetilde{\theta}_1^{(B)}\right]. \tag{12.52b}$$

Step 3: Show that the bias of the constrained oracle $\widetilde{\theta}^{(B)}$ on item 1 is bounded by $\mathbb{E}[\widetilde{\theta}_1^{(B)}] - \theta_1^* \leq -\frac{c}{\sqrt{dk}}$, by making a reduction to the 2-item case

In this step, we modify the proof of Proposition 12.8 in the 2-item case to lower bound the bias of the constrained oracle $\widetilde{\theta}^{(B)}$. Specifically, we show that given $\theta^* = \left[B, -\frac{B}{d-1}, \ldots, -\frac{B}{d-1}\right]$, the bias on item 1 is bounded as (cf. (12.40)):

$$\mathbb{E}[\widetilde{\theta}_1^{(B)}] - \theta^* \leq -\frac{c}{\sqrt{dk}},$$

for some constant $c > 0$.

First, we solve for the unconstrained oracle $\widetilde{\theta}^{(\infty)}$ and the constrained oracle $\widetilde{\theta}^{(B)}$ in closed form. Set $m = 1$ in the gradient expression (12.4). Plugging in the expressions for the unconstrained oracle $\widetilde{\theta}^{(\infty)}$ (12.52a) and the manipulated observations $\{\widetilde{\mu}_{ij}\}$ (12.49), we have

$$\left.\frac{\partial \ell}{\partial \theta_1}\right|_{\theta=\widetilde{\theta}^{(\infty)}} = k(d-1)\left[\frac{1}{1+e^{-\frac{d}{d-1}\widetilde{\theta}_1^{(\infty)}}} - \mu_1\right] \tag{12.53}$$

Setting the derivative (12.53) to 0, we have

$$\begin{aligned} \frac{1}{1+e^{-\frac{d}{d-1}\widetilde{\theta}_1^{(\infty)}}} &= \mu_1 \\ \widetilde{\theta}_1^{(\infty)} &= -\frac{d-1}{d}\log\left(\frac{1}{\mu_1} - 1\right). \end{aligned} \tag{12.54}$$

Denote $\mu_{d,+} = \mu_1^* = \frac{1}{1+e^{-\frac{d}{d-1}B}}$, and $\mu_{d,-} = 1 - \mu_{d,+} = \frac{1}{1+e^{\frac{d}{d-1}B}}$. In the notations $\mu_{d,+}$ and $\mu_{d,-}$, the dependency on d is made explicit. When the dependency on d does not need to be emphasized, we also use the shorthand notations μ_+ and μ_-. Now consider the constrained oracle $\widetilde{\theta}^{(B)}$. By straightforward analysis, one can derive the following closed-form expression for the constrained oracle:

$$\widetilde{\theta}_1^{(B)}(\mu_1) = \begin{cases} -B & \text{if } 0 \leq \mu_1 < \mu_{d,-} \\ -\frac{d-1}{d}\log\left(\frac{1}{\mu_1} - 1\right) & \text{if } \mu_{d,-} < \mu_1 < \mu_{d,+} \\ B & \text{if } \mu_{d,+} \leq \mu_1 \leq 1. \end{cases} \tag{12.55}$$

Note the similarity between $\widetilde{\theta}^{(B)}$ in (12.55) and the 2-item case $\widehat{\theta}_1^{(B)}$ in (12.41) from Section 12.1.3. Similar to the function h defined in (12.42) of the 2-item case, we denote a function $h_d : [0, 1] \to [-B, B]$ as:

$$h_d(t) = \begin{cases} -B & \text{if } 0 \leq t < \mu_{d,-} \\ -\frac{d-1}{d}\log\left(\frac{1}{t} - 1\right) & \text{if } \mu_{d,-} < t < \mu_{d,+} \\ B & \text{if } \mu_{d,+} \leq t \leq 1, \end{cases}$$

where $h_d(t) = \widetilde{\theta}_1^{(B)}(\mu_1 = t)$ for any $t \in [0, 1]$. Then the estimator $\widetilde{\theta}_1^{(B)}(\mu)$ can be equivalently written as $h_d(\mu)$. Similar to the function h^+ defined in (12.43) of the 2-item case, we define an auxiliary function $h_d^+ : [0,1] \to [-B, B]$ as:

$$h_d^+(t) = \begin{cases} \frac{2B}{\mu_{d,+}}(t - \mu_{d,+}) + B & \text{if } 0 \le t < \mu_{d,+} \\ B & \text{if } \mu_{d,+} \le t \le 1. \end{cases}$$

Note that in the proofs of Lemma 12.9 and Lemma 12.10, we have only relied on the following two facts:

- There exists a constant c such that
$$\frac{1}{2} < \mu_+ < c < 1.$$
- The random variable μ is sampled as $\mu \sim \frac{1}{k}\text{Binom}(k, \mu_+)$.

In the general case, it can be verified that

- There exists a constant c such that
$$\frac{1}{2} < \mu_{d,+} < c < 1, \qquad \text{for all } d \ge 2.$$
- The random variable μ_1 as defined in (12.49) is sampled as $\mu_1 \sim \frac{1}{k'}\text{Binom}(k', \mu_+)$, where $k' := (d-1)k$ denotes the total number of comparisons in which item 1 is involved.

To extend the arguments in the 2-item case to the general case, we replace μ by μ_1, replace μ_+ by $\mu_{d,+}$, replace h^+ by h_d^+, and replace k by k' in the proof of Proposition 12.8. It can be verified that the arguments in Lemma 12.9 and Lemma 12.10 still hold after these replacements. Therefore, extending the arguments in Proposition 12.8, we have that at $\theta^* = \left[B, -\frac{B}{d-1}, \ldots, -\frac{B}{d-1}\right]$,

$$\mathbb{E}[\widetilde{\theta}_1^{(B)}] - \theta_1^* \le -\frac{c}{\sqrt{k'}} = -\frac{c}{\sqrt{(d-1)k}} \le -\frac{c'}{\sqrt{dk}}, \tag{12.56}$$

for some constants $c, c' > 0$.

Step 4: Bound the difference between the unconstrained oracle $\widetilde{\theta}^{(\infty)}$ and the unconstrained MLE $\widehat{\theta}^{(\infty)}$, by modifying the proof of Theorem 5.4(b)

Recall that the random variable μ_1 denotes the fraction of wins by item 1. In this step, we fix any real number $v \in [\frac{1}{2}, \mu_+]$, and denote E_v as the event that we observe $\mu_1 = v$. Then we prove that conditioned on the event E_v, the difference between the unconstrained oracle $\widetilde{\theta}^{(\infty)}$ and the unconstrained MLE $\widehat{\theta}^{(\infty)}$ is small in expectation, by modifying Step 1 to Step 4 in the upper-bound proof of Theorem 5.4(b) in Section 12.1.2.

We first conceptually explain how to modify the proof of Theorem 5.4(b). Our goal is to bound the difference between $\widetilde{\theta}^{(\infty)}$ and $\widehat{\theta}^{(\infty)}$ in expectation conditioned on the event E_v. By the definition of $\{\widetilde{\mu}_{ij}\}$ in (12.49), the quantities $\{\widetilde{\mu}_{ij}\}$ are fixed (not random) conditioned on E_v, and hence the unconstrained oracle $\widetilde{\theta}^{(\infty)}$ is fixed conditioned on E_v.

We therefore replace the role of the true parameter vector θ^* in the proof of Theorem 5.4(b) by the unconstrained oracle $\widetilde{\theta}^{(\infty)}$. Then we think of the actual observations $\{\mu_{ij}\}$ as a noisy version of $\{\widetilde{\mu}_{ij}\}$, and think of $\widehat{\theta}^{(\infty)}$ as the estimate for $\widetilde{\theta}^{(\infty)}$. Now we modify the proof of Theorem 5.4(b) to bound the expected difference between $\widehat{\theta}^{(\infty)}$ and $\widetilde{\theta}^{(\infty)}$ conditioned on E_v. At the end of this step, we provide more intuition why we need to condition on the event E_v.

Formally, we denote $\{\widetilde{\mu}_{ij}^v\}$ as the values of $\{\widetilde{\mu}_{ij}\}$ conditional on E_v. We denote $\widetilde{\theta}^v$ as the unconstrained oracle $\widetilde{\theta}^{(\infty)}$ conditional on E_v. It can be verified that $\{\widetilde{\mu}_{ij}^v\}$ and $\widetilde{\theta}^v$ are fixed (not random) given any $v \in [\frac{1}{2}, \mu_+]$. Conditioned on E_v, we think of $\widetilde{\theta}^v$ *as if* it is the "true" parameter vector to be estimated (replacing the role of θ^*), and think of $\{\widetilde{\mu}_{ij}^v\}$ *as if* it is the "true" underlying probabilities (replacing the role of $\{\mu_{ij}^*\}$).

Given the definition of $\{\widetilde{\mu}_{ij}\}$ in (12.49), we have that conditioned on event E_v,

$$\widetilde{\mu}_{ij}^v = \begin{cases} v & \text{if } i = 1,\ j \in \{2, \ldots, d\} \\ 1 - v & \text{if } j = 1,\ i \in \{2, \ldots, d\} \\ \frac{1}{2} & \text{otherwise.} \end{cases} \tag{12.57}$$

From the expression (12.54) of the unconstrained oracle $\widetilde{\theta}^{(\infty)}$, it can be verified that $\widetilde{\theta}^{(\infty)}$ satisfies the deterministic equality

$$\frac{1}{1 + e^{-\left(\widetilde{\theta}_i^{(\infty)} - \widetilde{\theta}_j^{(\infty)}\right)}} = \widetilde{\mu}_{ij}, \qquad \text{for all } i \neq j. \tag{12.58}$$

Now we start to replicate Step 1 to Step 4 in the proof of Theorem 5.4(b) presented in Section 12.1.2.

To replicate *Step 1* of Theorem 5.4(b), recall that in the proof of Theorem 5.4(b), we condition on Lemma 12.3 and Lemma 12.4. We first establish the modified versions of these two lemmas, when conditioned on E_v.

Lemma 12.12 (Conditional version of Lemma 12.3)**.** *Conditioned on the event E_v, there exists a finite solution $\widehat{\theta}^{(\infty)}$ to the unconstrained MLE* (12.7) *w.h.p.($\frac{1}{dk} \mid E_v$).*

See Section 12.1.4 for the proof of Lemma 12.12.

Lemma 12.13 (Conditional version of Lemma 12.4)**.** *Conditioned on the event E_v, there exists a constant $c > 0$, such that*

$$\left| \sum_{i \neq m} \mu_{mi} - \sum_{i \neq m} \widetilde{\mu}_{mi}^v \right| \leq c \sqrt{\frac{d(\log d + \log k)}{k}}, \tag{12.59}$$

simultaneously for all $m \in [d]$ w.h.p.($\frac{1}{dk} \mid E_v$).

See Section 12.1.4 for the proof of Lemma 12.13.

Recall that we have conditioned on the event E_v. Denote E_0 as the event that Lemma 12.12 and Lemma 12.13 both hold. (Note that the event E_0 is defined for some fixed v, so to be precise, the event E_0 should be denoted as $E_{0,v}$. For ease of notation, we drop the subscript

v.) Taking a union bound of Lemma 12.12 and Lemma 12.13, we have that E_0 happens w.h.p.($\frac{1}{dk} \mid E_v$). For the rest of the proof, we condition on the events (E_0, E_v).

To replicate *Step 2* of Theorem 5.4(b), we subtract equality (12.58) from both sides of (12.24). We obtain the (unconditional) deterministic equality:

$$\sum_{i=1}^{d} \left(\frac{1}{1 + e^{-(\widehat{\theta}_m^{(\infty)} - \widehat{\theta}_i^{(\infty)})}} - \frac{1}{1 + e^{-(\widetilde{\theta}_m^{(\infty)} - \widetilde{\theta}_i^{(\infty)})}} \right) = \sum_{i \neq m} (\mu_{mi} - \widetilde{\mu}_{mi}), \qquad \text{for every } m \in [d]. \tag{12.60}$$

Conditioning (12.60) on (E_0, E_v), we have the following deterministic equality, as a modified version of (12.25):

$$\sum_{i=1}^{d} \left(\frac{1}{1 + e^{-(\widehat{\theta}_m^{(\infty)} - \widehat{\theta}_i^{(\infty)})}} - \frac{1}{1 + e^{-(\widetilde{\theta}_m^{v} - \widetilde{\theta}_i^{v})}} \right) = \sum_{i \neq m} (\mu_{mi} - \widetilde{\mu}_{mi}^{v}), \qquad \text{conditioned on } (E_0, E_v). \tag{12.61}$$

To replicate *Step 3* of Theorem 5.4(b), note that v is bounded as $v \in [\frac{1}{2}, \mu_+]$. By the expression (12.54) of $\widetilde{\theta}^{(\infty)}$ (and hence of $\widetilde{\theta}^v$), it can be verified that $\widetilde{\theta}^v$ is bounded as $|\widetilde{\theta}^v| \leq c$ for some constant c. Denote $\widetilde{\delta} = \widehat{\theta}^{(\infty)} - \widetilde{\theta}^v$. Using the same arguments as in Lemma 12.5, we have the deterministic relation that

$$\|\widehat{\theta}^{(\infty)} - \widetilde{\theta}^v\|_\infty = \|\widetilde{\delta}\|_\infty \lesssim \sqrt{\frac{\log d + \log k}{dk}}, \qquad \text{conditioned on } (E_0, E_v). \tag{12.62}$$

To replicate *Step 4* of Theorem 5.4(b), we first apply the second-order mean value theorem on (12.61), and then take an expectation conditional on (E_0, E_v). The following equation establishes a modified version of (12.29):

$$\sum_{i=1}^{d} f'(\widetilde{\theta}_m^v - \widetilde{\theta}_i^v) \cdot \mathbb{E}\left[\widetilde{\delta}_i - \widetilde{\delta}_m \mid E_0, E_v\right] = \sum_{i \neq m} (\mathbb{E}[\mu_{mi} \, | E_0, E_v] - \widetilde{\mu}_{mi}^v) - \frac{1}{2} \sum_{i=1}^{d} \mathbb{E}[f''(\lambda_{mi})(\widetilde{\delta}_m - \widetilde{\delta}_i)^2 \mid E_0, E_v], \tag{12.63}$$

where each λ_{mi} is a random variable that takes values between $\widetilde{\theta}_m^v - \widetilde{\theta}_i^v$ and $\widetilde{\theta}_m^v - \widetilde{\theta}_i^v + \widetilde{\delta}_m - \widetilde{\delta}_i$. To apply Lemma 12.6, we set E as E_v, and set E' as E_0 in (12.31):

$$|\mathbb{E}[\mu_{ij} \mid E_0, E_v] - \mathbb{E}[\mu_{ij} \mid E_v]| \lesssim \frac{1}{dk}. \tag{12.64}$$

It can be verified that

$$\mathbb{E}[\mu_{ij} \mid E_v] = \widetilde{\mu}_{ij}^v. \tag{12.65}$$

Plugging (12.65) into (12.64), we have

$$|\mathbb{E}[\mu_{ij} \mid E_0, E_v] - \widetilde{\mu}_{ij}^v| \lesssim \frac{1}{dk}.$$

Using the same arguments as in Lemma 12.7 to handle the remaining terms in (12.63), we have the following upper bound as a modified version of (12.33):

$$\|\mathbb{E}[\widehat{\theta}^{(\infty)} - \widetilde{\theta}^v \mid E_0, E_v]\|_\infty = \|\mathbb{E}[\widehat{\theta}^{(\infty)} - \widetilde{\theta}^{(\infty)} \mid E_0, E_v]\|_\infty \lesssim \frac{\log d + \log k}{dk}. \tag{12.66}$$

Now that we have established the desired result (12.66) of this step, we conclude this step with some intuition why we need to condition on E_v. Without conditioning on E_v, we could still have utilized the proof of Theorem 5.4(b), and could have established a result of the form (cf. (12.66)):

$$\|\mathbb{E}[\widehat{\theta}^{(\infty)} - \widetilde{\theta}^v \mid E_0]\|_\infty = \|\mathbb{E}[\widehat{\theta}^{(\infty)} - \widetilde{\theta}^{(\infty)} \mid E_0]\|_\infty \lesssim \frac{\log d + \log k}{dk}. \tag{12.67}$$

Our goal here is to bound the constrained oracle $\widehat{\theta}^{(B)}$ and the constrained MLE $\widetilde{\theta}^{(B)}$ in expectation. However, the fact that two *unconstrained* estimators are close in expectation does not imply that their *constrained* counterparts are close in expectation[4]. Therefore, a bound of the form (12.67) is not sufficient for our goal, and instead we need to establish some "pointwise" control between $\widehat{\theta}^{(\infty)}$ and $\widetilde{\theta}^{(\infty)}$. That is, whenever the box constraint has little effect on $\widetilde{\theta}^{(\infty)}$, we want to show that the box constraint also has little effect on $\widehat{\theta}^{(\infty)}$. Thus, we condition on the event E_v for any $v \in [\frac{1}{2}, \mu_+]$, and bound the difference between $\widehat{\theta}^{(\infty)}$ and $\widetilde{\theta}^{(\infty)}$ in expectation conditioned on E_v (that is, the bound in (12.66)). Given this pointwise result, we then integrate over v to establish the desired result that $\widehat{\theta}^{(B)}$ and $\widetilde{\theta}^{(B)}$ are close in expectation, to be presented in the subsequent step of the proof.

