Journal of Machine Learning Research 25 (2024) 1-120 Submitted 7/22; Revised 8/24; Published 11/24

Debiasing Evaluations That Are Biased by Evaluations

Jingyan Wang[◇] JINGYANW@TTIC.EDU
Ivan Stelmakh[§] ISTELMAKH@NES.RU
Yuting Wei[*] YTWEI@WHARTON.UPENN.EDU
Nihar Shah[†] NIHARS@CS.CMU.EDU

[◇] *Toyota Technological Institute at Chicago*
Chicago, IL 60637, USA
[§] *New Economic School*
Moscow, Russia
[*] *Department of Statistics and Data Science*
University of Pennsylvania
Philadelphia, PA 19104, USA
[†] *School of Computer Science*
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Manuel Gomez-Rodriguez

Abstract

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.

Keywords: Crowdsourcing, bias mitigation, fairness, statistical estimation, shape constraints

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".[1] We now 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 (Greenwald and Gillmore, 1997; Johnson, 2003; Boring et al., 2016). For instance, as noted in Johnson (2003, 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 (Boring et al., 2016), and student ratings serve as a poor predictor on the follow-on course achievement of the students (Carrell and West, 2010; Braga et al., 2014):

> *"...teachers who are associated with better subsequent performance receive worst evaluations from their students."* (Braga et al., 2014)

The outcome we consider in teaching evaluation is the grades that the students receive in the course under evaluation[2] 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 peer review, where it has been proposed (Crowcroft et al., 2009) and implemented (Journal of Systems Research, 2021; Goldberg et al., 2023) that authors rate their received reviews as a method to measure and improve the quality of the review process. It is well understood that authors are more likely to give higher ratings to a positive review than to a negative review (Weber et al., 2002; Papagiannaki, 2007; Khosla et al., 2013; Goldberg et al., 2023):

> *"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."* (Weber et al., 2002)

Due to this problem, an author feedback experiment (Papagiannaki, 2007) 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."*

1. Throughout the paper we restrict the scope of the bias to the outcome-induced bias. Note that this concept is not related to the statistical notion of the bias.
2. 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.

By comparing scores given to reviews by authors, (other) paper reviewers, external reviewers, and meta-reviewers at NeurIPS 2022, a similar conclusion was made that "caution must be taken when asking authors to evaluate reviews on their own papers" (Goldberg et al., 2023). 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 (Becker and Watts, 1999), 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 (Carrell and West, 2010). Furthermore, author-provided ratings can be a factor for selecting reviewer awards (Khosla et al., 2013), and reviewers with a history of poor reviews may risk being removed from the editorial board (Journal of Systems Research, 2021); student-provided ratings can be a heavily-weighted component for salary or promotion and tenure decision of the faculty members (Becker and Watts, 1999; Carrell and West, 2010; Boring et al., 2016). 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 (Molina et al., 2019).

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 (Greenwald and Gillmore, 1997; Johnson, 2003; Boring et al., 2016) 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.

The code to reproduce our results is available at `https://github.com/jingyanw/outcome-induced-debiasing`.

1.1 Our contributions

We identify and formulate a problem of mitigating biases in evaluations that are biased by evaluations (Section 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 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 3.2).

We then provide a theoretical analysis of the performance of our proposed algorithm (Section 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 the only-bias (Section 4.2) and only-noise (Section 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) 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 experiments on synthetic, semi-synthetic, and real-world data (Section 5), including various settings not covered by the theoretical results.

A short version of this paper is published at the AAAI 2021 conference (Wang et al., 2021). In comparison to the conference version, the current paper extends the theoretical results (Theorem 5(b)), conducts simulation studies (Sections 5.1-5.4), and conducts an experiment using real-world data from peer review (Section 5.6).

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 (Caulkins et al., 1996), and more complicated parametric generative models (Johnson, 1997). 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 (Ge et al., 2013; Lee, 2015; Tomkins et al., 2017; Noothigattu et al., 2021; Stelmakh et al., 2021; Wang and Shah, 2019; Stelmakh et al., 2019; Fiez et al., 2020; Jecmen et al., 2020; Manzoor and Shah, 2021) 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. Barlow et al., 1972; Zhang, 2002; Mammen and Yu, 2007; Groeneboom

and Jongbloed, 2014). Specifically, our problem becomes isotonic regression, if in our exact formulation (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., Hastie and Tibshirani, 1990; Cuzick, 1992; Wood, 2004; Yu et al., 2011) with shape constraints (Chen and Samworth, 2016). In particular, one class of semiparametric additive models involves linear components and components with ordering (isotonic) constraints (Huang, 2002; Cheng, 2009; Meyer, 2013; Rueda, 2013). Our optimization (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 (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 2) is partly inspired by permutation-based models (Shah et al., 2017; Shah, 2017) 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 (e.g., Stone, 1974; Kohavi, 1995; Hastie et al., 2009). Yet, this generic scheme cannot be directly applied to models where training samples are not exchangeable — in which case, both the sub-sampling 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 (Tibshirani et al., 2011). 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.

Finally, our work is complementary to prior work in psychology, human-computer interaction, and mechanism design that aims at designing interventions to reduce cognitive biases in human decision making. On the psychology front, while in this work we do not aim at establishing the cognitive mechanism of the outcome-induced bias, we note in part it can be seen as an anchoring bias (Tversky and Kahneman, 1974; Strack and Mussweiler, 1997; Mussweiler and Strack, 2001) — students being unable to move away from their satisfaction with their own performance — or substitution bias (Kahneman and Frederick, 2002) — students substituting the complex question of teaching evaluation with a simpler question of measuring their overall satisfaction with the course. It has been observed that when the task is framed in a rational manner and requires analytical judgement, the reliance on cognitive heuristics can be reduced (Stanovich, 1999; Kahneman and Frederick, 2002). On the human-computation interaction front, recent work (Rastogi et al., 2022) observes that when people are required to spend more time on the task, the impact of the anchoring bias decreases.

On the mechanism design front, an extensive line of literature concerns incentivizing truthful reporting, in settings where ground-truth answers are available (Brier, 1950; Shah and Zhou, 2016) or unavailable (Prelec, 2004; Wolfers and Zitzewitz, 2004; Miller et al., 2005; Dasgupta and Ghosh, 2013). The technique designed in this paper can be coupled with the aforementioned interventions to further reduce the impact of the outcome-induced bias on the evaluations.

2. Problem formulation

For ease of exposition, throughout the paper we describe our problem formulation using the running example of course evaluation, but we note that our problem formulation is general, and we comment on how the problem formulation maps to other applications such as paper review where appropriate. 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.[3] Denote $y_{ij} \in \mathbb{R}$ as the rating given by the j^{th} student in course i, for 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:

$$y_{ij} = x_i^* + b_{ij} + z_{ij}, \tag{1}$$

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 $\frac{1}{d}\|\widehat{x} - x^*\|_2^2$.

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

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:

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

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 (2), so that the objective (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 14 of Appendix C.2.1. Furthermore,as explained in Appendix B.1, the optimization (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 (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 (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.

3.1 Behavior of our estimator under some fixed choices of λ

To facilitate understandings of the estimator (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 (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 4.2, 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 (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 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 (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.

$\lambda \in (0, \infty)$: 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.

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 1. 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 (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

Algorithm 1: 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 π_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 π_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 Ω^{t} and Ω^{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:

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

  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 π, 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^π, j_1^π) and (i_2^π, j_2^π) 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 $\lambda_{\mathrm{cv}} \in \arg\min_{\lambda \in \Lambda} e^{(\lambda)}$. (Ties are broken arbitrarily)

Ω^{v}, by computing $\frac{1}{|\Omega^{\text{v}}|}\|Y - \widehat{x}^{(\lambda)}\mathbf{1}^T - \widehat{B}^{(\lambda)}\|^2_{\Omega^{\text{v}}}$. However, recall that the estimate $\widehat{B}^{(\lambda)}$ only estimates the bias on the training set meaningfully, and we have $\widehat{B}^{(\lambda)}_{ij} = 0$ 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.

Interpolation. We now discuss how the algorithm interpolates the bias $\widetilde{b}^{(\lambda)}_{ij}$ 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 $[\widetilde{b}^{(\lambda)}_\pi]_{ij}$) using the value of $\widehat{b}_{i^\pi j^\pi}$. That is, we set $[\widetilde{b}^{(\lambda)}_\pi]_{ij} = \widehat{b}^{(\lambda)}_{i^\pi j^\pi}$. 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 Tibshirani et al. (2011).

- **Taking the mean over all total orderings in $\mathcal{T}$ (Line 15):** After we find the interpolated bias $\widetilde{B}^{(\lambda)}_\pi$ 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.[4] 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 1: 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 (Matthews, 1991; Bubley and Dyer, 1999) or exact (Huber, 2006). For iterating over $\mathcal{T}$ which can be computationally intractable, we approximate the true mean over $\mathcal{T}$ by sampling from $\mathcal{T}$

4. 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.

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 4.1); this subclass of partial orderings is also evaluated in the experiments (Section 5).

4. Theoretical guarantees

We now present theoretical guarantees for our proposed estimator (2) along with our cross-validation algorithm (Algorithm 1). In Section 4.2 and 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, we analyze the cross-validation algorithm. The proofs of all results are in Appendix C.

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 paper. Specifically, we make the following assumptions on the model (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 (2) and the cross-validation algorithm (Algorithm 1) requires no knowledge of these distributions or standard deviation parameters σ and η. Throughout the theoretical results, we consider the solution $\widehat{x}^{(\lambda_{\mathrm{cv}})}$ as solution at $\lambda = \lambda_{\mathrm{cv}}$ 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 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 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., $\{\mathrm{A},\mathrm{B},\mathrm{C},\mathrm{D},\mathrm{F}\}$ or $\{\mathrm{Pass},\mathrm{Fail}\}$), or numeric scores (e.g., in the range of $[0,100]$) of the students. Intuitively a group ordering assumes that a student

receiving a strictly higher grade is more positively biased in rating than a student receiving a lower grade, irrespective of their course membership. 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. In paper review, the groups can be the decisions for the papers (e.g., strong accept, weak accept, borderline, etc.) or scores (e.g., on a scale of 1 to 10).

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 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 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. Such assumptions naturally arise in practice. For example, in the peer review process of NeurIPS 2016 (Shah et al., 2018), the criteria for a paper receiving scores 5, 4, or 3 are explicitly defined as the paper being the top 1/1000, 3%, or 30% among all submissions, respectively. The final regularity condition below is motivated from the scenario where student performances are totally ranked in the course.

Definition 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 $\mathcal{O}$.*

Let us understand this condition through some examples from teaching evaluation, where this condition is or is not satisfied. Assume the total ordering is derived from real-valued grades, where grades for each course is sampled from a distribution specific to this course. Then if any pair of distributions has overlapping density of constant mass (such as two Gaussian distributions whose means are a constant away from each other), then Definition 4 is satisfied. On the other hand, if the grades in one course are all higher than the grades in another course, Definition 4 is not satisfied. With these preliminaries in place, we now present our main theoretical results.

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 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 Appendix C.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 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 5:*

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

In Appendix A.1, we also evaluate the mean estimator. We show that under the conditions of Theorem 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 Appendix A.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 13 in Appendix A.2).

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 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*

$$\begin{aligned} \widehat{x}^{(\infty)}_i &= \frac{1}{n} \sum_{j=1}^{n} y_{ij}, \qquad \text{for each } i \in [d], \text{ and} \\ \widehat{B}^{(\infty)} &= 0. \end{aligned} \tag{3}$$

The proof of this result is provided in Appendix C.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 8 (Implication of Example 15.8 in Wainwright, 2019)**.** *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 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 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 Appendix C.5. From Theorem 5 we have that the estimator $\widehat{x}^{(0)}$ (at $\lambda = 0$) is also consistent under the only-bias setting. Combining Theorem 5 with Theorem 9, we have $\widehat{x}^{(\lambda_{\mathrm{cv}})}$ approaches $\widehat{x}^{(0)}$. Formally, under the conditions of Theorem 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 $\|Y - \widehat{x}^{(\lambda)}\mathbf{1}^T - \widetilde{B}^{(\lambda)}\|^2_{\Omega^{\mathrm{v}}}$ 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 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 Appendix C.6. By the consistency of $\widehat{x}^{(\infty)}$ implied from Proposition 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 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 4.2 and 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.

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 1 to obtain a value of the hyperparameter λ_{cv}; we then compute the estimate $\widehat{x}^{(\lambda_{\text{cv}})}$ as the solution to (2) at $\lambda = \lambda_{\text{cv}}$ (that is, we solve (2) on the entire data combining the training set and the validation set).[5] Implementation details for the cross-validation algorithm (Algorithm 1) are provided in Appendix B.1. Throughout the experiments, we use $\Lambda = \{2^i : -9 \le i \le 5, i \in \mathbb{Z}\} \cup \{0, \infty\}$. 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.

To generate the ratings $\{y_{ij}\}$, we follow model (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 4.1. Namely, the noise terms $\{z_{ij}\}$ are i.i.d. $\mathcal{N}(0, \eta^2)$ for some parameter $\eta \ge 0$, and the bias terms $\{b_{ij}\}$ are marginally distributed as $\mathcal{N}(0, \sigma^2)$ for some parameter $\sigma \ge 0$ while obeying a total ordering uniformly sampled from all total orderings that are consistent with the given partial ordering. 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 set $x^* = 0$ without loss of generality, because, as explained in Proposition 18 in Appendix C.2.1, the results remain the same for any value of x^*.

We compare our cross-validation algorithm with the mean, median, and also the reweighted mean estimator introduced in Appendix A.2. The mean estimator is the sample mean for each course (same as our estimator at $\lambda = \infty$) defined as $[\widehat{x}_{\text{mean}}]_i = \frac{1}{n}\sum_{j\in[n]} y_{ij}$ for each $i \in [d]$, and the median estimator is defined as $[\widehat{x}_{\text{med}}]_i = \text{median}(y_{i1}, \ldots, y_{in})$ for each $i \in [d]$. The reweighted mean estimator is not applicable to total orderings or general partial orderings. 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.

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.

5. Note that this is different from the theoretical results in Section 4.4, where we solve (2) at $\lambda = \lambda_{\text{cv}}$ only on the training set.

- **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.

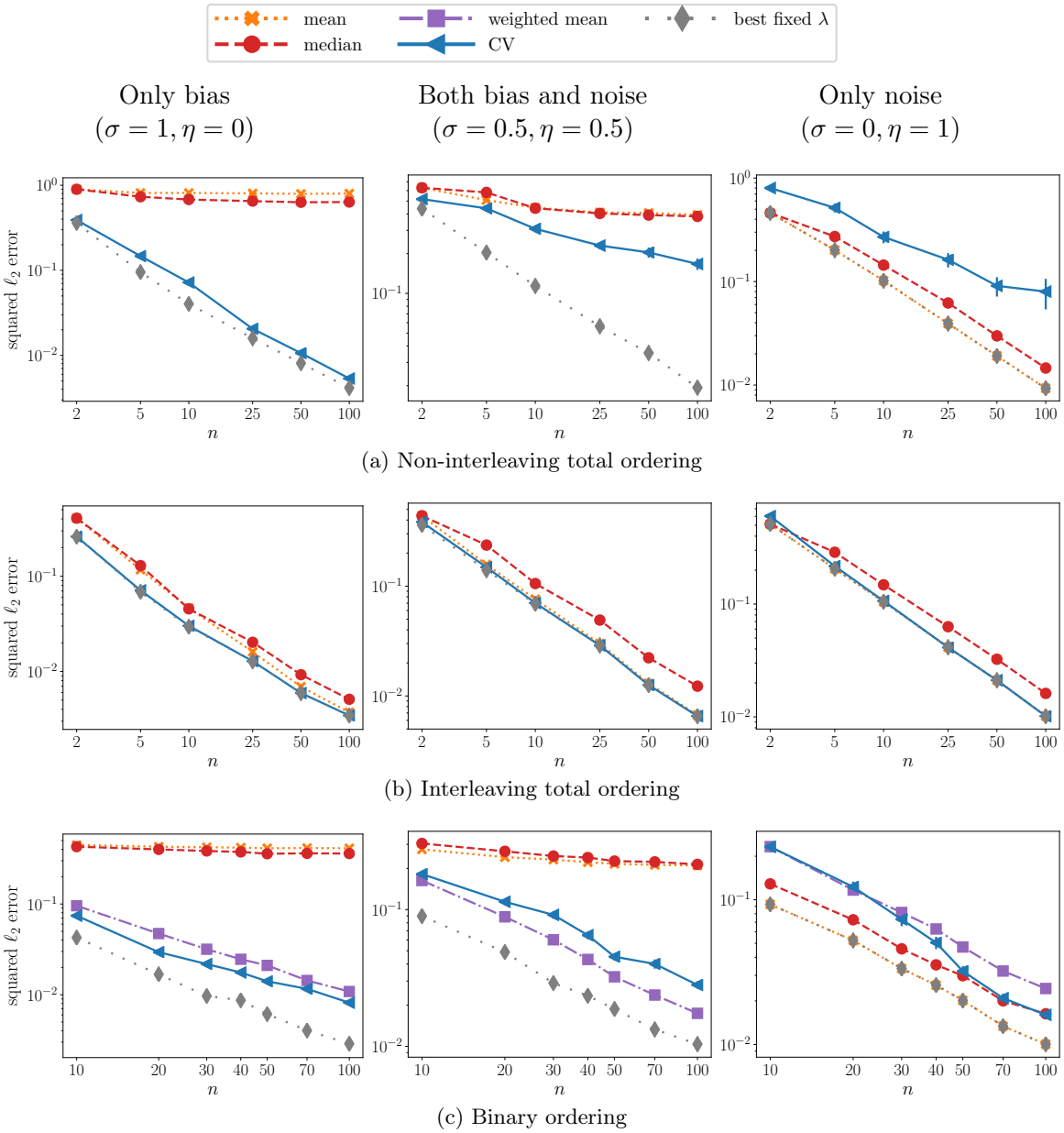
- **Binary ordering:** We call a group ordering a "binary" ordering, if there are $r = 2$ groups. Specifically, we consider a group distribution where $(\ell_{i1}, \ell_{i2}) = (0.9n, 0.1n)$ for half of the courses i, and $(\ell_{i1}, \ell_{i2}) = (0.1n, 0.9n)$ for the other half of the courses i.

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. 1. In the non-interleaving case (Fig. 1a) and the binary case (Fig. 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 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. 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 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.

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 5.1, with $n = 50$. The histograms in Fig. 2 plot the fraction of times that each value of $\lambda \in \Lambda$ is chosen by cross-validation. When there



(a) Non-interleaving total ordering

(b) Interleaving total ordering

(c) Binary ordering

Figure 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.

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, and show that cross-validation is indeed able to adapt to different amounts of bias and noise present in the data.

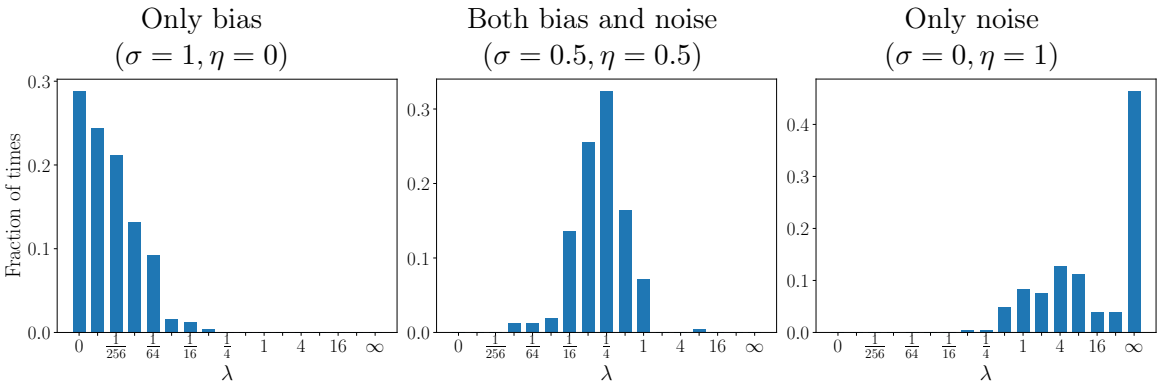
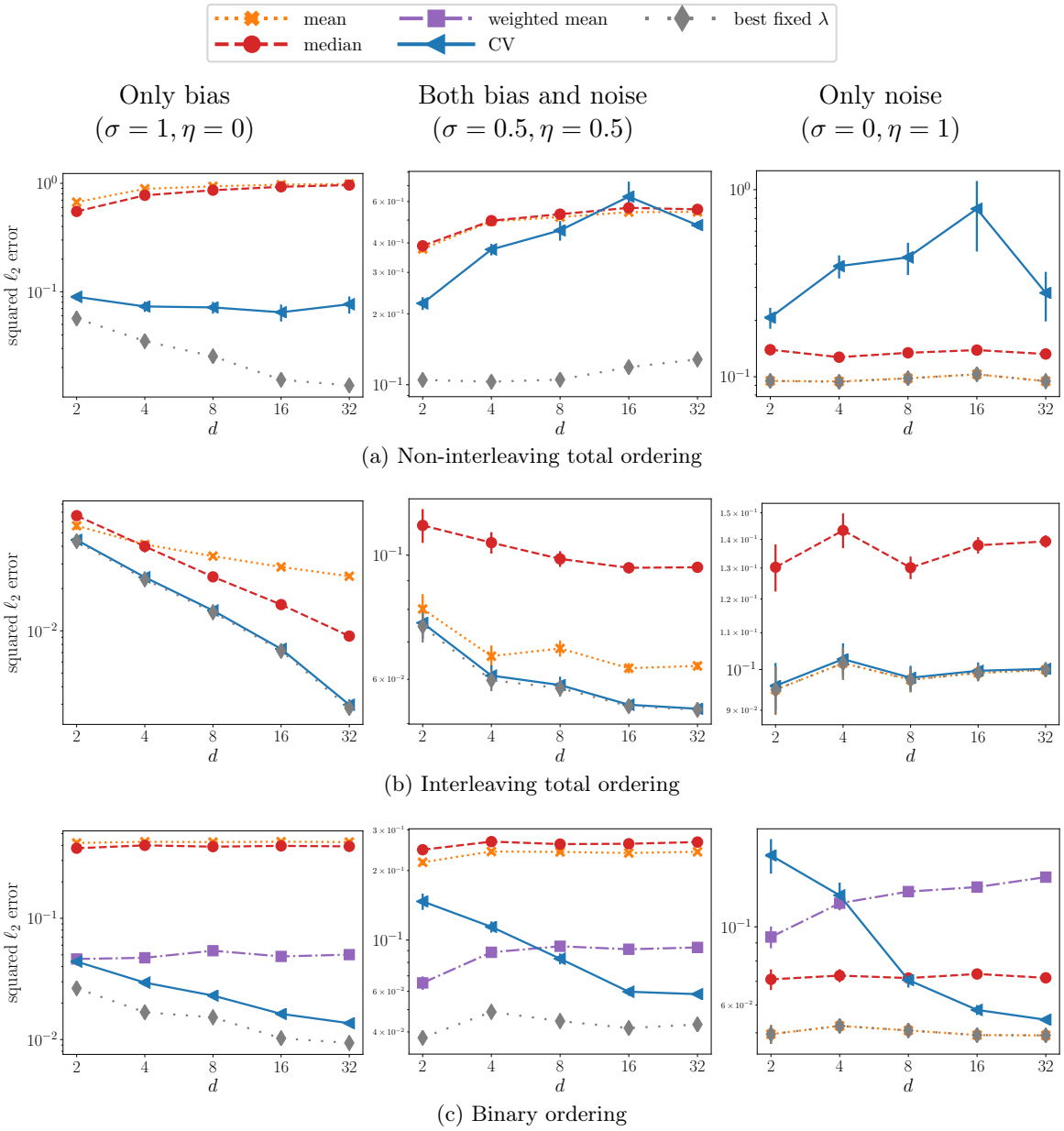


Figure 2: The histogram on the fraction of times each value of λ is chosen by cross-validation. Cross-validation is able to choose the value of λ adaptive to different amounts of bias and noise.

5.3 The regime of $d > n$

In our theoretical results from Section 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 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. 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 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.



(a) Non-interleaving total ordering

(b) Interleaving total ordering

(c) Binary ordering

Figure 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.

5.4 General partial orderings

In our theoretical results from Section 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.

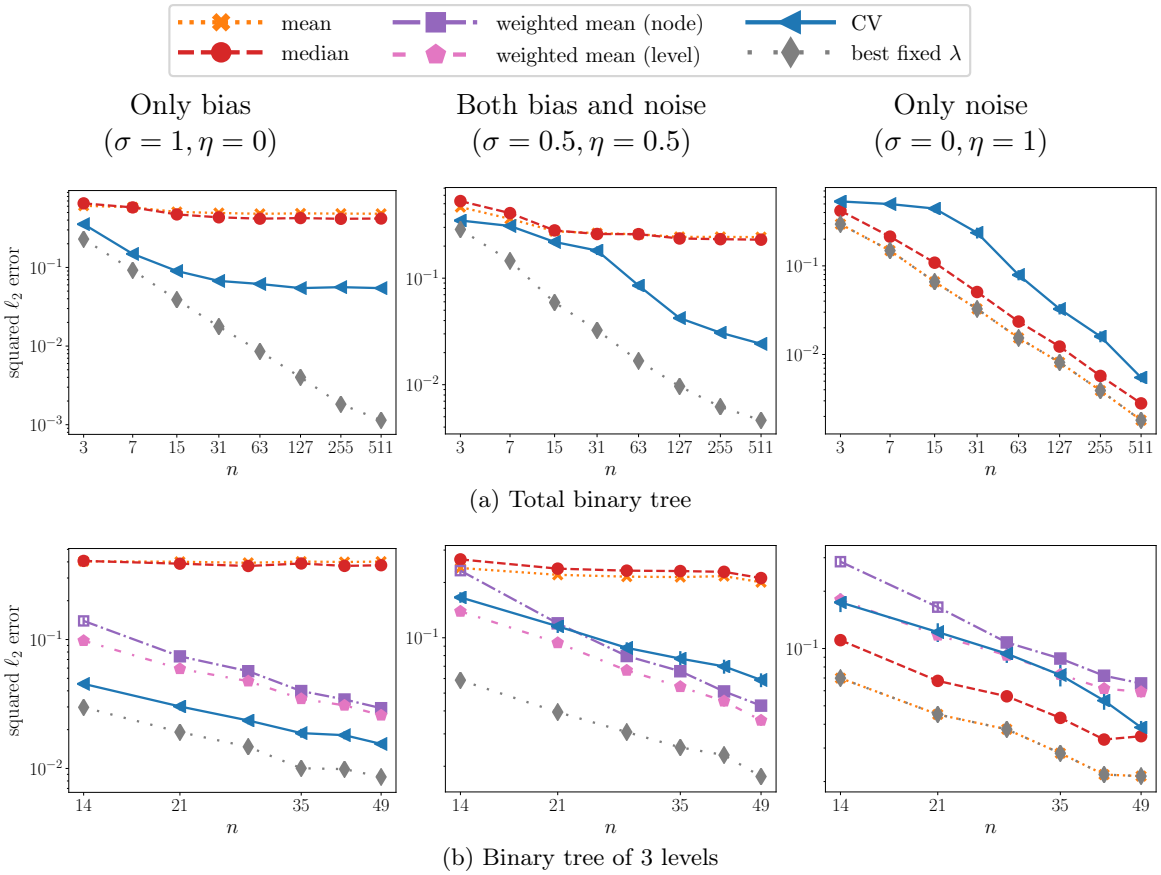
A tree ordering arises in elimination-based tournaments, where participants are divided into groups. Within each group, a winner is determined to proceed in the tournament, and all losing participants are eliminated. In such elimination-based tournaments, it is natural to assume that a participant is more negatively biased than the winner they lose to.

