A Robust Computerized Adaptive Testing Approach in Educational Question Retrieval

Yan Zhuang
Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
zykb@mail.ustc.edu.cn

Qi Liu*
Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China & Institute of Artificial Intelligence, Hefei Comprehensive National Science Center & State Key Laboratory of Cognitive Intelligence
Hefei, China
qiliuql@ustc.edu.cn

Zhenya Huang
Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
huangzhy@ustc.edu.cn

Zhi Li
Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
zhili03@mail.ustc.edu.cn

Binbin Jin
Huawei Cloud Computing Technologies Co., Ltd
Hangzhou, China
jinbinbin1@huawei.com

Haoyang Bi
Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
bhy0521@mail.ustc.edu.cn

Enhong Chen
Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
cheneh@ustc.edu.cn

Shijin Wang
State Key Laboratory of Cognitive Intelligence & iFLYTEK AI Research (Central China), iFLYTEK Co., Ltd
Hefei, China
sjwang3@iflytek.com

ABSTRACT

Computerized Adaptive Testing (CAT) is a promising testing mode in personalized online education (e.g., GRE), which aims at measuring student's proficiency accurately and reducing test length. The "adaptive" is reflected in its selection algorithm that can retrieve best-suited questions for student based on his/her estimated proficiency at each test step. Although there are many sophisticated selection algorithms for improving CAT's effectiveness, they are restricted and perturbed by the accuracy of current proficiency estimate, thus lacking robustness. To this end, we investigate a general method to enhance the robustness of existing algorithms by leveraging student's "multi-facet" nature during tests. Specifically, we present a generic optimization criterion Robust Adaptive Testing (RAT) for proficiency estimation via fusing multiple estimates at each step, which maintains a multi-facet description of student's potential proficiency. We further provide theoretical analyses of such estimator's desirable statistical properties: asymptotic unbiasedness, efficiency, and consistency. Extensive experiments on perturbed synthetic data and three real-world datasets show that selection algorithms in our RAT framework are robust and yield substantial improvements.

*Corresponding Author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '22, July 11–15, 2022, Madrid, Spain.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531928

CCS CONCEPTS

• **Applied computing** → **E-learning**.

KEYWORDS

computerized adaptive testing, educational resource search, educational measurement, cognitive diagnosis

ACM Reference Format:
Yan Zhuang, Qi Liu, Zhenya Huang, Zhi Li, Binbin Jin, Haoyang Bi, Enhong Chen, and Shijin Wang. 2022. A Robust Computerized Adaptive Testing

Approach in Educational Question Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3531928

1 INTRODUCTION

In the era of information explosion, Computerized Adaptive Testing (CAT) is a critical problem in educational resource search and educational measurement [39]. Being able to adaptively search for best-suited questions for student, CAT has been an increasingly popular way to measure student proficiency level in standardized tests (e.g., GMAT and GRE). Lan et al. [22] has proven that such adaptive tests need fewer questions than paper-and-pencil tests to reach the same measurement accuracy. Also, shorter tests reduce system load and are better for students, who may be frustrated or bored if they need to give too many answers [10, 41].

In real-life scenarios, CAT consists of two core components that work alternately until the end of test (see Figure 1): At test step t, **(1) Cognitive Diagnosis Model (CDM)**, as the user model, first estimates student's current proficiency $\hat{\theta}^t$ using his/her responses to previous t questions. The most famous model is Item Response Theory (IRT), which defines the probability of the correct response by aligning and comparing student's overall ability, $\theta \in \mathbb{R}$, with the question's difficulty [14]. More generally, θ is multidimensional [42, 43] and each dimension represents the proficiency of the corresponding concept (e.g., concept *Algebra* in Mathematics). **(2)** Next, the **selection algorithm** retrieves questions according to some next item criteria [3, 9, 27], which select the most informative one, e.g., whose difficulty is closest to student's proficiency estimated by CDM; hence, most algorithms adopt current estimate $\hat{\theta}^t$ as the *query* in selection. More informative and appropriate questions asked have been proven to reduce test length significantly [41].

While the above paradigm has achieved great success, its drawback is also apparent: *The selection algorithm is inefficient if the query (i.e., current estimate $\hat{\theta}^t$) is not close to student's true proficiency θ_0* [11]. Unfortunately, such deviation, $\|\hat{\theta}^t - \theta_0\|$, is *inevitable* under various perturbations, such as optimization dilemmas (e.g., overfitting), limited responses used in estimation at initial steps, student's guess and slip factors [26]. To alleviate such poor robustness in CAT, a series of criteria [35, 38, 40] have been proposed to introduce additional information in selection, whereas they still center on the current estimate to retrieve and hence bring limited improvements. Recently, many studies try to change this traditional paradigm – using reinforcement learning to train data-driven selection algorithms [16, 24, 30, 51]; as a result, these methods need to be retrained from scratch whenever a new question is added into action space, which is impractical in real education systems with enormous new questions uploaded or refreshed daily; they are also prone to biases in historical data [16]. Hence, a more realistic solution to enhance CAT's robustness remains underexplored.

Fortunately, we discover a previously overlooked fact: Student is "multi-facet" during the test, i.e., his/her previous responses often correspond to multiple estimates instead of the singleton $\hat{\theta}^t$:

- **Example 1**: During the test, if one student correctly answers a simpler question (e.g., *difficulty* = 0.3) but wrong answers a harder one (e.g., *difficulty* = 0.8), then his/her proficiency may locate at range [0.3, 0.8] instead of one certain value.
- **Example 2**: Usually, there are multiple solutions to a question (see Figure 2). A correct response to a question about concept *Algebra* can be inferred that he/she may indeed have a higher proficiency on *Algebra*, or *Geometry*, or none (i.e., guess factor), or both.

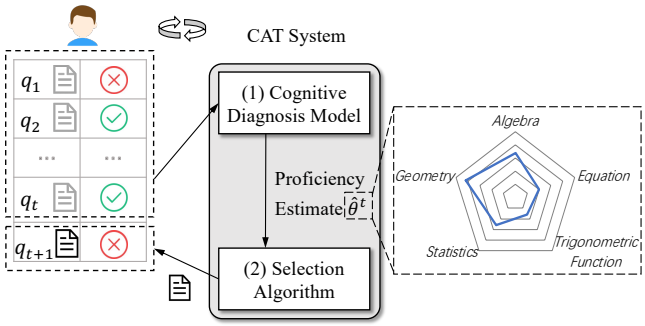


Figure 1: The workflow of CAT: At step t, the selection algorithm adaptively selects next question q_{t+1} based on student's current proficiency $\hat{\theta}^t$ estimated by CDM.

The above findings seem to provide a conceptually simple and intuitive solution towards CAT's poor robustness: *Two heads are better than one* – fusing a group of diverse estimates for question selection at each step. However, it is not easy to introduce multiple estimates and ensure their accuracy and diversity: **1)** accuracy: to better represent the above multi-facet nature of student (e.g., Example 1 and 2), all generated estimates should be in line with his/her previous responses; **2)** diversity: these estimates should diverse from each other of course to describe student's latent proficiency comprehensively, otherwise there is no need to introduce multiples. Nevertheless, they are estimated under the same response data, which might bring these estimates high correlation, even making them identical.

Based on these considerations, we propose a generic approach to enhance the robustness within existing CAT methods, which is called **R**obust **A**daptive **T**esting framework (RAT). Concretely, we design a new optimization formulation of proficiency estimation during CAT procedure based on traditional Maximum Likelihood Estimation. It generates multiple estimates and maintains their diversity and accuracy at each step to enable CAT in more robust settings. More importantly, we uncover the asymptotic *unbiasedness*, *efficiency*, and *consistency* of the estimator determined by these multiple estimates and provide theoretical proofs. Such desirable statistical properties of this new estimator ensure the effectiveness of using it as the new query for question selection and as student's final proficiency estimate. Note that RAT is simple yet more reliable, which requires no additional constraint on CAT's paradigm, and no modification to the existing selection algorithms.