Step 5: Bound the expected difference between $\widehat{\theta}^{(B)}$ and $\widetilde{\theta}^{(B)}$, by making a connection between $\widehat{\theta}^{(B)} - \widetilde{\theta}^{(B)}$ and $\widehat{\theta}^{(\infty)} - \widetilde{\theta}^{(\infty)}$

We decompose the bias of the standard MLE $\widehat{\theta}^{(B)}$ as

$$\mathbb{E}[\widehat{\theta}_1^{(B)}] - \theta_1^* = (\mathbb{E}[\widetilde{\theta}_1^{(B)}] - \theta_1^*) + \mathbb{E}[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)}]. \tag{12.68}$$

Recall from (12.56) that

$$\mathbb{E}[\widetilde{\theta}_1^{(B)}] - \theta_1^* \leq -\frac{c}{\sqrt{dk}}. \tag{12.69}$$

[4] For example, consider the following two univariate estimators. The first estimator always outputs a value within $[-B, B]$. The second estimator sometimes outputs a value within $[-B, B]$, and sometimes outputs a value greater than B. The two estimators could be constructed such that they are close (or equal) in expectation. However, now consider their constrained counterparts. The first estimator is not affected by a box constraint at B, whereas the expected value of second estimator can become significantly smaller due to the box constraint. Therefore, the constrained counterparts of these two estimators may not be close in expectation.

In what follows, we prove that

$$\mathbb{E}[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)}] \leq c' \frac{\log d + \log k}{dk}. \tag{12.70}$$

Then plugging (12.69) and (12.70) back into (12.68) yields

$$\mathbb{E}[\widehat{\theta}_1^{(B)}] - \theta_1^* \leq -\frac{c}{\sqrt{dk}} + c' \frac{\log d + \log k}{dk} \leq -\frac{c''}{\sqrt{dk}},$$

for all $d \geq d_0$ and $k \geq k_0$ where d_0 and k_0 are constants, completing the proof of Theorem 5.4(a).

The rest of this step is devoted to proving (12.70). To bound $\mathbb{E}[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)}]$, we make a connection between $\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)}$ and $\widetilde{\theta}_1^{(\infty)} - \widehat{\theta}_1^{(\infty)}$, and then we evoke the bound on $\widetilde{\theta}_1^{(\infty)} - \widehat{\theta}_1^{(\infty)}$ from (12.66) in Step 4.

Recall that μ_1 is a discrete random variable representing the fraction of wins by item 1. By the law of iterated expectation, we have

$$\begin{aligned}\mathbb{E}[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)}] =& \underbrace{\mathbb{E}\left[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)} \mid \frac{1}{2} < \mu_1 < \mu_1^*\right] \cdot \mathbb{P}\left(\frac{1}{2} < \mu_1 < \mu_1^*\right)}_{R_1} \\ &+ \underbrace{\mathbb{E}[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)} \mid \mu_1 \geq \mu_1^*] \cdot \mathbb{P}\left(\mu_1 \geq \mu_1^*\right)}_{R_2} + \underbrace{\mathbb{E}\left[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)} \mid \mu_1 < \frac{1}{2}\right] \cdot \mathbb{P}\left(\mu_1 < \frac{1}{2}\right)}_{R_3}.\end{aligned} \tag{12.71}$$

In what follows, we bound the terms R_1, R_2 and R_3 separately.

Consider the term R_2. From the expression of $\widetilde{\theta}^{(B)}$ in (12.55), we have $\widetilde{\theta}_1^{(B)} = B$ when $\mu_1 \geq \mu_1^*$. Therefore,

$$\mathbb{E}[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)} \mid \mu_1 \geq \mu_1^*] = \mathbb{E}[\widehat{\theta}_1^{(B)} \mid \mu_1 \geq \mu_1^*] - B \overset{\text{(i)}}{\leq} 0,$$

where (i) is true due to the box constraint $|\widehat{\theta}_1^{(B)}| \leq B$. Hence,

$$R_2 \leq 0. \tag{12.72}$$

Consider the term R_3, we have $\mathbb{E}[\mu_1] = \mu_1^* = \frac{1}{1+e^{-\frac{d}{d-1}B}}$, and therefore it can be verified that there exists a constant $\tau > 0$, such that $\mu_1^* > \frac{1}{2} + \tau$ for all $d \geq 2$. By Hoeffding's inequality, we have

$$\begin{aligned}\mathbb{P}\left(\mu_1 < \frac{1}{2}\right) &< \mathbb{P}\left(|\mu_1 - \mu_1^*| > \tau\right) \\ &\leq 2\exp\left(-2(d-1)k\tau^2\right) \lesssim \frac{1}{dk}.\end{aligned} \tag{12.73}$$

Therefore, we have

$$
\begin{aligned}
R_3 &= \mathbb{E}\left[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)} \mid \mu_1 < \frac{1}{2}\right] \cdot \mathbb{P}\left(\mu_1 < \frac{1}{2}\right) \\
&\overset{\text{(i)}}{\leq} 2B \cdot \mathbb{P}\left(\mu_1 < \frac{1}{2}\right) \\
&\overset{\text{(ii)}}{\lesssim} \frac{1}{dk},
\end{aligned} \tag{12.74}
$$

where (i) is true because $|\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)}| \leq |\widehat{\theta}_1^{(B)}| + |\widetilde{\theta}_1^{(B)}| \leq 2B$ by the box constraint, and (ii) is true due to (12.73).

Now consider the term R_1. Denote $\overline{E_0}$ as the complement of the event E_0. Using the law of iterated expectation again, we have

$$
\begin{aligned}
R_1 = \mathbb{E}\left[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)} \mid \frac{1}{2} < \mu_1 < \mu_1^*\right] \cdot \mathbb{P}\left(\frac{1}{2} < \mu_1 < \mu_1^*\right) = \\
\underbrace{\mathbb{E}\left[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)} \mid E_0, \frac{1}{2} < \mu_1 < \mu_1^*\right] \cdot \mathbb{P}\left(E_0, \frac{1}{2} < \mu_1 < \mu_1^*\right)}_{R_{11}} \\
+ \underbrace{\mathbb{E}\left[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)} \mid \overline{E_0}, \frac{1}{2} < \mu_1 < \mu_1^*\right] \cdot \mathbb{P}\left(\overline{E_0}, \frac{1}{2} < \mu_1 < \mu_1^*\right)}_{R_{12}}
\end{aligned} \tag{12.75}
$$

Consider the term R_{12}. We have

$$
\begin{aligned}
\mathbb{P}\left(\overline{E_0}, \frac{1}{2} < \mu_1 < \mu_1^*\right) &= \sum_{v \in (\frac{1}{2}, \mu_1^*)} \mathbb{P}(\overline{E_0} \mid E_v) \cdot \mathbb{P}(E_v) \\
&\overset{\text{(i)}}{\leq} \frac{c}{dk} \sum_{v \in (\frac{1}{2}, \mu_1^*)} \mathbb{P}(E_v) \\
&\lesssim \frac{1}{dk},
\end{aligned} \tag{12.76}
$$

where (i) is true because E_0 happens w.h.p.($\frac{1}{dk} \mid E_v$). Combining (12.76) with the fact that $|\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)}| \leq 2B$ due to the box constraint, we have

$$
R_{12} \lesssim \frac{1}{dk}. \tag{12.77}
$$

Now consider the term R_{11}. We first analyze the constrained oracle $\widetilde{\theta}^{(B)}$. By the expression of $\widetilde{\theta}^{(B)}$ in (12.55) and the expression of $\widetilde{\theta}^{(\infty)}$ in (12.54), we have

$$
\widetilde{\theta}^{(B)} = \widetilde{\theta}^{(\infty)}, \qquad \text{conditioned on } \frac{1}{2} < \mu_1 < \mu_1^*. \tag{12.78}
$$

Moreover, given $\frac{1}{2} < \mu_1 < \mu_1^*$, by the expression of $\widetilde{\theta}^{(B)}$ in (12.55), we have

$$0 < \widetilde{\theta}_1^{(B)} < B$$

and therefore by the parameterization of $\widetilde{\theta}^{(B)}$ in (12.52b),

$$|\widetilde{\theta}_i^{(B)}| \leq \frac{1}{d-1}B \qquad \text{for every } i \in \{2, \dots, d\}.$$

Hence, there exists a constant $\tau' > 0$ such that

$$\widetilde{\theta}_1^{(B)} > -B + \tau' \tag{12.80a}$$

and

$$-B + \tau' < \widetilde{\theta}_i^{(B)} < B - \tau' \qquad \text{for every } i \in \{2, \dots, d\}. \tag{12.80b}$$

Now we analyze the standard MLE $\widehat{\theta}^{(B)}$. Recall that E_v denotes the event that $\mu_1 = v$. We have that for every $v \in \left(\frac{1}{2}, \mu_1^*\right)$,

$$\|\widehat{\theta}_1^{(\infty)} - \widetilde{\theta}_1^{(B)}\|_\infty \overset{\text{(i)}}{=} \|\widehat{\theta}_1^{(\infty)} - \widetilde{\theta}^{(\infty)}\|_\infty \overset{\text{(ii)}}{\lesssim} \sqrt{\frac{\log d + \log k}{dk}}, \qquad \text{conditioned on } (E_0, E_v), \tag{12.81}$$

where (i) is true by (12.78), and (ii) is true by (12.62) from Step 4. By (12.81), we have that for every $v \in \left(\frac{1}{2}, \mu_1^*\right)$,

$$\|\widehat{\theta}_1^{(\infty)} - \widetilde{\theta}_1^{(B)}\|_\infty \leq \tau', \qquad \text{conditioned on } (E_0, E_v), \tag{12.82}$$

for all $d \geq d_0$ and all $k \geq k_0$, where d_0 and k_0 are constants. Combining (12.82) with (12.80), if the unconstrained MLE $\widehat{\theta}^{(\infty)}$ violates the box constraint, then only possible case is $\widehat{\theta}_1^{(\infty)} > B$. Then either $\widehat{\theta}_1^{(\infty)} = \widehat{\theta}_1^{(B)}$ (when $\widehat{\theta}^{(\infty)}$ does not violate the box constraint) or $\widehat{\theta}_1^{(\infty)} > B \geq \widehat{\theta}_1^{(B)}$ (when $\widehat{\theta}^{(\infty)}$ violates the box constraint). Hence, for every $v \in (\frac{1}{2}, \mu_1^*)$,

$$\widehat{\theta}_1^{(\infty)} \geq \widehat{\theta}_1^{(B)}, \qquad \text{conditioned on } (E_0, E_v). \tag{12.83}$$

Combining (12.78) and (12.83), we have that for every $v \in (\frac{1}{2}, \mu_1^*)$,

$$\widehat{\theta}^{(B)} - \widetilde{\theta}^{(B)} \leq \widehat{\theta}^{(\infty)} - \widetilde{\theta}^{(\infty)}, \qquad \text{conditioned on } (E_0, E_v). \tag{12.84}$$

By the law of iterated expectation again, we have

$$\begin{aligned}
R_{11} &= \sum_{v \in (\frac{1}{2}, \mu_1^*)} \mathbb{E}[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)} \mid E_0, \mu_1 = v] \cdot \mathbb{P}(E_0, \mu_1 = v) \\
&= \sum_{v \in (\frac{1}{2}, \mu_1^*)} \mathbb{E}[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)} \mid E_0, E_v] \cdot \mathbb{P}(E_0, E_v) \\
&\overset{\text{(i)}}{\leq} \sum_{v \in (\frac{1}{2}, \mu_1^*)} \mathbb{E}[\widehat{\theta}_1^{(\infty)} - \widetilde{\theta}_1^{(\infty)} \mid E_0, E_v] \cdot \mathbb{P}(E_0, E_v) \\
&\overset{\text{(ii)}}{\lesssim} \frac{\log d + \log k}{dk} \sum_{v \in (\frac{1}{2}, \mu_1^*)} \mathbb{P}(E_0, E_v) \\
&\lesssim \frac{\log d + \log k}{dk},
\end{aligned} \tag{12.85}$$

where (i) is true due to (12.84), and (ii) is true due to the bound (12.66) from Step 4. Plugging the term R_{11} from (12.85) and R_{12} from (12.77) back to (12.75), we have

$$R_1 = R_{11} + R_{12} \lesssim \frac{\log d + \log k}{dk}. \tag{12.86}$$

Finally, plugging the terms R_1 from (12.86), R_2 from (12.72), and R_3 from (12.74) back into (12.71) yields

$$\mathbb{E}[\widehat{\theta}_1^{(B)} - \widetilde{\theta}_1^{(B)}] \lesssim \frac{\log d + \log k}{dk},$$

completing the proof of (12.70).

12.1.4 Proofs of lemmas

In this section, we present the proofs of all the lemmas used for proving Theorem 5.4.

Proof of Lemma 12.2

We fix any constant $A > 0$.

The stretched-MLE (12.8) is an optimization over the compact set Θ_A, and the negative log-likelihood function ℓ is continuous. By the Extreme Value Theorem [149, Theorem 4.16], a solution $\widehat{\theta}^{(A)}$ is guaranteed to exist.

It remains to prove the uniqueness of $\widehat{\theta}^{(A)}$. Assume for contradiction that there exist two solutions $\widehat{\theta}, \widehat{\theta}' \in \Theta_A$ to the stretched-MLE (12.8) and $\widehat{\theta} \neq \widehat{\theta}'$. By Lemma 12.1, the negative log-likelihood function ℓ is strictly convex. Therefore,

$$\frac{1}{2}\left(\ell(\widehat{\theta}) + \ell(\widehat{\theta}')\right) > \ell\left(\frac{\widehat{\theta} + \widehat{\theta}'}{2}\right). \tag{12.87}$$

It can be verified that $\frac{\widehat{\theta}+\widehat{\theta}'}{2} \in \Theta_A$. Moreover, (12.87) along with the fact that $\ell(\widehat{\theta}) = \ell(\widehat{\theta}')$ implies that $\frac{\widehat{\theta}+\widehat{\theta}'}{2}$ attains a strictly smaller function value than both $\widehat{\theta}$ and $\widehat{\theta}'$. This contradicts the assumption that $\widehat{\theta}$ and $\widehat{\theta}'$ are both optimal solutions to the stretched-MLE (12.8).

Proof of Lemma 12.3

We first define a "comparison graph" $G(\{W_{ij}\})$ as a function of the pairwise-comparison outcomes $\{W_{ij}\}$. Let each item $i \in [d]$ be a node of the graph. Let there be a directed edge $(i \to j) \in G$, if and only if there exists a comparison where item i beats item j. A directed graph is called strongly-connected if and only if there exists a path from every node i to every other node j.

The following lemma from [63] relates the existence and uniqueness of a finite unconstrained MLE $\widehat{\theta}^{(\infty)}$ to the strong connectivity of the comparison graph G. This lemma is based on a different parameterization of the BTL model. In this parameterization, each item has a weight $w_i^* > 0$, and the probability that item i beats item j equals $\frac{w_i^*}{w_i^*+w_j^*}$.

Lemma 12.14 (Section 2 from [63]). *If the comparison graph $G(\{W_{ij}\})$ is strongly-connected, then there exists a unique solution to the following MLE:*

$$\widehat{w}_{\text{MLE}} = \underset{\substack{w\in\mathbb{R}^d \\ w_i>0,\ \sum_{i=1}^d w_i=1}}{\operatorname{argmin}} \quad \ell_w(\{W_{ij}\}; w),$$

where the negative log-likelihood function ℓ_w is defined as

$$\ell_w(w) = -\sum_{1\le i<j\le d} \left(W_{ij} \log\left(\frac{w_i}{w_i+w_j}\right) + W_{ji}\log\left(\frac{w_j}{w_i+w_j}\right)\right).$$

It can be seen that θ and w are simply different parameterizations of the same problem. There is a one-to-one mapping between θ and w, by taking $\theta_i = \log(w_i)$ and re-centering accordingly (or in the inverse direction, by taking $w_i = e^{\theta_i}$ and normalizing accordingly). Therefore, the existence and the uniqueness of the MLE $\widehat{w}_{\text{MLE}}$ in Lemma 12.3 carries over to our unconstrained MLE $\widehat{\theta}^{(\infty)}$ in (12.7). That is, if the comparison graph G is strongly-connected, then there exists a unique solution $\widehat{\theta}^{(\infty)}$ to the unconstrained MLE. It remains to show that the comparison graph G is strongly-connected w.h.p.($\frac{1}{dk}$).