- **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,
  - 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 Appendix B.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. 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.



(a) Total binary tree

(b) Binary tree of 3 levels

Figure 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.

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 Indiana University Bloomington (2020), where the possible grades that students receive are A+ through D-, and F. We consider three ways to construct the group orderings:

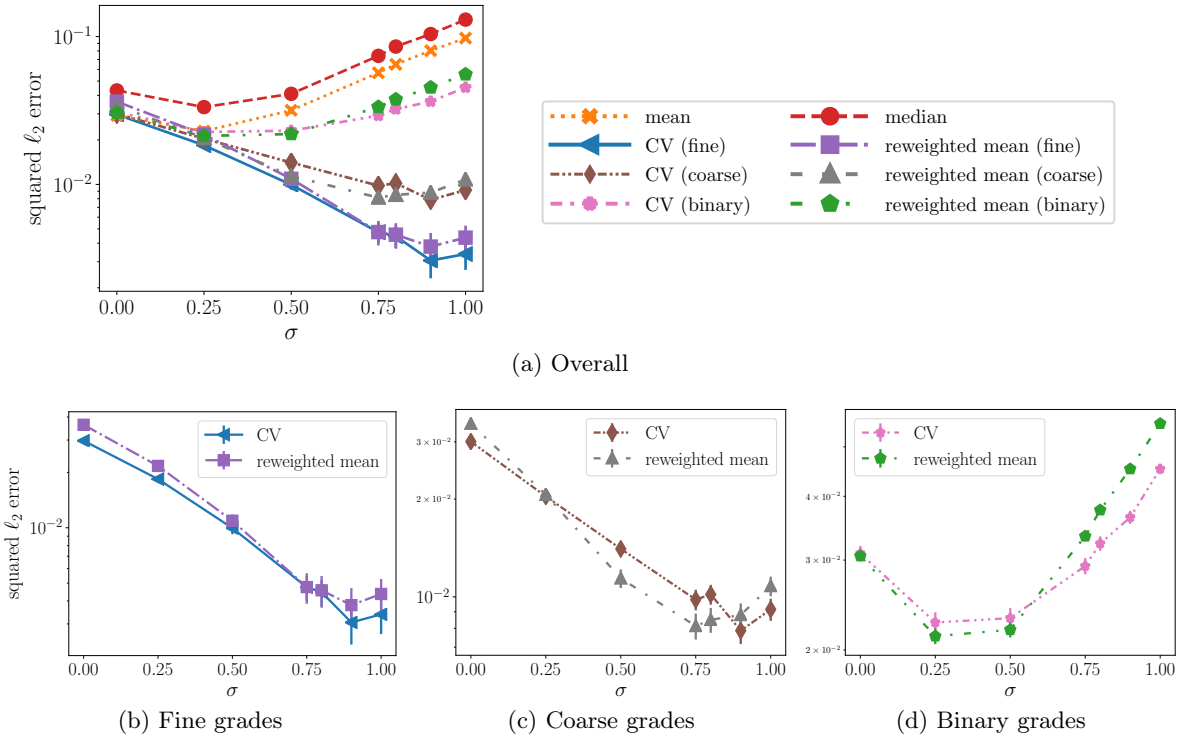
- **Fine grades:** The 13 groups correspond to the grades of A+ through D-, and F.
- **Coarse grades:** The fine grades are merged to 5 groups of A, B, C, D and F, where grades in $\{\text{A+}, \text{A}, \text{A-}\}$ are all considered A, etc.
- **Binary grades:** The grades are further merged to 2 groups 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 (1) under the Gaussian assumptions (A2) and (A1). 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 true quality is set as $x^* = 0$ (again the results are independent from the value of x^*, the results are independent from the value of x^*, by Proposition 18 in Appendix C.2.1). 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 Appendix B.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 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. 5b), and slightly better on a subset of values of σ for the coarse grades (Fig. 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. 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.



(a) Overall

(b) Fine grades

(c) Coarse grades

(d) Binary grades

Figure 5: The performance of our estimator (with cross-validation) on semi-synthetic grading data, compared to the mean, median and reweighted mean estimators.

5.6 Real-world data from proposal review

We now move to a real-world data (Kerzendorf et al., 2020) collected for proposal peer review at the European Southern Observatory (ESO). In their review process, each proposer submits one proposal, and each proposal is assigned to around 8 reviewers. Each reviewer provides a grade in the range of 1.0–5.0 (with one decimal point allowed) to each proposal that they are assigned, where 5.0 means the highest quality.[6] After the reviews are collected, each proposer is given access to each individual review comment in text and the mean score of all reviewers (without individual reviewers' scores). The proposer then scores each individual review in terms of its helpfulness by an integral number in the range of 1–4, where 4 means the most helpful. The data contains 706 helpfulness scores and the respective proposal grade, between 120 proposers and 136 reviewers. A positive correlation between the proposal grades given by the reviewers and the helpfulness scores given by the proposers has been qualitatively observed (see Figure 4d in Kerzendorf et al., 2020).

Our objective is to evaluate each reviewer's overall helpfulness across all their reviews, while reducing the outcome-induced bias due to this positive correlation. To apply our proposed method, we construct a set of ordering constraints as follows. Within each proposal, we assume that the proposer is more positively biased towards a review if the reviewer provides a higher grade to the proposal. That is, a proposer's biases towards all the reviewers

6. In the original data, a grade of 1.0 means the highest quality. We flip the scale by replacing each grade $g \in [1, 5]$ by the grade $(6 - g)$ to align with our notation.

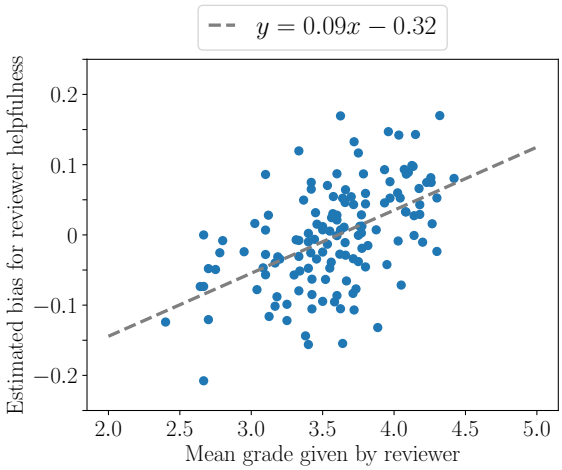


Figure 6: The estimated bias for each reviewer's helpfulness, as a function of the mean proposal grade given by the reviewer.

are in the same order as the one induced by the proposal grades given by these reviewers. Note that in the review process, the proposers do not have access to reviewers' individual proposal grades. Hence, this ordering constraint is constructed under the assumption that proposers are able to perceive how positive the reviewers are from the reviewers' text comments. We do not assume ordering constraints in terms of the bias terms across different proposers, because proposers may have different calibration and may also respond to the received proposal grades and text comments to different extent.

Note that there is no ground-truth available (i.e., a reviewer's "true" helpfulness) for this real-world data. Hence, we focus on providing qualitative interpretations on the output of our algorithm. The cross-validation algorithm identifies the value of the hyperparameter as $\lambda = 2$, suggesting a mixture of bias and noise in the data. In Figure 6, we plot the mean proposal grade given by each reviewer, and the amount of estimated bias in their mean helpfulness score. We observe a positive correlation (Pearson's $r = 0.519$) for the bias terms computed by our algorithm, where the estimated bias is more positive if the reviewer gives higher grades to the proposals.

We further qualitatively inspect the two reviewers whose bias terms are estimated to be the minimum and maximum in Tables 1 and 2. In Table 1, we show the reviewer ("u758" in the data) whose bias is estimated to be the minimum and accordingly receives the most increase in its estimated quality by the algorithm. Table 1 shows the three proposers who give helpfulness scores to this reviewer (in bold), along with other reviewers that these proposers give scores to. We observe that these three proposers give lower helpfulness scores of 1 and 2 primarily, compared to the mean helpfulness score 2.54 given by all proposers in the data. Since the ordering constraints are imposed only within each individual reviewer, the bias terms account for both proposer calibration (whether a proposer tends to give high or low scores in general) and outcome-induced bias within the proposer. In this case, we observe that bias primarily accounts for proposer calibration, as there is no clear positive correlation between the proposal grade received and the helpfulness score given by the proposer.

proposer	proposal grade	review helpfulness	estimated bias
p136	**2.5**	**1**	**-0.31**
	2.6	1	-0.31
	3.3	2	-0.20
	4.0	1	-0.20
	4.0	1	-0.20
	4.5	3	0.29
p334	2.2	2	-0.22
	2.5	**1**	**-0.19**
	3.0	2	-0.19
	3.4	2	-0.19
	3.4	2	-0.19
	3.5	1	-0.19
	3.5	2	-0.19
p341	2.0	2	-0.17
	2.5	1	-0.17
	3.0	**2**	**-0.13**
	3.3	2	-0.13
	4.5	1	-0.13
	4.6	1	-0.13

Table 1: Information about the proposers associated to the reviewer with the minimum estimated bias.

We also show in Table 2 the reviewer ("u178") whose estimated bias term is the maximum by the algorithm. We observe both proposer calibration and outcome-induced bias in the estimated bias terms. For example, proposers "p241" tend to give higher helpfulness scores in general, but also appears to favor reviewers who give high proposal grades to their proposal, where the two proposers who give the lowest proposal grade (3.3 and 3.5) in turn receive the lowest helpfulness score of 1. Hence, the positive bias terms suggest that the quality of this reviewer may be over-estimated.

In this experiment, we observe that the bias terms estimated by our algorithm capture both trends due to proposer calibration and outcome-induced bias. The outcome-induced bias presented in this data is not notably strong, potentially due to careful phrasing when eliciting proposer responses (by explaining to proposers that "positive comments like 'best proposal I ever read' can be ranked as not helpful as it does not improve the proposal further" (Kerzendorf et al., 2020)), and not presenting individual reviewer scores to proposers but only the mean score. Our qualitative observations align with our theoretical intuition developed for the algorithm, and it remains important future work to quantitatively validate our algorithm on other real data where ground-truth quality is available. While it may not be feasible to directly measured the "true quality", proxies to this ground-truth can be obtained by, for example, collecting helpfulness scores from program chairs or organization chairs who are experts in the field while being impartial to reviewers' grades given to individual proposals.

proposer	proposal grade	review helpfulness	estimated bias
p057	2.9	1	-0.28
	3.0	1	-0.28
	3.5	**1**	**-0.28**
	3.5	2	-0.28
	3.6	1	-0.28
	4.0	3	0.20
	4.1	3	0.20
p241	3.3	1	-0.10
	3.5	1	0.02
	3.5	4	0.02
	4.0	4	0.16
	4.4	**2**	**0.16**
	4.5	3	0.16
	4.5	4	0.16
p492	3.0	4	0.16
	3.7	3	0.16
	4.0	3	0.16
	4.2	**1**	**0.16**
	4.5	3	0.21
p632	2.4	1	-0.56
	2.8	1	-0.44
	3.9	1	-0.36
	4.5	**4**	**0.64**
	4.5	4	0.64

Table 2: Information about the proposers associated to the reviewer with the maximum estimated bias.

6. 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, in a non-asymptotic regime, in a high-dimensional regime with $d \gg n$, under other types of partial orderings, and under a model mismatch where the provided partial ordering $\mathcal{O}$ is inaccurate. 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, course difficulty, whether the course content is interesting to students, and whether the course is mandatory or elective. All of these confounding factors are not addressed by supplying an ordering due to grading bias. Second, the user should be careful in supplying the appropriate ordering constraints to the algorithm, ensuring these constraints have been validated separately. One potential caveat of using student grades as an ordering constraint in teaching evaluation is that teaching evaluation is often conducted before the grades are released to students, already mitigating such outcome-induced biases. However, it remains possible that students may still hold expectations about what final grades they will receive, based on their performance in the course so far. Another potential caveat is that students who find the course material interesting are more motivated and receive higher grades in the course as a result. In this case, their higher ratings to the instructors are an indication of teaching quality as opposed to bias. Third, introducing the debiasing algorithm may induce strategic behaviors. For example, would the instructors intentionally provide lower grades under the expectation that the debiasing algorithm may over-correct? Finally, 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. With these limitations in mind, we recommend using our algorithm 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

J.W., I.S., and N.S. were supported in part by NSF CAREER award 1942124 and in part by NSF CIF 1763734. Y.W. was supported in part by the NSF grants DMS-2147546/2015447, CAREER award DMS-2143215, CCF2106778, and the Google Research Scholar Award. We thank the anonymous reviewers for helpful comments and discussions.

References

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.

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.

William E. Becker and Michael Watts. How departments of economics evaluate teaching. *The American Economic Review*, 89(2):344–349, 1999.

D.P. Bertsekas. *Convex Optimization Theory.* Athena Scientific optimization and computation series. Athena Scientific, 2009.

Anne Boring, Kellie Ottoboni, and Philip B. Stark. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 2016.

Michela Braga, Marco Paccagnella, and Michele Pellizzari. Evaluating students' evaluations of professors. *Economics of Education Review*, 41:71–88, 2014.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO; 2. URL `https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml`.

Russ Bubley and Martin Dyer. Faster random generation of linear extensions. *Discrete Mathematics*, 201(1):81–88, 1999.

Scott E. Carrell and James E. West. Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3):409–432, 2010.

Jonathan P. Caulkins, Patrick D. Larkey, and Jifa Wei. Adjusting GPA to reflect course difficulty. Technical report, Carnegie Mellon University, 1996.

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.

Guang Cheng. Semiparametric additive isotonic regression. *Journal of Statistical Planning and Inference*, 139(6):1980–1991, 2009.

Jon Crowcroft, S. Keshav, and Nick McKeown. Viewpoint scaling the academic publication process to internet scale. *Communications of the ACM*, 52(1):27–30, 2009.

Jack Cuzick. Semiparametric additive regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):831–843, 1992.

Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *International Conference on World Wide Web (WWW)*, pages 319—-330, New York, NY, USA, 2013. Association for Computing Machinery.

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.

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.

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].

Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B. Shah. Peer reviews of peer reviews: A randomized controlled trial and other experiments, 2023. URL `https://arxiv.org/abs/2311.09497`.

Anthony G. Greenwald and Gerald M. Gillmore. Grading leniency is a removable contaminant of student ratings. *The American psychologist*, 52(11):1209–1217, November 1997.

Piet Groeneboom and Geurt Jongbloed. *Nonparametric estimation under shape constraints*, volume 38. Cambridge University Press, 2014.

Trevor Hastie and Robert Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jian Huang. A note on estimating a partly linear model under monotonicity constraints. *Journal of Statistical Planning and Inference*, 107(1):343–351, 2002.

Mark Huber. Fast perfect sampling from linear extensions. *Discrete Mathematics*, 306(4): 420–428, 2006.

Indiana University Bloomington. Grade distribution database, 2020. `https://gradedistribution.registrar.indiana.edu/index.php` [Online; accessed 30-Sep-2020].

Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar B. Shah, Vincent Conitzer, and Fei Fang. Mitigating manipulation in peer review via randomized reviewer assignments. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Valen E. Johnson. An alternative to traditional gpa for evaluating student performance. *Statistical Science*, 12(4):251–278, 11 1997.

Valen E. Johnson. *Grade Inflation: A Crisis in College Education*. Springer New York, 1 edition, 2003.

Journal of Systems Research. Call for papers, 2021. `https://www.jsys.org/cfp/` [Online; accessed 09-Jul-2022].

Daniel Kahneman and Shane Frederick. Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49: 49—81, 2002.

Wolfgang Kerzendorf, Ferdinando Patat, Dominic Bordelon, Glenn van de Ven, and Tyler Pritchard. Distributed peer review enhanced with natural language processing and machine learning. *Nature Astronomy*, 4:711—717, 2020.

Aditya Khosla, Derek Hoiem, and Serge Belongie. Analysis of reviews for CVPR 2012, 2013. `https://people.csail.mit.edu/khosla/papers/reviewer_analysis.pdf` [Online; accessed 09-Jul-2022]",.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 1995.

Carole J. Lee. Commensuration bias in peer review. *Philosophy of Science*, 82(5):1272–1283, 2015.

Enno Mammen and Kyusang Yu. Additive isotone regression. In *Asymptotics: particles, processes and inverse problems*, volume 55, pages 179–195. Institute of Mathematical Statistics, 2007.

Emaad Manzoor and Nihar B. Shah. Uncovering latent biases in text: Method and application to peer review. In *AAAI Conference on Artificial Intelligence*, 2021.

Peter Matthews. Generating a random linear extension of a partial order. *Annals of Probability*, 19(3):1367–1392, 1991.

Mary C. Meyer. Semi-parametric additive constrained regression. *Journal of nonparametric statistics*, 25(3):715–730, 2013.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

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.

Thomas Mussweiler and Fritz Strack. Considering the impossible: Explaining the effects of implausible anchors. *Social Cognition*, 19(2):145–160, 2001.

Ritesh Noothigattu, Nihar B. Shah, and Ariel Procaccia. Loss functions, axioms, and peer review. *Journal of Artificial Intelligence Research*, 70:1481–1515, 2021.

Konstantina Papagiannaki. Author feedback experiment at PAM 2007. *ACM SIGCOMM Computer Communication Review*, 37(3):73—78, 2007.

Dražen Prelec. A bayesian truth serum for subjective data. *Science*, 306(5695):462–466, 2004.

Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1), 2022.

Cristina Rueda. Degrees of freedom and model selection in semiparametric additive monotone regression. *Journal of Multivariate Analysis*, 117:88–99, 2013.

Nihar B. Shah and Dengyong Zhou. Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. *Journal of Machine Learning Research*, 17(165):1–52, 2016.

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.

Nihar B. Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. Design and analysis of the NIPS 2016 review process. *Journal of Machine Learning Research*, 19(1):1913—1946, 2018.

Nihar Bhadresh Shah. *Learning from people*. PhD thesis, UC Berkeley, 2017.

Keith Stanovich. *Who Is Rational? Studies of Individual Differences in Reasoning*. Lawrence Erlbaum Associates, 1999.

Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. On testing for biases in peer review. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. PeerReview4All: Fair and accurate reviewer assignment in peer review. *Journal of Machine Learning Research*, 22(163):1–66, 2021.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

Fritz Strack and Thomas Mussweiler. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, 73:437–446, 1997.

Ryan J. Tibshirani, Holger Hoefling, and Robert Tibshirani. Nearly-isotonic regression. *Technometrics*, 53(1):54–61, 2011.

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.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

A.W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Jingyan Wang and Nihar B. Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2019.

Jingyan Wang, Ivan Stelmakh, Yuting Wei, and Nihar B. Shah. Debiasing evaluations that are biased by evaluations. In *AAAI Conference on Artificial Intelligence*, 2021.

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.

Yuting Wei, Martin J. Wainwright, and Adityanand Guntuboyina. The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. *Annals of Statistics*, 47(2):994–1024, 2019.

Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, June 2004.

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.

Kyusang Yu, Enno Mammen, and Byeong U. Park. Semi-parametric regression: Efficiency gains from modeling the nonparametric part. *Bernoulli*, 17(2):736–748, 2011.

Cun-Hui Zhang. Risk bounds in isotonic regression. *Annals of Statistics*, 30(2):528–555, 2002.

Appendix A. Comparison with other estimators

In this section, we present auxiliary theoretical results on comparing our estimator with the mean estimator (Appendix A.1) and a reweighted mean estimator that we introduce (Appendix A.2).

A.1 Comparison with the mean estimator

Recall from Section 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 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 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 Appendix C.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 (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 12. *Suppose the assumptions (A1), (A2) and (A3) hold and there is no noise, or equivalently $\eta = 0$ in (A1). Consider any $x^* \in \mathbb{R}^d$. 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 6) and estimating the qualities (Theorem 5). In contrast, the negative results in Proposition 11 and Corollary 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.

A.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

$$[\widehat{x}_{\mathrm{rw}}]_i = \sum_{k\in R} \frac{\ell_{k,\min}}{\sum_{k'\in R} \ell_{k',\min}} \sum_{j:(i,j)\in Gk} \frac{y_{ij}}{\ell_{ik}}. \tag{4}$$

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) 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

$$\frac{1}{n} \sum_{i\in[d],j\in[n]} \mathbb{E}[y_{ij}] = \frac{1}{n} \sum_{i\in[d],j\in[n]} \mathbb{E}[x^*_i + b_{ij} + z_{ij}] = \sum_{i\in[d]} x^*_i. \tag{5}$$

Hence, we shift $\widehat{x}_{\mathrm{rw}}$ by a constant such that the empirical version of (5) holds, that is, $\sum_{i\in d}[\widehat{x}_{\mathrm{rw}}]_i = \frac{1}{n}\sum_{i\in[d],j\in[n]} y_{ij}$.

$$\widehat{x}_{\mathrm{rw}} \leftarrow \widehat{x}_{\mathrm{rw}} + \left(-\frac{1}{d} \sum_{i\in[d]} [\widehat{x}_{\mathrm{rw}}]_i + \frac{1}{dn} \sum_{i\in[d],j\in[n]} y_{ij} \right) \mathbf{1} \tag{6}$$

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 (22b) in Lemma 17, 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 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 Appendix C.8. Note that the risk of our estimator is at most half of the error of reweighted mean, if ignoring the higher-order term $O\left(\frac{1}{n^2}\right)$.

Appendix B. Additional experimental details

In this section, we provide additional details for the experiments in Section 5.

B.1 Implementation

We now discuss the implementation of our estimator.

Solving the optimization (Line 10 in Algorithm 1): We describe the implementation of solving the optimization (2) depending on the value of λ.

- $\boldsymbol{\lambda = \infty}$**:** The estimator is computed as taking the mean of each course according to Proposition 7.
- $\boldsymbol{\lambda \in (0, \infty)}$**:** In the proof of Proposition 14 we show that the objective 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 (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 (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 1): 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 1): We sample 100 total orderings to approximate the interpolation.

B.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 5.4. Note that the reweighted mean estimator defined in Appendix A.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.

B.3 Extending our estimator and the reweighted mean estimator to an unequal number of students per course

In the semi-synthetic experiment in Section 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 Appendix A.2, the reweighting step is the same (only using the observations in Ω). The recentering step restricted to Ω is defined as:

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

Similar to Appendix A.2, after this recentering step, the reweighted mean estimator satisfies the empirical version of an equality (Eq. (21b) in Appendix C.2.1) that our estimator also satisfies.

Our estimator: We extend Algorithm 1 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.

Appendix C. Proofs

In this section, we provide proofs for all the theoretical claims made earlier. We begin by introducing some additional notation in Section C.1 which is used throughout the proofs. In Section C.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.

C.1 Notation

Training-validation split $(\Omega^{\mathrm{t}}, \Omega^{\mathrm{v}})$**:** By Algorithm 1, 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

$$n^{\mathrm{t}} = n^{\mathrm{v}} = \frac{n}{2}. \tag{7}$$

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:

$$\Omega^{\mathrm{t}}_i := \{(i,j) \in \Omega^{\mathrm{t}}\}$$
$$\Omega^{\mathrm{v}}_i := \{(i,j) \in \Omega^{\mathrm{v}}\}.$$

We slightly abuse the notation and say $j \in \Omega^{\mathrm{t}}_i$ if $(i,j) \in \Omega^{\mathrm{t}}_i$. Likewise for Ω^{v}_i.

Group orderings: Recall that from Definition 1 that G_k denotes the set of elements in group $k \in [r]$. We define

$$G^{\mathrm{t}}_k := G_k \cap \Omega^{\mathrm{t}}$$
$$G^{\mathrm{v}}_k := G_k \cap \Omega^{\mathrm{v}}.$$

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

$$G^{\mathrm{t}}_{ik} := G^{\mathrm{t}}_k \cap \Omega^{\mathrm{t}}_i \qquad G^{\mathrm{v}}_{ik} := G^{\mathrm{v}}_k \cap \Omega^{\mathrm{v}}_i.$$

Again, we slightly abuse the notation and say $j \in G^{\mathrm{v}}_{ik}$ if $(i,j) \in G^{\mathrm{v}}_{ik}$.

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,

$$\ell_{ik} := |G_{ik}| \tag{8a}$$
$$\ell_k := |G_k| = \sum_{i \in [d]} \ell_{ik} \tag{8b}$$
$$\ell_{-i,k} := |G_k \setminus G_{ik}| = \sum_{i' \neq i} \ell_{i'k}. \tag{8c}$$

Furthermore, we define

$$\ell^{\mathrm{t}}_k := \left|G^{\mathrm{t}}_k\right| \qquad \ell^{\mathrm{v}}_k := |G^{\mathrm{v}}_k|, \tag{9a}$$
$$\ell^{\mathrm{t}}_{ik} := \left|G^{\mathrm{t}}_{ik}\right| \qquad \ell^{\mathrm{v}}_{ik} := |G^{\mathrm{v}}_{ik}|. \tag{9b}$$

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

$$\overline{y} = \frac{1}{dn} \sum_{i \in [d], j \in [n]} y_{ij}. \tag{10}$$

Denote the mean of the observations in any course $i \in [d]$ as

$$\overline{y}_i = \frac{1}{n} \sum_{j=1}^{n} y_{ij}. \tag{11}$$

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

$$y_{ik,\max} := \max_{j:(i,j)\in G_k} y_{ij} \tag{12a}$$

$$y_{ik,\min} := \min_{j:(i,j)\in G_k} y_{ij} \tag{12b}$$

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

$$\begin{aligned} b_{k,\min} &= \min_{(i,j)\in G_k} b_{ij} \\ b_{k,\max} &= \max_{(i,j)\in G_k} b_{ij}. \end{aligned} \tag{13}$$

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

$$\Pi_{\mathcal{M}}(x) := \arg\min_{u\in\mathcal{M}_\pi} \|x-u\|_2^2. \tag{14}$$

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 1 that our estimator restricted to any set of elements $\Omega \subseteq [d] \times [n]$ is defined as the solution to:

$$\arg\min_{x\in\mathbb{R}^d} \min_{\substack{B\in\mathbb{R}^{d\times n} \\ B \text{ satisfies } \mathcal{O}}} \left\|Y - x\mathbf{1}^T - B\right\|_\Omega^2 + \lambda\|B\|_\Omega^2, \tag{15}$$

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 (15) 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 1.

Recall from Line 13 in Algorithm 1 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 1 that $e^{(\lambda)}$ denotes the CV error at λ.

Define the random variable Λ_ϵ as the set

$$\Lambda_\epsilon := \{\lambda \in [0,\infty] : \|\widehat{x}^{(\lambda)}\|_2 > \epsilon\}. \tag{16}$$

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 5(a), we consider any fixed $\epsilon > 0$. Then the notation

$$\lim_{n\to\infty} \mathbb{P}\Big(\|\widehat{x}^{(0)} - x^*\|_2 < \epsilon\Big) = 1 \tag{17}$$

is considered equivalent to the original statement of Theorem 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 (17) 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 (17) is well-defined.

C.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.

C.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 (15) under its tie-breaking rule for any $\lambda \in [0, \infty)$. That is, the estimator is well-defined on $\lambda \in [0, \infty)$.

Proposition 14 (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* (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 C.9.1. Recall that the solution to (15) at $\lambda = \infty$ is defined by taking the limit of $\lambda \to \infty$ as:

$$\widehat{x}^{(\infty)} := \lim_{\lambda\to\infty} \widehat{x}^{(\lambda)} \tag{18a}$$

$$\widehat{B}^{(\infty)} := \lim_{\lambda\to\infty} \widehat{B}^{(\lambda)}. \tag{18b}$$

The following proposition shows the existence of the solution (18). That is, the limit in (18) is well-defined. This proposition is a generalization of Proposition 7 to any set $\Omega \subseteq [d] \times [n]$, and its proof is a straightforward generalization of the proof of Proposition 7 (Appendix C.4).