To validate the effectiveness of our proposed RAT framework, we conduct extensive experiments on both the synthetic dataset and three real-world datasets from different education systems. Empirical results show that RAT achieves state-of-the-art performance even at high noise rate. The surprise is when compared

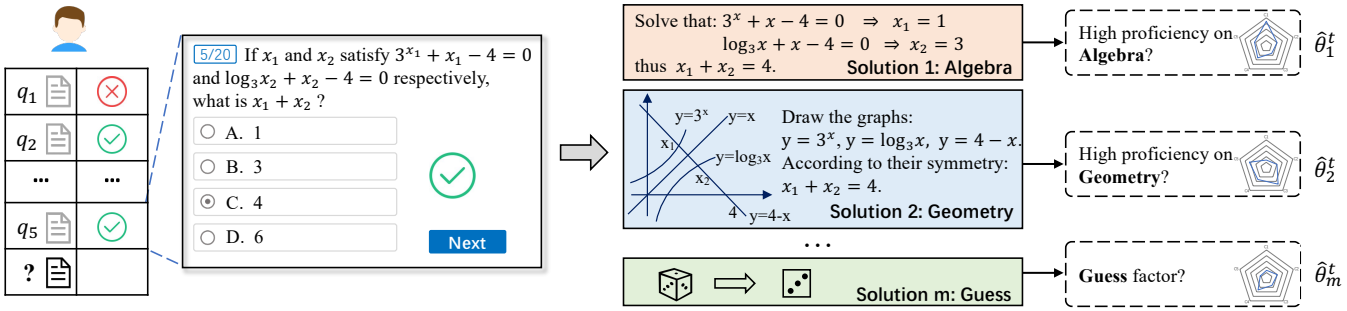


Figure 2: Usually, there are multiple solutions towards a question, and each solution corresponds to a possible proficiency, i.e., $\{\hat{\theta}_1^t, \hat{\theta}_2^t, ..., \hat{\theta}_m^t\}$, instead of singleton $\hat{\theta}^t$.

with strong deep-learning selection algorithms, traditional uncertainty information-based criteria in RAT framework achieve highly competitive performances, which require no additional training overhead towards the selection algorithm.

2 BACKGROUND AND RELATED WORKS

Computerized Adaptive Testing (CAT) is an iterative and dynamic process, including a Cognitive Diagnosis Model (CDM) and a selection algorithm. These two components work alternately until the termination criterion is satisfied and output the student's estimated proficiency, feeding back to themselves or instructors in a visual manner for facilitating individualized learning. The objective of CAT is to accurately estimate the proficiency of an upcoming student while minimizing the number of questions asked [8]. The following reviews these two components of CAT separately.

2.1 Cognitive Diagnosis Model

Cognitive Diagnosis means that student's cognitive state or proficiency is stable in a test thus can be inferred through their interactive behaviors (i.e., historical responses) [15, 25]. Hence, Cognitive Diagnosis Model (CDM) (or psychometric model) adopts question's and student's features to predict the response (correct or wrong). The most famous model is Item Response Theory (IRT) [1, 18] with their simplest form (1PL):

$$\Pr(\text{student answers question } j \text{ correctly}) = \sigma(\theta - b_j), \quad (1)$$

where $\sigma(\cdot)$ is the logistic function, $b_j \in \mathbb{R}$ represents each question's pre-calibrated parameter called *difficulty* [14], and $\theta \in \mathbb{R}$ is student's latent proficiency/ability to be estimated. Other representative ones include Matrix Factorization (MF) [20, 37], Deterministic Inputs, Noisy-And gate (DINA) [12, 42], and recently proposed Neural Cognitive Diagnosis Model (e.g., NCDM [43] and CDGK [45]) that leverages neural networks to model student-question interactions. Given the specific CDM and response data, Maximum Likelihood Estimation (binary cross-entropy loss) is generally used to estimate θ for subsequent selection.

2.2 Selection Algorithms

Traditional algorithms are based on some uncertainty or information metrics, specifically designed for different CDMs. For example, the most widely used is Fisher Information metric (FSI) [17, 27], designed for IRT, which selects the next question to maximize Fisher Information calculated at the current estimate; however, it is inefficient if the current proficiency estimate $\hat{\theta}$ is not close to the true [11] thus affecting the robustness of CAT system. To alleviate this, a series of methods based on FSI have been proposed, including Kullback-Leibler Information (KLI) [9, 35], Bayesian criterion [38] and weight criterion [40] which introduce additional integral, probability, and weight-assignment respectively. Unfortunately, these methods still center on the current estimate and ignore the fact that student's proficiency corresponding to his/her previous performance *is not unique naturally*. Recently, many researchers resort to changing CAT's paradigm and regard it as a reinforcement learning problem [16, 24, 30] to train selection algorithms directly from large-scale student response data, such as BOBCAT [16]; although the use of deep neural networks brings strong performance, they are computationally infeasible in practical applications and prone to bias in training data as mentioned in Section 1. In this paper, a more realistic approach is proposed to enhance the robustness of existing algorithms (e.g., FSI) without modifying CAT's paradigm.

3 PRELIMINARIES

The task of Computerized Adaptive Testing (CAT) is to provide a student with a best-fitting list of questions. An important assumption [8] of CAT is that student's true proficiency level $\theta_0 \in \mathbb{R}^d$ is constant throughout the test[1], where d refers to the proficiency's dimension (e.g., the number of knowledge concepts to be tested). The ultimate goal of CAT is to (1) select valuable and best-fitting questions for individual students, reducing test length; (2) utilize previous responses to estimate student's proficiency θ and ensure it's close to the true θ_0, when the test is over.

3.1 Task Formalization

At test step $t \in [1, 2, ..., T]$ in CAT, student's current proficiency estimate $\hat{\theta}^t$ is estimated using previous t responses; then leverage $\hat{\theta}^t$ to retrieve the next question q_{t+1} from question bank $\boldsymbol{Q}$ to ask student, and receive the response y_{t+1}. These interactions form a sequence $\{(q_1, y_1), (q_2, y_2), ..., (q_T, y_T)\}$, where $y_t = 1$ if the response to q_t is correct and 0 otherwise. Specifically, given quesiton

[1]This assumption makes CAT fundamentally different from other intelligent education systems (e.g., Knowledge Tracing [32]).

bank $Q = \{q_1, q_2, ..., q_{|Q|}\}$, a complete CAT system includes two components, and they work alternately and repeatedly for T steps:

(1) **Proficiency Estimation using CDM.** A CDM class f (e.g., IRT, NCDM) predicts the correctness of question responsed by student with proficiency θ, denoted as $f(q, \theta) = \Pr(y = 1|q, \theta)$. To accurately estimate his/her proficiency at each step, the well-known Maximum Likelihood Estimation (MLE) is utilized in CAT, ensuring the estimated $\hat{\theta}$ close to the true, i.e., $\|\hat{\theta} - \theta_0\| \to 0$. At step t, given previous responses $R = \{(q_1, y_1), (q_2, y_2), ..., (q_t, y_t)\}$, the corresponding MLE loss is formulated as follows, thus the estimate is updated as $\hat{\theta}^t = \arg\min_\theta \mathcal{L}_{MLE}(\theta)$.

$$\mathcal{L}_{MLE}(\theta) = -\frac{1}{t} \sum_{(q,y) \in R} y \log f(q, \theta) + (1 - y) \log(1 - f(q, \theta)). \tag{2}$$

(2) **Question Selection.** A question selection algorithm selects from Q based on student's current estimate $\hat{\theta}^t$. More specifically, it retrieves the next question q_{t+1} as

$$q_{t+1} = \arg\max_{q \in Q} \mathcal{I}_q(\hat{\theta}^t), \tag{3}$$

where $\mathcal{I}_q(\cdot)$ is the informativeness of question q. After receiving new response y_{t+1}, CDM updates and estimates proficiency $\hat{\theta}^{t+1}$.

3.2 Analysis of the problem

From the entire process of CAT above, it is not difficult to find that the proficiency estimate $\hat{\theta}^t$, summarizing this student's previous performance/responses, is the only basis/input for various question selection algorithms. As a result, if current estimate $\hat{\theta}^t$ is not close to θ_0, the selection algorithm is inefficient and thus interferes with the subsequent process [11]. As we have indicated before, such deviation is inevitable and exists objectively (e.g., student's guess and slip). What we can do is try to shrink it.

Therefore, instead of designing a new selection algorithm, in this paper, we focus on the query (i.e., proficiency estimate $\hat{\theta}$). Our method is conceptually simple: fusing a group of potential estimates $\{\hat{\theta}_i^t\}_{i=1}^m$ to improve its accuracy and robustness. The main reason is that: the proficiency inferred from previous responses is not unique. E.g. in Figure 2 a correct response to the fifth question can be inferred that he/she may have multiple potential proficiencies due to multiple solutions towards the question. The semantics of $\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m\}$ is that student's previous responses correspond to multiple estimates instead of single one due to uncertainty. This approach has two advantages:

- A group of estimates is leveraged to describe student's proficiency from various perspectives and resist perturbations. Further integrating them can improve the accuracy of estimation at each step (see next section).
- This approach is generic and based on the objective pattern within student-question, which contains no additional restrictions and assumptions towards CAT itself.

