Debiasing Evaluations That Are Biased by Evaluations

Jingyan Wang[1], **Ivan Stelmakh**[1], **Yuting Wei**[2] **and Nihar B. Shah**[1]
[1] School of Computer Science
[2] Department of Statistics & Data Science
Carnegie Mellon University
{jingyanw, stiv}@cs.cmu.edu, ytwei@cmu.edu, nihars@cs.cmu.edu

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.

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". 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, Ottoboni, and Stark 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, Ottoboni, and Stark 2016), and student ratings serve as a poor predictor on the follow-on course achievement of the students (Carrell and West 2008; Braga, Paccagnella, and Pellizzari 2014):

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

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

An analogous issue arises in conference peer review, where conference organizers survey authors to rate their received reviews in order to understand the quality of the review process. It is well understood that authors are more likely to give higher ratings to a positive review than a to negative review (Weber et al. 2002; Papagiannaki 2007; Khosla, Hoiem, and Belongie 2013):

> *"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."*

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

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

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 2008). Furthermore, author-provided ratings can be a factor for selecting reviewer awards (Khosla, Hoiem, and Belongie 2013), and 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 2008; Boring, Ottoboni, and Stark 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. We also 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, Bucca, and Macy 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, Ottoboni, and Stark 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.

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 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 the two aforementioned settings, under various conditions captured by our general formulation. We finally conduct a semi-synthetic experiment that establish the effectiveness of our proposed approach (Section 5).

An extended version of this paper is available on arXiv (Wang et al. 2020), including more extensive related work, more intuition of our approach, and additional theoretical and experimental results.

1.2 Related Work

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, Mammen, and Park 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 model differs from past work where the design matrix of the linear component exhibits a special $0/1$ structure and is not random.

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, Tibshirani, and Friedman 2009). Yet, this generic scheme cannot be directly applied to models where training samples are not exchangeable. 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, Hoefling, and Tibshirani 2011).

2 Problem Formulation

For ease of exposition, throughout the paper we describe our problem formulation using the running example of course evaluation. 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. 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 squared ℓ_2 error $\frac{1}{d}\|\widehat{x} - x^*\|_2^2$.

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 $\lambda \geq 0$, given by:

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

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 $\|B\|_F^2$. 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.

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.

$\boldsymbol{\lambda = \infty}$**:** As $\lambda \to \infty$, intuitively the bias term in (2) converges to zero. 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 of each course. 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.

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

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

Algorithm 1: Cross-validation. Inputs: observations Y, partial ordering $\mathcal{O}$, and set Λ.

```
   /* Step 1: Split the data */
1  Initialize the training and validation sets as $\Omega^{\text{t}} \leftarrow \{\}$, $\Omega^{\text{v}} \leftarrow \{\}$.
2  Sample a total ordering of $\pi_0$ uniformly at random from the set $\mathcal{T}$ of all total orderings (of the $dn$ elements) consistent with the partial ordering $\mathcal{O}$.
3  foreach $i \in [d]$ do
4      Find the sub-ordering of the $n$ elements in course $i$ according to $\pi_0$, denoted in increasing order as $(i, j^{(1)}), \ldots, (i, j^{(n)})$.
5      for $t = 1, \ldots, \frac{n}{2}$ do
6          Assign $(i, j^{(2t-1)}), (i, j^{(2t)})$ to $\Omega^{\text{t}}$ and $\Omega^{\text{v}}$, one each uniformly at random. If $n$ is odd, assign the last element $(i, j^{(n)})$ to the validation set.
7      end
8  end
   /* Step 2: Compute validation error */
9  foreach $\lambda \in \Lambda$ do
10     Obtain $(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$ as a solution to the following optimization problem:
           $\underset{x\in\mathbb{R}^d,\ B\in\mathbb{R}^{d\times n},\ B \text{ satisfies } \mathcal{O}}{\arg\min} \|Y - x\mathbf{1}^T - B\|_{\Omega^{\text{t}}}^2 + \lambda\|B\|_{\Omega^{\text{t}}}^2,$
       where ties are broken by minimizing $\|\widehat{B}^{(\lambda)}\|_F$.
11     foreach $(i, j) \in \Omega^{\text{v}}$ do
12         foreach $\pi \in \mathcal{T}$ do
13             Find the element $(i^\pi, j^\pi) \in \Omega^{\text{t}}$ that is closest to $(i, j)$ with respect to $\pi$, and set $[\widetilde{b}_\pi^{(\lambda)}]_{ij} = \widehat{b}_{i^\pi j^\pi}^{(\lambda)}$. There may be two closest elements at equal distance to $(i, j)$, in which case call them $(i_1^\pi, j_1^\pi)$ and $(i_2^\pi, j_2^\pi)$ and set $[\widetilde{b}_\pi^{(\lambda)}]_{ij} = \frac{\widehat{b}_{i_1^\pi j_1^\pi}^{(\lambda)} + \widehat{b}_{i_2^\pi j_2^\pi}^{(\lambda)}}{2}$.
14         end
15         Interpolate the bias as $\widetilde{B}^{(\lambda)} = \frac{1}{|\mathcal{T}|}\sum_{\pi\in\mathcal{T}} \widetilde{B}_\pi^{(\lambda)}$.
16     end
17     Compute the CV error $e^{(\lambda)} := \frac{1}{|\Omega^{\text{v}}|}\|Y - \widehat{x}_\lambda \mathbf{1}^T - \widetilde{B}^{(\lambda)}\|_{\Omega^{\text{v}}}^2$.
18 end
19 Output $\lambda_{\text{cv}} \in \arg\min_{\lambda\in\Lambda} e^{(\lambda)}$, where ties are broken arbitrarily.
```