Proposition 15 (Existence of the estimator at $\lambda = \infty$)**.** *For any $\Omega \subseteq [d] \times [n]$, the solution $(\widehat{x}^{(\infty)}, \widehat{B}^{(\infty)})$ defined in* (18) *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 16. *For any $\lambda \in [0, \infty]$, and any $\Omega \subseteq [d] \times [n]$, the solution $(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$ to the estimator* (15) *satisfies*

$$\widehat{x}_i^{(\lambda)} = \frac{1}{n_i} \sum_{j \in \Omega_i} \left(y_{ij} - \widehat{b}_{ij}^{(\lambda)} \right) \qquad \forall i \in [d]. \tag{19}$$

In particular, in the special case of $\Omega = [d] \times [n]$, we have

$$\widehat{x}_i^{(\lambda)} = \frac{1}{n} \sum_{j \in [n]} \left(y_{ij} - \widehat{b}_{ij}^{(\lambda)} \right) \qquad \forall i \in [d]. \tag{20}$$

The proof of this result is provided in Appendix C.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 17. *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*

$$\sum_{(i,j) \in \Omega} \widehat{b}_{ij}^{(\lambda)} = 0 \tag{21a}$$
$$\sum_{i \in [d]} n_i \widehat{x}_i^{(\lambda)} = \sum_{(i,j) \in \Omega} y_{ij}. \tag{21b}$$

In particular, in the special case of $\Omega = [d] \times [n]$, we have

$$\sum_{i \in [d], j \in [d]} \widehat{b}_{ij}^{(\lambda)} = 0 \tag{22a}$$
$$n \sum_{i \in [d]} \widehat{x}_i^{(\lambda)} = \sum_{i \in [d], j \in [n]} y_{ij}. \tag{22b}$$

The proof of this result is provided in Appendix C.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 18 (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 C.9.4. Note that the observation model (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 18 implies the following corollary.

Corollary 19. *Under the observation model* (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 19, it can be further verified that the cross-validation algorithm (Algorithm 1) 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 20 and Lemma 21) 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 20 is used in Theorem 9 and Theorem 10; Lemma 21 is used in Theorem 5. Recall the notation $\Lambda_\epsilon := \{\lambda \in [0, \infty] : \|\widehat{x}^{(\lambda)}\|_2 > \epsilon\}$.

Lemma 20. *Suppose $x^* = 0$. Consider random Ω^{t} obtained by Algorithm 1. 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

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{i, i' \in [d]} \left(\widehat{x}_i^{(\lambda)} - \widehat{x}_{i'}^{(\lambda)} \right) > \frac{\epsilon}{\sqrt{d}}, \quad \forall \lambda \in \Lambda_\epsilon \right) = 1,$$

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 C.9.5.

Lemma 21. *Suppose $x^* = 0$. Suppose the observations follow part (a) of Lemma 20. Suppose the estimator is restricted to the set of either*

(a) $\Omega = [d] \times [n]$, or

(b) random Ω^{t} obtained by Algorithm 1.

Fix any $\lambda \in [0, \infty]$ and any $\epsilon > 0$. Suppose we have

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{i, i' \in [n]} \left| \widehat{x}_i^{(\lambda)} - \widehat{x}_{i'}^{(\lambda)} \right| < \epsilon \right) = 1. \tag{23}$$

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 C.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 5(b) and Proposition 13. Recall the definitions of $\overline{y}, \overline{y}_i$, $y_{ik,\min}$ and $y_{ik,\max}$ from (10), (11) and (12).

Proposition 22. *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 (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*

$$\gamma = \begin{cases} y_{22,\min} - y_{11,\max} & \text{if } y_{22,\min} - y_{11,\max} < \overline{y}_2 - \overline{y}_1 \\ y_{21,\max} - y_{12,\min} & \text{if } y_{21,\max} - y_{12,\min} > \overline{y}_2 - \overline{y}_1 \\ \overline{y}_2 - \overline{y}_1 & \text{o.w.} \end{cases} \tag{24}$$

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 (24) is ignored.

The proof of this result is provided in Appendix C.9.7

C.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

$$X^{(1)} \le \ldots \le X^{(n)}.$$

Define the maximal spacing as

$$M_n := \max_{1\le i\le n-1} (X^{(i+1)} - X^{(i)}). \tag{25}$$

The following standard result from statistics states that the maximum difference between adjacent order statistics converges to 0 for the Gaussian distribution.

Lemma 23. *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 C.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 24. *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 C.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 25. *Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(0, \sigma^2)$. Then we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i\in[n]} X_i < 2\sigma\sqrt{\log n}\right) = 1$$
$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i\in[n]} X_i - \min_{i\in[n]} X_i < 4\sigma\sqrt{\log n}\right) = 1.$$

C.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 (8) and (9).

Lemma 26. *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 1. Then we have the deterministic relations*

$$\frac{\ell_{ik}}{4} \leq \ell^{\mathrm{v}}_{ik} \leq \frac{3\ell_{ik}}{4} \qquad \forall i \in [d], k \in [r] \tag{26a}$$
$$\frac{\ell_{ik}}{4} \leq \ell^{\mathrm{t}}_{ik} \leq \frac{3\ell_{ik}}{4} \qquad \forall i \in [d], k \in [r] \tag{26b}$$

and

$$\frac{\ell_k}{4} \leq \ell^{\mathrm{v}}_k \leq \frac{3\ell_k}{4} \qquad \forall k \in [r] \tag{27a}$$
$$\frac{\ell_k}{4} \leq \ell^{\mathrm{t}}_k \leq \frac{3\ell_k}{4} \qquad \forall k \in [r]. \tag{27b}$$

The proof of this result is provided in Appendix C.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 27. *For any partition $(\Omega^{\mathrm{t}}, \Omega^{\mathrm{v}})$ obtained by Algorithm 1, for any $1 \le k_1 < k_2 \le dn$, 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

$$k_2 - k_1 \le 2d + 1.$$

The proof of this result is provided in Appendix C.9.11. The following lemma bounds the mean of the bias terms using standard concentration inequalities.

Lemma 28. *Consider any partial ordering $\mathcal{O}$ and any random Ω^{t} obtained by Algorithm 1. Suppose that the bias is marginally distributed as $\mathcal{N}(0,1)$ following assumption (A2). For any $\epsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \frac{1}{n^{\mathrm{t}}} \sum_{j \in \Omega^{\mathrm{t}}_i} b_{ij} - \frac{1}{n} \sum_{j \in [n]} b_{ij} \right| < \epsilon \right) = 1 \qquad \forall i \in [d], \tag{28a}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \frac{1}{|\Omega^{\mathrm{t}}|} \sum_{(i,j) \in \Omega^{\mathrm{t}}} b_{ij} \right| < \epsilon \right) = 1, \tag{28b}$$

where the probabilities are over the randomness in B and in Ω^{t}.

The proof of this result is provided in Appendix C.9.12.

C.3 Proof of Theorem 5

The proof follows notation in Appendix C.1 and preliminaries in Appendix C.2. By Corollary 19, 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 (1) reduces to

$$Y = B. \tag{29}$$

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

$$S_c := \left\{ (i, i') \in [d]^2 : \exists k \in [r] \text{ such that } \frac{\ell_{ik}}{\ell_k}, \frac{\ell_{i',k+1}}{\ell_{k+1}} \ge c \right\}. \tag{30}$$

In words, the definition (30) 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 29. *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 $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\Big(\widehat{x}^{(0)}_{i'} - \widehat{x}^{(0)}_{i} < \epsilon\Big) = 1. \tag{31}$$

The proof of this result is provided in Appendix C.10.1. To state the next lemma, we first make the following definition of a "cycle" of courses.

Definition 30. *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*

$$(i_m, i_{m+1}) \in S_c \qquad \forall m \in [L-1], \tag{32a}$$

$$\text{and } (i_L, i_1) \in S_c. \tag{32b}$$

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 31. *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 C.10.2. Now we prove the three parts of Theorem 5 respectively.

C.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 3. Hence, by the definition (30) 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 30. Applying Lemma 31 followed by Lemma 21(a) completes the proof.

C.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

$$\ell_{11} \geq \ell_{21}. \tag{33}$$

Since we assume there are only two courses and two groups, we have

$$\ell_{12} = n - \ell_{11} \le n - \ell_{21} = \ell_{22}. \tag{34}$$

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 21 to complete the proof.

Bounding the probability of $\widehat{x}_2 - \widehat{x}_1 < \epsilon$: By the definition (30) of S_c, it can be verified that given (33) and (34) we have $(1, 2) \in S_{0.5}$ (taking $k = 1$). By Lemma 29, we have

$$\lim_{n\to\infty} \mathbb{P}(\widehat{x}_2 - \widehat{x}_1 < \epsilon) = 1. \tag{35}$$

Bounding the probability of $\widehat{x}_2 - \widehat{x}_1 > -\epsilon$: By the closed-form solution in Proposition 22, we have $\widehat{x}_2 - \widehat{x}_1 = \gamma$ where γ is defined in (24) as

$$\gamma = \begin{cases} y_{22,\min} - y_{11,\max} & \text{if } y_{22,\min} - y_{11,\max} < \overline{y}_2 - \overline{y}_1 \\ y_{21,\max} - y_{12,\min} & \text{if } y_{21,\max} - y_{12,\min} > \overline{y}_2 - \overline{y}_1 \\ \overline{y}_2 - \overline{y}_1 & \text{o.w.} \end{cases} \tag{36}$$

Recall from the model (29) that $Y = B$, and hence we have the deterministic relation $y_{22,\min} - y_{11,\max} = b_{22,\min} - b_{11,\max} \ge 0$ due to the assumption (A2) under the group ordering, and similarly we have the deterministic relation $y_{21,\max} - y_{12,\min} \le 0$. Consider the case of $\overline{y}_2 - \overline{y}_1 \ge 0$. In this case, only the first and the third cases in (36) are possible, and therefore we have $0 \le \gamma \le \overline{y}_2 - \overline{y}_1$. Now consider the case of $\overline{y}_2 - \overline{y}_1 < 0$. In this case, only the second and the third cases in (36) are possible, and we have $\overline{y}_2 - \overline{y}_1 \le \gamma \le 0$. Combining the two cases, we have the relation

$$\widehat{x}_2 - \widehat{x}_1 = \gamma > -\epsilon \qquad \text{if } \overline{y}_2 - \overline{y}_1 > -\epsilon. \tag{37}$$

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 $\mathbb{P}(\overline{y}_2 - \overline{y}_1 > -\epsilon) \ge 1 - \delta$ 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 $b^{*(1)} \le \ldots \le b^{*(2n)}$ (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 (Hoeffding, 1963, Section 6) on group 1 of course 1, we have

$$\mathbb{P}\left[\left|\sum_{j\in G_{11}} b_{1j} - \ell_{11}\overline{b}^*_{G_1}\right| \ge \Delta_{B^*}\sqrt{\ell_{11}\log\left(\frac{1}{\delta}\right)} \,\middle|\, B^*\right] \le 2\exp\left(-\frac{2\cdot\Delta^2_{B^*}\ell\log(\frac{1}{\delta})}{\ell\Delta^2_B}\right) = 2\delta^2 \overset{\text{(i)}}{\le} \frac{\delta}{8},$$

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 $\ell_{ik} \le n$ for any $i \in \{1,2\}$ and any $k \in \{1,2\}$, we have

$$\mathbb{P}\left[\left|\sum_{j \in G_{ik}} b_{ij} - \ell_{ik}\overline{b}^*_{G_k}\right| \ge \Delta_{B^*}\sqrt{n\log\left(\frac{1}{\delta}\right)} \;\middle|\; B^*\right] \le \frac{\delta}{8}. \tag{38}$$

Taking a union bound of (38) over $i \in \{1,2\}$ and $k \in \{1,2\}$, we have that with probability at least $1 - \frac{\delta}{2}$,

$$\begin{aligned}
\overline{y}_2 - \overline{y}_1 &= \frac{1}{n}\left(\sum_{j \in G_{21}} b_{2j} + \sum_{j \in G_{22}} b_{2j} - \sum_{j \in G_{11}} b_{1j} - \sum_{j \in G_{12}} b_{1j}\right) \\
&\overset{\text{(i)}}{\ge} \frac{1}{n}\left(\ell_{21}\overline{b}^*_{G_1} + \ell_{22}\overline{b}^*_{G_2} - \ell_{11}\overline{b}^*_{G_1} - \ell_{12}\overline{b}^*_{G_2} - 4\Delta_{B^*}\sqrt{n\log\left(\frac{1}{\delta}\right)}\right) \\
&= \frac{1}{n}\left((\ell_{21} - \ell_{11})\overline{b}^*_{G_1} + (\ell_{22} - \ell_{12})\overline{b}^*_{G_2} - 4\Delta_{B^*}\sqrt{n\log\left(\frac{1}{\delta}\right)}\right) \\
&\overset{\text{(ii)}}{=} \frac{1}{n}\left((\ell_{21} - \ell_{11})(\overline{b}^*_{G_1} - \overline{b}^*_{G_2}) - 4\Delta_{B^*}\sqrt{n\log\left(\frac{1}{\delta}\right)}\right) \\
&\overset{\text{(iii)}}{\ge} -4\Delta_{B^*}\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}},
\end{aligned} \tag{39}$$

where inequality (i) is true by (38), 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 $\overline{b}^*_{G_1} \le \overline{b}^*_{G_2}$ due to the assumption (A2) of the bias and the group orderings.

Now we analyze the term Δ_B in (39). By Lemma 25, there exists integer n_0 such that for any $n \ge n_0$,

$$\mathbb{P}\left(\Delta_B \le 4\sqrt{\log 2n}\right) \ge 1 - \frac{\delta}{2}. \tag{40}$$

Let n_1 be a sufficiently large such that $n_1 \ge n_0$ and $16\sqrt{\log 2n_1} \cdot \sqrt{\frac{\log(\frac{1}{\delta})}{n_1}} < \epsilon$. Then combining (40) with (39), we have that for any $n \ge n_0$,

$$\begin{aligned}
\mathbb{P}\left(\overline{y}_2 - \overline{y}_1 > -\epsilon\right) &= \int_{B \in \mathbb{R}^{2\times n}} \mathbb{P}\left(\overline{y}_2 - \overline{y}_1 > -\epsilon \mid B\right) \cdot \mathbb{P}(B)\,\mathrm{d}B \\
&\ge \int_{\substack{B \in \mathbb{R}^{2\times n}: \\ \Delta_B \le 4\sqrt{\log n}}} \mathbb{P}(\overline{y}_2 - \overline{y}_1 > -\epsilon \mid B) \cdot \mathbb{P}(B)\,\mathrm{d}B \\
&\overset{\text{(i)}}{\ge} \left(1 - \frac{\delta}{2}\right) \cdot \mathbb{P}(\Delta_B \le 4\sqrt{\log 2n}) \\
&\overset{\text{(ii)}}{\ge} \left(1 - \frac{\delta}{2}\right)^2 \ge 1 - \delta,
\end{aligned} \tag{41}$$

where inequality (i) is true by (39) due to the choice of n_1, and inequality (ii) is true by (40). Combining (41) with (37), 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,

$$\lim_{n\to\infty} \mathbb{P}(\widehat{x}_2 - \widehat{x}_1 > -\epsilon) = 1. \tag{42}$$

Finally, combining Step 1 and Step 2, we take a union bound of (35) and (42), we have

$$\lim_{n\to\infty} \mathbb{P}\Big(|\widehat{x}_2 - \widehat{x}_1| < \epsilon\Big) = 1. \tag{43}$$

Given (43), we invoke Lemma 21 and obtain

$$\lim_{n\to\infty} \mathbb{P}\Big(\|\widehat{x}\|_2 < \epsilon\Big) = 1,$$

completing the proof.

C.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 30 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.

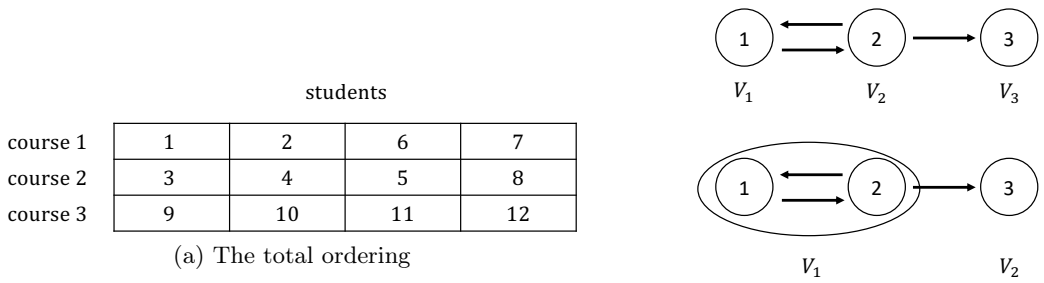
The proof consists of four steps:

- 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 21 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

$$\min_{j\in[n]} t_{ij} < \min_{j\in[n]} t_{i+1,j} \qquad \forall i \in [d-1]. \tag{44}$$

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.



(a) The total ordering

(b) The procedure of constructing the partition

Figure 7: An example for constructing the partition of hypernodes.

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".

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. 7. In this example we consider $d = 3$ courses and $n = 4$ students per course. We consider the total ordering in Fig. 7(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. 7(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. 7(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 32. *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 C.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 33. *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 C.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 33.

As an example, in Fig. 7 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

$$E_{\text{bad}} := \left\{ \max_{i, i' \in [n]} |\widehat{x}_{i'} - \widehat{x}_i| > 4d^2 \epsilon \right\}. \tag{45}$$

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 32, 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 31 we have that for each hypernode V,

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{i, i' \in V} |\widehat{x}_i - \widehat{x}_{i'}| < \epsilon \right) = 1. \tag{46}$$

Since the number of hypernodes is at most d, taking a union bound of (46) across all hypernodes in the partition, we have

$$\lim_{n \to \infty} \mathbb{P}\Bigg(\underbrace{\max_{i, i' \in V} |\widehat{x}_i - \widehat{x}_{i'}| < \epsilon, \quad \forall V \text{ hypernode in the partition}}_{E_1} \Bigg) = 1. \tag{47}$$

We denote this event in (47) 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 33, 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 29, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\widehat{x}_{i'} - \widehat{x}_i < \epsilon\right) = 1. \tag{48}$$

Since the number of hypernodes s is at most d, taking a union bound of (48) over all $m \in [s-1]$, we have

$$\lim_{n\to\infty} \mathbb{P}\Bigg(\underbrace{\min_{i\in V_m, i'\in V_{m+1}} \widehat{x}_{i'} - \widehat{x}_i < \epsilon, \quad \forall m \in [s-1]}_{E_2} \Bigg) = 1. \tag{49}$$

We denote this event in (49) 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:

$$\mathbb{P}(E_3) = 1. \tag{50}$$

Decompose E_{bad}: We decompose the bad event E_{bad} as

$$\begin{aligned}\mathbb{P}(E_{\text{bad}}) &= \mathbb{P}(E_{\text{bad}}, E_1 \cap E_2 \cap E_3) + \mathbb{P}(E_{\text{bad}}, \overline{E_1 \cap E_2 \cap E_3}) \\ &\leq \mathbb{P}(E_{\text{bad}}, E_1 \cap E_2 \cap E_3) + \mathbb{P}(\overline{E_1 \cap E_2 \cap E_3}).\end{aligned} \tag{51}$$

Combining (47), (49) and (50), we have

$$\lim_{n\to\infty} \mathbb{P}\Big(\overline{E_1 \cap E_2 \cap E_3}\Big) = \lim_{n\to\infty} \mathbb{P}(\overline{E_1} \cup \overline{E_2} \cup \overline{E_3}) \leq \lim_{n\to\infty} \big[\mathbb{P}(\overline{E_1}) + \mathbb{P}(\overline{E_2}) + \mathbb{P}(\overline{E_3})\big] = 0. \tag{52}$$

Combining (51) and (52), 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 $s \geq 2$. 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 (45), the event E_{bad} requires that there exists $i, i' \in [d]$ such that

$$|\widehat{x}_{i'} - \widehat{x}_i| > 4d^2\epsilon. \tag{53}$$

By the definition (47) 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

$$\begin{aligned}\widehat{x}_{i'} - \widehat{x}_i &= (\widehat{x}_{i'} - \widehat{x}_{i_{m'}}) + (\widehat{x}_{i_{m'}} - \widehat{x}_{i_{m'-1}}) + \ldots + (\widehat{x}_{i_{m+1}} - \widehat{x}_{i_m}) + (\widehat{x}_{i_m} - \widehat{x}_{i'}) \\ &\overset{\text{(i)}}{<} 2\epsilon + d\epsilon < 4d^2\epsilon, \end{aligned} \tag{54}$$

where (i) is true by events E_1 and E_2. Combining (53) and (54), we must have $\widehat{x}_{i'} - \widehat{x}_i < -4d^2\epsilon$, or equivalently

$$\widehat{x}_i - \widehat{x}_{i'} > 4d^2\epsilon. \tag{55}$$

We decompose $\widehat{x}_i - \widehat{x}_{i'}$ as

$$\begin{aligned}\widehat{x}_i - \widehat{x}_{i'} &= (\widehat{x}_i - \widehat{x}_{i_m}) + (\widehat{x}_{i_m} - \widehat{x}_{i_{m+1}}) + \ldots + (\widehat{x}_{i_{m'-1}} - \widehat{x}_{i_{m'}}) + (\widehat{x}_{i_{m'}} - \widehat{x}_{i'}) \\ &\overset{\text{(i)}}{<} 2\epsilon + (\widehat{x}_{i_m} - \widehat{x}_{i_{m+1}}) + \ldots + (\widehat{x}_{i_{m'-1}} - \widehat{x}_{i_{m'}}), \end{aligned} \tag{56}$$

where (i) is due to event E_1. Combining (55) and (56), 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

$$\begin{aligned} d \cdot \max\{(\widehat{x}_{i_m} - \widehat{x}_{i_{m+1}}), \ldots, (\widehat{x}_{i_{m'-1}} - \widehat{x}_{i_{m'}})\} > 3d^2\epsilon \\ \max\{(\widehat{x}_{i_m} - \widehat{x}_{i_{m+1}}), \ldots, (\widehat{x}_{i_{m'-1}} - \widehat{x}_{i_{m'}})\} > 3d\epsilon. \end{aligned} \tag{57}$$

Without loss of generality, we assume that in (57) we have integer m^* with $m \le m^* < m'$ such that

$$\widehat{x}_{i_{m^*}} - \widehat{x}_{i_{m^*+1}} > 3d\epsilon. \tag{58}$$

Now consider any $m, m' \in [s]$ such that $m \le m^* < m'$, 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 (58). Equivalently, denote $V_{\mathrm{L}} := V_1 \cup \ldots \cup V_{m^*}$ and $V_{\mathrm{H}} := V_{m^*+1} \cup \ldots \cup V_s$, we have

$$\widehat{x}_i - \widehat{x}_{i'} > \epsilon \qquad \forall i \in V_{\mathrm{L}},\ i' \in V_{\mathrm{H}}. \tag{59}$$

Step 4: Showing $\mathbb{P}(E_{\mathrm{bad}}, E_1 \cap E_2 \cap E_3) = 0$ by deriving a contradiction We consider any solution $(\widehat{x}, \widehat{B})$ of our estimator at $\lambda = 0$ conditional on $E_{\mathrm{bad}} \cap E_1 \cap E_2 \cap E_3$, and derive a contradiction. Hence, we have $\mathbb{P}(E_{\mathrm{bad}}, E_1 \cap E_2 \cap E_3) = 0$.

Analyzing properties of $\widehat{B}$ By Lemma 33, any bias term $\widehat{b}_{ij}$ for $i \in V_{\mathrm{L}}$ has a smaller rank than any bias term $\widehat{b}_{ij}$ for $i \in V_{\mathrm{H}}$. 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 $\widehat{b}_{\mathrm{L}}$ and $\widehat{b}_{\mathrm{H}}$ as

$$\widehat{b}_{\mathrm{L}} := \frac{1}{|V_{\mathrm{L}}| \cdot n} \sum_{i \in V_{\mathrm{L}}} \sum_{j \in [n]} \widehat{b}_{ij} \tag{60a}$$

$$\widehat{b}_{\mathrm{H}} := \frac{1}{|V_{\mathrm{H}}| \cdot n} \sum_{i \in V_{\mathrm{H}}} \sum_{j \in [n]} \widehat{b}_{ij}, \tag{60b}$$

We have the deterministic relation $\widehat{b}_{\mathrm{L}} \le \widehat{b}_{\mathrm{H}}$.

First consider the case of $\widehat{b}_{\mathrm{L}} = \widehat{b}_{\mathrm{H}}$. 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 (2) attained at $(\widehat{x}, \widehat{B})$ is strictly positive. Recall from the model (29) that $Y = B$. Hence, an objective (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 $\widehat{b}_{\mathrm{L}} < \widehat{b}_{\mathrm{H}}$. We have that either $\widehat{b}_{\mathrm{L}} < 0$ or $\widehat{b}_{\mathrm{H}} > 0$ (or both). Without loss of generality we assume $\widehat{b}_{\mathrm{H}} > 0$.

Constructing an alternative solution We now construct an alternative solution by increasing $\widehat{x}_i$ for every course $i \in V_{\mathrm{H}}$ 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

$$\begin{aligned} \widehat{x}'_i &= \begin{cases} \widehat{x}_i & \text{if } i \in V_{\mathrm{L}} \\ \widehat{x}_i + \Delta & \text{if } i \in V_{\mathrm{H}} \end{cases} \\ \widehat{B}' &= Y - \widehat{x}' \mathbf{1}^T, \end{aligned} \tag{61}$$

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 (61), 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

$$y_{ij} = \widehat{x}_i + \widehat{b}_{ij} = \widehat{x}'_i + \widehat{b}'_{ij} \qquad \forall i \in [d], j \in [n]. \tag{62}$$

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 $\widehat{b}_{ij} \le \widehat{b}_{i'j'}$. 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 33 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 $E_{\mathrm{bad}} \cap E_1 \cap E_2 \cap E_3$,

$$\begin{aligned}
\widehat{b}'_{ij} - \widehat{b}'_{i'j'} &= (y_{ij} - \widehat{x}'_i) - (y_{i'j'} - \widehat{x}'_{i'}) \\
&= (b_{ij} - \widehat{x}_i) - (b_{i'j'} - \widehat{x}_{i'} - \Delta) \\
&= (b_{ij} - b_{i'j'}) + (\widehat{x}_{i'} + \Delta - \widehat{x}_i) \overset{\text{(i)}}{<} 0,
\end{aligned}$$

where (i) is true because the ordering constraint $((i,j),(i',j'))$ gives $b_{ij} \le b_{i'j'}$. Moreover, we have $\widehat{x}_{i'} - \widehat{x}_i < -\epsilon$ due to (59). 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 (61), we compute $\|\widehat{B}'\|_F^2$ as

$$\begin{aligned}
\|\widehat{B}'\|_F^2 &= \sum_{i \in V_{\mathrm{L}}} \sum_{j \in [n]} (y_{ij} - \widehat{x}_i)^2 + \sum_{i \in V_{\mathrm{H}}} \sum_{j \in [n]} (y_{ij} - \widehat{x}_i - \Delta)^2 \\
&\overset{\text{(i)}}{=} \sum_{i \in V_{\mathrm{L}}} \sum_{j \in [n]} (\widehat{b}_{ij})^2 + \sum_{i \in V_{\mathrm{H}}} \sum_{j \in [n]} (\widehat{b}_{ij} - \Delta)^2,
\end{aligned} \tag{63}$$

where (i) is true by (62). Taking the partial derivative of (63) with respect to Δ, we have

$$\frac{\partial \|\widehat{B}'\|_F^2}{\partial \Delta} = 2 \left(|V_{\mathrm{H}}| \cdot n\Delta - \sum_{i \in V_{\mathrm{H}}} \sum_{j \in [n]} \widehat{b}_{ij} \right) = 2|V_{\mathrm{H}}| \cdot n(\Delta - \widehat{b}_{\mathrm{H}}). \tag{64}$$

By the assumption of $\widehat{b}_{\mathrm{H}} > 0$, the partial derivative (64) 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

$$\mathbb{P}(E_{\mathrm{bad}}, E_1 \cap E_2 \cap E_3) = 0.$$

Step 4: Invoking Lemma 21 Recall from Step 2 that $\lim_{n \to \infty} \mathbb{P}(E_{\mathrm{bad}}, E_1 \cap E_2 \cap E_3) = 0$ implies $\lim_{n \to \infty} \mathbb{P}(E_{\mathrm{bad}}) = 0$. 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 21 completes the proof.