4 PROPOSED APPROACH

In our Robust Adaptive Testing framework (RAT), as aforementioned, multiple estimates $\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m\}$ will be generated at each step as student's multi-facet perspective. Keeping the typical CAT paradigm unchanged and ensuring the versatility of our method, we leverage their average $\theta^* = \frac{1}{m} \sum_{i=1}^m \hat{\theta}_i^t$ as a new query (replacing $\hat{\theta}^t$ in Eq.(3)) for question selection:

$$q_{t+1} = \arg\max_{q \in Q} \mathcal{I}_q(\theta^*), \tag{4}$$

where $\mathcal{I}_q(\cdot)$ could be any existing information metric of question q, e.g., FSI [27] and KLI [9].

4.1 Motivation: Accuracy and Diversity

Similarly, as the guarantee of the selection algorithm's efficiency, the error of θ^* should also be restricted to a minimum at each step, i.e., $\|\theta^* - \theta_0\| \to 0$. Further, we find:

Lemma 1. *Assume $\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m\}$ are m estimates of the true value θ_0. Let $\theta^* = \frac{1}{m} \sum_{i=1}^m \hat{\theta}_i$. Then its error, $\|\theta^* - \theta_0\|^2$, what RAT attempts to minimize at each step can be decomposed as:*

$$\left\|\theta^* - \theta_0\right\|^2 = \underbrace{\frac{1}{m} \sum_{i=1}^m \left\|\hat{\theta}_i - \theta_0\right\|^2}_{\text{① Accuracy}} - \underbrace{\frac{1}{m} \sum_{i=1}^m \left\|\hat{\theta}_i - \theta^*\right\|^2}_{\text{② Diversity}}. \tag{5}$$

Proof. Expanding the ② Diversity term: $\frac{1}{m} \sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2$
$= \frac{1}{m} \sum_{i=1}^m \|\hat{\theta}_i - \theta_0 + \theta_0 - \theta^*\|^2$
$= \frac{1}{m} \sum_{i=1}^m \left\{\|\hat{\theta}_i - \theta_0\|^2 + 2(\hat{\theta}_i - \theta_0)^\top (\theta_0 - \theta^*) + \|\theta_0 - \theta^*\|^2\right\}$
$= \frac{1}{m} \sum_{i=1}^m \|\hat{\theta}_i - \theta_0\|^2 + 2(\theta_0 - \theta^*)^\top \frac{1}{m} \sum_{i=1}^m (\hat{\theta}_i - \theta_0) + \|\theta_0 - \theta^*\|^2$
$= \frac{1}{m} \sum_{i=1}^m \|\hat{\theta}_i - \theta_0\|^2 - 2\|\theta_0 - \theta^*\|^2 + \|\theta_0 - \theta^*\|^2$
$= \frac{1}{m} \sum_{i=1}^m \|\hat{\theta}_i - \theta_0\|^2 - \|\theta^* - \theta_0\|^2$. This completes the proof. □

At step t, the mean error of these estimates $\frac{1}{m} \sum_{i=1}^m \|\hat{\theta}_i - \theta_0\|^2$ qualifies their accuracy. Meanwhile, the difference between each estimate and their average, $\frac{1}{m} \sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2$, is used to measure their diversity. This decomposition reveals that: **(1)** The error of θ^* will never be larger than ① the mean error of $\{\hat{\theta}_i\}_{i=1}^m$ (upper bound), because ① and ② are non-negative. **(2)** Smaller ① (more accurate) and larger ② (more diverse) will lead to more accurate θ^*. Therefore, Lemma 1 inspires an intuitive way to improve the proficiency estimator by averaging $\{\hat{\theta}_i\}_{i=1}^m$ to get θ^*, and the key is to keep their *accuracy* and *diversity*. This decomposition can be regarded as an extension of error-ambiguity decomposition [50].

4.2 New Formulation of Proficiency Estimation

Based on the findings above, we now introduce our new formulation for estimating the proficiency of students. At step t, we adjust the optimization function of θ_i, $1 \le i \le m$, by adding diversity-regularization term $\psi(\theta_i)$ to the commonly used MLE target:

$$\hat{\theta}_i^t = \arg\min_{\theta_i} \mathcal{L}_{MLE}(\theta_i) - \lambda \psi(\theta_i) \quad \text{for } i = 1, ..., m. \tag{6}$$

$$\psi(\theta_i) = \frac{1}{2} \left\|\theta_i - \theta_i^*\right\|^2, \quad \theta_i^* = \frac{1}{i-1} \sum_{k=1}^{i-1} \hat{\theta}_k^t,$$

where $\mathcal{L}_{MLE}(\theta_i)$ is the MLE loss (Eq.(2)) ensuring the accuracy of estimation (i.e., ① in Eq.(5)). $\lambda \ge 0$ is the coefficient weight to control diversity-regularization term (i.e., ② in Eq.(5)). Since we can not get the average before all m estimates are generated when

Algorithm 1: RAT Flowchart

Input: $\mathcal{Q}$ - question bank, f - a specific CDM, $\mathcal{I}$ - a specific information metric;

Initialize:

Initialize m estiamtes $\{\hat{\theta}_1^0, \hat{\theta}_2^0, ..., \hat{\theta}_m^0\}$;

The responses data $R \leftarrow \emptyset$;

1 **for** $t = 1$ **to** T **do**

2 $\quad \theta^* \leftarrow \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i^{t-1}$;

3 $\quad q_t \leftarrow \arg\max_{q \in \mathcal{Q}} \mathcal{I}_q(\theta^*)$;

4 $\quad R \leftarrow R \cup \{(q_t, y_t)\}$; ▷ Student responses to it

5 $\quad \mathcal{Q} \leftarrow \mathcal{Q} - \{q_t\}$;

6 $\quad$ **for** $i = 1$ **to** m **do** ▷ Generate m estimates

7 $\quad\quad \theta_i^* \leftarrow \frac{1}{i-1}\sum_{k=1}^{i-1} \hat{\theta}_k^t$;

8 $\quad\quad \hat{\theta}_i^t \leftarrow \arg\min_{\theta_i} \mathcal{L}_{MLE}(\theta_i) - \frac{\lambda}{2}\left\|\theta_i - \theta_i^*\right\|^2$

Output: Student final proficiency estimate $\theta^* := \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i^T$

optimizing θ_i, the temporary average of previous $i-1$ estimates, θ_i^*, is used as an alternative to θ^* in Eq.(5). All estiamtes $\{\hat{\theta}_1^t, \hat{\theta}_2^t, ..., \hat{\theta}_m^t\}$ are generated sequentially, balancing their accuracy and diversity. We now comment on the rationale behind the new formulation:

- **Traditional MLE ($\lambda = 0$).** In this case, such optimization could be regarded as repeatedly estimating m times using MLE. That is, all estimates $\{\hat{\theta}_1^t, \hat{\theta}_2^t, ..., \hat{\theta}_m^t\}$ may share a high correlation or even be mutually identical, i.e., $\theta^* = \hat{\theta}_1^t = \cdots = \hat{\theta}_m^t$, thus Eq.(6) degenerates into traditional MLE.
- **Mixed estimation ($\lambda \in (0, \infty)$).** As λ increases, the diversity-regularization term $\lambda\psi(\theta_i)$ has an increasingly more substantial effect, ensuring that all estimates are diverse, as otherwise the penalty $\lambda\psi(\theta_i)$ would be too small.
- **Blind estimation ($\lambda = \infty$).** Intuitively, this limit case should force all estimates to be mutually diverse, regardless of accuracy and fit to the data, which is similar to randomly sampling all estimates on an infinite range.

The robustly optimized CAT procedure is shown in Algorithm 1, where we detail the interaction that occurs within each test step. Obviously, **(1)** RAT is generic and could be directly applied to existing selection algorithms and CDMs. **(2)** The average of student's multi-facet proficiency estimates (i.e., θ^*) is used as the new query for question selection (*line 3*) and student's final proficiency estimate.

Time Complexity. At each step, the time overhead of CAT falls in two folds: proficiency estimation (T_e) and the computation of each question's information for selection (T_s). Obviously, the introduction of m estimates will incur estimation overhead m times the original, mT_e. But compared to the computation of questions' information, the extra time of estimation using few responses in mini-batch can be ignored, especially when facing enormous candidate questions in real-world applications [39], i.e., $mT_e \ll T_s$. Also, empirical results in Section 6.4 suggested that $m = 5$ is sufficient for satisfactory performance. Therefore, the time complexity is acceptable and this new optimization criterion can be applied in actual deployments. We leave further explorations and optimizations as future work.

5 CHARACTERIZATION OF ESTIMATION

The new proficiency estimation method leverages θ^* as the robust enhancement estimator for CAT. In this section, we delve into its characterization and prove its rationality as the query for retrieving questions and as student's final proficiency estimate.