We first construct an undirected graph $G'(\{W_{ij}\})$ as follows. Let each item $i \in [d]$ be a node of the graph G'. Let there be an undirected edge $(i, j) \in G'$, if and only if in the directed graph G we have both $(i \to j) \in G$ and $(j \to i) \in G$. Equivalently, there exists an undirected edge $(i, j) \in G'$, if and only if $0 < \mu_{ij} < 1$. It can be verified that the connectivity of the undirected graph G' implies the strong connectivity of the directed graph G. Therefore,

$$\mathbb{P}(G \text{ strongly-connected}) \ge \mathbb{P}(G' \text{ connected}). \tag{12.88}$$

The probability that $(i, j) \in G'$ is $\mathbb{P}(0 < \mu_{ij} < 1)$. By Hoeffding's inequality, we have that for any $t > 0$,

$$\mathbb{P}(|\mu_{ij} - \mu_{ij}^*| > t) < 2e^{-kt^2}, \qquad \text{for all } 1 \le i < j \le d.$$

We have $0 < \frac{1}{1+e^{2B}} \leq \mu^*_{ij} \leq \frac{1}{1+e^{-2B}} < 1$, for any $i < j$. Since B is a constant, we have that μ^*_{ij} is bounded away from 0 and 1 by a constant. Set $t = \tau$ where τ is any constant such that $0 < \tau < \frac{1}{1+e^{2B}}$. Then for all $1 \leq i < j \leq d$, we have

$$\begin{aligned}\mathbb{P}(0 < \mu_{ij} < 1) &> \mathbb{P}(\mu^*_{ij} - \tau < \mu_{ij} < \mu^*_{ij} + \tau)\\ &\geq 1 - \mathbb{P}(|\mu_{ij} - \mu^*_{ij}| > \tau)\\ &> 1 - 2e^{-ck},\end{aligned}$$

for some constant $c > 0$.

Recall that the random variables $\{\mu_{ij}\}$ are independent across all $1 \leq i < j \leq d$. Hence, the probability of the undirected graph G' being connected is at least the probability of an (undirected) Erdős-Rényi random graph being connected, where each edge independently exists with probability $1 - 2e^{-ck}$.

The following lemma from [68] provides an upper bound on the probability of an (undirected) Erdős-Rényi random graph being disconnected (and hence a lower bound on the probability of the graph being connected).

Lemma 12.15 (Theorem 1 from [68]). *For an (undirected) Erdős-Rényi graph of d nodes, where each edge independently exists with probability p. Let $q := 1 - p$. Then the probability of the graph being disconnected is at most*

$$\left(1 - \frac{d-1}{2} q^{d-1}\right) dq^{d-1}.$$

To apply Lemma 12.15, we set $p = 1 - 2e^{-ck}$ and therefore $q = 2e^{-ck}$. Then we have

$$\begin{aligned}\mathbb{P}[G' \text{ disconnected}] &\leq \left(1 - \frac{d-1}{2} q^{d-1}\right) dq^{d-1}\\ &\leq dq^{d-1}\\ &= de^{-ck(d-1)}\\ &\leq \frac{c'}{dk}, \qquad \text{for some constant } c' > 0. \end{aligned} \tag{12.89}$$

Combining (12.88) and (12.89) completes the proof of the lemma.

Proof of Lemma 12.4

We first consider any fixed $m \in [d]$. By the definition of $\{\mu_{ij}\}$ in (12.2), we have

$$\sum_{i \neq m} \mu_{mi} = \frac{1}{k} \sum_{i \neq m} \sum_{r=1}^{k} X^{(r)}_{mi}. \tag{12.90}$$

There are $(d-1)k$ terms of the form $X^{(r)}_{mi}$ in (12.90). It can be verified that the terms $X^{(r)}_{mi}$ involved in (12.90) are independent. Moreover, since $X^{(r)}_{mi} \in \{0, 1\}$, changing the value of a single term $X_{mi}^{(r)}$ changes the value of (12.90) by $\frac{1}{k}$. By McDiarmid's inequality, we have that for any $t > 0$,

$$\mathbb{P}\left[\left|\sum_{i\neq m}\mu_{mi} - \sum_{i\neq m}\mu^*_{mi}\right| > t\right] \leq 2\exp\left(-\frac{2t^2}{(d-1)k\cdot(\frac{1}{k})^2}\right) = 2\exp\left(-\frac{2kt^2}{(d-1)}\right). \tag{12.91}$$

Setting $t = c\sqrt{\frac{d(\log d+\log k)}{k}}$ in (12.91), we have

$$\begin{aligned}\mathbb{P}\left[\left|\sum_{i\neq m}\mu_{mi} - \sum_{i\neq m}\mu^*_{mi}\right| \leq c\sqrt{\frac{d(\log d+\log k)}{k}}\right] &\geq 1-2\exp\left(-c'\frac{d}{d-1}(\log d+\log k)\right)\\ &\geq 1-\frac{c''}{d^2k},\end{aligned} \tag{12.92}$$

for some constants $c', c'' > 0$, provided that the constant $c > 0$ is sufficiently large.

Taking a union bound over $m \in [d]$ on (12.92) completes the proof.

Proof of Lemma 12.5

Denote the random variables $m^+ := \text{argmax}_{i\in[d]}\,\delta_i$ and $m^- := \text{argmin}_{i\in[d]}\,\delta_i$. When there are multiple maximizers or minimizers, we arbitrarily choose one.

Setting $m = m^+$ in the first-order optimality condition (12.27), we have

$$\underbrace{\sum_{i=1}^{d}[f(\theta^*_{m^+}-\theta^*_i+\delta_{m^+}-\delta_i) - f(\theta^*_{m^+}-\theta^*_i)]}_{T^+} = \sum_{i\neq m^+}(\mu_{mi}-\mu^*_{mi}) \overset{\text{(i)}}{\lesssim} \sqrt{\frac{d(\log d+\log k)}{k}}, \tag{12.93}$$

where (i) is true by Lemma 12.4 (recall that the lemma statement is conditioned on the event E_0 that both Lemma 12.3 and Lemma 12.4 hold).

Denote the function $g(x,t) := f(x+t) - f(x) = \frac{1}{1+e^{-(x+t)}} - \frac{1}{1+e^{-x}}$. The following lemma states three properties for the function g, which are used in later parts of the proof.

Lemma 12.16. *We have the following properties for the function g.*

$$g(x,t) = -g(-x,-t), \qquad \textit{for all } x,t\in\mathbb{R} \tag{12.94a}$$

$$g(x,t) \geq g(\tau,t) > 0, \qquad \textit{for all } \tau>0,\ t>0,\ \textit{and all } x \textit{ such that } -\tau\leq x\leq\tau \tag{12.94b}$$

$$g(\tau,t_1)+g(\tau,t_2) \geq g(\tau,t_1+t_2), \qquad \textit{for all } \tau>0, \textit{and all } t_1,t_2\geq 0. \tag{12.94c}$$

Lemma 12.16 can be verified by straightforward algebra. For completeness, we include the proof of Lemma 12.16 at the end of this section.

By the definition of m^+, we have $\delta_{m^+} = \max_{i\in[d]} \delta_i$, and therefore $\delta_{m^+} - \delta_i \geq 0$ for all $i \in [d]$. Hence, we have

$$\begin{aligned} T^+ &= \sum_{i=1}^d f(\theta^*_{m^+} - \theta^*_i + \delta_{m^+} - \delta_i) - f(\theta^*_{m^+} - \theta^*_i) \\ &= \sum_{i=1}^d g(\theta^*_{m^+} - \theta^*_i, \delta_{m^+} - \delta_i) \\ &\overset{\text{(i)}}{\geq} \sum_{i=1}^d g(2B, \delta_{m^+} - \delta_i), \end{aligned} \tag{12.95}$$

where (i) is true by (12.94b) combined with the fact that $|\theta^*_i - \theta^*_j| \leq |\theta^*_i| + |\theta^*_j| \leq 2B$ for all $i, j \in [d]$.

Similarly, setting $m = m^-$ in the first-order optimality condition (12.27), we have

$$\underbrace{\sum_{i=1}^d [f(\theta^*_{m^-} - \theta^*_i + \delta_{m^-} - \delta_i) - f(\theta^*_{m^-} - \theta^*_i)]}_{T^-} \lesssim \sqrt{\frac{d(\log d + \log k)}{k}}. \tag{12.96}$$

By the definition of m^-, we have $\delta_{m^-} = \min_{i\in[d]} \delta_i$, and therefore $\delta_i - \delta_{m^-} \geq 0$ for all $i \in [d]$. Hence, we have

$$\begin{aligned} T^- &= \sum_{i=1}^d f(\theta^*_{m^-} - \theta^*_i + \delta_{m^-} - \delta_i) - f(\theta^*_{m^-} - \theta^*_i) \\ &= \sum_{i=1}^d g(\theta^*_{m^-} - \theta^*_i, \delta_{m^-} - \delta_i) \\ &\overset{\text{(i)}}{=} \sum_{i=1}^d -g(\theta^*_i - \theta^*_{m^-}, \delta_i - \delta_{m^-}) \\ &\overset{\text{(ii)}}{\leq} \sum_{i=1}^d -g(2B, \delta_i - \delta_{m^-}), \end{aligned} \tag{12.97}$$

where (i) is true by (12.94a), and (ii) is true by (12.94b) combined with the fact that $|\theta^*_i - \theta^*_j| \leq 2B$ for all $i, j \in [d]$.

Combining (12.95) and (12.97), we have

$$\begin{aligned} T^+ - T^- &\geq \sum_{i=1}^d g(2B, \delta_{m^+} - \delta_i) + \sum_{i=1}^d g(2B, \delta_i - \delta_{m^-}) \\ &\overset{\text{(i)}}{\geq} \sum_{i=1}^d g(2B, \delta_{m^+} - \delta_{m^-}) \\ &= d \cdot g(2B, \delta_{m^+} - \delta_{m^-}) \overset{\text{(ii)}}{\geq} 0, \end{aligned} \tag{12.98}$$

where (i) is true due to (12.94c) since $\delta_{m^+} - \delta_i \geq 0$ and $\delta_i - \delta_{m^-} \geq 0$ for all $i \in [d]$, and (ii) is true since $\delta_{m^+} - \delta_{m^-} \geq 0$. On the other hand, combining (12.93) and (12.96), we have

$$T^+ - T^- \lesssim \sqrt{\frac{d(\log d + \log k)}{k}}. \tag{12.99}$$

Combining (12.98) and (12.99), we have

$$\begin{aligned} 0 \leq d \cdot g(2B, \delta_{m^+} - \delta_{m^-}) \leq T^+ - T^+ &\lesssim \sqrt{\frac{d(\log d + \log k)}{k}} \\ g(2B, \delta_{m^+} - \delta_{m^-}) &\lesssim \sqrt{\frac{\log d + \log k}{dk}} \\ f(2B + \delta_{m^+} - \delta_{m^-}) - f(2B) &\lesssim \sqrt{\frac{\log d + \log k}{dk}}. \end{aligned} \tag{12.100}$$

By the first-order mean value theorem on the LHS of (12.100), we have

$$f(2B + \delta_{m^+} - \delta_{m^-}) - f(2B) = f'(\lambda) \cdot (\delta_{m^+} - \delta_{m^-}) \leq c\sqrt{\frac{\log d + \log k}{dk}}, \tag{12.101}$$

where λ is a random variable that takes values in the interval $[2B, 2B + \delta_{m^+} - \delta_{m^-}]$.

Let ϵ be any constant such that $0 < \epsilon < 1 - f(2B)$. Then there exists a constant $\tau > 0$ such that $f(2B + \tau) - f(2B) = \epsilon$. On the other hand, there exist constants $d_0 > 0$ and $k_0 > 0$ such that

$$c\sqrt{\frac{\log d + \log k}{dk}} < \epsilon, \qquad \text{for any } d \geq d_0 \text{ and } k \geq k_0. \tag{12.102}$$

Combining (12.101) and (12.102), we have

$$\begin{aligned} f(2B + \delta_{m^+} - \delta_{m^-}) - f(2B) \leq c\sqrt{\frac{\log d + \log k}{dk}} &< \epsilon = f(2B + \tau) - f(2B) \\ f(2B + \delta_{m^+} - \delta_{m^-}) &\leq f(2B + \tau). \end{aligned} \tag{12.103}$$

By (12.5a), we have $f' > 0$ on $(-\infty, \infty)$, and hence the function f is monotonically increasing. Hence, from (12.103), we have $\delta_{m^+} - \delta_{m^-} \leq \tau$, and therefore the interval $[2B, 2B + \delta_{m^+} - \delta_{m^-}]$ is bounded. By the property (12.5a) of the sigmoid function f, we have $f' > c_3 > 0$ for some constant $c_3 > 0$ in the bounded interval $[2B, 2B + \delta_{m^+} - \delta_{m^-}]$. Recall that λ takes values in the interval $[2B, 2B + \delta_{m^+} - \delta_{m^-}]$. Therefore, we have

$$c_3(\delta_{m^+} - \delta_{m^-}) < f'(\lambda) \cdot (\delta_{m^+} - \delta_{m^-}). \tag{12.104}$$

Combining (12.101) and (12.104), we have

$$\begin{aligned} c_3(\delta_{m^+} - \delta_{m^-}) < f'(\lambda) \cdot (\delta_{m^+} - \delta_{m^-}) &\leq c\sqrt{\frac{\log d + \log k}{dk}} \\ \delta_{m^+} - \delta_{m^-} &\lesssim \sqrt{\frac{\log d + \log k}{dk}}. \end{aligned} \tag{12.105}$$

By the assumption that $\theta^* \in \Theta_B$, we have $\sum_{i=1}^d \theta_i^* = 0$. Similarly, by the centering constraint on the unconstrained MLE $\widehat{\theta}^{(\infty)}$ in (12.7), we have $\sum_{i=1}^d \widehat{\theta}_i^{(\infty)} = 0$. Hence, we have the deterministic relation

$$\sum_{i=1}^d \widehat{\theta}_i^{(\infty)} - \sum_{i=1}^d \theta_i^* = \sum_{i=1}^d \delta_i = 0. \tag{12.106}$$

Hence, $\delta_{m^+} \geq 0$ and $\delta_{m^-} \leq 0$. By (12.105), we have

$$\delta_{m^+} - \delta_{m^-} = |\delta_{m^+}| + |\delta_{m^-}| \lesssim \sqrt{\frac{\log d + \log k}{dk}}.$$

Hence, $|\delta_{m^+}| \lesssim \sqrt{\frac{\log d + \log k}{dk}}$ and $|\delta_{m^-}| \lesssim \sqrt{\frac{\log d + \log k}{dk}}$. Therefore,

$$|\delta_m| \lesssim \sqrt{\frac{\log d + \log k}{dk}}, \qquad \text{for all } m \in [d],$$

completing the proof of the lemma.

Proof of Lemma 12.16: We prove the three parts of the lemma separately.

(a) It can be verified that $f(x) = 1 - f(-x)$. Hence,

$$\begin{aligned} g(x,t) = f(x+t) - f(x) &= [1 - f(-x-t)] - [1 - f(-x)] \\ &= -[f(-x-t) - f(-x)] = -g(-x,-t). \end{aligned}$$

(b) We prove the two parts of the inequality separately.

We first prove that $g(\tau, t) > 0$. By (12.5a), the function f is strictly increasing. Therefore, for any $t > 0$, we have

$$g(\tau,t) = f(\tau + t) - f(\tau) > 0.$$

Now we prove that $g(x,t) \geq g(\tau,t)$. We have

$$\begin{aligned} g(x,t) - g(\tau,t) &= f(x+t) - f(x) - [f(\tau+t) - f(\tau)] \\ &= \int_x^{x+t} f'(u)\,\mathrm{d}u - \int_\tau^{\tau+t} f'(u)\,\mathrm{d}u \\ &= \int_0^t f'(x+u)\,\mathrm{d}u - \int_0^t f'(\tau+u)\,\mathrm{d}u \\ &= \int_0^t [f'(x+u) - f'(\tau+u)]\,\mathrm{d}u. \end{aligned} \tag{12.107}$$

By (12.107), it remains to prove that

$$f'(x+u) \geq f'(\tau+u), \qquad \text{for any } u \in [0,t]. \tag{12.108}$$

Fix any $u \in [0, t]$. By assumption we have $\tau > 0$. Hence, $\tau + u > 0$. Now we consider the sign of $(x + u)$.

If $x + u \geq 0$, then by the assumption that $x \leq \tau$, we have $0 \leq x + u \leq \tau + u$. It can be verified that f' is decreasing on $[0, \infty)$. Therefore,

$$f'(x + u) \geq f'(\tau + u). \tag{12.109}$$

If $x + u < 0$, we have

$$0 < -x - u \overset{\text{(i)}}{\leq} \tau - u \overset{\text{(ii)}}{\leq} \tau + u, \tag{12.110}$$

where (i) is true by the assumption that $x \geq -\tau$, and (ii) is true because $u \in [0, t]$ and therefore $u \geq 0$. We have

$$f'(x + u) \overset{\text{(i)}}{=} f'(-x - u) \overset{\text{(ii)}}{\geq} f'(\tau + u), \tag{12.111}$$

where (i) holds because it can be verified that $f'(x) = f'(-x)$ for any $x \in \mathbb{R}$, and (ii) is true by combining (12.110) with the fact that f' is decreasing on $[0, \infty)$.