C.4 Proof of Proposition 7

We denote $(\widehat{x}^{(\infty)}, B^{(\infty)})$ as the values given by expression (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 (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 (2). By Proposition 14, 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 (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 (20) in Lemma 16 (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.

C.5 Proof of Theorem 9

The proof follows notation in Appendix C.1 and preliminaries in Appendix C.2. By Corollary 19, 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 (1) reduces to:

$$Y = B. \tag{65}$$

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 $\Omega^{\mathrm{t}} \subseteq [d] \times [n]$ obtained by Algorithm 1. Note that Theorem 5(a) and Theorem 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 5(a) and Theorem 5(c) to any Ω^{t} given by Algorithm 1. The following theorem states that part (a) and part (c) of Theorem 5 still hold for the estimator (15) restricted to Ω^{t}. We use $(\widehat{x}^{(0)}, \widehat{B}^{(0)})$ to denote the solution to (15) restricted to Ω^{t} for the remaining of the proof of Theorem 9.

Theorem 34 (Generalization of Theorem 5 to any Ω^{t})**.** *Consider any fixed $\Omega^{\mathrm{t}} \subseteq [d] \times [n]$ obtained by Algorithm 1. 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 (15) *restricted to Ω^{t}) satisfies*

$$\mathbb{P}\Big(\|\widehat{x}^{(0)} - x^*\|_2 < \epsilon \Big) \geq 1 - \delta. \tag{66}$$

Equivalently, for any $\epsilon > 0$, we have

$$\lim_{n \to \infty} \mathbb{P}\Big(\|\widehat{x}^{(0)} - x^*\|_2 < \epsilon \Big) = 1. \tag{67}$$

The proof of this theorem is in Appendix C.11.1. Now we consider the consistency of the bias term $\widehat{B}$. Given the model (65), the objective (15) at $\lambda = 0$ equals 0 at the values of $(\widehat{x}, \widehat{B}) = (0, B)$. Hence, objective (15) attains a value of 0 at the solution $(\widehat{x}^{(0)}, \widehat{B}^{(0)})$. Therefore, we have the deterministic relation $Y_{\Omega^{\mathrm{t}}} = [\widehat{x}^{(0)} \mathbf{1}^T + \widehat{B}^{(0)}]_{\Omega^{\mathrm{t}}}$. For any $(i, j) \in \Omega^{\mathrm{t}}$, we have

$$\widehat{b}_{ij}^{(0)} = Y_{ij} - \widehat{x}_i^{(0)} \overset{\text{(i)}}{=} b_{ij} - \widehat{x}_i^{(0)}, \tag{68}$$

where equality (i) is true because of the model (65). Combining (68) with (67), we have that for any $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\Big(\Big| \widehat{b}_{ij}^{(0)} - b_{ij} \Big| < \epsilon, \quad \forall (i, j) \in \Omega^{\mathrm{t}} \Big) = 1. \tag{69}$$

This completes Step 1 of the proof. The remaining two steps are presented separately for the two parts.

C.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 1 and Definition 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

$$\widetilde{b}_{ij} = \widetilde{b}_{i'j'}, \text{ for any } (i,j), (i',j') \in G_k^{\mathrm{v}}. \tag{70}$$

Denote $\widetilde{b}_k := \widetilde{b}_{ij}$ for any $(i,j) \in G_k^{\mathrm{t}}$. By (70), 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

$$b_k^{\mathrm{t}} := \frac{1}{|G_k^{\mathrm{t}}|} \sum_{(i,j) \in G_k^{\mathrm{t}}} b_{ij} \tag{71}$$

$$b_k^{\mathrm{v}} := \frac{1}{|G_k^{\mathrm{v}}|} \sum_{(i,j) \in G_k^{\mathrm{v}}} b_{ij} \tag{72}$$

$$b_{ik}^{\mathrm{v}} := \frac{1}{|G_{ik}^{\mathrm{v}}|} \sum_{j \in G_{ik}^{\mathrm{v}}} b_{ij}. \tag{73}$$

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 (65), the validation error at $\lambda = 0$ is computed as:

$$\begin{aligned} e^{(0)} &= \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j) \in \Omega^{\mathrm{v}}} \left(y_{ij} - \widehat{x}_i^{(0)} - \widetilde{b}_{ij} \right)^2 \\ &= \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{i \in [d], k \in [r]} \sum_{j \in G_{ik}^{\mathrm{v}}} \left(b_{ij} - \widehat{x}_i^{(0)} - \widetilde{b}_k \right)^2. \end{aligned} \tag{74}$$

We first analyze the term $\widetilde{b}_k$ in (74). The following lemma shows that the interpolation procedure in Algorithm 1 ensures that $\widetilde{b}_k$ is close to $\widehat{b}_k^{\mathrm{t}}$, the mean of the estimated bias over G_k^{t}.

Lemma 35. *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 1. Then for any $\lambda \in [0, \infty]$ we have the deterministic relation:*

$$\left| \widetilde{b}_k - \widehat{b}_k^{\mathrm{t}} \right| \le \frac{12}{c_{\mathrm{f}} dn} \cdot \max_{(i,j) \in \Omega^{\mathrm{t}}} \left| \widehat{b}_{ij} \right| \qquad \forall k \in [r].$$

The proof of this result is provided in Appendix C.11.2. Combining Lemma 35 with the consistency (69) of $\widehat{B}^{(0)}$ from Step 1 and a bound on $\max_{(i,j) \in \Omega^{\mathrm{t}}} |b_{ij}|$ from Lemma 25, we have the following lemma.

Lemma 36. *Under the same condition as Lemma 35, 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 C.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 37. *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 1. For any $\epsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|b^{\mathrm{v}}_{ik} - \overline{b}_{G_k}\Big| < \epsilon, \quad \forall i \in [d], k \in [r]\Big) = 1 \tag{75a}$$

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|b^{\mathrm{t}}_{k} - \overline{b}_{G_k}\Big| < \epsilon, \quad \forall k \in [r]\Big) = 1. \tag{75b}$$

The proof of this result is provided in Appendix C.11.4. Combining Lemma 36 and (75) from Lemma 37 with a union bound, we have the following corollary.

Corollary 38. *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 1. 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 (74) 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 (73) 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 (67) of $\widehat{x}^{(0)}$ in Theorem 34(a), we have

$$\lim_{n\to\infty} \mathbb{P}\left(T_2 \le \frac{|G^{\mathrm{v}}_{ik}|}{|\Omega^{\mathrm{v}}|}\epsilon_1^2, \quad \forall i \in [d], k \in [r]\right) = 1. \tag{76}$$

Term T_3: We have

$$T_3 \le 2\frac{|G_{ik}^{\mathrm{v}}|}{|\Omega^{\mathrm{v}}|}\cdot\left|b_{ik}^{\mathrm{v}}-\widetilde{b}_k\right|\cdot|\widehat{x}_i| \le 2\left|b_{ik}^{\mathrm{v}}-\widetilde{b}_k\right|\cdot|\widehat{x}_i|.$$

By combining the convergence (67) of $\widehat{x}^{(0)}$ in Theorem 34(a) and Corollary 38 with a union bound, we have

$$\lim_{n\to\infty}\mathbb{P}\Big(T_3 \le \frac{2|G_{ik}^{\mathrm{v}}|}{|\Omega^{\mathrm{v}}|}\epsilon_1^2, \quad \forall i\in[d], k\in[r]\Big) = 1. \tag{77}$$

Term T_1: We have

$$\begin{aligned}
T_1 = \frac{1}{|\Omega^{\mathrm{v}}|}\sum_{j\in G_{ik}^{\mathrm{v}}}\left(b_{ij}-\widetilde{b}_k\right)^2 &= \frac{1}{|\Omega^{\mathrm{v}}|}\sum_{j\in G_{ik}^{\mathrm{v}}}\left(b_{ij}-b_{ik}^{\mathrm{v}}+b_{ik}^{\mathrm{v}}-\widetilde{b}_k\right)^2 \\
&= \frac{1}{|\Omega^{\mathrm{v}}|}\left[\sum_{j\in G_{ik}^{\mathrm{v}}}(b_{ij}-b_{ik}^{\mathrm{v}})^2 + |G_{ik}^{\mathrm{v}}|\cdot(b_{ik}^{\mathrm{v}}-\widetilde{b}_k)^2 + 2\sum_{j\in G_{ik}^{\mathrm{v}}}(b_{ij}-b_{ik}^{\mathrm{v}})(b_{ik}^{\mathrm{v}}-\widetilde{b}_k)\right] \\
&\overset{\text{(i)}}{=} \frac{1}{|\Omega^{\mathrm{v}}|}\left[\sum_{j\in G_{ik}^{\mathrm{v}}}(b_{ij}-b_{ik}^{\mathrm{v}})^2 + |G_{ik}^{\mathrm{v}}|\cdot(b_{ik}^{\mathrm{v}}-\widetilde{b}_k)^2\right]
\end{aligned}$$

where inequality (i) holds because $\sum_{j\in G_{ik}^{\mathrm{v}}}(b_{ij}-b_{ik}^{\mathrm{v}})=0$ by the definition (73) of b_{ik}^{v}. By Corollary 38, we have

$$\lim_{n\to\infty}\left(T_1 < \frac{1}{|\Omega^{\mathrm{v}}|}\sum_{j\in G_{ik}^{\mathrm{v}}}(b_{ij}-b_{ik}^{\mathrm{v}})^2 + \frac{|G_{ik}^{\mathrm{v}}|}{|\Omega^{\mathrm{v}}|}\epsilon_1^2, \quad \forall i\in[d], k\in[r]\right) = 1. \tag{78}$$

Combining the three terms from (76), (77) and (78), we bound $e_{ik}^{(0)}$ as

$$\lim_{n\to\infty}\left(e_{ik}^{(0)} = T_1+T_2+T_3 < \frac{1}{|\Omega^{\mathrm{v}}|}\sum_{j\in G_{ik}^{\mathrm{v}}}(b_{ij}-b_{ik}^{\mathrm{v}})^2 + \frac{4|G_{ik}^{\mathrm{v}}|}{|\Omega^{\mathrm{v}}|}\epsilon_1^2, \quad \forall i\in[d], k\in[r]\right) = 1. \tag{79}$$

By the all c_{f}-fraction assumption, the number of groups is upper-bounded by a constant as $r \le \frac{1}{c_{\mathrm{f}}}$. Taking a union bound of (79) over $i\in[d]$ and $k\in[r]$, we have

$$\begin{aligned}
\lim_{n\to\infty}\mathbb{P}\left(e^{(0)} = \sum_{i\in[d],k\in[r]} e_{ik}^{(0)} < \frac{1}{|\Omega^{\mathrm{v}}|}\sum_{i\in[d],k\in[r]}\left[\sum_{j\in G_{ik}^{\mathrm{v}}}(b_{ij}-b_{ik}^{\mathrm{v}})^2 + 4|G_{ik}^{\mathrm{v}}|\cdot\epsilon_1^2\right]\right) = 1 \\
\lim_{n\to\infty}\mathbb{P}\left(e^{(0)} < \frac{1}{|\Omega^{\mathrm{v}}|}\sum_{i\in[d],k\in[r]}\sum_{j\in G_{ik}^{\mathrm{v}}}(b_{ij}-b_{ik}^{\mathrm{v}})^2 + 4\epsilon_1^2\right) = 1.
\end{aligned} \tag{80}$$

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 (16) the definition of the random set $\Lambda_\epsilon := \{\lambda \in [0, \infty] : \|\widehat{x}^{(\lambda)}\|_2 > \epsilon\}$. In this step, we show that

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} > e^{(0)}, \quad \forall \lambda \in \Lambda_\epsilon\right) = 1. \tag{81}$$

From (81), 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 (81).

Proof of (81) 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:

$$\begin{aligned} e_{ik}^{(\lambda)} = \frac{1}{|\Omega^{\text{v}}|} \sum_{j \in G_{ik}^{\text{v}}} \left(b_{ij} - \widehat{x}_i^{(\lambda)} - \widetilde{b}_k^{(\lambda)}\right)^2 &= \frac{1}{|\Omega^{\text{v}}|} \sum_{j \in G_{ik}^{\text{v}}} \left(b_{ij} - b_{ik}^{\text{v}} + b_{ik}^{\text{v}} - \widehat{x}_i - \widetilde{b}_k\right)^2 \\ &\overset{\text{(i)}}{=} \frac{1}{|\Omega^{\text{v}}|} \sum_{j \in G_{ik}^{\text{v}}} (b_{ij} - b_{ik}^{\text{v}})^2 + \underbrace{\frac{|G_{ik}^{\text{v}}|}{|\Omega^{\text{v}}|} \left(b_{ik}^{\text{v}} - \widehat{x}_i - \widetilde{b}_k\right)^2}_{T_{ik}}, \end{aligned} \tag{82}$$

where (i) is true because $\sum_{j \in G_{ik}^{\text{v}}} (b_{ij} - b_{ik}^{\text{v}}) = 0$ by the definition (73) of b_{ik}^{v}. Note that the first term in (82) is identical to the first term in (79) from Step 2. We now analyze the second term T_{ik} in (82). On the one hand, by Lemma 20(a), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i,i' \in [d]} \widehat{x}_i - \widehat{x}_{i'} > \frac{\epsilon}{\sqrt{d}}, \quad \forall \lambda \in \Lambda_\epsilon\right) = 1. \tag{83}$$

On the other hand, taking a union bound of (75a) in Lemma 37 over $i, i' \in [d]$, we have

$$\lim_{n\to\infty} \mathbb{P}\Big(|b_{ik}^{\text{v}} - b_{i'k}^{\text{v}}| < \frac{\epsilon}{2\sqrt{d}}, \quad \forall i, i' \in [d], k \in [r]\Big) = 1. \tag{84}$$

Conditional on (83) and (84), 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 (83) and (84),

$$\max_{i,i' \in [d]} \left\{ (b^{\mathrm{v}}_{ik} - \widehat{x}_i - \widetilde{b}_k)^2, (b^{\mathrm{v}}_{i'k} - \widehat{x}_{i'} - \widetilde{b}_k)^2 \right\} \geq \frac{\epsilon^2}{16d} \qquad \forall k \in [r], \forall \lambda \in \Lambda_\epsilon. \tag{85}$$

Now consider the terms T_{ik}. By (26a) from Lemma 26 combined with the all c_{f}-fraction assumption, we have

$$\frac{|G^{\mathrm{v}}_{ik}|}{|\Omega^{\mathrm{v}}|} \geq \frac{1}{|\Omega^{\mathrm{v}}|} \cdot \frac{|G_{ik}|}{4} \geq \frac{c_{\mathrm{f}} n}{4|\Omega^{\mathrm{v}}|} = \frac{c_{\mathrm{f}}}{2d}. \tag{86}$$

Conditional on (83) and (84), for every $\lambda \in \Lambda_\epsilon$ and $i \in [d]$,

$$\begin{aligned} \max_{i,i' \in [d]} (T_{ik} + T_{i'k}) &\overset{\text{(i)}}{\geq} \frac{c_{\mathrm{f}}}{2d} \left[\left(b^{\mathrm{v}}_{ik} - \widehat{x}_i - \widehat{b}^{\mathrm{t}}_k \right)^2 + \left(b^{\mathrm{v}}_{i'k} - \widehat{x}_{i'} - \widehat{b}^{\mathrm{t}}_k \right)^2 \right] \\ &\overset{\text{(ii)}}{\geq} \frac{c_{\mathrm{f}}}{2d} \frac{\epsilon^2}{16d} = \frac{c_{\mathrm{f}} \epsilon^2}{32 d^2}, \end{aligned}$$

where inequality (i) is true by (86), and inequality (ii) is true by (85). Now consider the validation error $e^{(\lambda)}$. Conditional on (83) and (84), for every $\lambda \in \Lambda_\epsilon$,

$$\begin{aligned} e^{(\lambda)} = \sum_{i \in [d], k \in [r]} e^{(\lambda)}_{ik} &\overset{\text{(i)}}{\geq} \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{i \in [d], k \in [r]} \sum_{j \in G^{\mathrm{v}}_{ik}} (b_{ij} - b^{\mathrm{v}}_{ik})^2 + \sum_{i \in [d], k \in [r]} (T_{ik} + T_{i'k}) \\ &> \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{i \in [d], k \in [r]} \sum_{j \in G^{\mathrm{v}}_{ik}} (b_{ij} - b^{\mathrm{v}}_{ik})^2 + \frac{c_{\mathrm{f}} \epsilon^2}{32 d^2}, \end{aligned}$$

where inequality (i) is true by plugging in (82). Hence,

$$\lim_{n \to \infty} \left(e^{(\lambda)} > \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{i \in [d], k \in [r]} \sum_{j \in G^{\mathrm{v}}_{ik}} (b_{ij} - b^{\mathrm{v}}_{ik})^2 + \frac{c_{\mathrm{f}} \epsilon^2}{32 d^2}, \quad \forall \lambda \in \Lambda_\epsilon \right) = 1. \tag{87}$$

We set ϵ_1 to be sufficient small such that $4\epsilon_1^2 < \frac{c_{\mathrm{f}} \epsilon^2}{32 d^2}$. Taking a union bound of (87) with (80) 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 (81).

C.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 1, 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 (65) that $Y = B$. The validation error at $\lambda = 0$ is computed as:

$$\begin{aligned} e^{(0)} &= \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in\Omega^{\mathrm{v}}} \Big(b_{ij} - \widehat{b}^{(0)}_{\mathrm{NN}(i,j)} - \widehat{x}^{(0)}_i\Big)^2 \\ &\le \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in\Omega^{\mathrm{v}}} \Big(\big|b_{ij} - b_{\mathrm{NN}(i,j)}\big| + \Big|b_{\mathrm{NN}(i,j)} - \widehat{b}^{(0)}_{\mathrm{NN}(i,j)}\Big| + \Big|\widehat{x}^{(0)}_i\Big|\Big)^2. \end{aligned} \tag{88}$$

We consider the three terms inside the summation in (88) separately. For the first term $\big|b_{ij} - b_{\mathrm{NN}(i,j)}\big|$, combining Lemma 27(b) with Lemma 23, we have

$$\lim_{n\to\infty} \mathbb{P}\Big(\big|b_{ij} - b_{\mathrm{NN}(i,j)}\big| < \epsilon_1, \quad \forall (i,j) \in \Omega^{\mathrm{v}}\Big) = 1 \tag{89}$$

For the second term $|b_{\mathrm{NN}(i,j)} - \widehat{b}^{(0)}_{\mathrm{NN}(i,j)}|$, we have $|b_{\mathrm{NN}(i,j)} - \widehat{b}^{(0)}_{\mathrm{NN}(i,j)}| \le \max_{i\in[d],j\in[n]} |b_{ij} - \widehat{b}^{(0)}_{ij}|$. By the consistency (69) of $\widehat{B}^{(0)}$ from Step 1, we have

$$\lim_{n\to\infty} \mathbb{P}\left(|b_{\mathrm{NN}(i,j)} - \widehat{b}^{(0)}_{\mathrm{NN}(i,j)}| < \epsilon_1, \quad \forall (i,j) \in \Omega^{\mathrm{v}}\right) = 1. \tag{90}$$

For the third term $\widehat{x}^{(0)}_i$, by (67) in Theorem 34(b), we have

$$\lim_{n\to\infty} \mathbb{P}\Big(|\widehat{x}_i| < \epsilon_1, \quad \forall i \in [d]\Big) = 1. \tag{91}$$

Taking a union bound over the three terms (89), (90) and (91) and plugging them back to (88), the validation error at $\lambda = 0$ satisfies

$$\lim_{n\to\infty} \mathbb{P}\Big(e^{(0)} \le 9\epsilon_1^2\Big) = 1. \tag{92}$$

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

$$\lim_{n\to\infty} (\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon) = 1.$$

By Lemma 20(a) combined with the assumption that $d = 2$, we have

$$\lim_{n\to\infty} \mathbb{P}\Big(\underbrace{\Big|\widehat{x}_1^{(\lambda)} - \widehat{x}_2^{(\lambda)}\Big| > \frac{\epsilon}{\sqrt{2}}, \quad \forall \lambda \in \Lambda_\epsilon}_{E} \Big) = 1. \tag{93}$$

We denote the the event in (93) as E. We define

$$\Lambda_{2>1} := \left\{ \lambda \in [0,\infty] : \widehat{x}_2^{(\lambda)} - \widehat{x}_1^{(\lambda)} > \frac{\epsilon}{\sqrt{2}} \right\} \tag{94a}$$

$$\Lambda_{1>2} := \left\{ \lambda \in [0,\infty] : \widehat{x}_1^{(\lambda)} - \widehat{x}_2^{(\lambda)} > \frac{\epsilon}{\sqrt{2}} \right\}. \tag{94b}$$

Then we have

$$\Lambda_\epsilon \subseteq \Lambda_{2>1} \cup \Lambda_{1>2} \;\mid\; E. \tag{95}$$

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

$$\lim_{n\to\infty} \mathbb{P}\Big(\lambda_{\text{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E_1\Big) = \lim_{n\to\infty}(E_1). \tag{96}$$

To show (96) 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 (28a) in Lemma 28, we have

$$\lim_{n\to\infty} \mathbb{P}\left(b_1^{\text{t}} - \frac{1}{n}\sum_{j\in[n]} b_{1j} < -\frac{\epsilon}{8} \right) = 0 \tag{97a}$$

$$\lim_{n\to\infty} \mathbb{P}\left(b_2^{\text{t}} - \frac{1}{n}\sum_{j\in[n]} b_{2j} > \frac{\epsilon}{8} \right) = 0 \tag{97b}$$

Taking a union bound of (97), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\underbrace{b_1^{\text{t}} - b_2^{\text{t}} > \frac{1}{n}\sum_{j\in[n]} (b_{1j} - b_{2j}) - \frac{\epsilon}{4}}_{E'} \right) = 1. \tag{98}$$

Denote this event in (98) as E'. Hence, we have

$$b_1^{\mathrm{t}} - b_2^{\mathrm{t}} > -\frac{\epsilon}{4} \;\Big|\; (E', E_1) \tag{99}$$

Recall from Proposition 15 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

$$\infty \notin \Lambda_\epsilon \cap \Lambda_{2>1} \mid (E', E_1). \tag{100}$$

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 (21b) in Lemma 17, we have

$$\widehat{x}^{(\lambda_1)} + \widehat{x}^{(\lambda_2)} = 2b^{\mathrm{t}},$$

and hence $\widehat{x}^{(\lambda)}$ can be reparameterized as

$$\widehat{x}^{(\lambda)} = b^{\mathrm{t}} + \Delta \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \text{ for some } \Delta > \frac{\epsilon}{2\sqrt{2}}. \tag{101}$$

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 (14) as $\Pi_{\mathcal{M}}(y) = \arg\min_{u\in\mathcal{M}} \|y - u\|_2^2$.

Lemma 39. *Consider any $y \in \mathbb{R}^d$ and any $\lambda \in [0, \infty)$. Then we have*

$$\min_{u\in\mathcal{M}} \left(\|y - u\|_2^2 + \lambda \|u\|_2^2 \right) = \frac{1}{1+\lambda} \|y - \Pi_{\mathcal{M}}(y)\|_2^2 + \frac{\lambda}{1+\lambda} \|y\|_2^2. \tag{102}$$

The proof of this result is provided in Appendix C.11.5. We denote the objective (15) under any fixed $x \in \mathbb{R}^d$ as

$$\begin{aligned} L(x) &:= \min_{B \text{ obeys } \pi} \left\| Y - x\mathbf{1}^T - B \right\|_{\Omega^{\mathrm{t}}}^2 + \lambda \|B\|_{\Omega^{\mathrm{t}}}^2 \\ &\overset{\text{(i)}}{=} \frac{1}{1+\lambda} \underbrace{\left\| (Y - x\mathbf{1}^T) - \Pi_\pi(Y - x\mathbf{1}^T) \right\|_{\Omega^{\mathrm{t}}}^2}_{L_1(x)} + \frac{\lambda}{1+\lambda} \underbrace{\left\| Y - x\mathbf{1}^T \right\|_{\Omega^{\mathrm{t}}}^2}_{L_2(x)}, \end{aligned} \tag{103}$$

where equality (i) is true by (102) in Lemma 39. 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 (103) separately.

Term L_1: Recall from the model (65) 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,

$$0 = L_1(\widehat{x}') \leq L_1(\widehat{x}^{(\lambda)}) \qquad \forall \lambda \in [0, \infty]. \tag{104}$$

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 (99) with (101). Hence, we have

$$L_2(\widehat{x}) > L_2(\widehat{x}'), \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1} \setminus \{\infty\} \,\Big|\, (E', E_1). \tag{105}$$

Combining the term L_1 from (104) and the term L_2 from (105), 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

$$\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1} \setminus \{\infty\} \mid (E', E_1). \tag{106}$$

Combining the cases of $\lambda = \infty$ from (100) and $\lambda \neq \infty$ from (106), we have

$$\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1} \mid (E', E_1).$$

Hence,

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E_1\right) &\geq \mathbb{P}(E', E_1) \\
&= \mathbb{P}(E_1) - \mathbb{P}(E_1 \cap \overline{E'}) \\
&\geq \mathbb{P}(E_1) - \mathbb{P}(\overline{E'})
\end{aligned} \tag{107}$$

Taking the limit of (107), we have

$$\lim_{n \to \infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E_1\right) \overset{\text{(i)}}{=} \lim_{n \to \infty} \mathbb{P}(E_1), \tag{108}$$

where (i) is true by (98).

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:

$$S_c := \{(j, j') \in [n]^2 : 0 < b_{2j'} - b_{1j} < c\}. \tag{109}$$

Let $c' > 0$ be a constant. Denote $E^{\mathrm{v}}_{c',c}$ 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 $(j_k, j'_k) \in S_c \cap \Omega^{\mathrm{v}}$ for all $k \in [c'n]$. That is, the set $S_c \cap \Omega^{\mathrm{v}}$ 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 $E^{\mathrm{v}}_{c',c}$ happens under case E_2.