5.1 Theoretical Analysis: Unbiasedness, Efficiency and Consistency

In statistics, estimators are usually adopted because of their statistical properties, most notably *unbiasedness*, *efficiency*, and *consistency* [34]. For this, the expression of θ^* determined by Eq.(6) first needs to be parsed. We apply Taylor expansion on each dimension of $\nabla_{\theta_i}\left[\mathcal{L}_{MLE}(\theta_i) - \lambda\psi(\theta_i)\right] := \nabla_{\theta_i}\mathcal{L}(\theta_i)$ at the point θ_0, yielding

$$\nabla_{\theta_i}\mathcal{L}(\theta_i) \approx \nabla_{\theta_i}\mathcal{L}_{MLE}(\theta_0) - \lambda\left(\theta_0 - \theta_i^*\right) + \left(\theta_i - \theta_0\right)\left(H_D(\theta_0) - \lambda\right), \tag{7}$$

where $H_D = diag\left[\nabla^2_{\theta_i}\mathcal{L}_{MLE}\right]$ is the diagonal of $\mathcal{L}_{MLE}$'s Hessian Matrix. $\hat{\theta}_i$ is the solution to $\nabla_{\theta_i}\mathcal{L}(\theta_i) = 0$ as in [13, 34], and we put this into Eq.(7) to get $\hat{\theta}_i \approx \frac{\theta_0 H_D(\theta_0) - \lambda\theta_i^* - \nabla\mathcal{L}_{MLE}(\theta_0)}{H_D(\theta_0) - \lambda}$. Thus, we get the expression of estimator θ^* we want:

$$\theta^* = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i \approx \frac{\theta_0 H_D(\theta_0) - \frac{\lambda}{m}\sum_i \theta_i^* - \nabla\mathcal{L}_{MLE}(\theta_0)}{H_D(\theta_0) - \lambda} \tag{8}$$

Based on the above results, each property of estimator θ^* could be deduced in detail. To explain more intuitively, we only show the one-dimension scenario, and the multi-dimension is similar.

Theorem 1 (Asymptotic Unbiasedness and Efficiency). *The CDM's Fisher information on θ_0 is denoted as $I(\theta_0)$. When $\lambda < I(\theta_0)$ and $m \to \infty$, the estimator θ^* is asymptotically unbiased, that is,*

$$\mathbb{E}\left[\theta^*\right] = \theta_0. \tag{9}$$

Further, it is asymptotically efficient, with an asymptotic variance: $\mathbb{V}\mathrm{ar}\left[\theta^*\right] = \frac{1}{tI(\theta_0)}$, which is equal to Cramér–Rao lower bound [6].

Proof. Detailed proof can be found in Appendix A. □

Such statistical property of unbiasedness refers to the expected value of the sampling distribution of θ^* is equal to the true proficiency θ_0 of student. However, simply knowing that this estimator is unbiased is not very advantageous if the values of θ^* vary greatly from sample to sample and deviate from the true (Figure 3(a)). Fortunately, the second statement above suggests that the variance, $\frac{1}{tI(\theta_0)}$, decreases to zero as test step t grows, and hence the estimate θ^* are increasingly accuracy as t grows (Figure 3(b)).

Theorem 2 (Consistency). *Given any arbitrary small positive quantity ϵ, when $\lambda < I(\theta_0)$ and $m \to \infty$, the estimator θ^* is consistent, that is,*

$$\lim_{t\to\infty} \Pr\left\{\left|\theta^* - \theta_0\right| \geq \epsilon\right\} = 0. \tag{10}$$

Proof. Combining the conclusions in Theorem 1: $\mathbb{E}\left[\theta^*\right] = \theta_0$ and $\mathbb{V}\mathrm{ar} = \frac{1}{tI(\theta_0)}$, for all $\epsilon > 0$, we have $\lim_{t\to\infty} \Pr\left\{|\theta^* - \theta_0| \geq \epsilon\right\} = \lim_{t\to\infty} \Pr\left\{|\theta^* - \mathbb{E}\left[\theta^*\right]| \geq \epsilon\right\} \leq \lim_{t\to\infty} \frac{\mathbb{V}\mathrm{ar}[\theta^*]}{\epsilon^2}$ (Chebyshev's Inequality) $\approx \lim_{t\to\infty} \frac{1}{tI(\theta_0)\epsilon^2} = 0$. This completes the proof. □

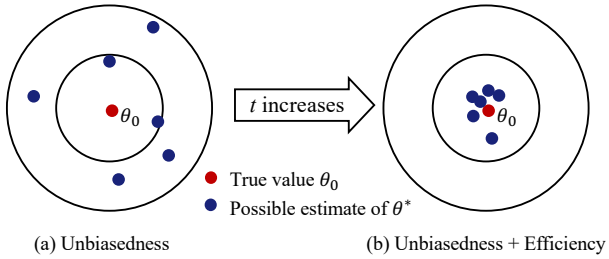


Figure 3: (a) The unbiasedness property of the new proficiency estimation method. Unfortunately, the estimate θ^* may still deviate far from the true. (b) The variance of the estimator θ^* keeps shrinking as the test step t increases (i.e., efficiency), and it surrounds the true (i.e., unbiasedness).

This consistency property intuitively means that the larger the step t, the less is the chance that the difference between θ^* and θ_0 will exceed any fixed value. Therefore, θ^*'s distribution becomes more and more concentrated around the true proficiency θ_0.

5.2 Relations to other methods

We discuss the relations of our proposed RAT with two other related methods, Ensemble Learning and Monte Carlo Method.

5.2.1 Ensemble Learning. Ensemble learning has long been acknowledged to be more robust and accurate than a single model on a wide range of tasks [21, 29, 47, 49]. Its classic techniques include Bagging [5] and Boosting [36]. To get a good ensemble, it is believed that the base model should be accurate and diverse [50], which is similar to the idea of our proposed method. The difference is reflected in two aspects. **(1)** Accuracy: ensemble learning focuses on the accurate prediction (i.e., model's output). Our method focuses on the accuracy of parameter estimation instead. **(2)** Diversity: the diversity in ensemble learning is maintained in *implicit* manner (e.g., instance sampling, instance weighting and different parameter initialization). Our method adds diversity-regularization term in the optimization Eq.(6) to *explicitly* control their diversity.

5.2.2 Monte Carlo Method. Monte Carlo method (or Sampling method) is a collection of techniques for approximating the solution of problems, which make fundamental use of *independent random sampling* [31]. From this perspective, at step t, proficiency estimation in CAT can be viewed as sampling $\hat{\theta}^t$ in the area determined by previous t responses used in estimation. If we sample $\{\hat{\theta}^t_1, \hat{\theta}^t_2, ..., \hat{\theta}^t_m\}$ at step t instead with m non-overlapping response data and $m \to \infty$, the true θ_0 can be approximated using their mean $\frac{1}{m}\sum_{i=1}^{m} \hat{\theta}^t_i \approx \theta_0$, by the law of large numbers [28]. However, *limited* response data incur high correlation/similarity within these estimates (not *independent* sampling), which results in poor approximation. Therefore, with the help of diversity-regularization in RAT, m estimates are forced to be mutual diverse, trying to approach ideal Monte Carlo Method.

Table 1: Statistics of the datasets

Dataset	Math	Junyi	Eedi
#Students	4,211	61,027	263,568
#Questions	472	22,726	27,613
#Response logs	100,833	13,578,787	19,752,063
#Response logs per student	24.0	222.5	74.9
#Response logs per question	213.6	597.5	715.3

6 EXPERIMENTS

In this section, we conduct both quantitative and qualitative experiments on three real-world datasets and perturbed synthetic data to evaluate the effectiveness of our generic method RAT.

6.1 Experiment Setup

6.1.1 Datasets. We conduct experiments on three educational benchmark datasets, namely Junyi, Eedi, and Math. Junyi [33] is from *junyiacademy.org* and consists of millions of exercise attempt logs on its platform over the course of a year (2018-2019). Eedi [46] refers to the dataset in the *NeurIPS 2020 Education Challenge*. And it is collected from two school years (2018-2020) of students' answers to questions from Eedi, an educational platform where millions of students interact daily around the globe. The EXAM dataset was supplied by iFLYTEK Co., Ltd., which contains mathematical exercises and logs of high school examinations. The complete statistical information for datasets is depicted in Table 1. The datasets can be found in https://github.com/bigdata-ustc/EduData.

6.1.2 Evaluation Method. Following the common strategy [16], we split the student–question interactions in the ratio of 7:2:1 for training, validation, and testing by student. The students in validation/testing set won't appear in training set, meeting the standard CAT settings. The training set is mainly used for initially learning some pre-fixed parameters of questions in CDMs (e.g., difficulty), and some data-driven selection algorithm baselines (e.g., BOBCAT).