$\Lambda \subseteq [0, \infty]$ 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).

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 Ω^{v}, by computing $\frac{1}{|\Omega^{\text{v}}|}\|Y - \widehat{x}^{(\lambda)}\mathbf{1}^T - \widehat{B}^{(\lambda)}\|_{\Omega^{\text{v}}}^2$. However, recall that the estimate $\widehat{B}^{(\lambda)}$ only estimates the bias on the training set meaningfully, and we have $\widehat{B}_{ij}^{(\lambda)} = 0$ for each element (i, j) in the validation set Ω^{v}. Therefore, we "synthesize" the estimated bias $\widetilde{B}^{(\lambda)}$ on the validation from the estimated bias $\widehat{B}^{(\lambda)}$ on the training set via an interpolation procedure (Lines 11-16), as explained below.

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

- **Interpolating with respect to a total ordering (Line 13):** Given some total ordering π, we find the element in the training set that is the closest to (i, j) in the total ordering π. We denote this closest element from the training set as (i^π, j^π), and simply interpolate the bias at (i, j) with respect to π (denoted $[\widetilde{b}_\pi^{(\lambda)}]_{ij}$) using the value of $\widehat{b}_{i^\pi j^\pi}$. That is, we set $[\widetilde{b}_\pi^{(\lambda)}]_{ij} = \widehat{b}_{i^\pi j^\pi}^{(\lambda)}$. If there are two closest elements of equal distance to (i, j) (one ranked higher than (i, j) and one lower than (i, j) in π), we use the mean of the estimated bias $\widehat{B}^{(\lambda)}$ of these two elements. This step is similar to the CV error computation in Tibshirani, Hoefling, and Tibshirani (2011).
- **Taking the mean over all total orderings in $\mathcal{T}$ (Line 15):** After we find the interpolated bias $\widetilde{B}_\pi^{(\lambda)}$ on the validation set with respect to each π, the final interpo-

lated 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 to reduce the variance of the CV error.

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).

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 of the extended version (Wang et al. 2020).

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 $\widehat{x}^{(\lambda_{\text{cv}})}$ as solution at $\lambda = \lambda_{\text{cv}}$ on the training set.

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

Definition 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., $\{\text{A},\text{B},\text{C},\text{D},\text{F}\}$ or $\{\text{Pass},\text{Fail}\}$), or numeric scores (e.g., in the range of $[0,100]$) of the students. 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.

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

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

Definition 4 (Constant-fraction interleaving assumption) *Let $\mathcal{O}$ be a total ordering (of the dn elements $\{(i,j)\}_{i\in[d],j\in[n]}$). We define an interleaving point as any number $t \in [dn-1]$, such that the t^{th} and the $(t+1)^{th}$ highest-ranked elements according to the total ordering $\mathcal{O}$ belong to different courses. A total ordering $\mathcal{O}$ is said to satisfy the c-fraction interleaving assumption for some constant $c \in (0,1)$, if there are at least cn interleaving points in $\mathcal{O}$.*

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

4.2 $\lambda = 0$ 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 *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 either:*

(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 total ordering.

Then for any $\epsilon > 0$ and $\delta > 0$, there exists an integer n_0 (dependent on ϵ,δ,c,d,η), such that for every $n \geq n_0$ and every partial ordering satisfying condition (a) or (b):

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

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 $i \neq i'$.

Corollary 6 *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 condition (a) or (b) in Theorem 5:*

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

In Appendix A.1 of the extended version (Wang et al. 2020), 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.

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 *The limit of $\lim_{\lambda\to\infty}(\widehat{x}^{(\lambda)}, \widehat{B}^{(\lambda)})$ exists, given by*

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

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.

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 $\lambda = 0$.

Theorem 8 *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):

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

From Theorem 5 we have that the estimator $\widehat{x}^{(0)}$ (at $\lambda = 0$) is also consistent under the only-bias setting. Thus, we also have $\widehat{x}^{(\lambda_{\text{cv}})}$ converges to $\widehat{x}^{(0)}$ in probability. 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).