Lemma 40. *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*

$$\lim_{n\to\infty} \mathbb{P}\left(E^{\mathrm{v}}_{c',c} \cap E_2\right) = \lim_{n\to\infty} \mathbb{P}(E_2).$$

The proof of this result is provided in Appendix C.11.6. Now consider the the validation error contributed by the pairs in the set S'. We have

$$e^{(\lambda)} \le \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(j,j')\in S'} \left[\left(b_{1j} - \widehat{b}^{(\lambda)}_{\mathrm{NN}(1,j)} - \widehat{x}^{(\lambda)}_1\right)^2 + \left(b_{2j'} - \widehat{b}^{(\lambda)}_{\mathrm{NN}(2,j')} - \widehat{x}^{(\lambda)}_2\right)^2\right]. \tag{110}$$

We consider each individual term $(j, j') \in S'$. On the one hand, we have $b_{1j} < b_{2j'}$ by the definition (109) of S_c. Therefore, the element $(1, j)$ is ranked lower than $(2, j')$ in the total ordering $\mathcal{T}$. According to Algorithm 1, it can be verified that their interpolated bias satisfies

$$\widetilde{b}^{(\lambda)}_{\mathrm{NN}(1,j)} \le \widetilde{b}^{(\lambda)}_{\mathrm{NN}(2,j')} \qquad \forall \lambda \in [0, \infty]. \tag{111}$$

On the other hand, we have

$$b_{1j} - \widehat{x}_1 - (b_{2j'} - \widehat{x}_2) = (b_{1j} - b_{2j'}) + (\widehat{x}_2 - \widehat{x}_1) \overset{\text{(i)}}{>} -\frac{\epsilon}{2} + \frac{\epsilon}{\sqrt{2}} = \frac{\epsilon}{5}, \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1} \,\Big|\, (E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}, E), \tag{112}$$

where (i) is true by the definition of S_c in (109) (setting $c = \frac{\epsilon}{2}$), and the definition 94 of $\Lambda_{2>1}$. Combining (111) and (112), we have that for all $(j, j') \in S'$:

$$\begin{aligned}\left(b_{1j} - \widetilde{b}^{(\lambda)}_{\mathrm{NN}(1,j)} - \widehat{x}^{(\lambda)}_1\right)^2 + \left(b_{2j'} - \widetilde{b}^{(\lambda)}_{\mathrm{NN}(2,j')} - \widehat{x}^{(\lambda)}_2\right)^2 &\ge \min_{\substack{u_1,u_2\in\mathbb{R}\\ u_1\le u_2}} \min_{\substack{v_1,v_2\in\mathbb{R}\\ v_1 - v_2 > \frac{\epsilon}{5}}} (v_1 - u_1)^2 + (v_2 - u_2)^2 \\ &> \frac{\epsilon^2}{50}, \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1} \,\Big|\, (E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}, E).\end{aligned} \tag{113}$$

Conditional on $E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}$, there are at least $c'n$ such non-overlapping pairs. Plugging (113) to (110), the validation error is lower-bounded as

$$e^{(\lambda)} \geq \frac{1}{|\Omega^{\mathrm{v}}|} c'n \cdot \frac{\epsilon^2}{50} \geq \frac{2}{dn} c'n \cdot \frac{\epsilon^2}{50} = \frac{c'\epsilon^2}{25d}, \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1} \,\Big|\, (E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}, E). \tag{114}$$

Setting the constant ϵ_1 to be a sufficiently small constant such that $9\epsilon_1^2 < \frac{c'\epsilon^2}{25d}$, we have

$$\begin{aligned}
\mathbb{P}\left(e^{(\lambda)} \geq e^{(0)}, \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1}, E_2\right) &\geq \mathbb{P}\left(e^{(\lambda)} > \frac{c'\epsilon^2}{25d} > 9\epsilon_1^2 > e^{(0)}, \quad \forall \lambda \in \Lambda_\epsilon \cap \Lambda_{2>1}, E_2\right) \\
&\geq \mathbb{P}\left(e^{(\lambda)} > \frac{c'\epsilon^2}{25d}, E_2\right) - \mathbb{P}\left(e^{(0)} > 9\epsilon_1^2, E_2\right) \\
&\overset{\text{(i)}}{\geq} \mathbb{P}\left(E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}, E, E_2\right) - \mathbb{P}\left(e^{(0)} > 9\epsilon_1^2\right) \qquad (115) \\
&= \mathbb{P}\left(E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}, E\right) - \mathbb{P}\left(E^{\mathrm{v}}_{c',\frac{\epsilon}{2}}, E, \overline{E_2}\right) - \mathbb{P}\left(e^{(0)} > 9\epsilon_1^2\right), \qquad (116)
\end{aligned}$$

where (i) is true by (114). Taking the limit of $n \to \infty$ in (116), 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 40, (93) and (92) from Step 2. Equivalently,

$$\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E_2\right) = 1. \tag{117}$$

Finally, combining the two cases from (108) and (117), we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}\right) &= \lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E_1\right) + \lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E_2\right) \\
&= \lim_{n\to\infty} \mathbb{P}(E_1) + \lim_{n\to\infty} \mathbb{P}(E_2) = 1. \qquad (118a)
\end{aligned}$$

By a symmetric argument on the set $\Lambda_{1>2}$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{1>2}\right) = 1. \tag{118b}$$

Hence, we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon\right) &\geq \lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon, E\right) \\
&\overset{\text{(i)}}{\geq} \lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{1>2}, E\right) + \lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}, E\right) \\
&\geq \lim_{n\to\infty} \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{1>2}\right) + \mathbb{P}\left(\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon \cap \Lambda_{2>1}\right) - 2 \lim_{n\to\infty} \mathbb{P}(\overline{E}) \overset{\text{(ii)}}{=} 1,
\end{aligned}$$

where inequality (i) is true by (95), and equality (ii) is true by combining (118) with (93). This completes the proof.

C.6 Proof of Theorem 10

The proof follows notation in Appendix C.1 and preliminaries in Appendix C.2. Similar to the proof of Theorem 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 (1) reduces to:

$$Y = Z. \tag{119}$$

The proof consists of 3 steps that are similar to the steps in Theorem 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 15, our estimator $\widehat{x}^{(\infty)}$ at $\lambda = \infty$ is identical to taking the sample mean of each course. By the model (119), conditional on any training-validation split $(\Omega^{\mathrm{t}}, \Omega^{\mathrm{v}})$ given by Algorithm 1, each observation is i.i.d. noise of $\mathcal{N}(0,1)$. Recall from (7) that the number of observations in each course restricted to the training set Ω^{t} is $n^{\mathrm{t}} = \frac{n}{2}$. 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,

$$\lim_{n\to\infty} \mathbb{P}\Big(\|\widehat{x}^{(\infty)}\|_\infty < \epsilon_1 \Big) = 1. \tag{120}$$

By Proposition 15, we have $\widehat{B}^{(\infty)} = 0$.

Step 2: Computing the validation error at $\lambda = \infty$

Recall from Algorithm 1 that the interpolated bias $\widetilde{b}_{ij}$ for any element $(i,j) \in \Omega^{\mathrm{v}}$ 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 (119) of $Y = Z$. The validation error at $\lambda = \infty$ is computed as

$$\begin{aligned} e^{(\infty)} = \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in\Omega^{\mathrm{v}}} \Big(y_{ij} - \widehat{x}_i^{(\infty)} - \widetilde{b}_{ij}^{(\infty)} \Big)^2 &= \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in\Omega^{\mathrm{v}}} \Big(z_{ij} - \widehat{x}_i^{(\infty)} \Big)^2 \\ &= \frac{1}{|\Omega^{\mathrm{v}}|} \Bigg[\underbrace{\sum_{(i,j)\in\Omega^{\mathrm{v}}} {z_{ij}}^2}_{T_1} - 2 \underbrace{\sum_{(i,j)\in\Omega^{\mathrm{v}}} z_{ij} \widehat{x}_i^{(\infty)}}_{T_2} + \underbrace{\sum_{(i,j)\in\Omega^{\mathrm{v}}} (\widehat{x}_i^{(\infty)})^2}_{T_3} \Bigg]. \end{aligned} \tag{121}$$

We consider the three terms T_1, T_2 and T_3 in (121) separately. For the term T_1, we have $\mathbb{E}[z_{ij}^2] = \eta^2 = 1$. The number of samples is $|\Omega^{\mathrm{v}}| = dn^{\mathrm{v}} = d\frac{n}{2}$. By Hoeffding's inequality we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in\Omega^{\mathrm{v}}} z_{ij}^2 < 1 + \epsilon_1 \right) = 1. \tag{122}$$

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

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|}\left|\sum_{j\in\Omega_i^{\mathrm{v}}} z_{ij}\right| < \epsilon_1, \quad \forall i \in [d]\right) = 1. \tag{123}$$

Combining (123) with the consistency result (120) on $\widehat{x}^{(\infty)}$ from Step 1, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|}|T_2| < d\epsilon_1^2\right) = 1. \tag{124}$$

For the term T_3, we have

$$\frac{1}{|\Omega^{\mathrm{v}}|}T_3 \le \max_{i\in[d]} |\widehat{x}_i|^2. \tag{125}$$

Combining (125) with the consistency result (120) on $\widehat{x}^{(\infty)}$ from Step 1, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|}T_3 < \epsilon_1^2\right) = 1. \tag{126}$$

Taking a union bound of the terms T_1, T_2 and T_3 from (122), (124) and (126) and plugging them back to (121), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\infty)} \le (1+\epsilon_1) + d\epsilon_1^2 + \epsilon_1^2 = 1 + \epsilon_1 + (d+1)\epsilon_1^2\right) = 1. \tag{127}$$

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 (16) the definition of $\Lambda_\epsilon := \{\lambda \in [0,\infty] : \|\widehat{x}^{(\lambda)}\|_2 > \epsilon\}$. In this step, we show that

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} > e^{(\infty)}, \quad \forall \lambda \in \Lambda_\epsilon\right) = 1. \tag{128}$$

Then from (128) we have

$$\lim_{n\to\infty} (\lambda_{\mathrm{cv}} \notin \Lambda_\epsilon) = 1,$$

yielding the result of Theorem 10. It is sufficient to establish (128).

We now give some additional preliminary results for this step. By Lemma 20, we have

$$\lim_{n\to\infty} \mathbb{P}\underbrace{\left(\max_{i,i'\in[d]} \widehat{x}_i - \widehat{x}_{i'} > \frac{\epsilon}{\sqrt{d}}, \quad \forall \lambda \in \Lambda_\epsilon\right)}_{E} = 1. \tag{129}$$

We denote this event in (129) 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 41. *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*

$$\max_{(i,j)\in\Omega} \left|\widehat{b}_{ij}^{(\lambda)}\right| \le \max_{(i,j)\in\Omega} |y_{ij}| + \|\widehat{x}^{(\lambda)}\|_\infty. \tag{130}$$

The proof of this result is provided in Appendix C.12.1. Now we proceed differently for Step 3 for part (a) and part (b).

C.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 9(a), by Algorithm 1 the interpolated bias for elements in each group $k \in [r]$ is identical for all $(i,j) \in G_k^{\mathrm{v}}$. That is,

$$\widetilde{b}_{ij} = \widetilde{b}_{i'j'} \qquad \forall (i,j), (i',j') \in G_k^{\mathrm{v}}. \tag{131}$$

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

$$\max_{i\in[d]} \left|\widehat{x}_i + \widetilde{b}_{k_{\mathrm{f}}}\right| > 4\sqrt{\frac{d}{c_{\mathrm{f}}}} \qquad \forall \lambda \in \Lambda_1. \tag{132}$$

Combining (26a) from Lemma 26 with the single c_{f}-fraction assumption, one can see

$$\ell_{ik_{\mathrm{f}}}^{\mathrm{v}} \ge \frac{\ell_{ik_{\mathrm{f}}}}{4} > \frac{c_{\mathrm{f}} n}{4}. \tag{133}$$

Given (133), by Hoeffding's inequality we have

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{j\in G_{ik_{\mathrm{f}}}^{\mathrm{v}}} \mathbb{1}\{z_{ij} > 0\} \ge \frac{c_{\mathrm{f}} n}{12}\right) = 1 \tag{134a}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{j\in G_{ik_{\mathrm{f}}}^{\mathrm{v}}} \mathbb{1}\{z_{ij} < 0\} \ge \frac{c_{\mathrm{f}} n}{12}\right) = 1. \tag{134b}$$

We denote the event

$$E_1 := \left\{ \sum_{j \in G^{\mathrm{v}}_{ik_{\mathrm{f}}}} \mathbb{1}\{z_{ij} > 0\} \geq \frac{c_{\mathrm{f}} n}{12}, \quad \forall i \in [d] \right\} \cap \left\{ \sum_{j \in G^{\mathrm{v}}_{ik_{\mathrm{f}}}} \mathbb{1}\{z_{ij} < 0\} \geq \frac{c_{\mathrm{f}} n}{12}, \quad \forall i \in [d] \right\}. \tag{135}$$

Given that d is a constant by the assumption (A3), taking (134) with a union bound over $i \in [d]$, we have

$$\lim_{n \to \infty} \mathbb{P}(E_1) = 1. \tag{136}$$

Let i^* be a random variable (as a function of λ) defined as $i^* := \arg\max_{i \in [d]} \left| \widehat{x}_i + \widetilde{b}_{k_{\mathrm{f}}} \right|$ where the tie is broken arbitrarily. Conditional on E_1, for any $\lambda \in \Lambda_1$ we have the deterministic relation

$$\begin{aligned} e^{(\lambda)} = \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{k \in [r]} \sum_{(i,j) \in G^{\mathrm{v}}_k} \left(z_{ij} - \widehat{x}^{(\lambda)}_i - \widetilde{b}^{(\lambda)}_k \right)^2 &\geq \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j) \in G^{\mathrm{v}}_{k_{\mathrm{f}}}} (z_{ij} - \widehat{x}_i - \widetilde{b}_{k_{\mathrm{f}}})^2 \\ &\geq \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{j \in G^{\mathrm{v}}_{i^* k_{\mathrm{f}}}} (z_{i^* j} - \widehat{x}_{i^*} - \widetilde{b}_{k_{\mathrm{f}}})^2 \\ &\overset{\text{(i)}}{\geq} \frac{1}{|\Omega^{\mathrm{v}}|} \frac{c_{\mathrm{f}} n}{12} \left(4\sqrt{\frac{d}{c_{\mathrm{f}}}} \right)^2 \\ &= \frac{2}{dn} \cdot \frac{c_{\mathrm{f}} n}{12} \frac{16 d}{c_{\mathrm{f}}} = \frac{8}{3}, \quad \forall \lambda \in \Lambda_1 \,\Big|\, E_1. \end{aligned} \tag{137}$$

where (i) is true by (132) and the definition (135) of E_1. Combining (137) with (136), we have

$$\lim_{n \to \infty} \mathbb{P}\left(e^{(\lambda)} \geq \frac{4}{3}, \quad \forall \lambda \in \Lambda_1 \right) \geq \mathbb{P}\left(E_1\right) = 1. \tag{138}$$

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:

$$\begin{aligned} e^{(\lambda)} &= \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{k \in [r]} \sum_{(i,j) \in G^{\mathrm{v}}_k} \left(z_{ij} - \widehat{x}^{(\lambda)}_i - \widetilde{b}^{(\lambda)}_k \right)^2 \\ &= \frac{1}{|\Omega^{\mathrm{v}}|} \left[\sum_{(i,j) \in \Omega^{\mathrm{v}}} z_{ij}^2 - 2 \sum_{k \in [r]} \sum_{(i,j) \in G^{\mathrm{v}}_k} z_{ij} \left(\widehat{x}^{(\lambda)}_i + \widetilde{b}^{(\lambda)}_k \right) + \sum_{k \in [r]} \sum_{(i,j) \in G^{\mathrm{v}}_k} \left(\widehat{x}^{(\lambda)}_i + \widetilde{b}^{(\lambda)}_k \right)^2 \right] \\ &= \frac{1}{|\Omega^{\mathrm{v}}|} \Bigg[\underbrace{\sum_{(i,j) \in \Omega^{\mathrm{v}}} z_{ij}^2}_{T_1} \underbrace{- 2 \sum_{(i,j) \in \Omega^{\mathrm{v}}} z_{ij} \widehat{x}^{(\lambda)}_i}_{T_2} \underbrace{+ 2 \sum_{k \in [r]} \sum_{(i,j) \in G^{\mathrm{v}}_k} z_{ij} \widetilde{b}^{(\lambda)}_k}_{T_3} + \underbrace{\sum_{k \in [r]} \sum_{(i,j) \in G^{\mathrm{v}}_k} \left(\widehat{x}^{(\lambda)}_i + \widetilde{b}^{(\lambda)}_k \right)^2}_{T_4} \Bigg]. \end{aligned} \tag{139}$$

We analyze the four terms T_1, T_2, T_3 and T_4 in (139) separately.

Term T_1: Similar to (122) from Step 2, by Hoeffding's inequality we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in\Omega^{\mathrm{v}}} z_{ij}^2 > 1 - \epsilon_2 \right) = 1. \tag{140}$$

Term T_2: Recall that d is a constant by the assumption (A3). Similar to (123) from Step 2, by Hoeffding with a union bound over $i \in [d]$, we have

$$\lim_{n\to\infty} \mathbb{P} \underbrace{\left(\frac{1}{|\Omega^{\mathrm{v}}|} \left| \sum_{j\in\Omega_i^{\mathrm{v}}} z_{ij} \right| < \epsilon, \quad \forall i \in [d] \right)}_{E_2} = 1. \tag{141}$$

Denote this event in (141) as E_2.

We now bound $\|\widehat{x}\|_\infty$. By Hoeffding's inequality, on the training Ω^{t} we have:

$$\lim_{n\to\infty} \mathbb{P} \underbrace{\left(\frac{1}{|\Omega^{\mathrm{t}}|} \left| \sum_{(i,j)\in\Omega^{\mathrm{t}}} z_{ij} \right| < \sqrt{\frac{1}{dc_{\mathrm{f}}}} \right)}_{E_2'} = 1. \tag{142}$$

Plugging (21b) in Lemma 17 to (142), we have

$$\left| \sum_{i\in[d]} \widehat{x}_i^{(\lambda)} \right| = \frac{1}{n^{\mathrm{t}}} \left| \sum_{(i,j)\in\Omega^{\mathrm{t}}} z_{ij} \right| < \sqrt{\frac{d}{c_{\mathrm{f}}}} \qquad \forall \lambda \in \Lambda_2, \quad \text{conditional on } E_2'. \tag{143}$$

Combining (143) with the definition of Λ_2, we have

$$\|\widehat{x}\|_\infty \le 8\sqrt{\frac{d}{c_{\mathrm{f}}}} \qquad \forall \lambda \in \Lambda_2 \;\Big|\; E_2'. \tag{144}$$

To see (144), assume for contradiction that (144) 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 (143). A similar argument applies if $\widehat{x}_{i^*} < -8\sqrt{\frac{d}{c_{\mathrm{f}}}}$. Hence, (144) holds.

Finally, combining (144) with (141), we have:

$$\frac{1}{|\Omega^{\mathrm{v}}|}|T_2| = \frac{1}{|\Omega^{\mathrm{v}}|} \left| \sum_{(i,j)\in\Omega^{\mathrm{v}}} z_{ij}\widehat{x}_i \right| \tag{145}$$

$$\le \frac{d}{|\Omega^{\mathrm{v}}|} \max_{i\in[d]} \left| \sum_{(i,j)\in\Omega^{\mathrm{v}}} z_{ij} \right| \cdot \|\widehat{x}\|_\infty < 8d\sqrt{\frac{d}{c_{\mathrm{f}}}}\epsilon_2 \qquad \forall \lambda \in \Lambda_2, \quad \text{conditional on } (E_2, E_2'). \tag{146}$$

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 (141) and (142).

Term T_3: We use the following standard result derived from statistics.

Lemma 42. *Consider any fixed $d\geq 1$. Let $Z\sim\mathcal{N}(0,I_d)$. Then we have*

$$\lim_{d\to\infty}\mathbb{P}\left(\sup_{\substack{\|\theta\|_2=1\\ \theta_1\leq\ldots\leq\theta_d}}\theta^T Z\leq d^{\frac{1}{4}}\right)=1.$$

For completeness, the proof of this lemma is in Appendix C.12.2. We now explain how to apply Lemma 42 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 (131) 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

$$\max_{k\in[r]}|\widetilde{b}_k|\leq\max_{(i,j)\in\Omega^{\mathrm{t}}}\left|\widehat{b}_{ij}^{(\lambda)}\right|\overset{\text{(i)}}{\leq}\max_{(i,j)\in\Omega^{\mathrm{t}}}|y_{ij}|+\|\widehat{x}^{(\lambda)}\|_\infty\qquad\forall\lambda\in[0,\infty],\tag{147}$$

where (i) is true by (130) in Lemma 41. We consider the term $\max_{(i,j)\in\Omega^{\mathrm{v}}}|y_{ij}|$ on the RHS of (147). Recall from the model (119) that $Y=Z$. Hence, we have

$$\lim_{n\to\infty}\mathbb{P}\underbrace{\left(\max_{(i,j)\in\Omega^{\mathrm{v}}}|y_{ij}|<2\sqrt{\log dn}\right)}_{E_2''}\overset{\text{(i)}}{=}1,\tag{148}$$

where (i) is true by Lemma 25. Plugging (148) and the bound on $\|\widehat{x}\|_\infty$ from (144) to (147), we have that conditional on E_2'' and E_2',

$$\begin{aligned}\max_{k\in[r]}|\widetilde{b}_k|&\leq\max_{(i,j)\in\Omega^{\mathrm{t}}}|y_{ij}|+\|\widehat{x}^{(\lambda)}\|_\infty\\&\leq 2\sqrt{\log dn}+8\sqrt{\frac{d}{c_{\mathrm{f}}}}\qquad\forall\lambda\in\Lambda_2\;\Big|\;(E_2',E_2'').\end{aligned}$$

Hence, we have

$$\|\widetilde{B}\|_{\Omega^{\mathrm{t}}}\leq\sqrt{|\Omega^{\mathrm{t}}|}\cdot\max_{k\in[r]}\left|\widetilde{b}_k\right|\leq\sqrt{dn^{\mathrm{v}}}\left(2\sqrt{\log dn}+8\sqrt{\frac{d}{c_{\mathrm{f}}}}\right)\qquad\forall\lambda\in\Lambda_2\;\Big|\;(E_2',E_2'').$$

and therefore

$$\lim_{n\to\infty}\mathbb{P}\left(\|\widetilde{B}\|_{\Omega^{\mathrm{t}}}\leq\sqrt{dn^{\mathrm{v}}}\left(2\sqrt{\log dn}+8\sqrt{\frac{d}{c_{\mathrm{f}}}}\right),\quad\forall\lambda\in\Lambda_2\right)\geq\lim_{n\to\infty}\mathbb{P}(E_2'\cap E_2'')=1.\tag{149}$$

Applying Lemma 42: For the term T_3, for any constant $C > 0$, we have

$$\mathbb{P}\left(|T_3| < C(dn^{\mathrm{t}})^{\frac{1}{4}}, \quad \forall \lambda \in \Lambda_2\right) \geq \mathbb{P}\Bigg(\underbrace{\left\{\left|\frac{T_3}{C}\right| < (dn^{\mathrm{t}})^{\frac{1}{4}}, \quad \forall \lambda \in \Lambda_2\right\}}_{E_3} \cap \underbrace{\left\{\left\|\frac{\widetilde{B}}{C}\right\|_{\Omega^{\mathrm{t}}} \leq 1, \quad \forall \lambda \in \Lambda_2\right\}}_{E_4}\Bigg) \tag{150}$$

We have

$$\mathbb{P}(\overline{E_3 \cap E_4}) = \mathbb{P}(\overline{E_4}) + \mathbb{P}(\overline{E_3} \cap E_4) \tag{151}$$

Setting $C = \sqrt{dn^{\mathrm{v}}}\left(2\sqrt{\log dn} + 8\sqrt{\frac{d}{c_{\mathrm{f}}}}\right)$, by (149) we have

$$\mathbb{P}(\overline{E_4}) = 0. \tag{152}$$

Applying Lemma 42 on $\frac{\widetilde{B}_{\Omega^{\mathrm{t}}}}{C}$, we have

$$\lim_{n\to\infty} \mathbb{P}(\overline{E_3} \cap E_4) = 0. \tag{153}$$

Plugging (152) and (153) to (151), we have

$$\lim_{n\to\infty} \mathbb{P}(\overline{E_3 \cap E_4}) = 0. \tag{154}$$

Combining (154) with (150), 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

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|}|T_3| < \epsilon_2\right) = 1. \tag{155}$$

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 (129), we have

$$\max_{i,i' \in [d]} (\widehat{x}_i + \widetilde{b}_{k_{\mathrm{f}}}) - (\widehat{x}_{i'} + \widetilde{b}_{k_{\mathrm{f}}}) > \frac{\epsilon}{\sqrt{d}} \qquad \forall \lambda \in \Lambda_2, \Bigg| E. \tag{156}$$

Therefore, we have

$$\max_{i,i' \in [d]} \left[(\widehat{x}_i + \widetilde{b}_{k_{\mathrm{f}}})^2 + (\widehat{x}_{i'} + \widetilde{b}_{k_{\mathrm{f}}})^2\right] > \frac{\epsilon^2}{4d} \qquad \forall \lambda \in \Lambda_2 \Bigg| E. \tag{157}$$

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 (133) and (157). Hence,

$$\mathbb{P}\left(T_4 \geq \frac{c_{\mathrm{f}}\epsilon^2}{8d^2} \quad \forall \lambda \in \Lambda_2\right) \geq \mathbb{P}(E) = 1. \tag{158}$$

Putting things together: Plugging the four terms from (140), (141), (155) and (158) respectively back to (139), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} > (1-\epsilon_2) + 8d\sqrt{\frac{d}{c_{\mathrm{f}}}}\epsilon_2 + \epsilon_2 + \frac{c_{\mathrm{f}}\epsilon^2}{8d^2}, \quad \forall \lambda \in \Lambda_2 \right) = 1. \tag{159}$$

Finally, combining the two cases from (138) and (159), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} \geq \frac{8}{3} \wedge \left(1 + 16d\sqrt{\frac{d}{c_{\mathrm{f}}}}\epsilon_2 + \frac{c_{\mathrm{f}}\epsilon^2}{8d^2} \right), \quad \forall \lambda \in \Lambda_\epsilon \right) = 1. \tag{160}$$

Recall from (127) that the validation error at $\lambda = \infty$ is bounded as

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\infty)} \leq 1 + \epsilon_1 + (d+1)\epsilon_1^2 \right) = 1. \tag{161}$$

Combining (160) and (161) 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.

C.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 (21b) with Hoeffding's inequality, we have

$$\lim_{n\to\infty} \mathbb{P}\Bigg(\underbrace{|\widehat{x}_1 + \widehat{x}_2| = \frac{1}{n^{\mathrm{t}}}\left| \sum_{(i,j)\in\Omega^{\mathrm{t}}} z_{ij} \right| < \epsilon \wedge \frac{16}{\sqrt{c_{\mathrm{f}}}}, \quad \forall \lambda \in \Lambda_\epsilon}_{E_1} \Bigg) = 1. \tag{162}$$

We denote this event in (162) 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:

$$S_{\text{pairs}} := \{(t, t') \in [2n-1]^2 : t \in S_1^{\text{v}},\ t' \in S_2^{\text{v}},\ t < t'\}.$$

Define E_c as the event that there exists distinct values $(t_1, t'_1, \ldots, t_{cn}, t'_{cn})$ 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 43. *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 C.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

$$S^+ := S \cap \{t \in [2n-1] : t > \text{median}(S)\}$$
$$S^- := S \cap \{t \in [2n-1] : t < \text{median}(S)\}.$$

Furthermore, for $i \in \{1, 2\}$, we define

$$S_i^{\text{v}+} := S^+ \cap S_i \cap \Omega^{\text{v}}$$
$$S_i^{\text{v}-} := S^- \cap S_i \cap \Omega^{\text{v}}.$$

The following lemma lower-bounds the size of $S_i^{\text{v}+}$ and $S_i^{\text{v}-}$.