In the validation/testing, the responses of each student i are further divided into the candidate ($\mathcal{Q}_i$) and meta ($\mathcal{M}_i$) question sets to simulate CAT process, following [3, 16]. Specifically, **(1)** different selection algorithms first select a question from $\mathcal{Q}_i$; **(2)** CDM then updates the proficiency estimate with the corresponding responses; **(3)** evaluate this estimate's accuracy by predicting binary-valued responses on the held-out meta set $\mathcal{M}_i$. In other words, *the better the selection algorithm, the more likely it is to select a best-suited question to improve the estimation accuracy.* Thus, from this binary classification perspective, we use Prediction Accuracy (ACC) [15] and Area Under ROC Curve (AUC) [4] for the evaluation of different selection algorithms. All the methods are developed and trained on two 2.20 GHz Intel Xeon E5-2650 v4 CPUs and a TITAN Xp GPU.

6.1.3 Compared Approaches. Since the selection algorithm must depend on Cognitive Diagnosis Model (CDM) as mentioned above, we choose two classic CDMs: Item Response Theory (**IRT**) [14] and the deep learning-based model (e.g, **NCDM** [43]). The codes of different CDM are available at https://github.com/bigdata-ustc/EduCDM. We use the following state-of-the-art selection algorithms as baselines:

Table 2: The performance of different methods on Student Performance Prediction with ACC and AUC metrics. "–" indicates the information/uncertainty-based selection algorithms (e.g., FSI) cannot be applied to the deep learning CDM (NCDM).

Dataset	Junyi						Eedi						Math					
CDM	IRT			NCDM			IRT			NCDM			IRT			NCDM		
Metric	ACC (%)						ACC (%)						ACC (%)					
Step	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20
Random	70.30	71.73	72.11	70.28	71.96	73.12	62.83	65.88	68.62	62.16	66.30	69.22	72.57	73.88	80.31	72.11	76.12	81.40
FSI	71.25	72.93	74.02	–	–	–	64.63	67.72	70.54	–	–	–	74.07	78.63	83.63	–	–	–
KLI	71.37	72.98	74.92	–	–	–	64.57	67.14	70.08	–	–	–	73.42	77.40	83.14	–	–	–
MAAT	72.31	73.31	75.22	72.44	73.17	75.47	64.86	67.38	71.42	64.22	68.13	71.70	75.84	77.37	82.53	76.36	79.87	82.81
BOBCAT	73.25	73.81	75.89	**73.54**	74.13	76.32	65.58	68.14	72.20	66.30	69.56	72.31	77.19	79.90	82.66	78.36	81.00	85.04
FSI+RAT	73.46	74.82	76.10	–	–	–	66.10	70.39	73.17	–	–	–	77.36	80.75	**84.92**	–	–	–
KLI+RAT	73.76	**75.88**	**77.19**	–	–	–	66.01	70.27	**73.55**	–	–	–	**78.09**	**81.19**	84.57	–	–	–
MAAT+RAT	**73.78**	75.35	76.92	73.10	**75.30**	**77.13**	**66.14**	**70.42**	73.25	**67.35**	**71.65**	**73.37**	77.14	79.71	83.87	**78.38**	**81.14**	**85.05**
Metric	AUC (%)						AUC (%)						AUC (%)					
Step	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20
Random	72.83	73.18	75.32	72.55	74.46	76.87	65.48	68.63	72.20	66.00	69.82	72.55	67.82	67.61	76.90	67.98	70.50	76.97
FSI	73.70	74.28	76.16	–	–	–	67.27	70.72	74.50	–	–	–	69.56	73.13	78.15	–	–	–
KLI	73.91	74.41	76.07	–	–	–	67.10	70.33	73.89	–	–	–	69.82	73.28	78.28	–	–	–
MAAT	74.16	75.32	77.35	75.27	75.91	78.32	67.19	70.32	74.74	67.13	71.36	74.73	69.10	73.90	78.89	69.67	75.15	78.90
BOBCAT	75.99	76.25	78.49	**75.81**	76.33	79.64	68.43	71.03	75.76	69.11	72.01	76.13	70.62	74.32	79.19	**71.17**	74.54	79.58
FSI+RAT	**76.56**	76.64	78.86	–	–	–	**68.93**	**73.12**	75.99	–	–	–	70.89	76.17	79.38	–	–	–
KLI+RAT	76.33	**77.94**	**79.67**	–	–	–	68.90	72.99	76.03	–	–	–	**71.03**	76.01	**80.66**	–	–	–
MAAT+RAT	75.67	77.74	79.41	75.33	**77.06**	**79.83**	**68.93**	73.05	**76.09**	**70.39**	**73.88**	**76.63**	70.44	**77.41**	79.14	70.44	**76.40**	**80.63**

- **Random**: The random selection strategy is a benchmark to quantify the improvement of other methods.
- **FSI** [27]: It selects the question with the maximum Fisher information, which is one of the most widely used selection algorithms. This method is specially designed for IRT.
- **KLI** [9]: It utilizes Kullback-Leibler information to measure the divergence between two consecutive posteriors of proficiency. It is also specially designed for IRT.
- **MAAT** [3]: It's an active learning-based method, which measures uncertainty towards individual students by calculating Expected Model Change (EMC) of CDM, caused by each candidate question. It is agnostic to the underlying CDM.
- **BOBCAT** [16]: It's a recently proposed method, which recasts CAT as a bilevel optimization problem and optimizes via meta/reinforcement learning. Thus, it learns a data-driven selection algorithm directly from large-scale student response data. Following their settings, we use a fully-connected network (with 2×256 hidden layers, Tanh nonlinearity, and a softmax output layer) as the question selection algorithm. It is also agnostic to the underlying CDM.

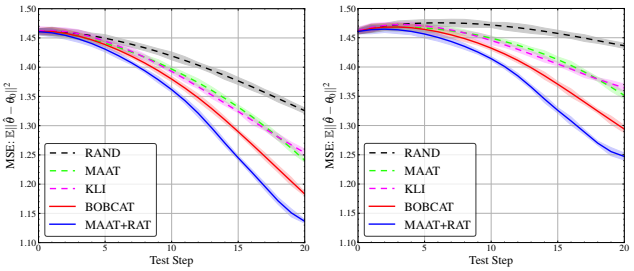
Note that our generic approach RAT can be applied to all selection algorithms (indicated by **X+RAT**) except for BOBCAT, since BOBCAT changes the original paradigm of CAT. Also, BOBCAT is the only deep learning-based baseline recently proposed, and other reinforcement learning methods in related work cannot be verified on real-world datasets. Meanwhile, to further verify RAT's effectiveness, we also compare with powerful and generic ensemble methods (i.e., **Bagging** [5] and **AdaBoost** [7]).

6.1.4 Implementation Details. We implement all the methods with PyTorch and optimize them with the Adam algorithm [19]. The batch size is fixed to 128 and the learning rate is fixed to 0.001 in the whole process. Since 20 is enough for a standard CAT, we set the maximum test length $T = 20$. The number of estimates m at each step is set to 5 and the hyperparameter λ in estimation formulation is set to 0.1. More detailed analysis of m and λ can be found in Section 6.4.

6.2 Comparison with Baselines

We calculate the ACC and AUC scores on three datasets for evaluation and the performances are shown in Table 2. From them, we have the following key findings.

- *Finding 1 – The selection algorithms that applied our proposed method (X+RAT) outperform almost all baselines on three datasets.* Table 2 shows the comparison results of all methods on these datasets. X+RAT achieves 1.0∼2.0 ACC/AUC score improvements compared to the baseline without RAT. E.g. MAAT+RAT achieves 1.61 AUC points (on average) improvement over the baseline MAAT, which clearly demonstrates the effectiveness and universality of RAT. Specifically, our MAAT+RAT also surpasses the strong model BOBCAT in Eedi, which requires additional training on large-scale student response data.
- *Finding 2 - The strength of the deep learning method (BOBCAT) cannot be underestimated.* Either of X+RAT and BOBCAT can not consistently beat the other; both of them show their distinct strengths. BOBCAT showed similar performance to the RAT method in the initial stage of test (i.e., step 5), and even surpassed all RAT methods on Junyi dataset with NCDM. This may be because it leverages the neural networks in meta-learning framework, which is good at alleviating the cold-start problem [23, 44]. In addition, the lack of response in the initial stage of test may limit all algorithms including RAT due to greater uncertainty. Although deep learning-based models are not practical for



(a) Guess: 25%, Slip: 5%. (b) Guess: 50%, Slip: 10%.

Figure 4: MSE curves along the simulated CAT process at two different noise levels: (a) shows the results in Guess: 25% (4-choice question), Slip: 5%. (b) shows the results in Guess: 50% (True or False), Slip: 10%.

real CAT scenarios (i.e., enormous questions are uploaded daily), adapting the proposed RAT to enhance them is a very promising future work.