Theorem 9 *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):

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

By the consistency of $\widehat{x}^{(\infty)}$ implied from Section 4.3 under the only-noise setting, we also have $\widehat{x}^{(\lambda_{\text{cv}})}$ converges to $\widehat{x}^{(\infty)}$ in probability. 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 a semi-synthetic experiment using real grading statistics to evaluate our estimator and our cross-validation algorithm. Additional experimental results are in Section 5 of the extended version (Wang et al. 2020). 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).[2] Throughout the experiments, we use $\Lambda = \{2^i : -9 \leq i \leq 5, i \in \mathbb{Z}\} \cup \{0, \infty\}$. Implementation details for the cross-validation algorithm (Algorithm 1) are provided in Appendix B.1 of the extended version (Wang et al. 2020).

We use the grading data from the course "Business Statistics" in Spring 2020 from Indiana University Bloomington (2020). This course consists of 10 sessions taught by multiple instructors. The average number of students per session is 50. The possible grades that students receive are A+ through D-, and F. We consider three ways to construct the group orderings:

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

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 vary the choices of σ. The true quality is set as $x^* = 0$ (the results are independent from the value of x^*). The estimators are given one of the three group orderings listed above.

We compare our cross-validation algorithm with the mean, median, and also the reweighted mean estimator . The mean and median baselines are defined as taking the mean and median of each course respectively. The reweighted mean estimator is introduced in Appendix A.2 and B.3 of the extended version (Wang et al. 2020). Each point is computed as the empirical mean over 250 runs. Error bars represent the standard error of the mean.

The results are shown in Fig 1. 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.

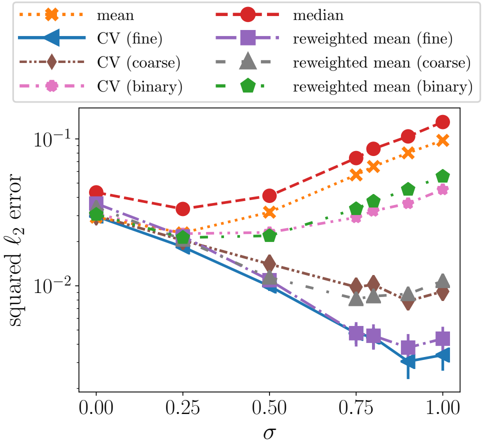

Figure 1: The performance of our estimator (with cross-validation) on semi-synthetic grading data, compared to the mean, median and reweighted mean estimators.

Our estimator performs slightly better than the reweighted mean estimator for the fine grades, and slightly better on a subset of values of σ for the coarse grades. For the binary grades, the error of both our estimator and the reweighted mean estimator increases as the relative amount of bias increases. 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.

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. 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.

[2] 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.

Acknowledgments

The work of J.W., I.S., and N.S. was supported in part by NSF CAREER award 1942124 and in part by NSF CIF 1763734. Y.W. was partially supported by the NSF grants CCF-2007911 and DMS-2015447.

Ethics Statement

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

References

Barlow, R.; Bartholomew, D.; Bremner, J.; and Brunk, H. 1972. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley.

Becker, W. E.; and Watts, M. 1999. How Departments of Economics Evaluate Teaching. *The American Economic Review* 89(2): 344–349.

Boring, A.; Ottoboni, K.; and Stark, P. B. 2016. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research* .

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

Carrell, S. E.; and West, J. E. 2008. Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. Working Paper 14081, National Bureau of Economic Research.

Chen, Y.; and Samworth, R. J. 2016. Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* .

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

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

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

Groeneboom, P.; and Jongbloed, G. 2014. *Nonparametric estimation under shape constraints*, volume 38. Cambridge University Press.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hastie, T. J.; and Tibshirani, R. J. 1990. *Generalized additive models*, volume 43. CRC press.

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

Indiana University Bloomington. 2020. Grade Distribution Database. https://gradedistribution.registrar.indiana.edu/index.php [Online; accessed 30-Sep-2020].

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

Khosla, A.; Hoiem, D.; and Belongie, S. 2013. Analysis of Reviews for CVPR 2012 .

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, 1137–1145. Montreal, Canada.

Mammen, E.; and Yu, K. 2007. Additive isotone regression. In *Asymptotics: particles, processes and inverse problems*, 179–195. Institute of Mathematical Statistics.

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

Molina, M. D.; Bucca, M.; and Macy, M. W. 2019. It's not just how the game is played, it's whether you win or lose. *Science Advances* 5(7).

Papagiannaki, K. 2007. Author Feedback Experiment at PAM 2007. *SIGCOMM Comput. Commun. Rev.* 37(3): 73–78.

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

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

Tibshirani, R. J.; Hoefling, H.; and Tibshirani, R. 2011. Nearly-Isotonic Regression. *Technometrics* 53(1): 54–61.

Wang, J.; Stelmakh, I.; Wei, Y.; and Shah, N. B. 2020. Debiasing Evaluations That are Biased by Evaluations. *arXiv preprint arXiv:2012.00714* .

Weber, E. J.; Katz, P. P.; Waeckerle, J. F.; and Callaham, M. L. 2002. Author perception of peer review: impact of review quality and acceptance on satisfaction. *JAMA* 287(21): 2790–2793.

Wood, S. N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* 99(467): 673–686.

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

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