Combining the two cases of (12.109) and (12.111) completes the proof of (12.108).

(c) We have

$$\begin{aligned}
g(\tau, t_1) + g(\tau, t_2) &= f(\tau + t_1) - f(\tau) + f(\tau + t_2) - f(\tau) \\
&= \int_{\tau}^{\tau + t_1} f'(u)\,\mathrm{d}u + \int_{\tau}^{\tau + t_2} f'(u)\,\mathrm{d}u \\
&\overset{\text{(i)}}{\geq} \int_{\tau}^{\tau + t_1} f'(u)\,\mathrm{d}u + \int_{\tau + t_1}^{\tau + t_1 + t_2} f'(u)\,\mathrm{d}u \\
&= \int_{\tau}^{\tau + t_1 + t_2} f'(u)\,\mathrm{d}u \\
&= f(\tau + t_1 + t_2) - f(\tau) = g(\tau, t_1 + t_2),
\end{aligned}$$

where (i) is true because f' is decreasing on $(0, \infty)$, and because $\tau > 0$ and $t_1, t_2 \geq 0$ by assumption.

Proof of Lemma 12.6

We fix any $i, j \in [d]$ where $i \neq j$. By the law of iterated expectation, we have

$$\mathbb{E}[\mu_{ij} \mid E] = \mathbb{E}[\mu_{ij} \mid E', E] \cdot \mathbb{P}(E' \mid E) + \mathbb{E}[\mu_{ij} \mid \overline{E}', E] \cdot \mathbb{P}(\overline{E}' \mid E). \tag{12.112}$$

Subtracting $\mathbb{E}[\mu_{ij} \mid E', E]$ from both sides of (12.112), we have

$$\begin{aligned}
\mathbb{E}[\mu_{ij} \mid E] - \mathbb{E}[\mu_{ij} \mid E', E] &= \mathbb{E}[\mu_{ij} \mid E', E] \cdot [\mathbb{P}(E' \mid E) - 1] + \mathbb{E}[\mu_{ij} \mid \overline{E}', E] \cdot \mathbb{P}(\overline{E}' \mid E) \\
&= (-\mathbb{E}[\mu_{ij} \mid E', E] + \mathbb{E}[\mu_{ij} \mid \overline{E}', E]) \cdot \mathbb{P}(\overline{E}' \mid E). \qquad (12.113)
\end{aligned}$$

Taking an absolute value on (12.113), we have

$$\begin{aligned}|\mathbb{E}[\mu_{ij} \mid E] - \mathbb{E}[\mu_{ij} \mid E', E]| &= \left|-\mathbb{E}[\mu_{ij} \mid E', E] + \mathbb{E}[\mu_{ij} \mid \overline{E}', E]\right| \cdot \mathbb{P}(\overline{E}' \mid E)\\ &\overset{\text{(i)}}{\lesssim} \frac{1}{dk},\end{aligned}$$

where (i) is true due to the deterministic inequality $0 \leq \mu_{ij} \leq 1$ and the fact that event E' happens w.h.p.($\frac{1}{dk} \mid E$).

Proof of Lemma 12.7

Denote $m^+ := \operatorname{argmax}_{i \in [d]} \Delta_i$ and $m^- := \operatorname{argmin}_{i \in [d]} \Delta_i$. When there are multiple maximizers or minimizers, we arbitrarily choose one. The proof works similarly in spirit to the proof of Lemma 12.5. We first show that $\Delta_{m^+} - \Delta_{m^-}$ satisfies the desired upper bound. Then we show that Δ_{m^+} and Δ_{m^-} have different signs, and therefore the desired upper bound holds on $|\Delta_m|$ uniformly across all $m \in [d]$.

Recall from (12.30) that for every $m \in [d]$,

$$\sum_{i=1}^{d} f'(\theta^*_m - \theta^*_i) \cdot (\Delta_m - \Delta_i) = \underbrace{\sum_{i \neq m} (\mathbb{E}[\mu_{mi} \mid E_0] - \mu^*_{mi})}_{R_1} - \underbrace{\frac{1}{2} \sum_{i=1}^{d} \mathbb{E}[f''(\lambda_{mi})(\delta_m - \delta_i)^2 \mid E_0]}_{R_2}, \tag{12.114}$$

where λ_{mi} is a random variable that takes values between $\theta^*_m - \theta^*_i$ and $\theta^*_m - \theta^*_i + (\delta_m - \delta_i)$.

We consider the two terms on the RHS of (12.30) separately. For the term R_1, recall from (12.32) that

$$|\mathbb{E}[\mu_{mi} \mid E_0] - \mu^*_{mi}| \lesssim \frac{1}{dk}.$$

Therefore,

$$|R_1| \lesssim (d-1) \cdot \frac{1}{dk} \lesssim \frac{1}{k}. \tag{12.115}$$

Now consider the term R_2. Recall that $\theta^* \in \Theta_B$. Therefore, for every $m \in [d]$, we have $|\theta^*_m| \leq B$. Recall from Lemma 12.5 that for every $m \in [d]$, we have

$$|\delta_m| \lesssim \sqrt{\frac{\log d + \log k}{dk}}, \qquad \text{conditioned on } E_0. \tag{12.116}$$

Let $c > 0$ be any constant. By (12.116), we have $|\delta_m| \leq c$, for all $d \geq d_0$ and $k \geq k_0$, where d_0 and k_0 are constants which may only depend on c. Hence, conditioned on E_0, the interval between $\theta^*_m - \theta^*_i$ and $\theta^*_m - \theta^*_i + (\delta_m - \delta_i)$ is contained in the interval $[-2B - 2c, 2B + 2c]$. By the property (12.5b) of the sigmoid function f, we have

$$|f''| < c_5, \qquad \text{on the bounded interval } [-2B - 2c, 2B + 2c].$$

Therefore,

$$\left|\mathbb{E}\left[f''(\lambda_{mi}) \cdot (\delta_m - \delta_i)^2 \mid E_0\right]\right| \leq c_5 \cdot \mathbb{E}[(\delta_m - \delta_i)^2 \mid E_0] \overset{\text{(i)}}{\lesssim} \frac{\log d + \log k}{dk}, \qquad \text{for all } i, m \in [d],$$

where (i) is again by (12.116). Therefore,

$$|R_2| \lesssim d \cdot \frac{\log d + \log k}{dk} = \frac{\log d + \log k}{k}. \tag{12.117}$$

Taking an absolute value on (12.114) and using the triangle inequality, we have

$$\left|\sum_{i=1}^{d} f'(\theta_m^* - \theta_i^*) \cdot (\Delta_m - \Delta_i)\right| \leq |R_1| + |R_2| \overset{\text{(i)}}{\lesssim} \frac{\log d + \log k}{k}, \tag{12.118}$$

where (i) is true by combining the term R_1 from (12.115) and the term R_2 from (12.117). Taking $m = m^+$ in (12.118), we have

$$\sum_{i=1}^{d} f'(\theta_{m^+}^* - \theta_i^*) \cdot (\Delta_{m^+} - \Delta_i) \leq c\frac{\log d + \log k}{k}. \tag{12.119}$$

Taking $m = m^-$ in (12.118), we have

$$\sum_{i=1}^{d} f'(\theta_{m^-}^* - \theta_i^*) \cdot (\Delta_{m^-} - \Delta_i) \geq -c\frac{\log d + \log k}{k}$$

and hence

$$\sum_{i=1}^{d} f'(\theta_{m^-}^* - \theta_i^*) \cdot (\Delta_i - \Delta_{m^-}) \leq c\frac{\log d + \log k}{k}. \tag{12.120}$$

Adding (12.119) and (12.120), we have

$$\underbrace{\sum_{i=1}^{d} f'(\theta_{m^+}^* - \theta_i^*) \cdot (\Delta_{m^+} - \Delta_i) + \sum_{i=1}^{d} f'(\theta_{m^-}^* - \theta_i^*) \cdot (\Delta_i - \Delta_{m^-})}_{R} \leq c\frac{\log d + \log k}{k}. \tag{12.121}$$

Consider the term R. We have $|\theta_m^* - \theta_i^*| \leq 2B$ for all $i, m \in [d]$. By the property (12.5a) of the sigmoid function, there exists some constant c_3, such that

$$f'(\theta_m^* - \theta_i^*) > c_3 > 0, \qquad \text{for all } i, m \in [d]. \tag{12.122}$$

By the definition of m^+ and m^-, we have $\Delta_{m^+} - \Delta_i \geq 0$ and $\Delta_i - \Delta_{m^-} \geq 0$ for every $i \in [d]$. Plugging (12.122) into (12.121), combined with the fact that $\Delta_{m^+} - \Delta_i \geq 0$ and $\Delta_i - \Delta_{m^-} \geq 0$, we have

$$\begin{aligned} c_3\left[\sum_{i=1}^{d}(\Delta_{m^+}-\Delta_i)+\sum_{i=1}^{d}(\Delta_i-\Delta_{m^-})\right] \leq R &\leq c\frac{\log d+\log k}{k} \\ c_3 d\cdot(\Delta_{m^+}-\Delta_{m^-}) &\leq c\frac{\log d+\log k}{k} \\ \Delta_{m^+}-\Delta_{m^-} &\lesssim \frac{\log d+\log k}{dk}. \end{aligned} \tag{12.123}$$

By (12.106) in the proof of Lemma 12.5, we have the deterministic relation

$$\sum_{i=1}^{d}\delta_i = 0. \tag{12.124}$$

Taking an expectation over (12.124) conditional on E_0, we have

$$\sum_{i=1}^{d}\Delta_i = 0.$$

Hence, $\Delta_{m^+} \geq 0$ and $\Delta_{m^-} \leq 0$. By (12.123), we have

$$\Delta_{m^+}-\Delta_{m^-} = |\Delta_{m^+}|+|\Delta_{m^-}| \lesssim \frac{\log d+\log k}{dk}.$$

Hence, $|\Delta_{m^+}| \lesssim \frac{\log d+\log k}{dk}$ and $|\Delta_{m^-}| \lesssim \frac{\log d+\log k}{dk}$. Therefore,

$$|\Delta_m| \lesssim \frac{\log d+\log k}{dk}, \qquad \text{for all } m\in[d].$$

Proof of Lemma 12.9

To compare the functions h and h^+, we introduce an auxiliary function $h_0 : [0,1] \to [-B, B]$:

$$h_0(t) = \begin{cases} -B & \text{if } 0 \leq t \leq \mu_- \\ \frac{B}{\mu_+-\frac{1}{2}}(t-\frac{1}{2}) & \text{if } \mu_- < t < \mu_+ \\ B & \text{if } \mu_+ \leq t \leq 1. \end{cases}$$

In words, the function h_0 is piecewise linear. On the interval $[0, \mu_-]$, its value equals the constant $-B$. On the interval $[\mu_-, \mu_+]$, it is a line passing through the points $(\mu_-, -B)$ and (μ_+, B). On the interval $[\mu_+, 1]$, its value equals the constant B. See Fig. 12.1 for a comparison of the three functions h, h^+ and h_0.

It can be verified that $h^+(t) \geq h_0(t)$ for any $t \in [0,1]$. Hence,

$$\mathbb{E}[h^+(\mu)] \geq \mathbb{E}[h_0(\mu)]. \tag{12.125}$$

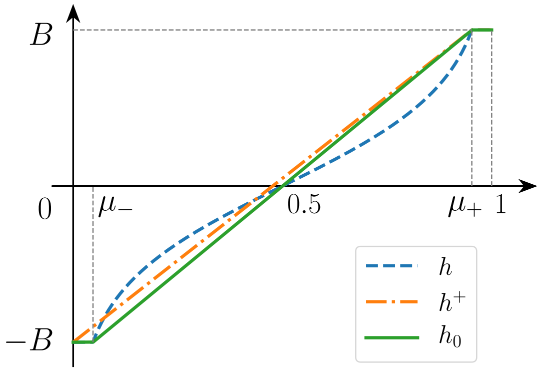


Figure 12.1: The functions h, h^+ and h_0.

Recall that our goal is to prove (12.44):

$$\mathbb{E}[h(\mu)] \leq \mathbb{E}[h^+(\mu)].$$

Given (12.125), it suffices to prove that

$$\mathbb{E}[h(\mu)] \leq \mathbb{E}[h_0(\mu)]. \tag{12.126}$$

The rest of the proof is devoted to proving (12.126).

It can be verified that h and h_0 are anti-symmetric around $\frac{1}{2}$. That is, for any $t \in [0,1]$, we have

$$h(t) = -h(1-t) \tag{12.127a}$$
$$h_0(t) = -h_0(1-t). \tag{12.127b}$$

In particular, we have

$$h\left(\frac{1}{2}\right) = h_0\left(\frac{1}{2}\right) = 0. \tag{12.128}$$

It can also be verified that

$$h(t) \geq h_0(t), \qquad \text{for all } t \in \left[0, \frac{1}{2}\right]. \tag{12.129}$$

Recall the notation of $W = k\mu$ representing the number of times that item 1 beats item 2

among the k comparisons between them. We have $W \sim \text{Binom}(k, \mu_+)$. Therefore,

$$\begin{aligned}
\mathbb{E}[h(\mu)] - \mathbb{E}\left[h_0(\mu)\right] &= \mathbb{E}_W\left[h\left(\frac{W}{k}\right)\right] - \mathbb{E}_W\left[h_0\left(\frac{W}{k}\right)\right] \\
&= \sum_{w=0}^{k}\left[h\left(\frac{w}{k}\right) - h_0\left(\frac{w}{k}\right)\right] \cdot \mathbb{P}(W = w) \\
&\overset{\text{(i)}}{=} \left(\sum_{w=0}^{\lfloor \frac{k}{2} \rfloor} + \sum_{w=\lceil \frac{k}{2} \rceil}^{k}\right)\left[(h - h_0)\left(\frac{w}{k}\right)\right] \cdot \mathbb{P}(W = w) \\
&\overset{\text{(ii)}}{=} \sum_{w=0}^{\lfloor \frac{k}{2} \rfloor}\left[(h - h_0)\left(\frac{w}{k}\right) \cdot \mathbb{P}(W = w) + (h - h_0)\left(1 - \frac{w}{k}\right) \cdot \mathbb{P}(W = k - w)\right] \\
&\overset{\text{(iii)}}{=} \sum_{w=0}^{\lfloor \frac{k}{2} \rfloor}(h - h_0)\left(\frac{w}{k}\right) \cdot [\mathbb{P}(W = w) - \mathbb{P}(W = k - w)], \qquad (12.130)
\end{aligned}$$

where (i) is true by (12.128). Specifically, when k is even, we double-count the term of $w = \frac{k}{2}$. This term equals $(h - h_0)(\frac{1}{2}) = 0$, so double-counting this term does not affect the equality. Moreover, step (ii) is true by a change of variable $w \leftarrow k - w$ in the second summation, and step (iii) is true by the anti-symmetry (12.127) of the functions h and h^+.

Now consider the terms in the summation (12.130). By (12.129), we have

$$(h - h_0)\left(\frac{w}{k}\right) \geq 0, \qquad \text{for all } 0 \leq w \leq \left\lfloor \frac{k}{2} \right\rfloor. \qquad (12.131)$$

Using the binomial probabilities of $W \sim \text{Binom}(k, \mu_+)$, we also have

$$\begin{aligned}
\mathbb{P}(W = w) - \mathbb{P}(W = k - w) &= \binom{k}{w}[(\mu_+)^w(1 - \mu_+)^{k-w} - (\mu_+)^{k-w}(1 - \mu_+)^w] \\
&= \binom{k}{w}(\mu_+)^w(1 - \mu_+)^w \cdot [(1 - \mu_+)^{k-2w} - (\mu_+)^{k-2w}] \\
&\overset{\text{(i)}}{\leq} 0, \qquad \text{for all } 0 \leq w \leq \left\lfloor \frac{k}{2} \right\rfloor, \qquad (12.132)
\end{aligned}$$

where (i) is true because $\mu_+ = \frac{1}{1+e^{-2B}} > \frac{1}{2}$, combined with the fact that $k - 2w \geq 0$, for all $0 \leq w \leq \left\lfloor \frac{k}{2} \right\rfloor$. Plugging (12.131) and (12.132) back into (12.130), we have

$$\mathbb{E}[h(\mu)] - \mathbb{E}[h_0(\mu)] \geq 0,$$

completing the proof of (12.126).