Lemma 44. *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 C.12.4. We denote this event in Lemma 44 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 (162), we have

$$\widehat{x}_1^{(\lambda)} < -\frac{32}{\sqrt{c_{\text{f}}}} < \frac{16}{\sqrt{c_{\text{f}}}} < \widehat{x}_2^{(\lambda)} \quad \forall \lambda \in \Lambda_1 \,\Big|\, E. \tag{163}$$

By Hoeffding's inequality combined with Lemma 44, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{(i,j)\in S_1^{\mathrm{v}-}} \mathbb{1}\{z_{ij} > 0\} > \frac{c_{\mathrm{f}} n}{96} \right) = 1 \tag{164a}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{(i,j)\in S_2^{\mathrm{v}+}} \mathbb{1}\{z_{ij} < 0\} > \frac{c_{\mathrm{f}} n}{96} \right) = 1. \tag{164b}$$

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 (164), we have

$$\lim_{n\to\infty} \mathbb{P}(E_3) = 1. \tag{165}$$

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 1. 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:

$$e^{(\lambda)} \ge \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in S_1^{\mathrm{v}-}} \left(z_{ij} - \widehat{x}_1^{(\lambda)} - \widetilde{b}_{ij}^{(\lambda)} \right)^2 \tag{166}$$

$$\overset{\text{(i)}}{\ge} \frac{1}{|\Omega^{\mathrm{v}}|} \cdot \left|S_1^{\mathrm{v}-}\right| \cdot \left(0 + \frac{16}{\sqrt{c_{\mathrm{f}}}} + 0\right)^2 \overset{\text{(i)}}{\ge} \frac{1}{n} \frac{c_{\mathrm{f}} n}{96} \frac{256}{c_{\mathrm{f}}} = \frac{8}{3}, \quad \forall \lambda \in \Lambda_1 \,\Big|\, (E_1, E_2, E_3), \tag{167}$$

where (i) is true by (163) and the definition of E_3, and (ii) is true by the definition of E_2. Hence, we have

$$\lim_{n\to\infty} \left(e^{(\lambda)} \ge \frac{8}{3} \quad \forall \lambda \in \Lambda_1, \{\widetilde{b}_t \le 0 \text{ for all } t \in S^-\} \right) \overset{\text{(i)}}{\ge} \mathbb{P}\left(\widetilde{b}_t \le 0 \text{ for all } t \in S^- \right), \tag{168a}$$

where (i) is true by (162), Lemma 44 and (165). By a similar argument, we have

$$\lim_{n\to\infty} \left(e^{(\lambda)} \ge \frac{8}{3} \quad \forall \lambda \in \Lambda_1, \{\widetilde{b}_t \ge 0 \text{ for all } t \in S^+\} \right) \ge \mathbb{P}\left(\widetilde{b}_t \ge 0 \text{ for all } t \in S^+ \right), \tag{168b}$$

Summing over (168), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} \ge \frac{8}{3}, \quad \forall \lambda \in \Lambda_1 \right) = 1. \tag{169}$$

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 (162), we have

$$-\frac{32}{\sqrt{c_{\mathrm{f}}}} < \{\widehat{x}_1, \widehat{x}_2\} < \frac{48}{\sqrt{c_{\mathrm{f}}}}. \tag{170}$$

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 (170), the analysis of the terms T_1, T_2 and T_3 follows the proof in part (a). We have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|} T_1 > 1 - \epsilon_2\right) = 1. \tag{171a}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(i,j)\in\Omega^{\mathrm{v}}} |T_2| < \frac{96}{\sqrt{c_{\mathrm{f}}}}\epsilon_2\right) = 1. \tag{171b}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|} |T_3| < \epsilon_2\right) = 1. \tag{171c}$$

Now we consider the last term T_4. Recall from (129) 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 $(t,t') \in S_{\text{pairs}}$. By the definition of S_{pairs} we have $t < t'$. Hence, we have $\widetilde{b}_t \le \widetilde{b}_{t'}$ due to the nearest-neighbor interpolation in Algorithm 1. 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:

$$\begin{aligned} \frac{1}{|\Omega^{\mathrm{v}}|} T_4 &\ge \frac{1}{|\Omega^{\mathrm{v}}|} \sum_{(t,t')\in S_{\text{pairs}}} \left[(\widehat{x}_1 + \widetilde{b}_t)^2 + (\widehat{x}_2 + \widetilde{b}_{t'})^2\right] \\ &\overset{\text{(i)}}{\ge} \frac{1}{2n} \cdot \frac{c_{\mathrm{f}} n}{36} \cdot \frac{\epsilon^2}{8} = \frac{c_{\mathrm{f}}\epsilon^2}{576} \qquad \forall \lambda \in \Lambda_2 \cap \Lambda_{2>1} \,\Big|\, (E_{\frac{1}{36}}, E), \end{aligned} \tag{172a}$$

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

$$\frac{1}{|\Omega^{\mathrm{v}}|} T_4 \geq \frac{c_{\mathrm{f}}\epsilon^2}{576}, \quad \forall \lambda \in \Lambda_2 \cap \Lambda_{1>2} \;\Big|\; (E'_{\frac{1}{36}}, E). \tag{172b}$$

Combining (172), 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 43 and (129), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{|\Omega^{\mathrm{v}}|} T_4 \geq \frac{c_{\mathrm{f}}\epsilon^2}{576}, \quad \forall \lambda \in \Lambda_2\right) \geq \lim_{n\to\infty} \mathbb{P}\left(E_{\frac{1}{36}}, E'_{\frac{1}{36}}, E\right) = 1. \tag{173}$$

Putting things together: Combining the four terms from (171) and (173), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} > 1 - \epsilon_2 - \frac{128}{\sqrt{c_{\mathrm{f}}}}\epsilon_2 - 2\epsilon_2 + \frac{c_{\mathrm{f}}\epsilon^2}{576} = 1 - \left(3 + \frac{128}{\sqrt{c_{\mathrm{f}}}}\right)\epsilon_2 + \frac{c_{\mathrm{f}}\epsilon^2}{576}, \quad \forall \lambda \in \Lambda_2\right) = 1. \tag{174}$$

Combining the two cases from (169) and (174), we have

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\lambda)} > \frac{8}{3} \wedge \left[1 - \left(3 + \frac{128}{\sqrt{c_{\mathrm{f}}}}\right)\epsilon_2 + \frac{c_{\mathrm{f}}\epsilon^2}{576}\right], \quad \forall \lambda \in \Lambda_2\right) = 1. \tag{175}$$

Recall from (127) that the validation error at $\lambda = \infty$ is bounded as (taking $d = 2$):

$$\lim_{n\to\infty} \mathbb{P}\left(e^{(\infty)} \leq 1 + \epsilon_1 + 3\epsilon_1^2, \quad \forall \lambda \in \Lambda_\epsilon\right) = 1. \tag{176}$$

Combining (175) and (176) 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.

C.7 Proof of Proposition 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 $\sigma = 1$.

C.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 3 that the all c_{f}-fraction assumption requires that each course $i \in [d]$ has at least $\ell_{ik} \geq c_{\mathrm{f}} n$ students in each group $k \in [r]$. Let $c_0 = 1 - c_{\mathrm{f}} r$. Due to the assumption that $c_{\mathrm{f}} \in (0, \frac{1}{r})$, 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 $(c_{\mathrm{f}} + c_0)n$ students from group 1, and $c_{\mathrm{f}} n$ students from each remaining group $k \in \{2, \ldots, r\}$. That is,

$$\ell_{1k} = \begin{cases} (c_{\mathrm{f}} + c_0)n & \text{if } k = 1 \\ c_{\mathrm{f}} n & \text{if } 2 \leq k \leq r. \end{cases} \tag{177a}$$

- **Course 2:** The course has $(c_{\mathrm{f}} + c_0)n$ students from group r, and $c_{\mathrm{f}} n$ students from each remaining group $k \in [r-1]$. That is,

$$\ell_{2k} = \begin{cases} (c_{\mathrm{f}} + c_0)n & \text{if } 1 \leq k \leq r-1 \\ c_{\mathrm{f}} n. & \text{if } k = r. \end{cases} \tag{177b}$$

- **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 37 (see Appendix C.3.1 for its statement and Appendix C.11.4 for its proof), due to assumptions (A2) and (A3) we establish the following lemma.

Lemma 45. *Consider any group ordering $\mathcal{O}$ that satisfies the all c_{f}-fraction assumption. For any $\epsilon > 0$, we have*

$$\lim_{n \to \infty} \mathbb{P}\Big(\underbrace{\big|b_{ik} - \bar{b}_{G_k}\big| < \epsilon, \quad \forall i \in [d], k \in [r]}_{E_1} \Big) = 1.$$

Denote this event in Lemma 45 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, $\{b^{(1)}, \dots, b^{(\ell_0)}\}$, and the group r contains the ℓ_r highest bias terms, $\{b^{(dn-\ell_0+1)}, \dots, b^{(dn)}\}$. Hence, we have

$$\begin{aligned}\overline{b}_{G_1} &< \frac{b^{(\frac{\ell_0}{2})} + b^{(\ell_0)}}{2} \\ \overline{b}_{G_r} &> \frac{b^{(dn-\ell_0)} + b^{(dn-\frac{\ell_0}{2})}}{2}.\end{aligned}$$

By the convergence of the order statistics from Lemma 24, it can be shown that there exists some constant $c > 0$ (dependent on d, r and c_{f}), such that

$$\lim_{n\to\infty} \mathbb{P}\Big(\underbrace{\overline{b}_{G_r} - \overline{b}_{G_1} > c}_{E_2} \Big) = 1. \tag{178}$$

Denote this event in (178) as E_2. The mean estimator is computed as

$$[\widehat{x}_{\mathrm{mean}}]_1 = x_1^* + \frac{1}{n}\sum_{k\in[r]} \ell_{1k} b_{1k} \tag{179a}$$

$$[\widehat{x}_{\mathrm{mean}}]_2 = x_2^* + \frac{1}{n}\sum_{k\in[r]} \ell_{2k} b_{2k} \tag{179b}$$

Taking the difference on (178), conditional on E_1 and E_2,

$$\begin{aligned}[\widehat{x}_{\mathrm{mean}}]_2 - [\widehat{x}_{\mathrm{mean}}]_1 &= (x_2^* - x_1^*) + \frac{1}{n}\sum_{k\in[r]} (\ell_{2k} b_{2k} - \ell_{1k} b_{1k}) \\ &\overset{\text{(i)}}{>} (x_2^* - x_1^*) + \frac{1}{n}\sum_{k\in[r]} (\ell_{2k} \overline{b}_{G_k} - \ell_{1k} \overline{b}_{G_k}) - 2\epsilon \\ &\overset{\text{(i)}}{=} (x_2^* - x_1^*) + c_0(b_r - \overline{b}_{G_1}) - 2\epsilon \\ &\overset{\text{(iii)}}{>} (x_2^* - x_1^*) + c_0 c - 2\epsilon. \end{aligned} \tag{180}$$

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 (177), and inequality (iii) is true by the definition (178) 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 (180) we have

$$\mathbb{P}([\widehat{x}_{\mathrm{mean}}]_2 - [\widehat{x}_{\mathrm{mean}}]_1 > 0) = 1. \tag{181}$$

Combining (181) with the fact that $x_2^* - x_1^* < 0$, completing the proof of part (a).

C.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).

C.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 5.1):

$$b_{11} \le \ldots \le b_{1n} \le b_{21} \le \ldots \le b_{2n} \le \ldots \le b_{d1} \le \ldots \le b_{dn}.$$

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

$$\overline{b}_1 = \frac{1}{n}\sum_{j\in[n]} b_{1j} < \frac{b^{(\frac{n}{2})} + b^{(n)}}{2}$$
$$\overline{b}_r = \frac{1}{n}\sum_{j\in[n]} b_{2j} > \frac{b^{(dn-\frac{n}{2})} + b^{(dn)}}{2}.$$

Similar to part (a), by Lemma 24, 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

$$\lim_{n\to\infty} \mathbb{P}([\widehat{x}_{\text{mean}}]_r - [\widehat{x}_{\text{mean}}]_1 = x_2^* - x_1^* + \overline{b}_2 - \overline{b}_1 > 0) = 1. \tag{182}$$

Combining (182) with the fact that $x_1^* > x_r^*$ completes the proof of part (c).

C.8 Proof of Proposition 13

By Corollary 19, 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 C.8.1) and for our estimator at $\lambda = 0$ (Appendix C.8.2). Both parts use the following standard result regarding the uniform distribution.

Lemma 46. *Let $X_1, \ldots, X_n$ be i.i.d. Unif$[0, 1]$, we have*

$$\mathbb{E}\left(\frac{\sum_{i=1}^n X_i}{n}\right)^2 = \frac{1}{4} + \frac{1}{12n}.$$

C.8.1 The reweighted mean estimator

We follow the definition of the reweighted mean estimator defined in Appendix A.2. In the reweighting step, by (4) we have

$$\widehat{x}_{\text{rw}} = \frac{1}{2}\begin{bmatrix}\overline{U} + \overline{V} \\ \overline{U}' + \overline{V}'\end{bmatrix}. \tag{183}$$

In the recentering step, by (6) we have

$$\begin{aligned}
\widehat{x}_{\mathrm{rw}} &\leftarrow \widehat{x}_{\mathrm{rw}} + \left(-\frac{1}{2}\sum_{i\in\{1,2\}}[\widehat{x}_{\mathrm{rw}}]_i + \frac{1}{2n}\sum_{i\in\{1,2\},j\in[n]} y_{ij}\right)\mathbf{1} \\
&= \widehat{x}_{\mathrm{rw}} + \left(-\frac{[\widehat{x}_{\mathrm{rw}}]_1 + [\widehat{x}_{\mathrm{rw}}]_2}{2} + \frac{rn\overline{U} + (1-r)n\overline{V} + (1-r)n\overline{U}' + rn\overline{V}'}{2n}\right)\mathbf{1} \\
&= \frac{[\widehat{x}_{\mathrm{rw}}]_1 - [\widehat{x}_{\mathrm{rw}}]_2}{2}\begin{bmatrix}1\\-1\end{bmatrix} + \left(\frac{r\overline{U} + (1-r)\overline{V} + (1-r)\overline{U}' + r\overline{V}'}{2}\right)\mathbf{1} \\
&\overset{\text{(i)}}{=} \frac{\overline{U} + \overline{V} - \overline{U}' - \overline{V}'}{4}\begin{bmatrix}1\\-1\end{bmatrix} + \left(\frac{r\overline{U} + (1-r)\overline{V} + (1-r)\overline{U}' + r\overline{V}'}{2}\right)\mathbf{1}, \qquad (184)
\end{aligned}$$

where equality (i) is true by plugging in (183) from the reweighting step. By symmetry, we have $\mathbb{E}[\widehat{x}_{\mathrm{rw}}]_1^2 = \mathbb{E}[\widehat{x}_{\mathrm{rw}}]_2^2$, so we only consider course 1. By (184), we have

$$\begin{aligned}
\mathbb{E}[\widehat{x}_{\mathrm{rw}}]_1^2 &\overset{\text{(i)}}{=} \mathbb{E}\left(\frac{\overline{U} + \overline{V} - \overline{U}' - \overline{V}'}{4}\right)^2 + \mathbb{E}\left(\frac{r\overline{U} + (1-r)\overline{V} + (1-r)\overline{U}' + r\overline{V}'}{2}\right)^2 \\
&= \frac{1}{16}\mathbb{E}\left[\overline{U}'^2 + \overline{V}'^2 + \overline{U}^2 + \overline{V}^2 - 4\cdot\frac{1}{2}\frac{1}{2}\right] \\
&\qquad + \frac{1}{4}\mathbb{E}\left[(1-r)^2\overline{U}'^2 + r^2\overline{V}'^2 + r^2\overline{U}^2 + (1-r)^2\overline{V}^2 - 2\left(\frac{r^2}{4} + \frac{(1-r)^2}{4}\right)\right] \\
&= \frac{1}{8}\mathbb{E}\left[\overline{U}^2 + \overline{V}^2 - \frac{1}{2}\right] + \frac{1}{2}\mathbb{E}\left[r^2\overline{U}^2 + (1-r)^2\overline{V}^2 - \frac{r^2 + (1-r)^2}{4}\right] \\
&\overset{\text{(ii)}}{=} \frac{1}{8}\left[\frac{1}{4} + \frac{1}{12rn} + \frac{1}{4} + \frac{1}{12(1-r)n} - \frac{1}{2}\right] + \frac{1}{2}\mathbb{E}\left[\frac{r^2}{4} + \frac{r^2}{12rn} + \frac{(1-r)^2}{4} + \frac{(1-r)^2}{12(1-r)n} - \frac{r^2 + (1-r)^2}{4}\right] \\
&= \frac{1}{96n}\left(\frac{1}{r} + \frac{1}{1-r}\right) + \frac{1}{24n} \\
&= \frac{1}{24n} + \frac{1}{96r(1-r)n}.
\end{aligned}$$

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 46. Finally, we have

$$\frac{1}{2}\mathbb{E}\|\widehat{x}_{\mathrm{rw}}\|]_2^2 = \frac{1}{2}\left(\mathbb{E}[\widehat{x}_{\mathrm{rw}}]_1^2] + \mathbb{E}[\widehat{x}_{\mathrm{rw}}]_2^2\right) = \mathbb{E}[\widehat{x}_{\mathrm{rw}}]_1^2 = \frac{1}{24n} + \frac{1}{96r(1-r)n} \geq \frac{1}{24n} + \frac{1}{24n} = \frac{1}{12n},$$

where the inequality holds because $r(1-r) \leq \frac{1}{4}$ for every $r \in (0,1)$.

C.8.2 Our estimator at $\lambda = 0$

Recall from Proposition 22 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

$$\gamma = \begin{cases} y_{22,\min} - y_{11,\max} & \text{if } y_{22,\min} - y_{11,\max} < \overline{y}_2 - \overline{y}_1 \\ y_{21,\max} - y_{12,\min} & \text{if } y_{21,\max} - y_{12,\min} > \overline{y}_2 - \overline{y}_1 \\ \overline{y}_2 - \overline{y}_1 & \text{o.w.} \end{cases} \tag{185}$$

By (185), we have

$$\frac{1}{2}\mathbb{E}\|\widehat{x}^{(0)}\|_2^2 = \frac{1}{2}\mathbb{E}\left[\left(\overline{y} - \frac{\gamma}{2}\right)^2 + \left(\overline{y} + \frac{\gamma}{2}\right)^2\right] = \mathbb{E}[\overline{y}^2] + \frac{1}{4}\mathbb{E}[\gamma^2]. \tag{186}$$

We analyze the two terms in (186) separately.

Term of $\mathbb{E}[\overline{y}^2]$: For ease of notation, we denote the random variables

$$\{\widetilde{U}_j\}_{j\in[n]} := \{U_j\}_{j\in[rn]} \cup \{U'_j\}_{j\in[(1-r)n]}$$
$$\{\widetilde{V}_j\}_{j\in[n]} := \{V_j\}_{j\in[(1-r)n]} \cup \{V'_j\}_{j\in[rn]}$$

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

$$\begin{aligned}
\mathbb{E}[\overline{y}^2] &= \mathbb{E}\left(\frac{\sum_{i\in[n]}\widetilde{U}_i + \sum_{i\in[n]}\widetilde{V}_i}{2n}\right)^2 \\
&= \frac{1}{4n^2}\mathbb{E}\left[\sum_{i\in[n]}\widetilde{U}_i^2 + \sum_{i\in[n]}\widetilde{V}_i^2 + 2\sum_{i\in[n],j\in[n]}\widetilde{U}_i\widetilde{V}_j + \sum_{i\in[n]}\sum_{j\neq i}\widetilde{U}_i\widetilde{U}_j + \sum_{i\in[n]}\sum_{j\neq i}\widetilde{V}_i\widetilde{V}_j\right] \\
&= \frac{1}{4n^2}\left[\frac{n}{3} + \frac{n}{3} + 2n^2\left(-\frac{1}{4}\right) + n(n-1)\frac{1}{4} + n(n-1)\frac{1}{4}\right] \\
&= \frac{1}{24n}.
\end{aligned} \tag{187}$$

Term of $\mathbb{E}[\gamma^2]$: To analyze the term $\mathbb{E}[\gamma^2]$, we use the following standard result from statistics.

Lemma 47. *Let $X_1, \ldots, X_n \sim$ Unif$[0, 1]$. Let $X_{\min} = \min_{i\in[n]} X_i$. We have*

$$\mathbb{E}[X_{\min}] = \frac{1}{n+1}$$
$$\mathbb{E}[X_{\min}^2] = \frac{2}{(n+1)(n+2)}.$$

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 (185) it can be verified that we have the deterministic relation

$$\begin{aligned}|\gamma| &\le (y_{22,\min} - y_{11,\max}) \vee (y_{12,\min} - y_{21,\max}) \\ &\overset{\text{(i)}}{=} (V'_{\min} - U_{\max}) \vee (V_{\min} - U'_{\max}) \\ &\le V'_{\min} - U_{\max} + V_{\min} - U'_{\max},\end{aligned}$$

where equality (i) is true by the assumption that there is no noise and the assumption of $x^* = 0$. Therefore,

$$\begin{aligned}\mathbb{E}[\gamma^2] &\le \mathbb{E}\left[(V'_{\min} - U_{\max}) + (V_{\min} - U'_{\max})\right]^2 \\ &= \underbrace{\mathbb{E}(V'_{\min} - U_{\max})^2}_{T_1} + \underbrace{\mathbb{E}(V_{\min} - U'_{\max})^2}_{T_2} + 2\underbrace{\mathbb{E}(V'_{\min} - U_{\max})(V_{\min} - U'_{\max})}_{T_3}. \end{aligned} \tag{188}$$

We consider the three terms T_1, T_2 and T_3 separately. For the term T_1, by Lemma 47 we have

$$\begin{aligned}T_1 &= \mathbb{E}[V'_{\min}]^2 + \mathbb{E}[U^2_{\max}] - 2\mathbb{E}[V'_{\min}U_{\max}] \\ &= 2 \cdot \frac{2}{(rn+1)(rn+2)} + 2 \cdot \frac{1}{(rn+1)^2} \le \frac{6}{r^2n^2}.\end{aligned}$$

Likewise, for the term T_2 we have

$$T_2 \le \frac{6}{(1-r)^2n^2}.$$

For the term T_3, by Lemma 47 we have

$$T_3 = \frac{2}{rn+1} \cdot \frac{2}{(1-r)n+1} \le \frac{4}{r(1-r)n^2}.$$

Plugging the three terms back to (188), we have

$$\mathbb{E}[\gamma^2] \le \frac{6}{r^2n^2} + \frac{6}{(1-r)^2n^2} + \frac{8}{r(1-r)n^2} = \frac{c}{n^2}, \tag{189}$$

for some constant $c > 0$.

Finally, plugging (187) and (189) back to (186), we have

$$\frac{1}{2}\mathbb{E}\|\widehat{x}^{(0)}\|_2 \le \frac{1}{24n} + \frac{c}{4n^2},$$

completing the proof.

C.9 Proof of preliminaries

In this section, we present the proofs of the preliminary results presented in Appendix C.2.

C.9.1 Proof of Proposition 14

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 (15) reduces to (2). We separately consider the cases of $\lambda = 0$ and $\lambda \in (0, \infty)$.

Case of $\lambda = 0$ The objective (2) becomes

$$\min_{x \in \mathbb{R}^d} \min_{\substack{B \in \mathbb{R}^{d \times n} \\ B \text{ satisfies } \mathcal{O}}} \left\| Y - x\mathbf{1}^T - B \right\|_F^2 = \min_{\substack{W \in \mathbb{R}^{d \times n} \\ W \in \{x\mathbf{1}^T + B \mid x \in \mathbb{R}^d,\, B \in \mathbb{R}^{d \times n},\, B \text{ satisfies } \mathcal{O}\}}} \|Y - W\|_F^2. \quad (190)$$

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 (Bertsekas, 2009, Proposition 1.1.9), a unique minimizer W_0 to the RHS of (190) exists. Therefore, the set of minimizers to the LHS of (190) 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

$$\min_{x \in \mathbb{R}^d} \left\| W_0 - x\mathbf{1}^T \right\|_F^2. \quad (191)$$

It can be verified that a unique solution to (191) 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 (2) is strictly convex in (x, B). Therefore, there exists at most one minimizer (Bertsekas, 2009, 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

$$\|x\|_2^2 + \|B\|_F^2 > R_M \quad (192)$$

Given coercivity, invoking Weierstrass' Theorem (Bertsekas, 2009, Proposition 3.2.1) completes the proof.

We set

$$R_M = d \left[\left(1 + \frac{1}{\sqrt{\lambda}} \right) \sqrt{M} + \max_{i \in [d], j \in [n]} Y \right]^2 + \frac{1}{\lambda} M. \quad (193)$$

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 (15) is lower-bounded as $\lambda \|B\|_F^2 \geq M$. Hence, the objective (2) is at least M.

Case of $\|B\|_F^2 < \frac{M}{\lambda}$: Combining (192) and (193), 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

$$|x_{i^*}| > (1+\frac{1}{\sqrt{\lambda}})\sqrt{M} + \max_{i\in[d],j\in[n]} y_{ij}. \tag{194}$$

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 (194) and the assumption of the case that $\|B\|_F^2 < \frac{1}{\lambda}M$. Hence, the second term in the objective (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 (2) is greater than M.

Combining the two cases depending on $\|B\|_F^2$ completes the proof of the coercivity of the objective (2) in terms of (x, B). Invoking Weierstrass' Theorem (Bertsekas, 2009, 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 (15) 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 (15) exists for general Ω.

C.9.2 Proof of Lemma 16

It is sufficient to prove the general version (19). First consider $\lambda = \infty$. It can be verified that the closed-form expression (3) for the solution at $\lambda = \infty$ satisfies the claimed relation (19).

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 (2) as

$$\min_{x\in\mathbb{R}^d} \|Y - x\mathbf{1}^T - \widehat{B}^{(\lambda)}\|_F^2. \tag{195}$$

Writing out all the terms in (195) and completing the square yields the claimed relation (19).

C.9.3 Proof of Lemma 17

It is sufficient to prove the general version (21). First consider the case of $\lambda = \infty$. It can be verified that the closed-form expression expressions (3) for the solution at $\lambda = \infty$ satisfies the claimed relations (22).

It remains to consider the case of $\lambda \in [0, \infty)$. First we prove (21a). 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

$$\widehat{x}_\gamma = \widehat{x} + \gamma \mathbf{1}_d \tag{196a}$$

$$\widehat{B}_\gamma = \widehat{B} - \gamma \mathbf{1}_d \mathbf{1}_n^T. \tag{196b}$$

Note that the original solution $(\widehat{x}, \widehat{B})$ corresponds to $\gamma = 0$.

Since $\widehat{B}_\gamma$ in (196) 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 (196) 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 (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 (21a).

Now we prove (21b). By (19) from Lemma 16 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 (21a), completing the proof of (21b).

C.9.4 Proof of Proposition 18

First consider the case of $\lambda = \infty$, the claimed result can be verified using the closed-form expressions (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

$$\|(Y + \Delta x \mathbf{1}^T) - (\widehat{x} + \Delta x + u)\mathbf{1}^T - \widehat{B}'\|_\Omega^2 + \lambda\|\widehat{B}'\|_\Omega^2 \leq \|(Y + \Delta x \mathbf{1}^T) - (\widehat{x} + \Delta x)\mathbf{1}^T - \widehat{B}\|_\Omega^2 + \lambda\|\widehat{B}\|_\Omega^2 \tag{197}$$

$$\|Y - (\widehat{x} + u)\mathbf{1}^T - \widehat{B}'\|_\Omega^2 + \lambda\|\widehat{B}'\|_\Omega^2 \leq \|Y - \widehat{x}\mathbf{1}^T - \widehat{B}\|_\Omega^2 + \lambda\|\widehat{B}\|_\Omega^2. \tag{198}$$

If strict inequality in (198) 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 (198) and hence in (197). By the tie-breaking rule of the equality (197) on the observations $(Y + \Delta x \mathbf{1}^T)$, we have

$$\|\widehat{B}'\|_\Omega^2 < \|\widehat{B}\|_\Omega^2, \tag{199}$$

Combining (199) with the equality of (198) 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 $(\widehat{x} + u, \widehat{B}')$.

