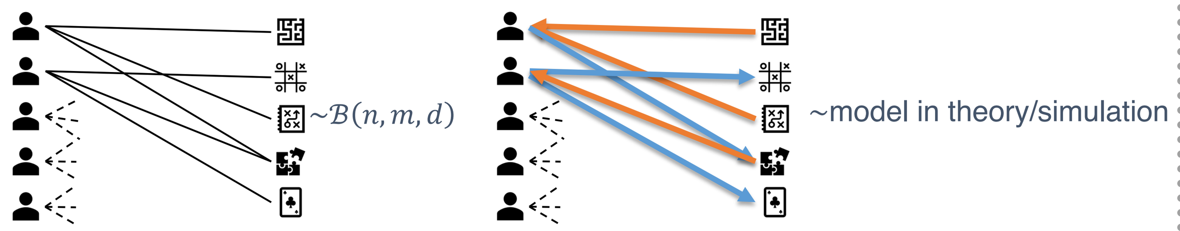
Fair Grading Algorithms for Randomized Exams

Jiale Chen, Department of Management Science and Engineering, Stanford University
Jason Hartline, Department of Computer Science, Northwestern University
Onno Zoeter, Booking.com

This paper studies grading algorithms for randomized exams. In a randomized exam, each student is asked a small number of random questions from a large question bank. The predominant grading rule is simple averaging, i.e., calculating grades by averaging scores on the questions each student is asked, which is fair ex-ante, over the randomized questions, but not fair ex-post, on the realized questions. The fair grading problem is to estimate the average grade of each student on the full question bank. The maximum-likelihood estimator for the Bradley-Terry-Luce model on the bipartite student-question graph is shown to be consistent with high probability when the number of questions asked to each student is at least the cubed-logarithm of the number of students. In an empirical study on exam data and in simulations, our algorithm based on the maximum-likelihood estimator significantly outperforms simple averaging in prediction accuracy and ex-post fairness even with a small class and exam size.

Randomized Exam

1. Assign a small number of random questions to each student (task assignment graph G)



2. Grade students according to the exam result (exam result graph G')

Model [Bradley and Terry 1952, Rasch 1993]

One-dimensional model: an unknown parameter vector u, where u_i for student $i \in S$ represents her ability and u_j for question $j \in Q$ represents its difficulty.

Result of answering process: a Bernoulli random variable w_{ij} for student i and question j, where $w_{ij} = 1$ represents a correct answer and $w_{ij} = 0$ represents an incorrect answer

Probability distribution of w_{ij}: softmax of the student ability u_i and the question difficulty u_j,

$$\Pr[w_{ij} = 1] = 1 - \Pr[w_{ij} = 0] = \frac{\exp(u_i)}{\exp(u_i) + \exp(u_j)} = f(u_i - u_j),$$

where $f(x) = \frac{1}{1+\exp(-x)}$.

Fairness of the Algorithm

Algorithm: an arbitrary mapping from the exam result to student grades alg

Benchmark: the expected grade student gets from the traditional exam design

$$\forall i \in S, opt_i = \frac{1}{|Q|} \sum_{j \in Q} \mathbb{E}[w_{ij}]$$

Different ways to compare the algorithm to the benchmark due to two sources of randomness:

1. random task assignments
2. students' random mistakes.

Ex-ante Bias: Compare the students' expected grade over the random task assignments and their random mistakes to the benchmark.

$$(E_G E_w[alg_i] - opt_i)^2$$

Ex-post Bias: Given a task assignment, compare the students' expected grade over their random mistakes to the benchmark.

$$(E_w[alg_i] - opt_i)^2$$

Ex-post Error: Directly compare the final grade to the benchmark

$$(alg_i - opt_i)^2$$

Simple Averaging

Definition: Given an exam result graph G', simple averaging grades student i by

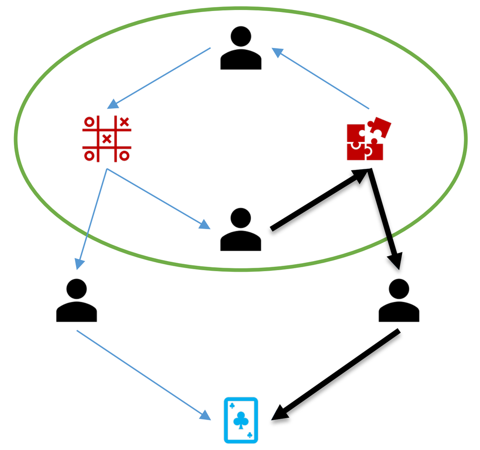
$$\text{avg}_i = \frac{\deg_i^+}{\deg_i^- + \deg_i^+} = \frac{\#correct}{\#asked},$$