- *Finding 3 - Traditional methods have similar potential and capacity, which may be hindered by the traditional estimation method before.* From Table 2, our RAT framework improves the performance on all types of selection algorithms baselines, e.g., 1.43 points and 1.34 points ACC improvement (step 20, IRT) on Math over the two baselines KLI and MAAT, respectively. Interestingly, more improvement can be achieved when the baseline model is not so strong, e.g., 2.07 ACC gain (on average) above the FSI baseline, which is very classic and widely used. As a result, the (FSI, KLI, MAAT)+RAT methods enhance their baselines towards almost the same level, which is obvious from their competitive performances. These findings inspire us: **1)** The query generated by traditional proficiency estimation methods hinders these algorithms, while they essentially have similar capacity and potential. **2)** We can utilize RAT to strengthen these less powerful methods, especially when they have to be deployed in some scenarios with limited data and time delay.

6.3 Further Study

Following the comparison with baselines, we will further introduce the simulation experiments of proficiency estimation (quantitative), visualization of estimates (qualitative), and the comparison with powerful ensemble methods.

6.3.1 Simulation of Proficiency Estimation under Guess & Slip. Following CAT's traditional evaluation methods [41], since the ground truth of student proficiency θ_0 is not available, we artificially generate their θ_0 and simulate student-question interaction process using simple IRT. To further verify the robustness of RAT, we expose this simulated CAT to various perturbations:

- Guess factors: When faced with a multiple-choice question with 4 options, even if the student doesn't master it, there is a 25% chance of answering it correctly. (The label is changed from 0 to 1 with 25% probability.)

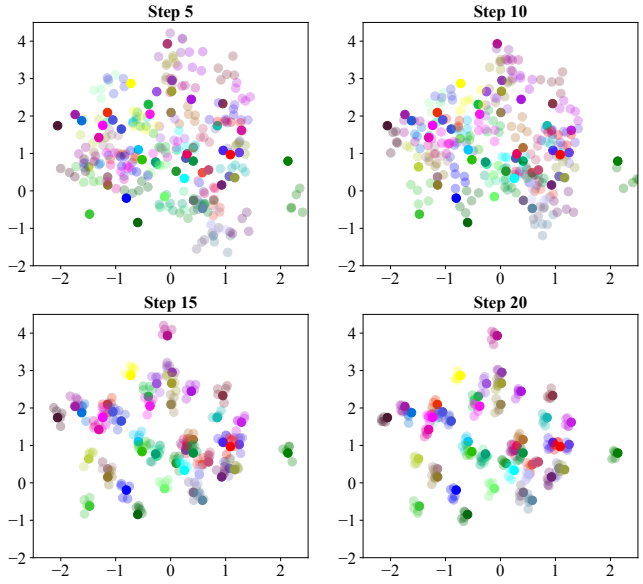


Figure 5: Visualization of 50 students' proficiency estimates using RAT at step 5, 10, 15, and 20. Different colors represent different students: opaque colors represent their true proficiency θ_0 and the translucent represent m estimates ($m = 5$).

- Slip factors: There may be a small chance (e.g., 5%) to accidentally or carelessly fail an item that students could have solved. (The label is changed from 1 to 0 with 5% probability.)

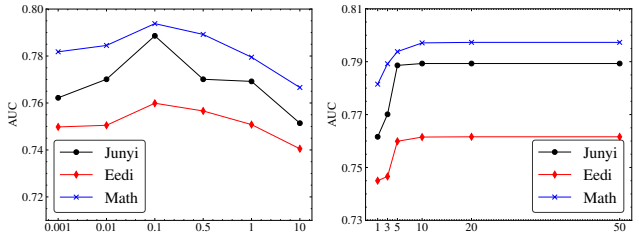
We utilize the parameters of questions and students trained on the real-world datasets Eedi as the ground truth, instead of generating them. The results at two noise levels are reported in Figure 4. When step increases, the MSE metric $\mathbb{E}||\hat{\theta} - \theta_0||^2$ of all methods decrease steadily, and the RAT achieves a much superior performance. The performance gap between baseline and RAT also shows a growing trend. At higher noise level (Figure 4(b)), the MSE of all methods rise slightly in the first steps, but consistent improvements are finally obtained. These observations demonstrate that our RAT can still benefit the selection performance even under various perturbations. Also, it motivates us to pay more attention to the complex relationship in students-questions, which may be the key to boosting CAT's efficiency. In a word, it clearly shows RAT provides a robust estimation of student proficiency by fusing m estimates at each step.

6.3.2 Estimate Analysis. To make a deep analysis of the student's multi-facet proficiency estimation, we visualize all estimates generated at each step in the simulation process. For intuitively observing their relevance and the dynamic process in RAT, we utilize the two-dimensional IRT and randomly choose 50 students to output their estimates ($m = 5$) along such process. As illustrated in Figure 5, we represent each student in different colors: the opaque represent their true proficiency θ_0 and the translucent represent m estimates.

For one thing, with the help of diversity-regularization term in Eq.(6), the multiple estimates at each step are diverse and their distribution is relatively uniform. For another, in the initial stage, the range of all possible proficiency is large; as the step increases,

Table 3: Comparison with ensemble methods on Math dataset using NCDM at step 20.

Metric	ACC	IMP(%)	AUC	IMP(%)	MSE	IMP(%)
MAAT	82.81	2.70	78.90	2.19	0.47	23.40
MAAT+AdaBoost	82.91	2.58	79.23	1.77	0.42	14.29
MAAT+Bagging	83.32	2.08	79.55	1.36	0.42	14.29
MAAT+RAT	**85.05**	–	**80.63**	–	**0.36**	–



(a) Weight of λ. (b) Number of estimates m.

Figure 6: Parameter Sensitivity w.r.t the weight of λ and number of estimates m.

the scope of them gradually shrinks and approaches the true θ_0. Therefore, these observations further demonstrate the effectiveness of the new proficiency estimation method in RAT.

6.3.3 RAT vs Ensemble Methods. As mentioned in Section 5.2, our proposed method and ensemble learning have some similarities in form. Thus, Table 3 shows the comparisons with two generic and powerful ensemble methods: Bagging [5] and AdaBoost [7]. The baseline algorithm (MAAT) in RAT consistently surpassed the other two ensemble methods. The relative improvements to MAAT+Bagging are 2.08% with respect to ACC while the relative improvements to MAAT+AdaBoost can be up to 2.58%. The advantages in MSE are even more impressive. MAAT+RAT achieves the state-of-the-art MSE score and performance improvements are at least 14.29%. Surprisingly, the ensemble methods that have outstanding performance in other tasks (e.g., CV [2] and RS [48]) have little effect on CAT. It may be due to the following two reasons. **1)** Scarce response/training data: The advantage of ensemble learning is based on a large-scale training samples [50]. When faced with each student's responses in the CAT (up to 20), they are prone to overfitting as a result. **2)** Different goals: The goal of traditional ensemble learning is the accuracy of prediction, while CAT is to pursue accurate parameter estimation (see Section 5.2 for details); and this can be clearly verified from the MSE comparisons in Table 3. In short, compared to ensemble methods, RAT is less prone to overfitting and more suitable for CAT application scenarios.

6.4 Parameter Sensitivity

Here, we explore the sensitivity of two important hyperparameters in our method: diversity-regularization weight λ and estimation size m at each step.

6.4.1 Effect of Weight λ. Here we vary the λ in {0.001, 0.01, 0.1, 0.5, 1, 10} and conduct experiments. As shown in Figure 6(a), small λ (e.g., 0.001) can not perform as good as large λ (e.g., 0.1), which means we should pay more attention to the diversity regularization during estimation. However, the performance decreases when λ continuously increases. The reason is that too large diversity-regularization may introduce more randomness which will reduce accuracy, verifying the condition about λ in Theorem 1. The best balanced choice is $\lambda = 0.1$.

6.4.2 Effect of the Number of Estimates m. Figure 6(b) reports the effect of the number of estimates. We vary m in the set {1, 3, 5, 10, 20, 50}. We find that RAT is not sensitive to this hyper-parameter when m is greater than 5. This observation ensures that m does not need to be large to achieve better performance, although theorems are based on $m \to \infty$. In addition, this conclusion also reveals RAT could get a further balance between performance and computation overhead by utilizing relatively small m.

7 CONCLUSION

This paper focuses on the foundation of personalized online education and presents a generic optimization criterion called RAT for educational measurement. It is a simple yet very effective method, which alleviates the lack of robustness in CAT by fusing multiple proficiency estimates for individual students. Theoretically, we show that the new estimator in RAT possesses highly desirable statistical properties [34]: asymptotic unbiasedness, efficiency, and consistency. Experimental results demonstrate that RAT maintains strong performance at reducing test length even under high noise rates. Furthermore, the thorough comparisons prove that compared with designing another sophisticated selection algorithm, enhancing the query in selection can also achieve competitive performances. In future work, we will apply our approaches to other datasets and applications.