C.9.5 Proof of Lemma 20

The proof relies on (21b) from Lemma 17. Assume without loss of generality that $x^* = 0$. We first show that on the RHS of (21b), we have that $\sum_{(i,j)\in\Omega^{\mathrm{t}}} y_{ij}$ converges to 0 for random Ω^{t} obtained by Algorithm 1.

Fix some constant $\epsilon_1 > 0$ whose value is determined later.

Part (b): For any fixed Ω^{t}, by Hoeffding's inequality, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{|\Omega^{\mathrm{t}}|}\sum_{(i,j)\in\Omega^{\mathrm{t}}} y_{ij}\right| < \epsilon_1\right) = 1. \tag{200a}$$

Part (a): Given the assumption that $x^* = 0$ and the assumption that there is no noise, we have $Y = B$. By (28b) from Lemma 28, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{|\Omega^{\mathrm{t}}|}\sum_{(i,j)\in\Omega^{\mathrm{t}}} y_{ij}\right| < \epsilon_1\right) = 1. \tag{200b}$$

The rest of the proof is the same for both parts. Denote the event in (200) as E. We now condition on E and consider the LHS of (21b). By (7), 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 (21b), 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

$$\max_{i,i'\in[d]}(\widehat{x}_i - \widehat{x}_{i'}) > \frac{\epsilon}{\sqrt{d}}, \quad \forall\lambda \in \Lambda_\epsilon \;\middle|\; E. \tag{201}$$

Combining (201) with (200), 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.

C.9.6 Proof of Lemma 21

We follow the proof of Lemma 20, we assume $x^* = 0$ without loss of generality. Then fix some constant $\epsilon_1 > 0$, and estalish concentration inequalities on the RHS of (21b).

Part (b): Same as (200b) from Lemma 20, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{|\Omega^{\mathrm{t}}|}\sum_{(i,j)\in\Omega^{\mathrm{t}}} y_{ij}\right| < \epsilon_1\right) = 1. \tag{202a}$$

Part (a): By Hoeffding's inequality, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{dn}\left|\sum_{i\in[d],j\in[n]} y_{ij}\right| < \epsilon_1\right) = 1. \tag{202b}$$

The rest of the proof is the same for both parts. Combining (202) with (21b), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{d}\sum_{i\in[d]} \widehat{x}_i\right| < \epsilon_1\right) = 1. \tag{203}$$

Fix any value $\epsilon > 0$. Denote E as the event that the events in both (23) and (203) hold. By a union bound of (23) and (203), we have

$$\lim_{n\to\infty}(E) = 1. \tag{204}$$

Condition on E and consider the value of $\widehat{x}_1^{(\lambda)}$. First consider the case of $\widehat{x}_1 > \epsilon$, then by (23) we have $\widehat{x}_i > 0$ for each $i \in [d]$. Then

$$\frac{1}{d}\left|\sum_{i\in[d]} \widehat{x}_i\right| = \frac{1}{d}\sum_{i\in[d]} \widehat{x}_i > \frac{\epsilon}{d} \quad \Big| \ \widehat{x}_1 > \epsilon, E$$

A similar argument applies to the case of e $\widehat{x}_1 < -\epsilon$, and we have

$$\frac{1}{d}\left|\sum_{i\in[d]} \widehat{x}_i\right| > \frac{\epsilon}{d} \quad \Big| \ |\widehat{x}_1| > \epsilon, E$$

The same argument applies to each $i \in [d]$. We have

$$\frac{1}{d}\left|\sum_{i\in[d]} \widehat{x}_i\right| > \frac{\epsilon}{d} \quad \Big| \ \|\widehat{x}\|_\infty > \epsilon, E$$

Taking a sufficiently small ϵ_1 such that $\epsilon_1 < \frac{\epsilon}{d}$ in (203) yields a contradiction. Hence, we have

$$\lim_{n\to\infty} \mathbb{P}(\|\widehat{x}\|_\infty > \epsilon, E) = 0. \tag{205}$$

Hence,

$$\lim_{n\to\infty} \mathbb{P}\left(\|\widehat{x}\|_2 > \sqrt{d}\epsilon\right) \le \lim_{n\to\infty} \mathbb{P}\left(\|\widehat{x}\|_\infty > \epsilon\right) \overset{\text{(i)}}{=} \lim_{n\to\infty} \left(\|\widehat{x}\|_\infty > \epsilon, \overline{E}\right) \le \lim_{n\to\infty} \mathbb{P}(\overline{E}) \overset{\text{(ii)}}{=} 0,$$

where inequality (i) is true by (205) and (ii) is true by (204), completing the proof.

C.9.7 Proof of Proposition 22

Without loss of generality we assume $x^* = 0$. By (22b) from Lemma 17 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

$$\widehat{x}_\gamma = \overline{y} + \begin{bmatrix} -1 \\ 1 \end{bmatrix} \cdot \frac{\gamma}{2} \tag{206}$$

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 (2). Hence, at the optimal solution $(\widehat{x}_\gamma, \widehat{B}_\gamma)$, the objective (2) equals 0. At the optimal solution, we have

$$\widehat{B}_\gamma = Y - \widehat{x}_\gamma \mathbf{1}^T. \tag{207}$$

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

$$\begin{cases} y_{11,\max} - \widehat{x}_1 \le y_{22,\min} - \widehat{x}_2 \\ y_{21,\max} - \widehat{x}_2 \le y_{12,\min} - \widehat{x}_1. \end{cases} \tag{208}$$

Plugging the parameterization (206) of $\widehat{x}_\gamma$ into (208), we have

$$y_{21,\max} - y_{12,\min} \le \gamma \le y_{22,\min} - y_{11,\max}. \tag{209}$$

Note that the range in (209) 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 (209) **minimizing** $\|\widehat{B}_\gamma\|_F^2$ Using the parameterizations (206) and (207), we write $\|\widehat{B}_\gamma\|_F^2$ as

$$\begin{aligned} \|\widehat{B}_\gamma\|_F^2 &= \|Y - \widehat{x}_\gamma \mathbf{1}^T\|_F^2 \\ &\overset{\text{(i)}}{=} \sum_{j\in[n]} \left(y_{1j} - \overline{y} + \frac{\gamma}{2}\right)^2 + \sum_{j\in[n]} \left(y_{2j} - \overline{y} - \frac{\gamma}{2}\right)^2. \end{aligned} \tag{210}$$

Writing out the terms in (210) and completing the square, we have that minimizing $\|\widehat{b}_\gamma\|_F^2$ is equivalent to minimizing the term:

$$\frac{n}{2}\left(\gamma - (\overline{y}_2 - \overline{y}_1)\right)^2 \tag{211}$$

Combining (209) and (211) gives the yields expression (24) for the optimal γ.

C.9.8 Proof of Lemma 23

The lemma is a direct consequence of the following result (given that almost-sure convergence implying convergence in probability).

Lemma 48 (Theorem 2 in Deheuvels, 1985)**.** *Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(0,1)$. We have*

$$\limsup_{n\to\infty} \frac{\sqrt{2\log n}}{\log\log n} M_n = 1 \quad \text{almost surely,}$$

where log *is the logarithm of base* 2*.*

C.9.9 Proof of Lemma 24

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 49 (Example 3.9.21 of van der Vaart and Wellner, 1996; Corollary 21.5 of van der Vaart, 1998)**.** *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 50 (Lemma 21.7 in van der Vaart, 1998)**.** *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*

$$\sqrt{n}\left[X^{(k_n:n)} - G_n^{-1}(p)\right] \xrightarrow{P} \frac{c}{g(G^{-1}(p))}$$

C.9.10 Proof of Lemma 26

We consider any fixed $i \in [d], k \in [r]$, and any fixed total ordering π_0 generated by Line 2 of Algorithm 1. 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 1. Then it can be verified from Line 5-7 of Algorithm 1 that

$$\frac{\ell_{ik}}{2} - 1 \le \ell_{ik}^{\mathrm{v}} \le \frac{\ell_{ik}}{2} + 1, \tag{212}$$

It can be verified that (212) along with the assumption that $\ell_{ik} \ge 4$ yields (26a). Summing (26a) over $i \in [d]$ yields (27a). Finally, replacing the validation set Ω^{v} by the training set Ω^{t} in the proof of (26a) and (27a) yields (26b) and (27b), respectively.

C.9.11 Proof of Lemma 27

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 1. According to Line 7 of Algorithm 1, 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.

C.9.12 Proof of Lemma 28

Proof of (28a)**:** 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 1). Recall from Line 4 of Algorithm 1 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 1 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 (7) 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 $n \geq n_1$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n^{\mathrm{t}}}\sum_{j\in\Omega^{\mathrm{t}}_i} b_{ij} - \frac{1}{n}\sum_{j\in[n]} b_{ij}\right| < \Delta_{B^*}\sqrt{\frac{\log n}{n}} \;\middle|\; B = B^*\right) \geq 1 - \frac{\delta}{2}. \tag{213}$$

Now we analyze the term Δ_B. By Lemma 25, we have that there exists n_2 such that for all $n \geq n_2$,

$$\mathbb{P}\Big(\Delta_B \leq 4\sqrt{\log n}\Big) \geq 1 - \frac{\delta}{2}. \tag{214}$$

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 (213) and inequality (ii) is true by (214), completing the proof.

Proof of (28b)**:** By Hoeffding's inequality, we have that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{dn}\left|\sum_{i\in[d], j\in[n]} b_{ij}\right| < \epsilon\right) = 1. \tag{215}$$

Recall from assumption (A3) that d is assumed to be a constant. Taking a union bound of (28a) over $i \in [d]$ and (215), folloed by using the triangle inequality yields the claimed result.

C.10 Proof of auxiliary results for Theorem 5

In this section, we present the proofs of the auxiliary results for Theorem 5.

C.10.1 Proof of Lemma 29

Fix any $c > 0$ and fix any $(i, i') \in S_c$. Suppose $k \in [r]$ satisfies the definition (30) 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 (31) on $\widehat{x}_{i'}^{(0)} - \widehat{x}_i^{(0)}$.

Step 1: Bounding the difference of a pair of bias terms Recall from (13) 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

$$\mathbb{P}(T_k \geq t_0) = \prod_{m=0}^{t_0-1} \frac{\ell_{-i,k} - m}{\ell_k - m} < \left(\frac{\ell_{-i,k}}{\ell_k}\right)^{t_0} \overset{\text{(i)}}{\leq} (1-c)^{t_0}, \tag{216}$$

where step (i) is true by the definition (30) of S_c. Similarly we have

$$\mathbb{P}(T_{k+1} \geq t_0) \leq (1-c)^{t_0}. \tag{217}$$

Taking $t_0 = \frac{\log(\frac{4}{\delta})}{\log(1-c)}$ and taking a union bound of (216) and (217), we have

$$\mathbb{P}\Big(T_k + T_{k+1} < 2t_0\Big) \geq \mathbb{P}\Big(T_k < t_0, T_{k+1} < t_0\Big) \geq 1 - 2(1-c)^{t_0} = 1 - \frac{\delta}{2}. \tag{218}$$

By Lemma 23, there exists n_0 such that for all $n \geq n_0$, we have

$$\mathbb{P}\left(M < \frac{\epsilon}{2t_0+1}\right) > 1 - \frac{\delta}{2}, \tag{219}$$

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 (219) with (218), we have that for all $n \geq n_0$

$$\begin{aligned} b_{i',k+1,\min} - b_{ik,\max} &< [(t+1+T_{k+1}) - (t - T_k) + 1] \cdot M \\ &\leq (2t_0+1)M < \epsilon, \quad \text{with probability at least } 1-\delta. \end{aligned} \tag{220}$$

Due to the assumption of no noise and the assumption of $x^* = 0$, the observation model (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 (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

$$\begin{aligned}\widehat{b}^{(0)}_{i',k+1,\min} - \widehat{b}^{(0)}_{ik,\max} &= (y_{i',k+1,\min} - \widehat{x}^{(0)}_{i'}) - (y_{ik,\max} - \widehat{x}^{(0)}_{i}) \\ &= (b_{i',k+1,\min} - \widehat{x}^{(0)}_{i'}) - (b_{ik,\max} - \widehat{x}^{(0)}_{i}) \geq 0 \end{aligned} \tag{221}$$

Rearranging (221)and combining it with (220), we have that for all $n \geq n_0$,

$$\mathbb{P}\Big(\widehat{x}^{(0)}_{i'} - \widehat{x}^{(0)}_{i} \leq b_{i',k+1,\min} - b_{ik,\max} < \epsilon\Big) \geq 1 - \delta,$$

completing the proof.

C.10.2 Proof of Lemma 31

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 29 on the L pairs in (32) 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

$$\begin{aligned} \widehat{x}_{m_2} - \widehat{x}_{m_1} &< \frac{\epsilon}{d}, \\ \widehat{x}_{m_3} - \widehat{x}_{m_2} &< \frac{\epsilon}{d}, \\ &\vdots \\ \widehat{x}_{m_L} - \widehat{x}_{m_{L-1}} &< \frac{\epsilon}{d}, \\ \widehat{x}_{m_1} - \widehat{x}_{m_L} &< \frac{\epsilon}{d}. \end{aligned} \tag{222}$$

Consider any $m < m'$ with $m, m' \in [L]$. Conditional on (222) we have

$$\widehat{x}_{i_{m'}} - \widehat{x}_{i_m} = (\widehat{x}_{i_{m'}} - \widehat{x}_{i_{m'-1}}) + \ldots + (\widehat{x}_{i_{m+1}} - \widehat{x}_{i_m}) < \epsilon. \tag{223}$$

On the other hand, conditional on (222) we also have

$$\widehat{x}_{i_m} - \widehat{x}_{i_{m'}} = (\widehat{x}_{i_m} - \widehat{x}_{i_{m-1}}) + \ldots + (\widehat{x}_{i_2} - \widehat{x}_{i_1}) + (\widehat{x}_{i_1} - \widehat{x}_{i_L}) + \ldots + (\widehat{x}_{i_{m'+1}} - \widehat{x}_{i_{m'}}) < \epsilon \tag{224}$$

Combining (223) and (224), 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.

C.10.3 Proof of Lemma 32

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'.

C.10.4 Proof of Lemma 33

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 (44), we have $t < t'$. We consider the courses corresponding to the elements of ranks t through t', denoted as $(i_t, \dots, i_{t'})$. For any integer $k \in \{t, \dots, t'-1\}$ if $i_k \neq i_{k+1}$, then by the definition of S_c from (30) 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 $\{i_k \to i_{k+1}\}_{k \in \{t,\dots,t'-1\}: i_k \neq i_{k+1}\}}$ 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 $V = \{i, i+1, \dots, i'\}$. 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 32 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.

C.11 Proof of auxiliary results for Theorem 9

In this section, we present the proofs of the auxiliary results for Theorem 9.

C.11.1 Proof of Theorem 34

The proof closely follows part (a) and part (c) of Theorem 5 (see Appendix C.3). Therefore, we outline the modifications to the proof of Theorem 5, in order to extend to any $\Omega^{\mathrm{t}} \subseteq [d] \times [n]$ obtained by Algorithm 1.

Proof Theorem 34(a) The proof closely follows the proof of Theorem 5(a) (see Appendix C.3.1) with the modifications discussed in what follows.

Extending S_c to S_c^{t} Recall from (9) 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 (30) 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 29 to S_c^{t} restricted to the training set Ω^{t} We show that Lemma 29 holds for any $(i,i') \in S_c^{\mathrm{t}}$, and the estimator (15) $\widehat{x}^{(0)}$ restricted to Ω^{t}.

Denote $b_{ik,\max}^{\mathrm{t}}$ as the largest bias in course i of group k restricted to the training set Ω^{t}, and denote $b_{k,\max}^{\mathrm{t}}$ as the largest bias of group k restricted to the training set Ω^{t}. We extend (216) to show that the difference between the ranks of $b_{ik,\max}^{\mathrm{t}}$ and $b_{k,\max}^{\mathrm{t}}$ is bounded by some constant with high probability.

Moreover, it can be verified that the difference between the ranks of $b_{k,\max}^{\mathrm{t}}$ and $b_{k,\max}$ is bounded by a constant with high probability. Combining these two bounds, the difference between the ranks of $b_{ik,\max}^{\mathrm{t}}$ and $b_{k,\max}$ is bounded by a constant with high probability. We define $b_{i'k+1,\min}^{\mathrm{t}}$ and $b_{k+1,\min}$ likewise, and extend (217) to show that the difference between the ranks of $b_{i'k+1,\min}^{\mathrm{t}}$ and $b_{k+1,\min}$ is bounded by a constant with high probability. Therefore, we extend 220 to:

$$b_{i'k+1,\min}^{\mathrm{t}} - b_{ik,\max}^{\mathrm{t}} < \epsilon, \quad \text{with probability at least } 1-\delta.$$

Following the rest of the original arguments for Lemma 29 (see Appendix C.10) completes the extension of Lemma 29 to being restricted to Ω^{t}.

Extending Lemma 31 to S_c^{t} restricted to Ω^{t} We replace the set S_c in Lemma 31 by the set S_c^{t}. It can be verified that Lemma 31 holds under this extension following its original proof (see Appendix C.10).

Extending the rest of the arguments For any $i \in [d], k \in [r]$, by (26b) and (27b) from Lemma 26 we have

$$\frac{\ell_{ik}^{\mathrm{t}}}{\ell_k^{\mathrm{t}}} \geq \frac{\frac{\ell_{ik}}{4}}{\frac{3\ell_k}{4}} = \frac{\ell_{ik}}{3\ell_k}.$$

Hence, any $(i,i') \in S_{\frac{c_{\mathrm{f}}}{d}}$, we have $(i,i') \in S_{\frac{c_{\mathrm{f}}}{3d}}^{\mathrm{t}}$. The rest of the arguments follow from the original proof of Theorem 5(a) (see Appendix C.3.1).

Proof of Theorem 34(b) The proof closely follows the proof of Theorem 5(c) (see Appendix C.3.3) with the modifications discussed in what follows.

Extending S_c to $S_c^{\mathrm{t}'}$ 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 (30) of S_c and define

$$S^{\mathrm{t}'} := \left\{ \begin{array}{rl} (i,i') \in [d]^2 : \exists 1 \le k < k' \le |\Omega^{\mathrm{t}}| & \\ \text{such that} & \text{the element of rank } k \text{ is in } \Omega_i^{\mathrm{t}}, \\ & \text{the element of rank } k' \text{ is in } \Omega_{i+1}^{\mathrm{t}}, \\ & \text{the elements of ranks } (k+1) \text{ through } (k'-1) \text{ are in } \Omega^{\mathrm{v}} \end{array} \right\}. \tag{225}$$

Extending Lemma 29 By Lemma 27(a) we have that for any $(i, i') \in S^{\mathrm{t}'}$, the corresponding values of k and k' in (225) satisfy $k' - k \le 2d + 1$. We define M' as the maximal difference between elements that are adjacent within Ω^{t}. Then by Lemma 23 we extend the bound of M in (219) to M' as

$$\mathbb{P}\left(M' < \epsilon\right) > 1 - \frac{\delta}{2}.$$

Following the rest of the arguments in Appendix C.10.1, we have that Lemma 29 holds restricted to the training set Ω^{t}.

Extending Lemma 31 to S_c^{t} restricted to Ω^{t} We replace the set S_c in Lemma 31 by the set $S^{\mathrm{t}'}$. It can be verified that Lemma 31 holds under this extension following its original proof (see Appendix C.10).

Extending the rest of the arguments The rest of the arguments follow from the original proof of Theorem 5(c) (see Appendix C.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 (60) of $(\widehat{b}_{\mathrm{L}}, \widehat{b}_{\mathrm{H}})$ to $(\widehat{b}'_{\mathrm{L}}, \widehat{b}'_{\mathrm{H}})$ defined as:

$$\widehat{b}'_{\mathrm{L}} := \frac{1}{\sum_{i \in V_{\mathrm{L}}} |\Omega_i^{\mathrm{t}}|} \sum_{i \in V_{\mathrm{L}}} \sum_{j \in \Omega_i^{\mathrm{t}}} \widehat{b}_{ij}$$
$$\widehat{b}'_{\mathrm{H}} := \frac{1}{\sum_{i \in V_{\mathrm{H}}} |\Omega_i^{\mathrm{t}}|} \sum_{i \in V_{\mathrm{H}}} \sum_{j \in \Omega_i^{\mathrm{t}}} \widehat{b}_{ij}.$$

C.11.2 Proof of Lemma 35

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 1. 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 1 that the interpolated bias is computed as:

$$\widetilde{B}^{(\lambda)} = \frac{1}{|\mathcal{T}|} \sum_{\pi \in \mathcal{T}} \widetilde{B}_{\pi}^{(\lambda)}, \tag{226}$$

where recall from Line 13 of Algorithm 1 that $[\widetilde{B}^{(\lambda)}_{\pi}]_{ij}$ 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

$$[\widetilde{B}^{(\lambda)}_{\pi}]_{ij} = \frac{1}{|\mathrm{NN}(i,j;\pi)|} \sum_{(i^\pi, j^\pi) \in \mathrm{NN}} \widehat{B}^{(\lambda)}_{i^\pi j^\pi}. \tag{227}$$

Plugging (227) to (226), we have

$$\widetilde{B}^{(\lambda)}_{ij} = \frac{1}{|\mathcal{T}|} \sum_{\pi \in \mathcal{T}} \frac{1}{|\mathrm{NN}(i,j;\pi)|} \sum_{(i^\pi, j^\pi) \in \mathrm{NN}} \widehat{B}^{(\lambda)}_{i^\pi j^\pi}.$$

The remaining of the proof is outlined as follows. We decompose the summation over $\pi \in \mathcal{T}$ on the RHS of (226) 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,

$$\frac{1}{|\mathcal{T}_{\mathrm{in}}|} \sum_{\pi \in \mathcal{T}_{\mathrm{in}}} [\widetilde{B}^{(\lambda)}_{\pi}]_{ij} = \frac{1}{\left|G^{\mathrm{t}}_k\right|} \sum_{(i',j') \in G^{\mathrm{t}}_k} \widehat{b}^{(\lambda)}_{i'j'} \overset{\text{(i)}}{=} \widehat{b}^{\mathrm{t}}_k, \tag{228}$$

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:

$$\prod_{i=1}^{t-1} \frac{\ell_k^{\mathrm{v}} - i}{\ell_k - i} \leq \left(\frac{\ell_k^{\mathrm{v}}}{\ell_k}\right)^{t-1} \overset{\text{(i)}}{<} \left(\frac{3}{4}\right)^t,$$

where (i) is true due to (27a) from Lemma 26. 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:

$$\frac{1}{\ell_k} \sum_{t=1}^{\ell_k} \left(\frac{3}{4}\right)^t \leq \frac{3}{\ell_k} \overset{\text{(i)}}{<} \frac{3}{dc_{\mathrm{f}}n},$$

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

$$\frac{|\mathcal{T}_{\mathrm{out}}|}{|\mathcal{T}|} < \frac{6}{dc_{\mathrm{f}}n}. \tag{229}$$

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 (228). Hence, we have

$$\begin{aligned}
\left|\widetilde{b}_k - \widehat{b}_k^{\mathrm{t}}\right| &= \frac{1}{|\mathcal{T}|}\left|\sum_{\pi\in\mathcal{T}_{\mathrm{out}}} [\widetilde{B}_\pi^{(\lambda)}]_{ij} - \widehat{b}_k^{\mathrm{t}}\right| \\
&\leq \frac{1}{|\mathcal{T}|}\sum_{\pi\in\mathcal{T}_{\mathrm{out}}} \left(\left|[\widetilde{B}_\pi^{(\lambda)}]_{ij}\right| + \left|\widehat{b}_k^{\mathrm{t}}\right|\right) \\
&\overset{\text{(i)}}{\leq} \frac{2|\mathcal{T}_{\mathrm{out}}|}{|\mathcal{T}|} \max_{i\in[d],j\in[n]} \left|\widehat{b}_{ij}\right| \overset{\text{(ii)}}{\leq} \frac{12}{c_{\mathrm{f}}dn} \cdot \max_{i\in[d],j\in[n]} \left|\widehat{b}_{ij}\right|,
\end{aligned}$$

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 $\left|[\widetilde{B}_\pi^{(\lambda)}]_{ij}\right| \leq \max_{i\in[d],j\in[n]} \left|\widehat{b}_{ij}\right|$ and $\left|\widehat{b}_k^{\mathrm{t}}\right| \leq \max_{i\in[d],j\in[n]} \left|\widehat{b}_{ij}\right|$. Then step (ii) is true by plugging in (229). This completes the proof.

C.11.3 Proof of Corollary 36

Fix any $\epsilon > 0$. By the consistency of $\widehat{B}^{(0)}$ from (69), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{B}_{ij}^{(0)} - B_{ij}\right| < \frac{\epsilon}{2}, \quad \forall (i,j) \in \Omega^{\mathrm{t}}\right) = 1. \tag{230}$$

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

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|\widehat{b}_k^{\mathrm{t}} - b_k^{\mathrm{t}}\Big| < \frac{\epsilon}{2}, \quad \forall k \in [r]\Big) = 1. \tag{231}$$

For each $k \in [r]$, we have

$$\begin{aligned}
\Big|\widetilde{b}_k - b_k^{\mathrm{t}}\Big| &\le \Big|\widetilde{b}_k - \widehat{b}_k^{\mathrm{t}}\Big| + \Big|\widehat{b}_k^{\mathrm{t}} - b_k^{\mathrm{t}}\Big| \\
&\overset{\text{(i)}}{\le} \frac{12}{c_{\mathrm{f}} dn} \cdot \max_{i\in[d],j\in[n]} \Big|\widehat{b}_{ij}\Big| + \Big|\widehat{b}_k^{\mathrm{t}} - b_k^{\mathrm{t}}\Big| \\
&\le \frac{12}{c_{\mathrm{f}} dn} \left(\max_{i\in[d],j\in[n]} |b_{ij}| + \max_{i\in[d],j\in[n]} \Big|b_{ij} - \widehat{b}_{ij}\Big| \right) + \Big|\widehat{b}_k^{\mathrm{t}} - b_k^{\mathrm{t}}\Big|,
\end{aligned} \tag{232}$$

where (i) is true by combining Lemma 35. In (232), we bound the term $\max_{i\in[d],j\in[n]}|b_{ij}|$ by Lemma 25 as

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i\in[d],j\in[n]} |b_{ij}| < 2\sqrt{\log dn} \right) = 1. \tag{233}$$

We bound the term $\max_{i\in[d],j\in[n]} \Big|b_{ij} - \widehat{b}_{ij}\Big|$ by (230), and the term $\Big|\widehat{b}_k^{\mathrm{t}} - b_k^{\mathrm{t}}\Big|$ by (231). Hence, plugging (233), (230) and (231) into (232), we have

$$\lim_{n\to\infty} \mathbb{P}\left(\Big|\widetilde{b}_k - b_k^{\mathrm{t}}\Big| \le \frac{12}{c_{\mathrm{f}} dn} \left(2\sqrt{\log dn} + \frac{\epsilon}{2} \right) + \frac{\epsilon}{2}, \quad \forall k \in [r] \right) = 1.$$

Equivalently,

$$\lim_{n\to\infty} \mathbb{P}\left(\Big|\widetilde{b}_k - b_k^{\mathrm{t}}\Big| \le \epsilon, \quad \forall k \in [r] \right) = 1,$$

completing the proof.