Proof of Lemma 12.10

We have

$$\begin{aligned}
\mathbb{E}[h^+(\mu)] - \theta_1^* &= \mathbb{E}_W\left[h^+\left(\frac{W}{k}\right)\right] - B \\
&= \sum_{w=0}^{k} h^+\left(\frac{w}{k}\right) \cdot \mathbb{P}(W = w) - B \\
&\overset{\text{(i)}}{=} \sum_{w=0}^{\lfloor k\mu_+ \rfloor} \frac{2B}{\mu_+}\left(\frac{w}{k} - \mu_+\right) \cdot \mathbb{P}(W = w) \\
&= c\left(\underbrace{\sum_{w=0}^{\lfloor k\mu_+ \rfloor} \frac{w}{k} \cdot \mathbb{P}(W = w)}_{R_1} - \mu_+ \underbrace{\sum_{w=0}^{\lfloor k\mu_+ \rfloor} \mathbb{P}(W = w)}_{R_2}\right), \qquad (12.133)
\end{aligned}$$

where (i) is true by plugging in the definition of the function h^+ from (12.43).

Now we consider the two terms R_1 and R_2 separately. For any integer $n \geq 1$, any integer s such that $0 \leq s \leq n$, and any real number $p \in [0, 1]$, we define $\mathcal{P}_{\text{le}}(n, p, s)$ (resp. $\mathcal{P}_{\text{eq}}(n, p, s)$) as the probability that the value of the random variable Binom(n, p) is at most (resp. equal to) s. That is,

$$\begin{aligned}
\mathcal{P}_{\text{le}}(n, p, s) &= \mathbb{P}[\text{Binom}(n, p) \leq s], \\
\mathcal{P}_{\text{eq}}(n, p, s) &= \mathbb{P}[\text{Binom}(n, p) = s].
\end{aligned}$$

Then the term R_2 can be written as

$$R_2 = \mathcal{P}_{\text{le}}(k, \mu_+, \lfloor k\mu_+ \rfloor). \qquad (12.134)$$

For the term R_1, we have

$$\begin{aligned}
R_1 = \sum_{w=0}^{\lfloor k\mu_+ \rfloor} \frac{w}{k} \cdot \mathbb{P}(W = w) &= \sum_{w=0}^{\lfloor k\mu_+ \rfloor} \frac{w}{k} \cdot \binom{k}{w} \mu_+^w (1 - \mu_+)^{(k-w)} \\
&= \sum_{w=1}^{\lfloor k\mu_+ \rfloor} \frac{w}{k} \cdot \frac{k!}{w!(k-w)!} \mu_+^w (1 - \mu_+)^{(k-w)} \\
&= \mu_+ \sum_{w=1}^{\lfloor k\mu_+ \rfloor} \frac{(k-1)!}{(w-1)!(k-w)!} \mu_+^{w-1} (1 - \mu_+)^{(k-w)} \\
&\overset{\text{(i)}}{=} \mu_+ \sum_{w=0}^{\lfloor k\mu_+ \rfloor - 1} \frac{(k-1)!}{(w)!(k-w-1)!} \mu_+^w (1 - \mu_+)^{(k-1-w)} \\
&= \mu_+ \sum_{w=0}^{\lfloor k\mu_+ \rfloor - 1} \binom{k-1}{w} \mu_+^w (1 - \mu_+)^{(k-1-w)} \\
&= \mu_+ \cdot \mathcal{P}_{\text{le}}(k - 1, \mu_+, \lfloor k\mu_+ \rfloor - 1), \qquad (12.135)
\end{aligned}$$

where (i) is true by a change of variable $w \leftarrow w - 1$. Plugging (12.134) and (12.135) back into (12.133), we have

$$\mathbb{E}[h^+(\mu)] - \theta_1^* = c\mu_+ \cdot [\mathcal{P}_{\text{le}}(k-1, \mu_+, \lfloor k\mu_+ \rfloor - 1) - \mathcal{P}_{\text{le}}(k, \mu_+, \lfloor k\mu_+ \rfloor)]. \tag{12.136}$$

For any integer $n \geq 1$, any integer s such that $0 \leq s \leq n$, and any $p \in [0, 1]$, we claim the combinatorial equality

$$\mathcal{P}_{\text{le}}(n, p, s) = \mathcal{P}_{\text{le}}(n-1, p, s-1) + (1-p) \cdot \mathcal{P}_{\text{eq}}(n-1, p, s). \tag{12.137}$$

To prove (12.137), we use a standard combinatorial argument. Consider n balls, and we select each ball independently with probability p. Then the LHS of (12.137) equals the probability that at most s balls are selected. This event can be decomposed into two cases. Either there are at most $(s-1)$ balls selected from the first $(n-1)$ balls; or there are exactly s balls selected from the first $(n-1)$ balls, and the last ball is not selected. These two cases correspond to the two terms on the RHS of (12.137).

Now setting $n = k, p = \mu_+$, and $s = \lfloor k\mu_+ \rfloor$ in (12.137), we have

$$\mathcal{P}_{\text{le}}(k, \mu_+, \lfloor k\mu_+ \rfloor) = \mathcal{P}_{\text{le}}(k-1, \mu_+, \lfloor k\mu_+ \rfloor - 1) + (1-\mu_+) \cdot \mathcal{P}_{\text{eq}}(k-1, \mu_+, \lfloor k\mu_+ \rfloor]). \tag{12.138}$$

Combining (12.136) and (12.138), we have

$$\mathbb{E}[h^+(\mu)] - \theta_1^* = -c(1-\mu_+) \cdot \mathcal{P}_{\text{eq}}(k-1, \mu_+, \lfloor k\mu_+ \rfloor). \tag{12.139}$$

It remains to bound the term $\mathcal{P}_{\text{eq}}(k-1, \mu_+, \lfloor k\mu_+ \rfloor)$ on the RHS of (12.139). Writing out the binomial probability, we have

$$\mathcal{P}_{\text{eq}}(k-1, \mu_+, \lfloor k\mu_+ \rfloor) = \binom{k-1}{\lfloor k\mu_+ \rfloor} \mu_+^{\lfloor k\mu_+ \rfloor} (1-\mu_+)^{k-1-\lfloor k\mu_+ \rfloor}. \tag{12.140}$$

By the Stirling's approximation, we have

$$\sqrt{2\pi} \cdot k^{k+\frac{1}{2}} e^{-k} \leq k! \leq e \cdot k^{k+\frac{1}{2}} e^{-k}, \qquad \text{for any integer } k \geq 0.$$

Then for any integer $n \geq 1$, and any integer k such that $0 \leq k \leq n$, we have

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \geq c \frac{n^{n+\frac{1}{2}}}{k^{k+\frac{1}{2}}(n-k)^{n-k+\frac{1}{2}}}. \tag{12.141}$$

Plugging (12.141) into (12.140), we have

$$\begin{aligned}
\mathcal{P}_{\text{eq}}(k-1,\mu_+,\lfloor k\mu_+\rfloor) &\geq c\frac{(k-1)^{k-\frac{1}{2}}}{(k-1-\lfloor k\mu_+\rfloor)^{k-\frac{1}{2}-\lfloor k\mu_+\rfloor}\cdot(\lfloor k\mu_+\rfloor)^{\lfloor k\mu_+\rfloor+\frac{1}{2}}}\cdot\mu_+^{\lfloor k\mu_+\rfloor}(1-\mu_+)^{k-1-\lfloor k\mu_+\rfloor}\\
&\geq c\frac{(k-1)^{k-\frac{1}{2}}}{(k-k\mu_+)^{k-\frac{1}{2}-\lfloor k\mu_+\rfloor}\cdot(k\mu_+)^{\lfloor k\mu_+\rfloor+\frac{1}{2}}}\cdot\mu_+^{\lfloor k\mu_+\rfloor}(1-\mu_+)^{k-1-\lfloor k\mu_+\rfloor}\\
&\geq c\frac{(k-1)^{k-\frac{1}{2}}}{k^k\cdot(1-\mu_+)^{k-\frac{1}{2}-\lfloor k\mu_+\rfloor}\cdot(\mu_+)^{\lfloor k\mu_+\rfloor+\frac{1}{2}}}\cdot\mu_+^{\lfloor k\mu_+\rfloor}(1-\mu_+)^{k-1-\lfloor k\mu_+\rfloor}\\
&= c\frac{(k-1)^{k-\frac{1}{2}}}{k^k}\cdot\mu_+^{-\frac{1}{2}}(1-\mu_+)^{-\frac{1}{2}}\\
&\overset{\text{(i)}}{=} c\frac{(k-1)^{k-\frac{1}{2}}}{k^k}\geq c\frac{1}{\sqrt{k-1}}(1-\frac{1}{k})^k\gtrsim\frac{1}{\sqrt{k}},
\end{aligned}\tag{12.142}$$

where (i) is true because $\mu_+=\frac{1}{1+e^{-2B}}$ is bounded away from 0 and 1 by a constant.

Combining (12.139) and (12.142), we have

$$\mathbb{E}[h^+(\mu)]-\theta_1^*\leq-\frac{c}{\sqrt{k}},\qquad\text{for some constant } c>0.$$

Proof of Lemma 12.11

First consider the unconstrained oracle $\widetilde{\theta}^{(\infty)}$. We prove that for any $\theta\notin\Theta_{\text{oracle}}$, there exists some $\theta'\in\Theta_{\text{oracle}}$ such that $\ell(\theta')<\ell(\theta)$, where both θ and θ' satisfy the centering constraint.

Consider any $\theta\notin\Theta_{\text{oracle}}$. By the definition of Θ_{oracle}, there exist some integers i and j where $2\leq i<j\leq d$, such that $\theta_i\neq\theta_j$. By the symmetry of the manipulated observations $\{\widetilde{\mu}_{ij}\}$ defined in (12.49) with respect to item 2 through item d, we have that for any $\theta\in\mathbb{R}^d$,

$$\ell(\{\widetilde{\mu}_{i,j};\theta\})=\ell(\{\widetilde{\mu}_{i,j};\theta_\pi\}),\tag{12.143}$$

where $\pi:\{2,\ldots,d\}\to\{2,\ldots,d\}$ is any permutation of item 2 through item d, and $\theta_\pi=[\theta_1,\theta_{\pi(2)},\ldots,\theta_{\pi(d)}]$. For every $s\in\{0,1,\ldots,d-2\}$, define π_s as the permutation where item 2 through item d are shifted s positions to the left in a circle. That is, for every $i\in\{2,\ldots,d\}$, we have

$$\pi_s(i)=2+[(i-2+s)\mod(d-1)].$$

Now define $\theta'=\frac{1}{d-1}\sum_{s=0}^{d-2}\theta_{\pi_s}$. It can be verified that

$$\theta'=\left[\theta_1,\frac{1}{d-1}\sum_{i=2}^{d}\theta_i,\ldots,\frac{1}{d-1}\sum_{i=2}^{d}\theta_i\right]\in\Theta_{\text{oracle}}.\tag{12.144}$$

Moreover, we have

$$\ell(\theta')=\ell\left(\frac{1}{d-1}\sum_{s=0}^{d-2}\theta_{\pi_s}\right)\overset{\text{(i)}}{<}\frac{1}{d-1}\sum_{s=0}^{d-2}\ell(\theta_{\pi_s})\overset{\text{(ii)}}{=}\ell(\theta),$$

where (i) is due to the strict convexity of the negative log-likelihood function ℓ in Lemma 12.1, and (ii) is due to (12.143).

Now we argue the equivalence of the unconstrained oracle $\widetilde{\theta}^{(\infty)}$ defined in (12.50a) and (12.51a). If a solution $\widetilde{\theta}^{(\infty)}$ to (12.50a) exists, then we have $\widetilde{\theta}^{(\infty)} \in \Theta_{\text{oracle}}$ and it is trivially also the solution to (12.51a). On the other hand, if a solution $\widetilde{\theta}^{(\infty)}$ to (12.51a) exists, assume for contradiction that $\widetilde{\theta}^{(\infty)}$ is not a solution to (12.50a). Then either there exists no solution to (12.50a), or the solution to (12.50a) is not $\widetilde{\theta}^{(\infty)}$. In either case, there exists some θ such that $\ell(\theta) < \ell(\widehat{\theta}^{(\infty)})$. By (12.144), we construct some $\theta' \in \Theta_{\text{oracle}}$ such that $\ell(\theta') < \ell(\theta) < \ell(\widehat{\theta}^{(\infty)})$. This contradicts the assumption that $\widehat{\theta}^{(\infty)}$ is the optimal solution to (12.51a). Hence, Eq. (12.50a) and (12.51a) are equivalent definitions of the unconstrained oracle $\widehat{\theta}^{(\infty)}$.

The same argument can be extended to the constrained oracle $\widehat{\theta}^{(B)}$, by noting that if $\theta \in \Theta_B$, then in the construction (12.144) we have $\theta' \in \Theta_B$.

Proof of Lemma 12.12

Note that the lemma statement is conditioned on the event E_v. That is, we observe $\mu_1 = v$ for some $\frac{1}{2} \leq v \leq \mu_+ < 1$. In particular, we have $0 < \mu_1 < 1$. Then there exists at least one directed edge from node 1 to nodes $\{2, \ldots, d\}$, and at least one directed edge from nodes $\{2, \ldots, d\}$ to node 1. Then it suffices to prove that the subgraph consisting of nodes $\{2, \ldots, d\}$ is strongly-connected w.h.p.($\frac{1}{dk}$).

Note that the observations $\{\mu_{ij}\}$ for any $2 \leq i < j \leq d$ are all independent of μ_1, and therefore independent of the event E_v. Using the arguments in Lemma 12.3, we have that the subgraph consisting of nodes $\{2, \ldots, d\}$ is strongly-connected w.h.p.($\frac{1}{dk}$).

Proof of Lemma 12.13

Note that the lemma statement is conditioned on the event E_v. That is, we observe $\mu_1 = v$ for some $\frac{1}{2} \leq v \leq \mu_+ < 1$.

When $m = 1$, the desired inequality (12.59) holds trivially, because conditioned on E_v, we have

$$\sum_{i \neq 1} \mu_{1i} - \sum_{i \neq 1} \widetilde{\mu}^v_{1i} = (d-1)v - (d-1)v = 0.$$

Now consider every $m \in \{2 \ldots, d\}$. Consider the (unconditional) McDiarmid's inequality of (12.92) in the proof of Lemma 12.4. Replacing the summation sign $\sum_{i \neq m}$ on the LHS of (12.92) by the summation sign $\sum_{\substack{i \geq 2 \\ i \neq m}}$ (that is, we further exclude $i = 1$ from the summation) yields the unconditional inequality:

$$\mathbb{P}\left[\left|\sum_{\substack{2 \leq i \leq d \\ i \neq m}} \mu_{mi} - \sum_{\substack{2 \leq i \leq d \\ i \neq m}} \mu^*_{mi}\right| \leq c\sqrt{\frac{d(\log d + \log k)}{k}}\right] \geq 1 - \frac{c'}{d^2 k}, \tag{12.145}$$

where $c, c' > 0$ are constants. Now we condition (12.145) on the event E_v. Note that for all $i, m \in \{2, \ldots, d\}$ with $i \neq m$, the terms $\{\mu_{mi}\}$ are independent of E_v. Moreover, by the expression of $\widetilde{\mu}^v_{mi}$ in (12.57), we have $\mu^*_{mi} = \frac{1}{2} = \widetilde{\mu}^v_{mi}$ conditioned on E_v. Hence, we have

$$\mathbb{P}\left[\left|\sum_{\substack{2\leq i\leq d\\ i\neq m}} \mu_{mi} - \sum_{\substack{2\leq i\leq d\\ i\neq m}} \widetilde{\mu}^v_{mi}\right| \leq c\sqrt{\frac{d(\log d + \log k)}{k}} \;\middle|\; E_v\right] \geq 1 - \frac{c'}{d^2k}. \tag{12.146}$$

Now we bound the quantity $|\mu_{m1} - \widetilde{\mu}^v_{m1}|$ conditioned on E_v. By the definition of μ_1, we have that among the $(d-1)k$ comparisons $\{X^{(r)}_{1j}\}_{j\in\{2,\ldots,d\}, r\in[k]}$ in which item 1 is involved, there are $(d-1)k\mu_1$ terms that have value 1, and the rest have value 0. Hence, each μ_{1j} can be thought of as the mean of k comparisons sampled without replacement from the $(d-1)k$ comparisons $\{X^{(r)}_{1j}\}_{j\in\{2,\ldots,d\}, r\in[k]}$. By Hoeffding's inequality (sampling without replacement), we have that for every $j \in \{2, \ldots, d\}$,

$$\begin{aligned}\mathbb{P}\left[\left|\mu_{1j} - \widetilde{\mu}^v_{1j}\right| \leq c\sqrt{\frac{\log d + \log k}{k}} \;\middle|\; E_v\right] &\geq 1 - 2\exp\left(-c'(\log d + \log k)\right)\\ &\geq 1 - \frac{c''}{d^2k},\end{aligned}$$

where $c, c', c'' > 0$ are constants. Equivalently, by a change of variables, we have that for every $j \in \{2, \ldots, d\}$,

$$\mathbb{P}\left[|\mu_{m1} - \widetilde{\mu}^v_{m1}| \leq c\sqrt{\frac{\log d + \log k}{k}} \;\middle|\; E_v\right] \geq 1 - \frac{c''}{d^2k}. \tag{12.147}$$

Combining (12.146) and (12.147) by the triangle inequality, and taking a union bound over $m \in \{2, \ldots, d\}$ completes the proof.