where $\deg_i^+$ and $\deg_i^-$ represents the outdegree and indegree.

Fact (Ex-ante Fairness): Simple averaging is ex-ante fair over any family of task assignment graphs $\mathcal{G}$ that is symmetric w.r.t. the questions.

Our Algorithm

Our algorithm is based on the maximum likelihood estimators (MLEs), with additional strategies to solve the case when MLEs do not exist. Our algorithm makes different predictions h_{ij} for four types of student-question pairs.



Existing Edge (when w_{ij} is revealed in the exam results graph): We make prediction the same $h_{ij} = w_{ij}$.

Same Component (when $i \in S$ and $j \in Q$ are in the same strongly connected component): It can be proved in theory that the MLEs u^* exist within the component. And we use the MLEs for predition, i.e., $h_{ij} = f(u_i^* - u_j^*)$.

Comparable Components (when $i \in S$ and $j \in Q$ are in different strongly connected components and there is a directed path linking them): From the property of the directed graph, all directed paths linking them have the same direction. if it goes from the student to the question, we regard it as a strong evidence that the student has a much higher level of ability than the question's difficulty, so we make a prediction $h_{ij} = 1$; for the opposite direction, we make a prediction $h_{ij} = 0$.

Incomparable Components (when $i \in S$ and $j \in Q$ are in different strongly connected components and there is no directed path linking them): We take the average of the above three types of predictions on student i as the prediction for this edge. It is the same strategy as simple averaging.

Theoretical Results

Theorem (Existence and Uniqueness of MLEs). If

$$\frac{\exp(\alpha_{n,m})\,(n+m)\log(n+m)}{nd_{n,m}} \to 0 \ \ (n, m \to \infty),$$

where $\alpha_{n,m} = \max_{i,j \in S \cup Q} u_i - u_j$ is the largest difference between all possible pairs of parameters, then $\Pr[u^* \ exists\ and\ is\ unique] \to 1$, where u^* is the MLE vector.

Theorem (Uniform Consistency of MLEs). If

$$\exp\left(2(\alpha_{n,m}+1)\right)\sqrt{\frac{m\log^3(n+m)}{nd_{n,m}\log^2\left(\frac{n}{m}d_{n,m}\right)}} \to 0 \ \ (n, m \to \infty),$$

then u^* is uniformly consistent, i.e., $\|u^* - u\|_\infty \xrightarrow{\mathbb{P}} 0$.