ACKNOWLEDGMENTS

This research was partially supported by grants from the National Key Research and Development Program of China (Grant No. 2021 YFF0901003), the National Natural Science Foundation of China (Grants No. 61922073, U20A20229, and No. 62106244), and the Fundamental Research Funds for the Central Universities (Grants No. WK2150110021).

REFERENCES

[1] Terry A Ackerman, Mark J Gierl, and Cindy M Walker. 2003. Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice* 22, 3 (2003), 37–51.
[2] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. 2018. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9368–9377.
[3] Haoyang Bi, Haiping Ma, Zhenya Huang, Yu Yin, Qi Liu, Enhong Chen, Yu Su, and Shijin Wang. 2020. Quality meets Diversity: A Model-Agnostic Framework for Computerized Adaptive Testing. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 42–51.
[4] Andrew P Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 30, 7 (1997), 1145–1159.
[5] Peter Bühlmann. 2012. Bagging, boosting and ensemble methods. In *Handbook of computational statistics*. Springer, 985–1022.
[6] S Cavassila, S Deval, C Huegen, D Van Ormondt, and D Graveron-Demilly. 2001. Cramér–Rao bounds: an evaluation tool for quantitation. *NMR in Biomedicine:*

An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo 14, 4 (2001), 278–283.
[7] Jonathan Cheung-Wai Chan and Desiré Paelinckx. 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment* 112, 6 (2008), 2999–3011.
[8] Hua-Hua Chang. 2015. Psychometrics behind computerized adaptive testing. *Psychometrika* 80, 1 (2015), 1–20.
[9] Hua-Hua Chang and Zhiliang Ying. 1996. A global information approach to computerized adaptive testing. *Applied Psychological Measurement* 20, 3 (1996), 213–229.
[10] Suming Jeremiah Chen, Arthur Choi, and Adnan Darwiche. 2015. Computer Adaptive Testing Using the Same-Decision Probability.. In *BMA@ UAI*. 34–43.
[11] Ying Cheng. 2008. *Computerized adaptive testing—new developments and applications.* University of Illinois at Urbana-Champaign.
[12] Jimmy De La Torre. 2009. DINA model and parameter estimation: A didactic. *Journal of educational and behavioral statistics* 34, 1 (2009), 115–130.
[13] Bradley Efron and David V Hinkley. 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* 65, 3 (1978), 457–483.
[14] Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.
[15] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. 2021. RCD: Relation Map Driven Cognitive Diagnosis for Intelligent Education Systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 501–510.
[16] Aritra Ghosh and Andrew Lan. 2021. BOBCAT: Bilevel Optimization-Based Computerized Adaptive Testing. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 2410–2417.
[17] Giles Hooker, Matthew Finkelman, and Armin Schwartzman. 2009. Paradoxical results in multidimensional item response theory. *Psychometrika* 74, 3 (2009), 419–442.
[18] Zhenya Huang, Qi Liu, Yuying Chen, Le Wu, Keli Xiao, Enhong Chen, Haiping Ma, and Guoping Hu. 2020. Learning or forgetting? A dynamic approach for tracking the knowledge proficiency of students. *ACM Transactions on Information Systems (TOIS)* 38, 2 (2020), 1–33.
[19] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*. http://arxiv.org/abs/1412.6980
[20] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
[21] Oren Kurland and J Shane Culpepper. 2018. Fusion in information retrieval: Sigir 2018 half-day tutorial. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1383–1386.
[22] Andrew S Lan, Andrew E Waters, Christoph Studer, and Richard G Baraniuk. 2014. Sparse factor analysis for learning and content analytics. *Journal of Machine Learning Research (JMLR)* (2014).
[23] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. 2019. Melu: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1073–1082.
[24] Xiao Li, Hanchen Xu, Jinming Zhang, and Hua-hua Chang. 2020. Deep reinforcement learning for adaptive learning systems. *arXiv preprint arXiv:2004.08410* (2020).
[25] Qi Liu. 2021. Towards a New Generation of Cognitive Diagnosis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization.
[26] Ting Long, Yunfei Liu, Jian Shen, Weinan Zhang, and Yong Yu. 2021. Tracing Knowledge State with Individual Cognition and Acquisition Estimation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 173–182.
[27] Frederic M Lord. 2012. *Applications of item response theory to practical testing problems*. Routledge.
[28] R. B. Lund. 1997. *Markov Processes for Stochastic Modeling*. Markov processes for stochastic modeling.
[29] Deyuan Meng, Yingmin Jia, Junping Du, and Fashan Yu. 2008. Robust design of a class of time-delay iterative learning control systems with initial shifts. *IEEE Transactions on Circuits and Systems I: Regular Papers* 56, 8 (2008), 1744–1757.
[30] Darkhan Nurakhmetov. 2019. Reinforcement learning applied to adaptive classification testing. In *Theoretical and Practical Advances in Computer-based Educational Measurement*. Springer, Cham, 325–336.
[31] Penelope L Peterson, Eva Baker, and Barry McGaw. 2010. *International encyclopedia of education*. Elsevier Ltd.
[32] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, and Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. In *NIPS*. 505–513. http://papers.nips.cc/paper/5654-deep-knowledge-tracing
[33] Chen Pojen, Hsieh Mingen, and Tsai Tzuyang. 2020. Junyi Academy Online Learning Activity Dataset: A large-scale public online learning activity dataset from elementary to senior high school students. *Dataset available from https://www.kaggle.com/junyiacademy/learning-activity-public-dataset-by-junyi-academy* (2020).
[34] Sheldon M Ross. 2014. *A first course in probability*. Pearson.
[35] Lawrence M Rudner. 2002. An examination of decision-theory adaptive testing procedures. In *annual meeting of the American Educational Research Association*.
[36] Robert E Schapire. 2013. Explaining adaboost. In *Empirical inference*. Springer, 37–52.
[37] Andreas Toscher and Michael Jahrer. 2010. Collaborative filtering applied to educational data mining. *KDD cup* (2010).
[38] Wim J van der Linden. 1998. Bayesian item selection criteria for adaptive testing. *Psychometrika* 63, 2 (1998), 201–216.
[39] Wim J Van der Linden and Peter J Pashley. 2000. Item selection and ability estimation in adaptive testing. In *Computerized adaptive testing: Theory and practice*. Springer, 1–25.
[40] Wim JJ Veerkamp and Martijn PF Berger. 1997. Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics* 22, 2 (1997), 203–226.
[41] Jill-Jênn Vie, Fabrice Popineau, Éric Bruillard, and Yolaine Bourda. 2017. A review of recent advances in adaptive assessment. *Learning analytics: fundaments, applications, and trends* (2017), 113–142.
[42] Matthias Von Davier. 2014. The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *Brit. J. Math. Statist. Psych.* 67, 1 (2014), 49–71.
[43] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6153–6161.
[44] Jianling Wang, Kaize Ding, and James Caverlee. 2021. Sequential Recommendation for Cold-start Users with Meta Transitional Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1783–1787.
[45] Xinping Wang, Caidie Huang, Jinfang Cai, and Liangyu Chen. 2021. Using Knowledge Concept Aggregation towards Accurate Cognitive Diagnosis. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2010–2019.
[46] Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. 2020. Diagnostic questions: The neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061* (2020).
[47] Zhe Xue, Junping Du, Dawei Du, and Siwei Lyu. 2019. Deep low-rank subspace ensemble for multi-view clustering. *Information Sciences* 482 (2019), 210–227.
[48] Shijie Zhang, Hongzhi Yin, Tong Chen, Quoc Viet Nguyen Hung, Zi Huang, and Lizhen Cui. 2020. Gcn-based user representation learning for unifying robust recommendation and fraudster detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 689–698.
[49] Kuo Zhong, Ying Wei, Chun Yuan, Haoli Bai, and Junzhou Huang. 2020. TranSlider: Transfer Ensemble Learning from Exploitation to Exploration. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 368–378.
[50] Zhi-Hua Zhou. 2012. *Ensemble methods: foundations and algorithms*. CRC press. https://doi.org/10.1201/b12207
[51] Yan Zhuang, Qi Liu, Zhenya Huang, Zhi Li, Shuanghong Shen, and Haiping Ma. 2022. Fully Adaptive Framework: Neural Computerized Adaptive Testing for Online Education. (2022).

A PROOFS OF THEOREM 1

In this section, we introduce detailed proofs of asymptotic unbiasedness and efficiency of the estimator θ^* in RAT. To be intuitive, these proofs only consider one-dimensional case (the multi-dimensional is similar).