12.2 Proof of Theorem 5.5

In this section, we present the proof of Theorem 5.5. Both Theorem 5.5(a) and Theorem 5.5(b) are closely related to Theorem 2 from [156]. Under our setting, the quantity σ defined in [156] is a universal constant, and the quantities ζ and γ defined in [156] are constants that depend only on the constant B.

12.2.1 Proof of Theorem 5.5(a)

Theorem 5.5(a) is a direct consequence of Theorem 2(a) from [156]. We now provide some details on how to apply Theorem 2(a) from [156]. Under our setting, each pair of items is compared k times. Therefore, the sample size n is

$$n = \binom{d}{2}k = \Theta(d^2k). \tag{12.148}$$

Moreover, under our setting the underlying topology is a complete graph. Let L denote the scaled Laplacian as defined in Eq. (4) from [156], and let $L^\dagger$ denote the Moore-Penrose pseudoinverse of L. From [156], the spectrum of L for a complete graph is $0, \frac{2}{d-1}, \ldots, \frac{2}{d-1}$. Therefore, we have

$$\lambda_2(L) = \frac{2}{d-1}, \tag{12.149a}$$

$$\operatorname{tr}(L^\dagger) = (d-1) \cdot \frac{d-1}{2} = \frac{(d-1)^2}{2}. \tag{12.149b}$$

Plugging (12.148) and (12.149) into Theorem 2(a) from [156] shows that the Theorem 5.5(a) holds for all $k \geq k_0$, where k_0 is a constant.

12.2.2 Proof of Theorem 5.5(b)

The proof of Theorem 5.5(b) closely mimics the proof of Theorem 2(b) from [156] (which is in turn based on Theorem 1(b) from [156]). In what follows, we state a minor modification to be made in order to extend the proof from [156] to Theorem 5.5(b).

In the proof from [156], the box constraint for the MLE $\widehat{\theta}^{(B)}$ is only used to obtain the following bound (see Section A.2 from [156]):

$$v^T \nabla^2 \ell(w) v \geq \frac{\gamma}{n\sigma^2} \|Xv\|_2^2, \qquad \text{for all } v, w \in \Theta_B. \tag{12.150}$$

Now we fix any constant A such that $A > B$. It can be verified that (12.150) still holds when replacing Θ_B by Θ_A, where we now allow γ to depend on both A and B. Since A is assumed to be a constant, we have that γ is still a constant. Then the rest of the arguments from [156] carry to the proof of Theorem 5.5(b).

Chapter 13

Proofs of Chapter 6

In this section, we present all proofs for results in Section 6.

13.1 Proof of Proposition 6.1

In this section, we present the proof of Proposition 6.1. For any event E, we let $\overline{E}$ denote the complement of E. We first derive equality (13.2) that is common across parts (a) and (b), and then separately prove for parts (a) and (b) based on equality (13.2).

Consider both parts (a) and (b), where one or both attributes are protected. Let E_{hol} and E_{seg} denote the events that the estimator makes an error in the top-1 metric, for the holistic approach and the segmented approach, respectively. First, we note that for either the holistic approach or the segmented approach, the estimated quality of the advantaged candidates always equals their true quality, because there is no quality discounting on the advantaged candidates. For the disadvantaged candidates, due to the discounting, their estimated quality is always lower than or at most equal to their true quality. Hence, when the top candidate is an advantaged candidate, the estimator does not make an error. That is, for $E \in \{E_{\text{hol}}, E_{\text{seg}}\}$, we have

$$\mathbb{P}(E \mid X^{\max} < Y^{\max}) = 0. \tag{13.1}$$

Therefore, for $E \in \{E_{\text{hol}}, E_{\text{seg}}\}$,

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(E \mid X^{\max} > Y^{\max}) \cdot \mathbb{P}(X^{\max} > Y^{\max}) + \mathbb{P}(E \mid X^{\max} < Y^{\max}) \cdot \mathbb{P}(X^{\max} < Y^{\max}) \\ &\overset{\text{(i)}}{=} \mathbb{P}(E \mid X^{\max} > Y^{\max}) \cdot \mathbb{P}(X^{\max} > Y^{\max}) \\ &\overset{\text{(ii)}}{=} \frac{1}{2}\mathbb{P}(E \mid X^{\max} > Y^{\max}), \end{aligned} \tag{13.2}$$

where (i) is true by (13.1), and (ii) is true because $\mathbb{P}(X^{\max} > Y^{\max}) = \frac{1}{2}$ by symmetry.

Hence, it sufficies to consider the case where the top applicant is a disadvantaged candidate. We now analyze the term $\mathbb{P}(E \mid X^{\max} > Y^{\max})$, separately for part (a) and part (b).

13.1.1 Proof of Proposition 6.1(a)

Error for the segmented approach Recall that in the segmented approach, each of the two reviewers is assigned one attribute each. Let R denote the event that the protected attribute is assigned to a biased reviewer. We have $\mathbb{P}(R) = \gamma$. We have

$$\begin{aligned}\mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}) &= \mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}, R) \cdot \mathbb{P}(R \mid X^{\max} > Y^{\max}) \\ &\quad + \mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}, \overline{R}) \cdot \mathbb{P}(\overline{R} \mid X^{\max} > Y^{\max}) \\ &\overset{\text{(i)}}{=} \gamma \cdot \mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}, R) + (1-\gamma) \cdot \mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}, \overline{R}),\end{aligned} \tag{13.3}$$

where (i) is true because R is independent of the event $\{X^{\max} > Y^{\max}\}$. Now we consider the two probabilities in (13.3). If the protected attribute is assigned the unbiased reviewer, then the estimator correctly identifies the best candidate. That is,

$$\mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}, \overline{R}) = 1. \tag{13.4}$$

If the protected attribute is assigned the biased reviewer, then the estimated quality of the best disadvantaged candidate becomes $\frac{1+\beta}{2} X^{\max}$. We have

$$\mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}, R) = \mathbb{P}\Big(\frac{1+\beta}{2} X^{\max} > Y^{\max} \mid X^{\max} > Y^{\max}\Big). \tag{13.5}$$

Plugging (13.4) and (13.5) to (13.3), we have

$$\mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}) = \gamma \cdot \mathbb{P}\Big(\frac{1+\beta}{2} X^{\max} > Y^{\max} \mid X^{\max} > Y^{\max}\Big) + (1-\gamma). \tag{13.6}$$

Error for the holistic approach Recall that in the holistic approach, each of the two reviewers is assigned half of the applicants. Let R' denotes the event that the best disadvantaged candidate is assigned to a biased reviewer. Using analysis similar to the segmented approach, we have

$$\mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}) = \gamma \cdot \mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R') + (1-\gamma).$$

Now we analyze the term $\mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R')$. Denote R'' as the event that the second reviewer is also biased. We have

$$\begin{aligned}\mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R') &= \mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R', R'') \cdot \mathbb{P}(R'' \mid X^{\max} > Y^{\max}, R') \\ &\quad + \mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R', \overline{R''}) \cdot \mathbb{P}(\overline{R''} \mid X^{\max} > Y^{\max}, R') \\ &\overset{\text{(i)}}{=} \gamma \cdot \mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R', R'') + (1-\gamma) \cdot \mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R', \overline{R''}),\end{aligned} \tag{13.7}$$

where (i) is true because R'' is independent of R' and the event $\{X^{\max} > Y^{\max}\}$. Now we consider the two probabilities in (13.7). When both reviewers are biased, we have

$$\mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R', R'') = \mathbb{P}\left(\frac{1+\beta}{2} X^{\max} > Y^{\max} \mid X^{\max} > Y^{\max}\right). \tag{13.8}$$

When the second reviewer is unbiased, the estimator correctly identifies the best disadvantaged candidate, if and only if its estimated quality exceeds both the best advantaged candidate, and also the disadvantaged candidates that are assigned to the second (unbiased) reviewer. Denote the random variable $A \subseteq [\alpha n]$ the set of disadvantaged applicants assigned to the unbiased reviewer. We have

$$\mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R', \overline{R''}) = \mathbb{P}\Big(\left\{\frac{1+\beta}{2} X^{\max} > Y^{\max}\right\} \cap \left\{\frac{1+\beta}{2} X^{\max} > \max_{i \in A} X_i\right\} \mid X^{\max} > Y^{\max}\Big). \tag{13.9}$$

Plugging (13.8) and (13.9) back to (13.7), we have

$$\begin{aligned}\mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}) &= \gamma^2 \cdot \mathbb{P}\left(\frac{1+\beta}{2} X^{\max} > Y^{\max} \mid X^{\max} > Y^{\max}\right) \\ &+\gamma(1-\gamma)\cdot\mathbb{P}\Big(\left\{\frac{1+\beta}{2} X^{\max} > Y^{\max}\right\} \cap \left\{\frac{1+\beta}{2} X^{\max} > \max_{i \in A} X_i\right\} \mid X^{\max} > Y^{\max}\Big) + (1-\gamma).\end{aligned} \tag{13.10}$$

Finally, subtracting (13.6) and (13.10), we have

$$\mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}) - \mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}) = \tag{13.11}$$

$$\gamma(1-\gamma) \cdot \mathbb{P}\Big(\left\{\frac{1+\beta}{2} X^{\max} > Y^{\max}\right\} \cap \left\{\frac{1+\beta}{2} X^{\max} > \max_{i \in A} X_i\right\} \mid X^{\max} > Y^{\max}\Big) > 0. \tag{13.12}$$

Combining (13.12) with (13.2), we have

$$e_{\text{hol}} - e_{\text{seg}} > 0,$$

completing the proof.

13.1.2 Proof of Proposition 6.1(b)

We decompose the error based on the number of biased reviewers being $0, 1$, or 2. Denote R_i as the event that the number of biased reviewers is i, for $i \in \{0, 1, 2\}$.

Expression for the error For $E \in \{E_{\text{seg}}, E_{\text{hol}}\}$, we have

$$\begin{aligned}\mathbb{P}(\overline{E} \mid X^{\max} > Y^{\max}) &= \mathbb{P}(\overline{E} \mid X^{\max} > Y^{\max}, R_0) \cdot \mathbb{P}(R_0 \mid X^{\max} > Y^{\max}) \\ &+ \mathbb{P}(\overline{E} \mid X^{\max} > Y^{\max}, R_1) \cdot \mathbb{P}(R_1 \mid X^{\max} > Y^{\max}) \\ &+ \mathbb{P}(\overline{E} \mid X^{\max} > Y^{\max}, R_2) \cdot \mathbb{P}(R_2 \mid X^{\max} > Y^{\max}) \\ &\overset{\text{(i)}}{=} (1-\gamma)^2 \cdot \mathbb{P}(\overline{E} \mid X^{\max} > Y^{\max}, R_0) + 2\gamma(1-\gamma) \cdot \mathbb{P}(\overline{E} \mid X^{\max} > Y^{\max}, R_1) \\ &+ \gamma^2 \cdot \mathbb{P}(\overline{E} \mid X^{\max} > Y^{\max}, R_2),\end{aligned} \tag{13.13}$$

where (i) is true because for each $k \in \{0, 1, 2\}$, we have R_k is independent from the event $\{X^{\max} > Y^{\max}\}$. Hence, we have $\mathbb{P}(R_k \mid X^{\max} > Y^{\max}) = \mathbb{P}(R_k)$ and then compute the probabilities by the Bernoulli model of the reviewers.

Now we analyze the three terms $\mathbb{P}(\overline{E} \mid X^{\max} > Y^{\max}, R_k)$ for $k \in \{0, 1, 2\}$. If no reviewer is biased, then it can be verified that for both the holistic and segmented approaches, the estimator correctly identifies the top applicant. That is,

$$\mathbb{P}(\overline{E} \mid X^{\max} > Y^{\max}, R_0) = 1. \tag{13.14}$$

If both reviewers are biased, then both attributes of the disadvantaged candidate are discounted, and the estimated quality of the best disadvantaged candidate becomes $\beta X^{\max}$. The best disadvantaged candidate remains the best among the disadvantaged candidates. Hence,

$$\mathbb{P}(\overline{E} \mid X^{\max} > Y^{\max}, R_0) = \mathbb{P}\Big(\beta X^{\max} > Y^{\max} \mid X^{\max} > Y^{\max}\Big). \tag{13.15}$$

Now we analyze the remaining term $\mathbb{P}(\overline{E} \mid X^{\max} > Y^{\max}, R_1)$, for the segmented and the holistic approaches separately.

Term for the segmented approach If exactly one reviewer is biased, then one attribute of the disadvantaged candidate is discounted, and the estimated quality of the best disadvantaged candidate becomes $\frac{1+\beta}{2} X^{\max}$. In this case, the estimated quality of the best disadvantaged candidate remains the best among the disadvantaged candidates. An error occurs if and only if the estimated quality of the best advantaged candidate exceeds the best disadvantaged candidate. That is,

$$\begin{aligned} \mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}, R_1) &= \mathbb{P}\Big(\frac{1+\beta}{2} X^{\max} > Y^{\max} \mid X^{\max} > Y^{\max}\Big) && (13.16)\\ &\overset{\text{(i)}}{=} 2 \cdot \mathbb{P}\Big(\frac{1+\beta}{2} X^{\max} > Y^{\max}\Big), && (13.17) \end{aligned}$$

where (i) is true because by the definition of the conditional we have $\mathbb{P}\Big(\frac{1+\beta}{2} X^{\max} > Y^{\max} \mid X^{\max} > Y^{\max}\Big) = \frac{\mathbb{P}(\frac{1+\beta}{2} X^{\max} > Y^{\max}, X^{\max} > Y^{\max})}{\mathbb{P}(X^{\max} > Y^{\max})} = \frac{\mathbb{P}(\frac{1+\beta}{2} X^{\max} > Y^{\max})}{\mathbb{P}(X^{\max} > Y^{\max})}$, and also $\mathbb{P}(X^{\max} > Y^{\max}) = \frac{1}{2}$ by symmetry.

Term for the holistic approach Now we consider the term $\mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R_1)$ in (13.13) for the holistic approach. By symmetry, with probability $\frac{1}{2}$, the best disadvantaged applicant is assigned the unbiased reviewer. In this case, the estimator correctly identifies it as the best applicant. With probability $\frac{1}{2}$, the best disadvantaged applicant is assigned the biased reviewer. In this case, this candidate remains the best among all disadvantaged candidates assigned to the biased reviewer. Hence, this candidate is correctly identified, if and only if its estimated quality is higher than the best advantaged candidate, and also higher than the best disadvantaged candidate assigned to the unbiased reviewer. Let the random variable $A \subseteq [\alpha n]$ denote the set of disadvantaged applicants assigned to the unbiased reviewer. We have

$$\mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R_1) = \tag{13.18}$$

$$\frac{1}{2} + \frac{1}{2} \cdot \mathbb{P}\left(\{\beta X^{\max} > Y^{\max}\} \cap \left\{\beta X^{\max} > \max_{i \in A} X_i\right\} \mid X^{\max} > Y^{\max}, R_1\right). \tag{13.19}$$

Now setting $\beta = 0$ in (13.19), we have

$$\mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R_1) = \frac{1}{2}. \tag{13.20}$$

Comparing the error for the segmented and holistic approaches Now setting $E \in \{E_{\text{seg}}, E_{\text{hol}}\}$ in (13.13), subtracting the two, and using (13.14) and (13.15), we have

$$\begin{aligned}\mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}) - \mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}) &\overset{\text{(i)}}{=} \\ 2\gamma(1-\gamma) \cdot \left[\mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}, R_1) - \mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}, R_1)\right]\end{aligned} \tag{13.21}$$

Setting $\beta = 0$, and plugging (13.17) and (13.20) to (13.21), we have

$$\begin{aligned}\mathbb{P}(\overline{E}_{\text{seg}} \mid X^{\max} > Y^{\max}) - \mathbb{P}(\overline{E}_{\text{hol}} \mid X^{\max} > Y^{\max}) &= 2\gamma(1-\gamma) \cdot \left[2\mathbb{P}\Big(\frac{1}{2}X^{\max} > Y^{\max}\Big) - \frac{1}{2}\right] \\ &= \gamma(1-\gamma) \cdot \left[4\mathbb{P}\Big(\frac{1}{2}X^{\max} > Y^{\max}\Big) - 1\right].\end{aligned}$$

Finally, setting $E \in \{E_{\text{seg}}, E_{\text{hol}}\}$ in (13.2) and subtracting the two expressions, we have

$$\begin{aligned}e_{\text{hol}} - e_{\text{seg}} &= \frac{1}{2}\left[\mathbb{P}(E_{\text{hol}} \mid X^{\max} > Y^{\max}) - \mathbb{P}(E_{\text{seg}} \mid X^{\max} > Y^{\max})\right] \\ &= \frac{1}{2}\left[\mathbb{P}(E_{\text{seg}} \mid X^{\max} > Y^{\max}) - \mathbb{P}(E_{\text{hol}} \mid X^{\max} > Y^{\max})\right] \\ &= \frac{\gamma(1-\gamma)}{2}\left[4\mathbb{P}\Big(\frac{1}{2}X^{\max} > Y^{\max}\Big) - 1\right],\end{aligned}$$

completing the proof of (6.1) and (6.2).