Corollary (Upper Bound on the Exam Length when $n = m$ and $\alpha = O(1)$). If

$$\frac{\log n}{d_{n,m}} \to 0 \ \ (n, m \to \infty),$$

the MLEs exist and are unique. If

$$\frac{\log^3 n}{d_{n,m}\log^2 d_{n,m}} \to 0 \ \ (n, m \to \infty),$$

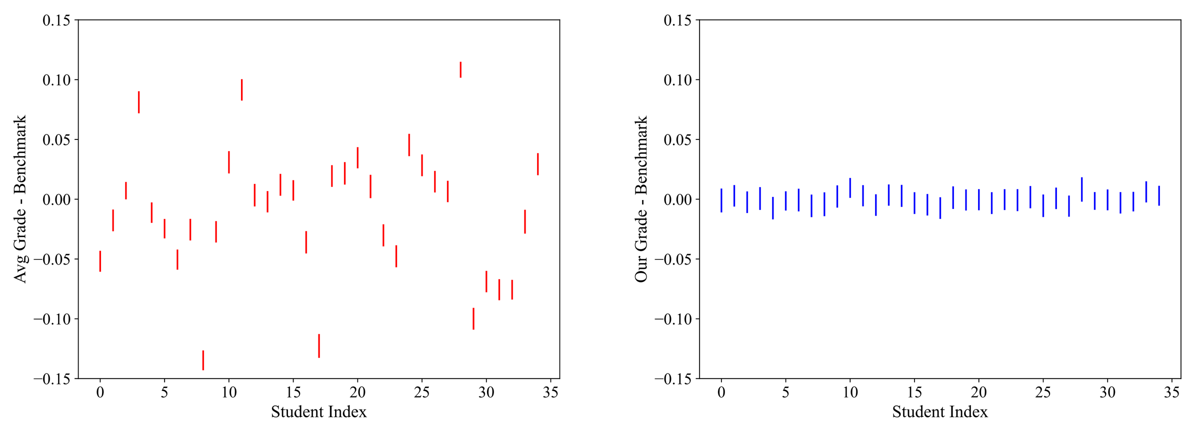
the MLEs are uniformly consistent.

Theorem (Ex-post Error of Our Algorithm). In the case where the MLEs exist and are unique, we have $\forall i \in S, (alg_i - opt_i)^2 \le \frac{1}{4}\|u^* - u\|_\infty^2$.

Visualization of Simple Averaging's Ex-post Unfairness

Setting: 35 students, each asked 10 out of 22 questions.

Ex-post grade deviation: $E_w[alg_i] - opt_i$



Ex-post Error and Bias-Variance Decomposition

The ex-post error can be decomposed into ex-post bias and the variance of the algorithm.

Theorem (Bias-Variance Decomposition).

$$\mathbb{E}_G\mathbb{E}_i\mathbb{E}_w[(alg_i - opt_i)^2]$$
$$= \mathbb{E}_G\mathbb{E}_i[(\mathbb{E}_w[alg_i] - opt_i)^2] + \mathbb{E}_G\mathbb{E}_i\mathbb{E}_w[(alg_i - \mathbb{E}_w[alg_i])^2]$$

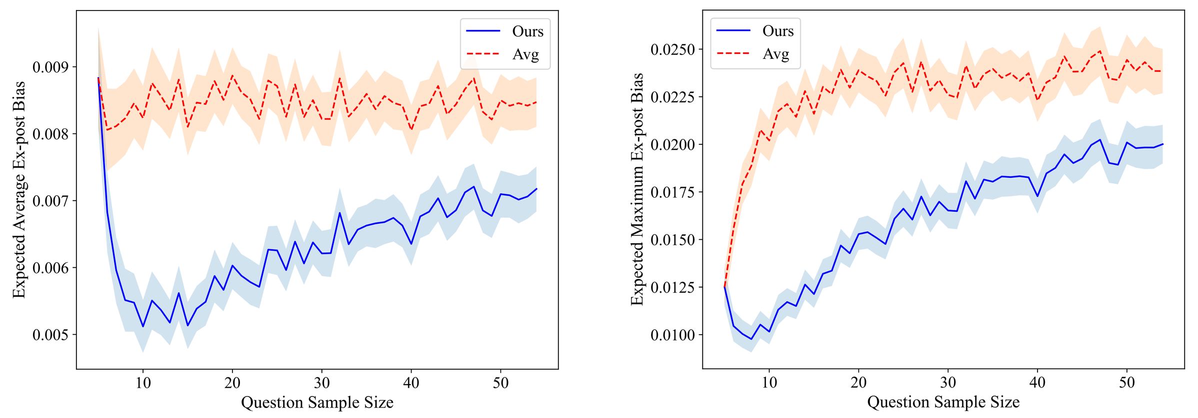
Setting: 35 students, each asked 10 out of 22 questions.

	Ex-post Bias	Variance	Ex-post Error
Ours	0.00004	0.0188	0.0188
Avg	0.00331	0.0170	0.0203
Ours-Avg	-0.00327 -99%	0.0018 +10%	-0.0015 -8%

Optimal Exam Design

We consider the problem of choosing the size of the question set from an infinite question bank.

Setting: 5 students, each asked 5 questions.



Real World Data Cross Validation

We randomly split the training set and the test set and measure the logarithm of mean square error. We also give a reference on when to choose our algorithm.

Setting: 35 students, 22 questions.