C.11.4 Proof of Lemma 37

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_{ik}^{v} (whose mean is b_{ik}^{v}) can be considered as randomly sampling ℓ_{ik}^{v} 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 (Hoeffding, 1963, Section 6), we have

$$\mathbb{P}\left(|b_{ik}^{\mathrm{v}} - b_k^*| > \Delta_{B^*} \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\ell_{ik}^{\mathrm{v}}}} \;\middle|\; B = B^* \right) \le 2\exp\left(-\frac{2\ell_{ik}^{\mathrm{v}} \Delta_{B^*}^2 \log\left(\frac{1}{\delta}\right)}{\ell_{ik}^{\mathrm{v}} \Delta_{B^*}^2} \right) = 2\delta^2 \overset{\text{(i)}}{<} \frac{\delta}{2}, \tag{234}$$

where inequality (i) is true for any $\delta \in (0, \frac{1}{4})$. Invoking (26a) from Lemma 26 and using the all c_{f}-fraction assumption, we have

$$\ell_{ik}^{\mathrm{v}} \ge \frac{\ell_{ik}}{4} > \frac{c_{\mathrm{f}} n}{4}. \tag{235}$$

Combining (234) with (235), we have that for any $\delta \in (0, \frac{1}{4})$,

$$\mathbb{P}\left(|b_{ik}^{\mathrm{v}} - b_k^*| > 2\Delta_{B^*} \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{c_{\mathrm{f}} n}} \;\middle|\; B = B^* \right) < \frac{\delta}{2}. \tag{236}$$

Now we analyze the term Δ_B in (236). By Lemma 25, there exists integer n_0 such that for any $n \geq n_0$,

$$\mathbb{P}\left(\Delta_B \leq 4\sqrt{\log dn} \right) \geq 1 - \frac{\delta}{2}. \tag{237}$$

Let n_1 be a sufficiently large constant such that $n_1 \geq n_0$ and $8\sqrt{\log dn} \cdot \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{c_{\mathrm{f}} n}} < \epsilon$. Then combining (237) with (236), for any $n \geq n_1$,

$$\begin{aligned}
\mathbb{P}\Big(|b_{ik}^{\mathrm{v}} - b_k| < \epsilon \Big) &= \int_{B^* \in \mathbb{R}^{d \times n}} \mathbb{P}\Big(|b_{ik}^{\mathrm{v}} - b_k| < \epsilon \;\Big|\; B = B^* \Big) \cdot \mathbb{P}(B^*) \, \mathrm{d}B^* \\
&\geq \int_{\substack{B^* \in \mathbb{R}^{d \times n} \\ \Delta_{B^*} \leq 4\sqrt{\log dn}}} \mathbb{P}\Big(|b_{ik}^{\mathrm{v}} - b_k| < \epsilon \;\Big|\; B \Big) \cdot \mathbb{P}(B) \, \mathrm{d}B^* \\
&\overset{\text{(i)}}{\geq} \left(1 - \frac{\delta}{2}\right) \cdot \mathbb{P}\Big(\Delta_B \leq \sqrt{4 \log dn} \Big) \\
&\overset{\text{(ii)}}{\geq} \left(1 - \frac{\delta}{2}\right)^2 \geq 1 - \delta,
\end{aligned}$$

where inequality (i) is true by (236) and inequality (ii) is true by (237). Equivalently, we have

$$\lim_{n \to \infty} \mathbb{P}\Big(|b_{ik}^{\mathrm{v}} - b_k| < \epsilon \Big) = 1. \tag{238}$$

Due to the all c-fraction assumption, the number of groups is upper-bounded as $r \leq \frac{1}{c_{\mathrm{f}}}$. Taking a union bound of (238) over $i \in [d], k \in [r]$, we have

$$\lim_{n \to \infty} \mathbb{P}\Big(|b_{ik}^{\mathrm{v}} - b_k| < \epsilon, \quad \forall i \in [d], k \in [r] \Big) = 1,$$

completing the proof of (75a). A similar argument yields (75b), where in (235) we invoke (27b) from Lemma 26 instead of (26a).

C.11.5 Proof of Lemma 39

In the proof, we use the following lemma.

Lemma 51. *Let $d \geq 1$ be an integer. For any $y \in \mathbb{R}^d$, we have*

$$\arg\min_{u \in \mathcal{M}} \|y - u\|_2^2 + \lambda \|u\|_2^2 = \arg\min_{u \in \mathcal{M}} \|\Pi_{\mathcal{M}}(y) - u\|_2^2 + \lambda \|u\|_2^2 \tag{239}$$

Algorithm 2: 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

The proof of Lemma 51 is presented at the end of this section. We now derive a the closed-form solution to (239). Consider the optimization problem on the RHS of (239). 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 (239) is:

$$u^*_{\text{un}} = \frac{1}{1+\lambda}\Pi_{\mathcal{M}}(y). \tag{240}$$

Note that this unconstrained solution u^*_{un} satisfies $u^*_{\text{un}} \in \mathcal{M}$, so u^*_{un} is also the (constrained) solution to (239). Plugging (240) to the objective on the LHS of (239) and rearranging the terms complete the proof.

Proof of Lemma 51 We apply induction on the Pool-Adjacent-Violators algorithm (PAVA) (Barlow et al., 1972, Section 1.2). For completeness, the Pool-Adjacent-Violators algorithm is shown in Algorithm 2. For any integer $d \geq 1$ and any input $y \in \mathbb{R}^d$, PAVA returns $\arg\min_{u \in \mathcal{M}} \|y - u\|_2^2$.

Assume that the while loop in Algorithm 2 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 2. We have $u^{(0)} = y$ and $u^{(T)} = \Pi_{\mathcal{M}} y$. In what follows, we show that for any $0 \leq t \leq T-1$,

$$\arg\min_{u \in \mathcal{M}} \|u^{(t)} - u\|_2^2 + \lambda\|u\|_2^2 = \arg\min_{u \in \mathcal{M}} \|u^{(t+1)} - u\|_2^2 + \lambda\|u\|_2^2. \tag{241}$$

By induction on (241), we have

$$\arg\min_{u \in \mathcal{M}} \|u^{(0)} - u\|_2^2 + \lambda\|u\|_2^2 = \arg\min_{u \in \mathcal{M}} \|u^{(T)} - u\|_2^2 + \lambda\|u\|_2^2. \tag{242}$$

Combining (242) with the fact that $u^{(0)} = y$ and $u^{(T)} = \Pi_{\mathcal{M}} y$ completes the proof.

Proof of (241)**:** Consider any t such that $0 \leq t \leq T-1$. We consider Line 4-6 of PAVA in Algorithm 2. 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 $1 \le m_1 \le i < m_2 \le d$, such that

$$S = \{m_1, \dots, i\}$$
$$S' = \{i+1, \dots, m_2\}.$$

Furthermore, by PAVA it can be verified that

$$a := u_i^{(t)} = u_{i'}^{(t)} \qquad \forall i, i' \in S \tag{243a}$$
$$b := u_i^{(t)} = u_{i'}^{(t)} \qquad \forall i, i' \in S' \tag{243b}$$
$$z := u_i^{(t+1)} = u_{i'}^{(t+1)} \qquad \forall i, i' \in S \cup S'. \tag{243c}$$

Denote these values in (243) as a, b and z, respectively. By the update of u in Line 6 of Algorithm 2, we have the relation

$$z = \frac{1}{|S| + |S'|} \left(|S| \cdot a + \left|S'\right| \cdot b \right). \tag{244}$$

Denote $u^{*(t)}$ and $u^{*(t+1)}$ as the minimizer to the LHS and RHS of (241), respectively. Using (243), it can be verified that

$$a^* := u_i^{*(t)} = u_{i'}^{*(t)} \qquad \forall i, i' \in S \tag{245a}$$
$$b^* := u_i^{*(t)} = u_{i'}^{*(t)} \qquad \forall i, i' \in S' \tag{245b}$$
$$u_i^{*(t+1)} = u_{i'}^{*(t+1)} \qquad \forall i, i' \in S \cup S'. \tag{245c}$$

Denote the values in (245a) and (245b) 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 $a^* \le b^*$. Hence, we have $a^* < b^*$. By Line 4 of Algorithm 2, 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 (241). Contradiction to the assumption that $u^{*(t)}$ is the minimizer to the LHS of (241). Hence, we have $a^* = b^*$, implying

$$u_i^{*(t)} = u_{i'}^{*(t)} \qquad \forall i, i' \in S \cup S'.$$

The LHS of (241) is equivalent to

$$\begin{aligned} &\operatorname*{arg\,min}_{\substack{u \in \mathcal{M}, t \in \mathbb{R} \\ t = u_i,\, \forall i, i' \in S \cup S'}} \sum_{i \notin S \cup S'} (u_i^{(t)} - x_i)^2 + \sum_{i \in S \cup S'} (u_i^{(t)} - x_i)^2 + \lambda \|u\|_2^2 \\ &\operatorname*{arg\,min}_{\substack{u \in \mathcal{M} \\ t = u_i,\, \forall i, i' \in S \cup S'}} \sum_{i \notin S \cup S'} (u_i^{(t)} - x_i)^2 + \underbrace{|S| \cdot (a-t)^2 + \left|S'\right| \cdot (b-t)^2}_{T} + \lambda \|u\|_2^2. \end{aligned} \tag{246}$$

We write the term T as

$$\begin{aligned} T &= |S| \cdot a^2 + \left|S'\right| \cdot b^2 - 2\left(|S| \cdot a + \left|S'\right| \cdot b\right) \cdot t + (|S| + \left|S'\right|) \cdot t^2 \\ &= (|S| + \left|S'\right|) \cdot \left(\frac{|S| \cdot a + |S'|b}{|S| + |S'|} - t\right)^2 + \text{term}(a, b, S, S') \\ &\overset{\text{(i)}}{=} (|S| + \left|S'\right|) \cdot (z - t)^2 + \text{term}(a, b, S, S'), \end{aligned} \tag{247}$$

where equality (i) is true by (244).

Using the relation $u_i^{(t)} = u_i^{(t+1)}$ for every $i \notin S \cup S'$, the RHS of (241) is equivalent to

$$\begin{aligned} &\underset{\substack{u \in \mathcal{M}, t \in \mathbb{R} \\ t = u_i,\ \forall i \in S \cup S'}}{\arg\min} \sum_{i \notin S \cup S'} (u_i^{(t+1)} - x_i)^2 + \sum_{i \in S \cup S'} (u_i^{(t+1)} - x_i)^2 + \lambda \|u\|_2^2 \\ &\underset{\substack{u \in \mathcal{M}, t \in \mathbb{R} \\ t = u_i,\ \forall i \in S \cup S'}}{\arg\min} \sum_{i \notin S \cup S'} (u_i^{(t)} - x_i)^2 + (|S| + \left|S'\right|) \cdot (z - t)^2 + \lambda \|u\|_2^2. \end{aligned} \tag{248}$$

The equivalence of the LHS and RHS of (241) can be verified by combining (246), (247), and (248).

C.11.6 Proof of Lemma 40

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 52. *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 40 given Lemma 52. The proof of Lemma 52 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 1, consider the elements $(1, j^{(2t-1)})$ and $(1, j^{(2t)})$ in Line 6 of Algorithm 1. 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 1. 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

$$\lim_{n\to\infty} \mathbb{P}\left(\left| S'' \cap \Omega^{\mathrm{v}} \right| \geq \frac{c'n}{32} \;\middle|\; E_{c',c} \right) = 1.$$

That is,

$$\lim_{n\to\infty} \mathbb{P}\left(E^{\mathrm{v}}_{\frac{c'}{32},c} \;\middle|\; E_{c',c} \right) = 1. \tag{249}$$

Hence, we have

$$\begin{aligned}
\mathbb{P}(E^{\mathrm{v}}_{\frac{c'}{32},c} \cap E_2) &\geq \mathbb{P}(E^{\mathrm{v}}_{\frac{c'}{32},c} \cap E_{c',c} \cap E_2) \\
&= \mathbb{P}(E_{c',c} \cap E_2) - \mathbb{P}(\overline{E^{\mathrm{v}}_{\frac{c'}{32},c}} \cap E_{c',c} \cap E_2) \\
&\geq \mathbb{P}(E_{c',c} \cap E_2) - \mathbb{P}(\overline{E^{\mathrm{v}}_{\frac{c'}{32},c}} \cap E_{c',c}). \qquad (250)
\end{aligned}$$

Taking the limit of $n \to \infty$ in (250), we have

$$\lim_{n\to\infty} \mathbb{P}(E^{\mathrm{v}}_{\frac{c'}{32},c} \cap E_2) \overset{\text{(i)}}{\geq} \lim_{n\to\infty} \mathbb{P}(E_2),$$

where inequality (i) is true by combining Lemma 52 and (249), completing the proof of Lemma 40. It remains to prove Lemma 52.

Proof of Lemma 52 Recall the definition (109) 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 53. *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 C.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 53 implies that for any $\gamma \in \left(0, \frac{1}{2}\right)$ and any $c' \in (0, 1)$,

$$\lim_{n\to\infty} \mathbb{P}\Big(E_{\gamma,c',c} \cap \overline{E_{c',2c}} \Big) = 0. \tag{251}$$

In what follows, we establish that there exists $\gamma > 0$ and $c' > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\Big(\overline{E_{\gamma,c',c}} \cap E_2 \Big) = 0, \tag{252}$$

where the choices of γ and c' are specified later.

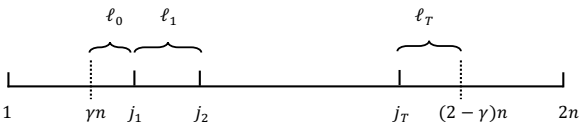


Figure 8: The definition (255) of ℓ.

Proof of (252)**:** 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

$$t < \min\left\{\frac{cg(\frac{\gamma}{2})}{2}, \gamma\right\} \cdot n. \tag{253}$$

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

$$T \leq n - t. \tag{254}$$

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

$$\ell_i = \begin{cases} j_1 - \gamma n - 1 & \text{if } i = 0 \\ j_{i+1} - j_i - 1 & \text{if } 1 \leq i \leq T-1 \\ (2-\gamma)n - j_i & \text{if } i = T. \end{cases} \tag{255}$$

The definition (255) of ℓ is also visualized in Fig. 8.

By in the definition of (255), 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 (254).

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

$$\sum_{i=0}^{T-1} \max\left\{\ell_i - cg\left(\frac{\gamma}{2}\right)n, 0\right\} + \ell_T \geq (1-2\gamma)n - t. \tag{256}$$

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 $t \geq \frac{cg\left(\frac{\gamma}{2}\right)n}{2}$. Contradiction to the assumption (253).

Case 2: $\ell_i < cg\left(\frac{\gamma}{2}\right)n$ for all $0 \leq i \leq T-1$. Then inequality (256) reduces to

$$\ell_T \geq (1-2\gamma)n - t \overset{\text{(i)}}{\geq} (1-3\gamma)n, \tag{257}$$

where inequality (i) is true by the assumption (253) 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 (253), 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 (255), and inequality (ii) is true by (257). We consider the number of elements from course 2 in this range, remaining or removed. By the definition of j_T from (255) 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 (253). 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.

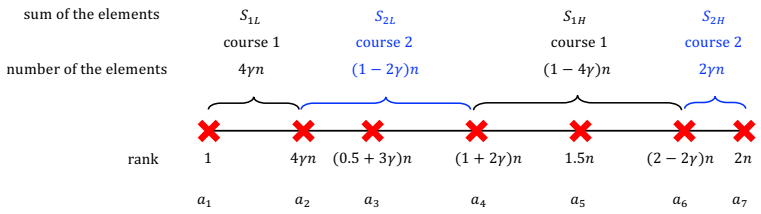


Figure 9: Assignment of biases to the 2 courses.

- **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. 9. We denote $S_{1L}, S_{2L}, S_{1H}, S_{2H}$ respectively as the sums of the subset of elements as shown in Fig. 9.

The following lemma now bounds the difference between the sums of the bias in the two courses, under this construction.

Lemma 54. *Consider $2n$ i.i.d. samples from $\mathcal{N}(0,1)$, ordered as $X^{(1)} \le \ldots \le X^{(2n)}$. 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 C.11.8. Denote the constant γ in Lemma 54 as γ_0. By Lemma 54, we have that under the assumption (253) of $t < \min\left\{ \frac{cg(\frac{\gamma_0}{2})}{2}, \gamma_0 \right\} n$, then

$$\lim_{n\to\infty} \mathbb{P}\left(\sum_{j\in[n]} (b_{2j} - b_{1j}) < 0 \right) = 1.$$

Equivalently, let $c_0' = \min\left\{ \frac{cg(\frac{\gamma_0}{2})}{\gamma_0} \right\}$, 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 (252).

Combining (251) **and** (252)**:** We have

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}\left(E_{c_0',c}\cap E_2\right) &= \mathbb{P}(E_2)-\mathbb{P}(E_2\cap\overline{E_{c',c}})\\ &= \mathbb{P}(E_2)-\mathbb{P}(E_2\cap\overline{E_{c',c}})\\ &= \mathbb{P}(E_2)-\mathbb{P}(E_2\cap\overline{E_{c',c}}\cap E_{\gamma_0,c_0',c})-\mathbb{P}(E_2\cap\overline{E_{c_0',c}}\cap\overline{E_{\gamma_0,c_0',c}}). \quad (258)\end{aligned}$$

Taking the limit of $n\to\infty$ in (258), 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 (251) and (252). This completes the proof of Lemma 52.

C.11.7 Proof of Lemma 53

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 24 we have

$$b^{(t_in:2n)}\xrightarrow{P}G^{-1}\left(\frac{t_i}{2}\right)\qquad\forall 0\le i\le T. \quad (259)$$

Taking a union bound of (259) over $0\le i\le T$, we have

$$\lim_{n\to\infty}\Bigg(\underbrace{\left|b^{(t_in:2n)}-G^{-1}\left(\frac{t_i}{2}\right)\right|<\frac{c}{2}\quad\forall 0\le i\le T}_{E}\Bigg)=1. \quad (260)$$

Denote this event in (260) 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 $t_in\le t_{1j}<t_{2j'}\le t_{i+2}n$. Conditional on the event E from (260), for any $(j,j')\in R_{\gamma,c}$,

$$\begin{aligned}b_{2j'}-b_{1j}\le b^{(t_{i+2}n:2n)}-b_{(t_in:2n)} &< G^{-1}\left(\frac{t_{i+2}}{2}\right)-G^{-1}\left(\frac{t_i}{2}\right)+c\\ &< \frac{(t_{i+2}-t_i)}{2}\cdot\max_{x\in(\frac{\gamma}{2},1-\frac{\gamma}{2})}(G^{-1})'(x)+c\\ &\overset{\text{(i)}}{=} cg\left(\frac{\gamma}{2}\right)\cdot\max_{x\in\left(\frac{\gamma}{2},1-\frac{\gamma}{2}\right)}\frac{1}{g(x)}+c\\ &= cg\left(\frac{\gamma}{2}\right)\cdot\frac{1}{g\left(\frac{\gamma}{2}\right)}+c=2c\Bigg|E.\end{aligned}$$

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 (260), completing the proof.

C.11.8 Proof of Lemma 54

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. 9. 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

$$\lim_{n\to\infty} \mathbb{P}\left(\left|S_{1L} - 4\gamma n \mu_{1L|a}\right| < (a_7 - a_1)\sqrt{n\log n} \;\middle|\; a\right) = 1 \tag{261a}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\left|S_{2L} - (1-2\gamma) n \mu_{2L|a}\right| < (a_7 - a_1)\sqrt{n\log n} \;\middle|\; a\right) = 1 \tag{261b}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\left|S_{1H} - (1-4\gamma) n \mu_{1H|a}\right| < (a_7 - a_1)\sqrt{n\log n} \;\middle|\; a\right) = 1 \tag{261c}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\left|S_{2H} - 2\gamma n \mu_{2H|a}\right| < (a_7 - a_1)\sqrt{n\log n} \;\middle|\; a\right) = 1. \tag{261d}$$

Taking a union bound of (261) 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

$$T = (1-4\gamma)(\mu_{2L|a} - \mu_{1H|a}) + 4\gamma(\mu_{2H|a} - \mu_{1L|a}) + 2\gamma(\mu_{2L|a} - \mu_{2H|a}) + 4(a_7 - a_1)\sqrt{\frac{\log n}{n}}. \tag{262}$$

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

$$A := \left\{ \begin{array}{l} a_1 < G^{-1}(1.5\gamma) \\ a_2 > G^{-1}(1.99\gamma) \\ a_3 < G^{-1}(0.25+1.5\gamma)+0.01 \\ a_5 > G^{-1}(0.75)-0.01 \\ a_6 < G^{-1}(1-0.99\gamma) \\ a_7 > G^{-1}(1-0.5\gamma) \end{array} \right\} \cap \left\{ \begin{array}{l} a_1 > -2\sqrt{\log 2n} \\ a_7 < 2\sqrt{\log 2n} \end{array} \right\}. \tag{263}$$

By Lemma 24, we have

$$a_2 \xrightarrow{P} G^{-1}(2\gamma) \tag{264a}$$
$$a_3 \xrightarrow{P} G^{-1}(0.25+1.5\gamma) \tag{264b}$$
$$a_5 \xrightarrow{P} G^{-1}(0.75) \tag{264c}$$
$$a_6 \xrightarrow{P} G^{-1}(1-\gamma). \tag{264d}$$

Moreover, for the extremal values a_1 and a_7, we have that for any $c\in\mathbb{R}$,

$$\lim_{n\to\infty}\mathbb{P}(a_1<c)=1 \tag{265a}$$
$$\lim_{n\to\infty}\mathbb{P}(a_7>c)=1. \tag{265b}$$

Combining (264), (265) and Lemma 25, 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 (262).

Term $(\mu_{2L|a}-\mu_{1H|a})$: We have $\mu_{2L}\le\frac{a_3+a_4}{2}$ and $\mu_{1H}\ge\frac{a_4+a_5}{2}$. Therefore, conditional on any $a\in A$, for any $\gamma<0.1$,

$$\mu_{2L|a}-\mu_{1H|a}\le\frac{a_3-a_5}{2}\overset{\text{(i)}}{\le}-0.5, \tag{266}$$

where inequality (i) is true by the definition (263) of A.

Term $(\mu_{2H}-\mu_{1L})$: Let X denote a random variable of $\mathcal{N}(0,1)$. Conditional on any $a\in A$,

$$\begin{aligned}
\mu_{2H|a} &= \frac{1}{\sqrt{2\pi}}\frac{1}{\mathbb{P}\left(a_6<X<a_7\right)}\int_{a_6}^{a_7} xe^{-\frac{x^2}{2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}}\frac{1}{\mathbb{P}\left(a_6<X<a_7\right)}\left[-e^{-\frac{x^2}{2}}\right]_{x=a_6}^{a_7} \\
&\le \frac{1}{\sqrt{2\pi}}\frac{1}{\mathbb{P}\left(a_6<X<a_7\right)}e^{-\frac{a_6^2}{2}} \\
&\overset{\text{(i)}}{\le} \frac{1}{\sqrt{2\pi}}\frac{1}{0.49\gamma}e^{-\frac{\left[G^{-1}(1-0.99\gamma)\right]^2}{2}},
\end{aligned} \tag{267a}$$

where (i) is true by the definition (263) of A. Similarly, conditional on the event E and on any a,

$$\mu_{1L|a} > -\frac{1}{\sqrt{2\pi}}\frac{1}{0.49\gamma}e^{-\frac{\left[G^{-1}(1.99\gamma)\right]^2}{2}}. \tag{267b}$$

Term: $(\mu_{2L|a} - \mu_{2H|a})$**:** For any $a \in \mathbb{R}^7$, we have

$$(\mu_{2L|a} - \mu_{2H|a}) < 0. \tag{268}$$

Showing $T < 0$**:** Plugging the three terms from (266), (267) and (268) back to (262), 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

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left(\sum_{i\in I_2} X^{(i)} - \sum_{i\in I_1} X^{(i)} \le 0\right) &\ge \lim_{n\to\infty} \int_{a\in\mathbb{R}^7} \mathbb{P}\left(T < 0 \mid a\right)\mathbb{P}(a) \\ &\ge \lim_{n\to\infty} \mathbb{P}(a \in A) = 1,\end{aligned}$$

completing the proof.

C.12 Proof of auxiliary results for Theorem 10

In this section, we present the proofs of the auxiliary results for Theorem 10.

C.12.1 Proof of Lemma 41

First, at $\lambda = \infty$ we have $\widehat{B}^{(\infty)} = 0$ by Proposition 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 $\max_{(i,j)\in\Omega}|u_{ij}| \le \max_{(i,j)\in\Omega}|y_{ij}| + \|x\|_\infty$. It then suffices to establish the inequality

$$\max_{(i,j)\in\Omega}|b_{ij}^{(\lambda)}| \le \max_{(i,j)\in\Omega}|u_{ij}|,$$

where $B^{(\lambda)}$ is the solution to the optimization

$$\operatorname*{arg\,min}_{B \text{ satisfies } \mathcal{O}} \|U - B\|_\Omega^2 + \lambda\|B\|_\Omega^2, \tag{269}$$

with ties broken by minimizing $\|B\|_\Omega^2$. Assume for contradiction that we have

$$\max_{(i,j)\in\Omega} |b_{ij}^{(\lambda)}| > \max_{(i,j)\in\Omega} |u_{ij}|. \tag{270}$$

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 (270), 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 (269) than $B^{(\lambda)}$. Contradiction to the assumption that $\widehat{B}^{(\lambda)}$ is the optimal solution of (269).

C.12.2 Proof of Lemma 42

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 (14) onto M.

From known results on the monotone cone (see (Amelunxen et al., 2014, Section 3.5)), we have $\mathbb{E}\|\Pi_M Z\|_2 \leq c\sqrt{\log d}$ for some fixed constant $c > 0$. Using the Moreau decomposition, we have (see (Wei et al., 2019, 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.

C.12.3 Proof of Lemma 43

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

$$-1 \leq |S_1| - |S_2| \leq 1. \tag{271}$$

By Definition 4 of the c_{f}-fraction interleaving assumption, we have

$$|S_1| + |S_2| = |S| \geq c_{\text{f}} n. \tag{272}$$

Combining (271) and (272), 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 1 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.

C.12.4 Proof of Lemma 44

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 43, using the fact that the interleaving points alternate between S_1 and S_2, we have

$$\left|T'\right| > \frac{c_{\text{f}} n}{6}.$$

We write the elements in T' in the increasing order as $k_1 < \ldots < k_{\frac{c_f n}{6}} < \ldots < k_{|T'|}$. It can be verified that the elements in $\{t_{2k}\}_{k \in \lceil \frac{c_f n}{12} \rceil}$ appear in different pairs when generating the training-validation split (Ω^t, Ω^v) in Line 5-7 of Algorithm 1. Hence, each element in $\{t_{2k}\}_{k \in \lceil \frac{c_f n}{12} \rceil}$ is assigned to Ω^v independently with probability $\frac{1}{2}$. Using Hoeffding's inequality, we lower-bound the size of $T' \cap \Omega^v$ as:

$$\lim_{n \to \infty} \mathbb{P}\left(\left| T' \cap \Omega^v \right| > \frac{c_f n}{36} \right) = 1. \tag{273}$$

Taking a union bound of (273) over $T' \in \{S^+ \cap S_1, S^- \cap S_1, S^+ \cap S_2, S^- \cap S_2\}$ completes the proof.