Power law distribution Following Definition 3 from [100], for non-negative functions $f(n)$ and $g(n)$, we define

$$f(n) \approxeq g(n)$$

if and only if $f(n) = g(n)\left(1 \pm O\left(\frac{(\ln n)^2}{n}\right)\right)$. Now consider the power law distribution with parameter δ. Setting $\alpha = 1, \beta = 2, c = \alpha\beta^{-(1+\delta)}$ and $k = 1$ in Theorem B.3 from [100] yields

$$\mathbb{P}(X^{\max} < 2Y^{\max}) \approxeq \left(1 + 2^{-(1+\delta)}\right)^{-1}. \tag{13.22}$$

By (6.2), the segmented approach is better if and only if

$$\mathbb{P}(X^{\max} > 2Y^{\max}) > 0.25,$$

or equivalently

$$\mathbb{P}(X^{\max} < 2Y^{\max}) < 0.75. \tag{13.23}$$

Combining (13.22) and (13.23), for sufficiently large n, the segmented approach is better if and only if

$$\left(1 + 2^{-(1+\delta)}\right)^{-1} < 0.75,$$

or equivalently

$$\delta < \frac{\log 3}{\log 2} - 1,$$

completing the proof of (6.3).

Part V

Conclusion and Discussion

In this thesis, we study various aspects of biases involved in decision-making problems. The thesis presents theoretical and experimental analysis on different sources of biases arising from people, estimation and policies. We conclude with a discussion on the open directions that are important in extending the scope of this thesis.

Modeling biases In this thesis, we have developed conceptual and technical methodologies in identifying and addressing certain types of biases. These high-level methodologies lay the foundation in tackling other sources of biases that are prominent in real-life applications. One concrete problem is to model the temporal dependency of the bias. People's evaluation standards often fluctuate over time. For example, if a person sees a number of below-the-average applications followed by a good one, the person is subject to overrating the good one due to the contrast. There is also fluctuation over the longer term. In many applications the evaluation takes place annually (e.g., admissions and paper review), and the quality of the items (applicants and papers) changes over the years. Furthermore, the quality within a single item can also change over time (e.g., student performance throughout a semester). On the human side, it is of interest to study whether and how people are delayed in recognizing and making adjustments for these changes. On the estimation side, it is of interest to study how a model mismatch of neglecting such time dependency affects the statistical estimation of the item qualities. Relevant technical tools include online learning and time series.

A second concrete problem is to study how people influence each other. In many applications people have knowledge about other peoples' evaluation, directly (e.g., in sports, judges immediately see the scores given by each other in each round of the competition) or indirectly (e.g., in certain peer review schemes, the reviewers do not see evaluation from each other but ultimately know whether the papers they review are accepted or rejected). It is of interest to understand how people respond to such knowledge from others: do people tend to follow others, or do people express their opinions more strongly if they anticipate others to disagree? One may start with identifying a realistic and tractable problem formulation, and then provide theoretical guarantees along with empirical validation.

Algorithmic fairness The research in this thesis focuses on the modeling aspect, where we propose formulations to describe the bias, and then design algorithms to minimize standard error metrics, such as the mean-squared error, under the proposed models of the bias. One interesting direction is to connect the research in this thesis more closely with the area of algorithmic fairness, which focuses on carefully defining and comparing different types of error metrics and performance guarantees. Drawing inspirations from prior work on various fairness constraints and metrics, it would be interesting to combine these concepts with the approach to bias modeling presented in this thesis. For example, do the proposed algorithms for the presented bias models improve certain fairness metrics in addition to estimation accuracy? Moreover, a lot of prior work in algorithmic fairness has focused on the classification setting, whereas other settings such as ranking are commonly used in domains such as admissions and peer review. These settings have been studied (e.g., [195]) but remain relatively unexplored. Combining algorithmic fairness, bias modeling, and application-specific challenges provides a more holistic picture in thinking about the bias involved in different parts of the decision-making pipeline.

Mechanism Design The majority of this thesis focuses on how to correct the bias after it is introduced. A complementary research question is how to prevent or reduce the bias before it is introduced. The comparison of the holistic and the segmented approaches in multi-attribute evaluation presented in Chapter 6 is one example, but a lot more can be done in terms of improving the design of the evaluation systems. It is an interesting direction to think about utilizing tools from mechanism design. For example, remotely inspired by recent work on inducing a specific allocation of effort (e.g., [99]), one specific direction is to design incentives to encourage desirable behaviors (e.g., incentivizing authors to submit high-quality papers under self-selection, and incentivizing reviewers to provide impartial evaluation). One challenge is that different applications impose different constraints on the available tools of incentives (e.g., revenue, penalty, or non-monetary). Research in this area is both of theoretical interest, and also lay a more convincing foundation for outreach and policy recommendations in practice.

Computational social choice One key assumption we make in this thesis is that every item has a true quality represented by a scalar value. A true ordering of items is also subsequently derived from these true qualities. However, such assumptions may not be always applicable. For example, in peer review, papers in different fields cannot be straightforwardly compared, meaning a total ordering of all the items may not exist. Furthermore, even if evaluators individually perceive a true total ordering of all the items, their perceived orderings may not necessarily be the same due to subjectivity. The area of (computational) social choice operates in such settings where no ground-truth is assumed, and the goal is to capture the consensus of people described by axiomatic properties. In peer review, there are objective yardsticks for theoretical and empirical results, while there are also subjective judgments, such as the relative weighting of theoretical and empirical results. It is a regime in between the two extremes – there may not exist a universally-agreed true ordering, but reviewers still share some common sense as in what defines a good paper. We envisage that bringing together tools from social choice and statistical estimation bridges the two regimes with new insights, and influences subsequent research in the two communities.

Bibliography

[1] FIDE rating regulations effective from 1 July 2017, 2017. `https://www.fide.com/fide/handbook.html?id=197&view=article` [Online; accessed May 21, 2019]. 66

[2] IEEE Information Theory Society newsletter, 2018. `https://www.itsoc.org/publications/newsletters/march-2018-issue/at_download/file/` [Online; accessed 18-Jun-2019]. 103

[3] Elo ratings - English Premier League, 2019. `https://sinceawin.com/data/elo/league/div/e0` [Online; accessed May 21, 2019]. 66

[4] Arpit Agarwal, Prathamesh Patil, and Shivani Agarwal. Accelerated spectral ranking. In *International Conference on Machine Learning*, 2018. 71

[5] David Aldous. Elo ratings and the sports model: A neglected topic in applied probability? *Statistical Science*, 32(4):616–629, 2017. 67

[6] Dennis Amelunxen, Martin B. Lotz, Michael McCoy, and Joel A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference*, 3:224–294, 2014. 235

[7] Ammar Ammar and Devavrat Shah. Efficient rank aggregation using partial data. In *SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, 2012. 7, 8, 87

[8] J. A. Anderson and S. C. Richardson. Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, 21(1):71–78, 1979. 71

[9] Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, March 2011. 28

[10] Yukino Baba and Hisashi Kashima. Statistical quality estimation for general crowdsourcing tasks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013. 8

[11] A. J. Baranchik. A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Statist.*, 41(2):642–645, 1970. 13

[12] R.E. Barlow, D.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley, 1972. 24, 40, 224

[13] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`. 3

[14] Regina Barzilay and Min-Yen Kan. Outstanding and best papers and the decision process, 2017. `https://acl2017.wordpress.com/2017/08/03/outstanding-and-best-papers-and-the-decision-process/` [Online; accessed 30-May-2021]. 104

[15] Jacob P. Baskin and Shriram Krishnamurthi. Preference aggregation in group recommender systems for committee decision-making. In *ACM Conference on Recommender Systems, RecSys*, 2009. 8

[16] Amir Beck. *Introduction to Nonlinear Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2014. 134, 135

[17] William E. Becker and Michael Watts. How departments of economics evaluate teaching. *The American Economic Review*, 89(2):344–349, 1999. 38

[18] Marc Bendick Jr and Ana Paula Nunes. Developing the research basis for controlling bias in hiring. *Journal of Social Issues*, 68:238–262, 2013. 88

[19] Lorenzo Beretta and Alessandro Santaniello. Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, 16, 07 2016. 28

[20] Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, September 2004. 2, 84

[21] D.P. Bertsekas. *Convex Optimization Theory*. Athena Scientific optimization and computation series. Athena Scientific, 2009. 204, 205

[22] M. E. Bock. Minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.*, 3(1):209–218, 1975. 13

[23] Anne Boring, Kellie Ottoboni, and Philip B. Stark. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 2016. 37, 38, 39

[24] Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 67

[25] Michela Braga, Marco Paccagnella, and Michele Pellizzari. Evaluating students' evaluations of professors. *Economics of Education Review*, 41:71 – 88, 2014. 38

[26] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016. 89

[27] Mark Braverman and Elchanan Mossel. Noisy sorting without resampling. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '08, page 268–276, USA, 2008. Society for Industrial and Applied Mathematics. 31

[28] Lyle Brenner, Dale Griffin, and Derek J Koehler. Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes*, 97(1):64–81, 2005. 88

[29] Russ Bubley and Martin Dyer. Faster random generation of linear extensions. *Discrete Mathematics*, 201(1):81 – 88, 1999. 45

[30] Michael A Campion, Elliott D Pursell, and Barbara K Brown. Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel psychology*, 41(1):25–42, 1988. 88

[31] Gilles Caraux and Sylvie Pinloche. PermutMatrix: a graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics*, 21(7):1280–1281, 11 2004. 23

[32] Dana R. Carney and Mahzarin R. Banaji. First is best. *PLOS ONE*, 7(6):1–5, 06 2012. 101

[33] Scott E. Carrell and James E. West. Does professor quality matter? Evidence from random assignment of students to professors. Working Paper 14081, National Bureau of Economic Research, June 2008. 38

[34] Jonathan P. Caulkins, Patrick D. Larkey, and Jifa Wei. Adjusting gpa to reflect course difficulty, Jun 1995. 39

[35] Laurent Charlin and Richard Zemel. The toronto paper matching system: An automated paper-reviewer assignment system. *ICML Workshop on Peer Reviewing and Publishing Models (PEER)*, 2013. 2

[36] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43(1):177–214, 2015. 32

[37] Baiyu Chen, Sergio Escalera, Isabelle Guyon, Víctor Ponce-López, Nihar Shah, and Marc Oliu Simón. Overcoming calibration problems in pattern labeling with pairwise ratings: application to personality traits. In *European Conference on Computer Vision*, 2016. 67

[38] Eric Chen, Gábor Simonovits, Jon A. Krosnick, and Josh Pasek. The impact of candidate name order on election outcomes in north dakota. *Electoral Studies*, 35:115 – 122, 2014. 101

[39] Yining Chen and Richard J. Samworth. Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 2016. 40

[40] Yuxin Chen and Changho Suh. Spectral MLE: top-K rank aggregation from pairwise comparisons. In *International Conference on Machine Learning*, 2015. 71

[41] Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized MLE are both optimal for top-K ranking. *Ann. Statist.*, 47(4):2204–2235, 08 2019. 71, 73, 77, 79

[42] Guang Cheng. Semiparametric additive isotonic regression. *Journal of Statistical Planning and Inference*, 139(6):1980–1991, 2009. 40

[43] Katherine B Coffman, Christine L Exley, and Muriel Niederle. The role of beliefs in driving gender discrimination. *Management Science*, 2021. 82

[44] Wade D. Cook, Boaz Golany, Michal Penn, and Tal Raviv. Creating a consensus ranking of proposals from reviewers' partial ordinal rankings. *Computers & Operations Research*, 34(4):954–965, 2007. ISSN 0305-0548. doi: https://doi.org/10.1016/j.cor.2005.05.030. 8

[45] Thomas M. Cover. *Pick the Largest Number*, pages 152–152. Springer New York, New York, NY, 1987. 9, 12

[46] Bo Cowgill. Bias and productivity in humans and algorithms: Theory and evidence from résumé screening. 2018. 87

[47] Bo Cowgill and Catherine E Tucker. Algorithmic fairness and economics. *Columbia Business School Research Paper*, 2020. 87

[48] D. R. Cox and E. J. Snell. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):248–275, 1968. 71

[49] Jack Cuzick. Semiparametric additive regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):831–843, 1992. 40

[50] Sanjoy Dasgupta, Christos H. Papadimitriou, and Umesh Vazirani. *Algorithms*. McGraw-Hill, Inc., 1 edition, 2008. ISBN 0073523402, 9780073523408. 15

[51] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, Oct 2018. `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G` [Online; accessed 10-Aug-2021]. 86, 87

[52] Paul Deheuvels. The limiting behaviour of the maximal spacing generated by an i.i.d. sequence of gaussian random variables. *Journal of Applied Probability*, 22(4):816–827, 1985. 210

[53] Komal Dhull, Jingyan Wang, Nihar Shah, Yuanzhi Li, and R. Ravi. A heuristic for statistical seriation. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021. 4

[54] John R Douceur. Paper rating vs. paper ranking. *ACM SIGOPS Operating Systems Review*, 43(2):117–121, 2009. 15

[55] John F Dovidio, Nancy Evans, and Richard B Tyler. Racial stereotypes: The contents of their cognitive representations. *Journal of Experimental Social Psychology*, 22(1):22 – 37, 1986. 2

[56] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *International Conference on World Wide Web*, 2001. 15

[57] Liran Einav and Leeat Yariv. What's in a surname? the effects of surname initials on academic success. *Journal of Economic Perspectives*, 20(1):175–187, March 2006. 100

[58] Yingying Fan, Emre Demirkaya, and Jinchi Lv. Nonuniformity of p-values can occur early in diverging dimensions. *Journal of Machine Learning Research*, 20(77):1–33, 2019. 71

[59] Tanner Fiez, Nihar B. Shah, and Lillian Ratliff. A SUPER* algorithm to optimize paper bidding in peer review. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020. 39

[60] David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993. 71

[61] Peter A. Flach, Sebastian Spiegler, Bruno Golénia, Simon Price, John Guiver, Ralf Herbrich, Thore Graepel, and Mohammed J. Zaki. Novel tools to streamline the conference review process: Experiences from SIGKDD'09. *SIGKDD Explor. Newsl.*, 11(2):63–67, 2010. 8

[62] Nicolas Flammarion, Cheng Mao, and Philippe Rigollet. Optimal rates of statistical seriation. *Bernoulli*, 25(1):623–653, 02 2019. 23, 24, 27

[63] L. R. Ford, Jr. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8P2):28–33, 1957. 262

[64] Yoav Freund, Raj D. Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003. 8, 15, 87

[65] Hong Ge, Max Welling, and Zoubin Ghahramani. A Bayesian model for calibrating conference review scores, 2013. `http://mlg.eng.cam.ac.uk/hong/unpublished/nips-review-model.pdf` [Online; accessed 23-Dec-2019]. 8, 15, 39

[66] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, volume 29, pages 2973–2981. Curran Associates, Inc., 2016. 27

[67] Rong Ge, Jason D. Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 27

[68] E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959. 263

[69] Mark E. Glickman and Thomas Doan. The US chess rating system, 2017. `http://www.glicko.net/ratings/rating.system.pdf` [Online; accessed May 21, 2019]. 66

[70] Mark E Glickman and Albyn C Jones. Rating the chess rating system. *Chance*, 12:21–28, 1999. 67

[71] Alexander Gnedin. Guess the larger number. *preprint arXiv:1608.01899*, 2016. 9, 12

[72] Alexander V. Gnedin. A solution to the game of googol. *Ann. Probab.*, 22(3):1588–1595, 07 1994. 13

[73] Alexander V. Gnedin and Ulrich Krengel. Optimal selection problems based on exchangeable trials. *Ann. Appl. Probab.*, 6(3):862–882, 08 1996. 13

[74] Paul E. Green, J. Douglas Carroll, and Wayne S. DeSarbo. Estimating choice probabilities in multiattribute decision making. *Journal of Consumer Research*, 8(1):76–84, 1981. 67

[75] Anthony G Greenwald and Gerald M Gillmore. Grading leniency is a removable contam-

inant of student ratings. *The American psychologist*, 52(11):1209–1217, November 1997. 37, 39

[76] Anthony G Greenwald, Brian A Nosek, and Mahzarin R Banaji. Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of personality and social psychology*, 85(2):197, 2003. 84

[77] Dale Griffin and Lyle Brenner. *Perspectives on Probability Judgment Calibration*, chapter 9, pages 177–199. Wiley-Blackwell, 2008. 8

[78] Piet Groeneboom and Geurt Jongbloed. *Nonparametric estimation under shape constraints*, volume 38. Cambridge University Press, 2014. 40

[79] Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems*, 2014. 67, 68, 70, 74

[80] Anne-Wil Harzing, Joyce Baldueza, Wilhelm Barner-Rasmussen, Cordula Barzantny, Anne Canabal, Anabella Davila, Alvaro Espejo, Rita Ferreira, Axele Giroud, Kathrin Koester, et al. Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International Business Review*, 18(4):417–432, 2009. 8

[81] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009. 40

[82] Trevor J. Hastie and Robert J. Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990. 40

[83] Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120 – 135, 2000. 71

[84] Reinhard Heckel, Nihar B Shah, Kannan Ramchandran, Martin J Wainwright, et al. Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*, 47(6):3099–3126, 2019. 21, 23

[85] Christiana E. Hilmer and Michael J. Hilmer. How Do Journal Quality, Co-Authorship, and Author Order Affect Agricultural Economists' Salaries? *American Journal of Agricultural Economics*, 87(2):509–523, 05 2005. 100

[86] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. 163, 222

[87] Jian Huang. A note on estimating a partly linear model under monotonicity constraints. *Journal of Statistical Planning and Inference*, 107(1):343 – 351, 2002. 40

[88] Mark Huber. Fast perfect sampling from linear extensions. *Discrete Mathematics*, 306(4): 420 – 428, 2006. 45

[89] David R. Hunter. MM algorithms for generalized Bradley-Terry models. *Ann. Statist.*, 32 (1):384–406, 02 2004. doi: 10.1214/aos/1079120141. 67, 240

[90] Indiana University Bloomington. Grade distribution database, 2020. `https://gradedistribution.registrar.indiana.edu/index.php` [Online; ac-

cessed 30-Sep-2020]. 4, 56

[91] William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961. 9, 13

[92] Minje Jang, Sunghyun Kim, Changho Suh, and Sewoong Oh. Optimal sample complexity of m-wise data for top-K ranking. In *Advances in Neural Information Processing Systems*, 2017. 71

[93] Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar B. Shah, Vincent Conitzer, and Fei Fang. Mitigating manipulation in peer review via randomized reviewer assignments. In *NeurIPS*, 2020. 39

[94] Valen E. Johnson. An alternative to traditional GPA for evaluating student performance. *Statist. Sci.*, 12(4):251–278, 11 1997. 39

[95] Valen E. Johnson. *Grade Inflation: A Crisis in College Education*. Springer New York, 1 edition, 2003. 37, 39

[96] Bennet B. Murdock Jr. The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5):482, 1962. 101

[97] Aditya Khosla, Derek Hoiem, and Serge Belongie. Analysis of reviews for CVPR 2012. 2013. 38

[98] Franz J. Király and Zhaozhi Qian. Modelling competitive sports: Bradley-Terry-Élő models for supervised and on-line learning of paired competition outcomes. *preprint arXiv:1701.08055*, 2017. 66, 72

[99] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, EC '19, page 825–844, New York, NY, USA, 2019. Association for Computing Machinery. 288

[100] Jon M. Kleinberg and Manish Raghavan. Selection problems in the presence of implicit bias. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, pages 33:1–33:17, 2018. 85, 88, 93, 284

[101] Dundar F Kocaoglu. A participative approach to program evaluation. *IEEE Transactions on Engineering Management*, EM-30(3):112–118, 1983. 88

[102] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145. Montreal, Canada, 1995. 40

[103] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009. 13

[104] Alec Lamon, Dave Comroe, Peter Fader, Daniel McCarthy, Rob Ditto, and Don Huesman. Making WHOOPPEE: A collaborative approach to creating the modern student peer assessment ecosystem. In *EDUCAUSE*, 2016. 67

[105] John Langford. ICML acceptance statistics, 2012. `http://hunch.net/?p=2517` [Online; accessed 14-May-2018]. 8

[106] Carole J. Lee. Commensuration bias in peer review. *Philosophy of Science*, 82(5):1272–1283, 2015. 39, 89

[107] Innar Liiv. Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 3(2):70–91, 2010. 23

[108] Allen Liu and Ankur Moitra. Better algorithms for estimating non-parametric models in crowd-sourcing and rank aggregation. In *Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 2780–2829. PMLR, 2020. 23, 27

[109] R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, NY, USA, 1959. 67

[110] Yao Ma, Alexander Olshevsky, Csaba Szepesvari, and Venkatesh Saligrama. Gradient descent for sparse rank-one matrix completion for crowd-sourced aggregation of sparsely interacting workers. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3335–3344. PMLR, 10–15 Jul 2018. 27, 28, 29

[111] R. S. MacKay, R. Kenna, R. J. Low, and S. Parker. Calibration with confidence: a principled method for panel assessment. *Royal Society Open Science*, 4(2), 2017. doi: 10.1098/rsos.160760. 8

[112] Michael J. Mahoney. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, 1(2):161–175, Jun 1977. 2

[113] Enno Mammen and Kyusang Yu. Additive isotone regression. In *Asymptotics: particles, processes and inverse problems*, pages 179–195. Institute of Mathematical Statistics, 2007. 40

[114] Heikki Mannila. Finding total and partial orders from data for seriation. In *Discovery Science*, pages 16–25. Springer Berlin Heidelberg, 2008. 23

[115] Emaad Manzoor and Nihar B. Shah. Uncovering latent biases in text: Method and application to peer review. *INFORMS Workshop on Data Science*, 2020. 39

[116] Cheng Mao, Jonathan Weed, and Philippe Rigollet. Minimax rates and efficient algorithms for noisy sorting. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 821–847. PMLR, 07–09 Apr 2018. 31

[117] Cheng Mao, Ashwin Pananjady, and Martin J. Wainwright. Towards optimal estimation of bivariate isotonic matrices with unknown permutations. *Ann. Statist.*, 48(6):3183–3205, 12 2020. 23, 27

[118] William H. Marquardt. Advances in archaeological seriation. *Advances in Archaeological Method and Theory*, 1:257–314, 12 1978. 22, 23

[119] Peter Matthews. Generating a random linear extension of a partial order. *The Annals of Probability*, 19(3):1367–1392, 1991. 45

[120] William T. McCormick, Paul J. Schweitzer, and Thomas W. White. Problem decompo-

sition and data reorganization by a clustering technique. *Operations Research*, 20(5): 993–1009, 1972. 23

[121] Mark D. McDonnell and Derek Abbott. Randomized switching in the two-envelope problem. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 2009. 13

[122] Mary C. Meyer. Semi-parametric additive constrained regression. *Journal of nonparametric statistics*, 25(3):715–730, 2013. 40

[123] Ioannis Mitliagkas, Aditya Gopalan, Constantine Caramanis, and Sriram Vishwanath. User rankings from comparisons: Learning permutations in high dimensions. In *Allerton Conference on Communication, Control, and Computing*, 2011. 7, 8, 87

[124] Mario D. Molina, Mauricio Bucca, and Michael W. Macy. It's not just how the game is played, it's whether you win or lose. *Science Advances*, 5(7), 2019. 38

[125] Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479, 2012. 2

[126] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, 2012. 8

[127] Sahand Negahban, Sewoong Oh, and Devavrat Shah. RankCentrality: Ranking from pairwise comparisons. *Operations Research*, 65:266–287, 2016. 67, 77, 79

[128] Richard E Nisbett and Timothy D Wilson. The halo effect: Evidence for unconscious alteration of judgments. *Journal of personality and social psychology*, 35(4):250, 1977. 89

[129] Ritesh Noothigattu, Snehalkumar (Neil) S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A voting-based system for ethical decision making. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 1587–1594. AAAI Press, 2018. 89

[130] Ritesh Noothigattu, Nihar B. Shah, and Ariel Procaccia. Loss functions, axioms, and peer review. *Journal of Artificial Intelligence Research*, 2021. 21, 39, 89, 99

[131] S. Ontañón, G. Synnaeve, A. Uriarte, F. Richoux, D. Churchill, and M. Preuss. A survey of real-time strategy game AI research and competition in StarCraft. *IEEE Transactions on Computational Intelligence and AI in Games*, 5(4):293–311, 2013. doi: 10.1109/TCIAIG.2013.2286295. 66

[132] Ashwin Pananjady and Richard J. Samworth. Isotonic regression with unknown permutations: Statistics, computation, and adaptation, 2020. 27

[133] Konstantina Papagiannaki. Author feedback experiment at PAM 2007. *SIGCOMM Comput. Commun. Rev.*, 37(3):73–78, July 2007. 38

[134] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An

imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 31

[135] S. R. Paul. Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology*, 34(2):213–223, 1981. 8

[136] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in MOOCs. *preprint arXiv:1307.2579*, 2013. 15

[137] Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. ChaLearn LAP 2016: First round challenge on first impressions-dataset and results. In *European Conference on Computer Vision*, 2016. 67

[138] Stephen Portnoy. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.*, 16(1):356–366, 03 1988. doi: 10.1214/aos/1176350710. 71

[139] Ariel D. Procaccia, Nisarg Shah, and Yair Zick. Voting rules as error-correcting codes. *Artif. Intell.*, 231:1–16, 2016. 15

[140] M. H. Quenouille. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):68–84, 1949. 71

[141] Arun Rajkumar, Suprovat Ghoshal, Lek-Heng Lim, and Shivani Agarwal. Ranking from stochastic pairwise preferences: Recovering Condorcet winners and tournament solution sets at the top. In *International Conference on Machine Learning*, pages 665–673, 2015. 8

[142] Debraj Ray and Arthur Robson. Certified random: A new order for coauthorship. *American Economic Review*, 108(2):489–520, February 2018. 101

[143] Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa Valentine, and Michael S Bernstein. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 75–85, 2014. 87

[144] Lauren A Rivera. Hiring as cultural matching: The case of elite professional service firms. *American sociological review*, 77(6):999–1022, 2012. 88

[145] Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 157–163, 1956. 9, 13

[146] Milton Rokeach. The role of values in public opinion research. *Public Opinion Quarterly*, 32(4):547–559, 1968. 8

[147] Magnus Roos, Jörg Rothe, and Björn Scheuermann. How to calibrate the scores of biased reviewers by quadratic programming. In *AAAI Conference on Artificial Intelligence*, 2011. 8

[148] Zick Rubin. On measuring productivity by the length of one's vita. *Personality and Social Psychology Bulletin*, 4(2):197–198, 1978. 2

[149] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976. ISBN 9780070856134. 28, 261

[150] Cristina Rueda. Degrees of freedom and model selection in semiparametric additive monotone regression. *Journal of Multivariate Analysis*, 117:88–99, 2013. 40

[151] Aadirupa Saha and Arun Rajkumar. Ranking with features: Algorithm and a graph theoretic analysis. *arXiv preprint arXiv:1808.03857*, 2018. 89

[152] Frank L Schmidt and John E Hunter. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, 124(2):262, 1998. 88

[153] Walter Dill Scott, WV Bingham, and GM Whipple. The scientific selection of salesmen. *Advertising and Selling*, 25(5):5–6, 1915. 88

[154] Nihar B. Shah and Martin J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(199):1–38, 2018. 8, 15, 21, 23

[155] Nihar B. Shah, Joseph K Bradley, Abhay Parekh, Martin Wainwright, and Kannan Ramchandran. A case for ordinal peer-evaluation in MOOCs. In *NIPS Workshop on Data Driven Education*, 2013. 15, 67

[156] Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58): 1–47, 2016. 8, 15, 67, 68, 70, 74, 75, 278, 279

[157] Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *preprint arXiv:1606.09632*, 2016. 21

[158] Nihar B. Shah, Sivaraman Balakrishnan, Adityanand Guntuboyina, and Martin J. Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Transactions on Information Theory*, 63(2):934–959, 2017. 21, 23, 25, 27, 31, 32, 34, 35, 40

[159] Nihar B. Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. Design and analysis of the NIPS 2016 review process. *JMLR*, 19(1):1913–1946, 2018. 15, 82, 88

[160] Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. Low permutation-rank matrices: Structural properties and noisy completion. In *Journal of Machine Learning Research*, 2019. 21, 23

[161] Nihar B Shah, Sivaraman Balakrishnan, and Martin J Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *IEEE Transactions on Information Theory*, 2020. 23, 27

[162] Nihar Bhadresh Shah. *Learning from people*. PhD thesis, UC Berkeley, 2017. 21, 40

[163] P. C. Sham and D. Curtis. An extended transmission/disequilibrium test (TDT) for multiallele marker loci. *Annals of Human Genetics*, 59(3):323–336, 1995. 67

[164] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, L Robert Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Fong Celine Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550:354–359, 2017. 66

[165] Stephen Silverman and Arthur Nádas. On the game of googol as the secretary problem. *Contemporary Mathematics*, 125:77–83, 1992. 13

[166] Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206, 1956. 9, 13

[167] Ivan Stelmakh, Nihar Shah, and Aarti Singh. PeerReview4All: Fair and accurate reviewer assignment in peer review. *arXiv preprint arxiv:1806.06237*, 2018. 21, 39

[168] Ivan Stelmakh, Nihar Shah, and Aarti Singh. On testing for biases in peer review. In *NeurIPS*, 2019. 39

[169] Stephen M. Stigler. Citation patterns in the journals of statistics and probability. *Statistical Science*, 9(1):94–108, 1994. 66, 67

[170] Stephen M. Stigler. Regression towards the mean, historically considered. *Statistical methods in medical research*, 6(2):103–14, 02 1997. 71

[171] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974. 40

[172] Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116 (29):14516–14525, 2019. 71

[173] Balázs Szörényi, Róbert Busa-Fekete, Adil Paul, and Eyke Hüllermeier. Online rank elicitation for Plackett-Luce: A dueling bandits approach. In *Advances in Neural Information Processing Systems*, 2015. 67

[174] Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4):15–42, 2019. 86, 87, 88

[175] Robyn Tamblyn, Nadyne Girard, Christina J. Qian, and James Hanley. Assessment of potential bias in research grant peer review in Canada. *CMAJ*, 190(16):E489–E499, 2018. 2

[176] Edward L Thorndike. A constant error in psychological ratings. *Journal of applied psychology*, 4(1):25–29, 1920. 88

[177] L. L. Thurstone. A law of comparative judgement. *Psychological Review*, 34:278–286, 1927. 79

[178] Kevin Tian, Weihao Kong, and Gregory Valiant. Learning populations of parameters. In *Advances in Neural Information Processing Systems*, 2017. 13

[179] Ryan J. Tibshirani, Holger Hoefling, and Robert Tibshirani. Nearly-isotonic regression.

Technometrics, 53(1):54–61, 2011. 28, 40, 45

[180] Andrew Tomkins, Min Zhang, and William D. Heavlin. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48): 12708–12713, 2017. 39

[181] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. 210, 211

[182] A.W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. 210

[183] Karen I Van der Zee, Arnold B Bakker, and Paulien Bakker. Why are structured interviews so rarely used in personnel selection? *Journal of Applied Psychology*, 87(1):176, 2002. 88

[184] C. Mirjam van Praag and Bernard M.S. van Praag. The benefits of being economics professor A (rather than Z). *Economica*, 75(300):782–796, 2008. 100

[185] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. 48

[186] Jingyan Wang and Nihar B. Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 864–872. International Foundation for Autonomous Agents and Multiagent Systems, 2019. 4, 23, 39

[187] Jingyan Wang, Nihar B. Shah, and R. Ravi. Stretching the effectiveness of MLE from accuracy to bias for pairwise comparisons. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 66–76. PMLR, 2020. 4

[188] Jingyan Wang, Ivan Stelmakh, Yuting Wei, and Nihar B. Shah. Debiasing evaluations that are biased by evaluations. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*. AAAI Press, 2021. 4

[189] Ellen J Weber, Patricia P Katz, Joseph F Waeckerle, and Michael L Callaham. Author perception of peer review: impact of review quality and acceptance on satisfaction. *JAMA*, 287(21):2790–2793, 2002. 38

[190] Yuting Wei, Martin J. Wainwright, and Adityanand Guntuboyina. The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. *Ann. Statist.*, 47(2):994–1024, 04 2019. 235

[191] Simon N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686, 2004. 40

[192] Yichong Xu, Han Zhao, Xiaofei Shi, and Nihar Shah. On strategyproof conference review. In *IJCAI*, 2019. 21

[193] H. P. Young. Condorcet's theory of voting. *American Political Science Review*, 82(4): 1231–1244, 1988. 15

[194] Kyusang Yu, Enno Mammen, and Byeong U Park. Semi-parametric regression: Efficiency gains from modeling the nonparametric part. *Bernoulli*, 17(2):736–748, 2011. 40

[195] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. *FA*IR: A Fair Top-k Ranking Algorithm*, page 1569–1578. Association for Computing Machinery, New York, NY, USA, 2017. 287

[196] Cun-Hui Zhang. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2): 528–555, 2002. 40