A.1 Asymtoptic Unbiasedness

We consider a random variable X (i.e., student's response) for which the pdf or pmf is $f(X|\theta)$ (i.e., the distribution of correctness determined by CDM), where θ represents the latent proficiency, t is the number of instances (i.e., test rounds). Therefore, the MLE loss

(binary cross-entropy loss) is

$$\mathcal{L}_{MLE}(\theta) = -\frac{1}{t}\sum_{i=1}^{t}\log f(X_i|\theta).$$

When we optimize the i-th estimate, $i = 1, 2, ..., m$ and m is the total number of estimates at each step, our target function Eq.(6) is

$$\mathcal{L}(\theta_i) = \mathcal{L}_{MLE}(\theta_i) - \frac{\lambda}{2}(\theta_i - \theta_i^*)^2,$$

where $\theta_i^* = \frac{1}{i-1}\sum_{k=1}^{i-1}\hat{\theta}_k$. We apply Taylor expansion on $\nabla\mathcal{L}(\theta_i)$ at the point θ_0, and $\hat{\theta}_i$ is the solution to $\nabla\mathcal{L}(\theta_i) = 0$ [13, 34], yielding

$$0 = \nabla\mathcal{L}(\hat{\theta}_i) \approx \nabla\mathcal{L}_{MLE}(\theta_0) - \lambda(\theta_0 - \theta_i^*) + (\hat{\theta}_i - \theta_0)(\nabla^2\mathcal{L}_{MLE}(\theta_0) - \lambda).$$

Therefore,

$$\theta^* = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i \approx \frac{\theta_0\nabla^2\mathcal{L}_{MLE}(\theta_0) - \frac{\lambda}{m}\sum_i\theta_i^* - \nabla\mathcal{L}_{MLE}(\theta_0)}{\nabla^2\mathcal{L}_{MLE}(\theta_0) - \lambda}. \quad (11)$$

We intend to prove $\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i\right] = \theta_0$. So we focus on the $\frac{1}{m}\sum_{i=1}^{m}\theta_i^*$ in Eq.(11): we have $\theta_1^* = 0$ and $\theta_k^* = \frac{\hat{\theta}_1+\hat{\theta}_2+...+\hat{\theta}_{k-1}}{k-1}$. Therefore, we find the following relationship: $(k-1)\theta_k^* = (k-2)\theta_{k-1}^* + \hat{\theta}_{k-1} = (k-2)\theta_{k-1}^* + \frac{\theta_0\nabla^2\mathcal{L}_{MLE}(\theta_0) - \lambda\theta_{k-1}^* - \nabla\mathcal{L}_{MLE}(\theta_0)}{\nabla^2\mathcal{L}_{MLE}(\theta_0) - \lambda}$. Further, let $a = \frac{-\lambda}{\nabla^2\mathcal{L}_{MLE}(\theta_0) - \lambda}$ and $b = \frac{\theta_0\nabla^2\mathcal{L}_{MLE}(\theta_0) - \nabla\mathcal{L}_{MLE}(\theta_0)}{\nabla^2\mathcal{L}_{MLE}(\theta_0) - \lambda}$ then

$$\begin{cases} (k-1)\theta_k^* &= (k-2)\theta_{k-1}^* + a\theta_{k-1}^* + b \\ \theta_1^* &= 0 \end{cases}$$

Solve the recurrence relation above:

$$\begin{aligned}\theta_k^* &= \frac{b\Gamma(a+k-1)}{\Gamma(k)}\sum_{t=1}^{k-1}\frac{\Gamma(k-t)}{\Gamma(a+k-t)} \\ &= \frac{b\Gamma(a+k-1)}{(a-1)\Gamma(k)}\sum_{t=1}^{k-1}\frac{\Gamma(k-t)}{\Gamma(a+k-t-1)} - \frac{(k-t)\Gamma(k-t)}{(a+k-t-1)\Gamma(a+k-t-1)} \\ &= \frac{b}{a-1}\frac{\Gamma(a+k-1)}{\Gamma(a)\Gamma(k)} - \frac{b}{a-1}\end{aligned}$$

and θ_i^* has such a relationship: $(m-1)\theta_m^* = a\sum_{i=1}^{m-1}\theta_i^* + (m-1)b$. Therefore,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\theta_i^* &= \frac{m-1+a}{am}\theta_m^* - \frac{m-1}{am}b \\ &= \left(\frac{b}{a(a-1)} + \frac{b}{am}\right)\frac{\Gamma(a+m-1)}{\Gamma(a)\Gamma(m)} + \frac{b}{1-a}. \quad (12)\end{aligned}$$

Next, we consider $\nabla^2\mathcal{L}_{MLE}(\theta_0) = -\frac{1}{t}\sum_{i=1}^{t}\nabla^2\log f(X_i|\theta)$. By the law of large numbers, this expression converges to

$$\mathbb{E}\left[\nabla^2\log f(X_i|\theta)\right] = -I(\theta_0),$$

where $I(\theta_0) \geq 0$ is the Fisher information, and we have

$$\nabla^2\mathcal{L}_{MLE}(\theta_0) \approx I(\theta_0).$$

thus $a \approx \frac{\lambda}{\lambda - I(\theta_0)}$, $b \approx \frac{\theta_0 I(\theta_0) - \nabla\mathcal{L}_{MLE}(\theta_0)}{I(\theta_0) - \lambda}$.

When $a - 1 < 0$ (i.e., $\lambda < I(\theta_0)$) and $m \to \infty$, in Eq.(12), we have

$$\frac{\Gamma(a+m-1)}{\Gamma(a)\Gamma(m)} \approx 0.$$

Hence,

$$\frac{1}{m}\sum_{i=1}^{m}\theta_i^* \approx \frac{b}{1-a} = \theta_0 - \frac{\nabla\mathcal{L}_{MLE}(\theta_0)}{\nabla^2\mathcal{L}_{MLE}(\theta_0)}.$$

According to [13]

$$\mathbb{E}[\nabla\mathcal{L}_{MLE}(\theta_0)] = -\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[\nabla\log f(X_i|\theta_0)\right] = 0.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\theta^*\right] &\approx \mathbb{E}\left[\frac{\theta_0\nabla^2\mathcal{L}_{MLE}(\theta_0) - \frac{\lambda}{m}\sum_i\theta_i^* - \nabla\mathcal{L}_{MLE}(\theta_0)}{\nabla^2\mathcal{L}_{MLE}(\theta_0) - \lambda}\right] \\ &\approx \frac{1}{I(\theta_0) - \lambda}\mathbb{E}\left[\theta_0\nabla^2\mathcal{L}_{MLE}(\theta_0) - \nabla\mathcal{L}_{MLE}(\theta_0) - \lambda\theta_0 + \lambda\frac{\nabla\mathcal{L}_{MLE}(\theta_0)}{\nabla^2\mathcal{L}_{MLE}(\theta_0)}\right] \\ &= \frac{1}{I(\theta_0) - \lambda}\mathbb{E}\left[\theta_0 I(\theta_0) - \lambda\theta_0\right] = \theta_0.\end{aligned}$$

This completes the proof of unbiasedness.

A.2 Asymptotic Efficiency

Combining the above conclusions and let $Y = \frac{1}{m}\sum_{i=1}^{m}\theta_i^*$ then

$$\mathbb{E}[Y] \approx \mathbb{E}\left[\theta_0 - \frac{\nabla\mathcal{L}_{MLE}(\theta_0)}{\nabla^2\mathcal{L}_{MLE}(\theta_0)}\right] = \theta_0.$$

Because $\mathbb{E}\left\{[\nabla\mathcal{L}_{MLE}(\theta_0)]^2\right\} = \frac{I(\theta_0)}{t}$ (FIsher Information's definition),

$$\mathbb{E}\left[Y^2\right] \approx \mathbb{E}\left[\theta_0 - \frac{\nabla\mathcal{L}_{MLE}(\theta_0)}{\nabla^2\mathcal{L}_{MLE}(\theta_0)}\right] = \theta_0^2 + \frac{1}{tI(\theta_0)},$$

$$\mathbb{E}[Y\nabla\mathcal{L}_{MLE}(\theta_0)] \approx \mathbb{E}\left[\theta_0\nabla\mathcal{L}_{MLE}(\theta_0) - \frac{\nabla\mathcal{L}_{MLE}(\theta_0)^2}{\nabla^2\mathcal{L}_{MLE}(\theta_0)}\right] = -\frac{1}{t}.$$

Hence,

$$\begin{aligned}\mathbb{E}\left[(\theta^*)^2\right] &= \mathbb{E}\left[\left(\frac{\theta_0\nabla^2\mathcal{L}_{MLE}(\theta_0) - \lambda Y - \nabla\mathcal{L}_{MLE}(\theta_0)}{\nabla^2\mathcal{L}_{MLE}(\theta_0) - \lambda}\right)^2\right] \\ &= \frac{1}{(I(\theta_0) - \lambda)^2}\left[\theta_0^2 I(\theta_0)^2 + \lambda^2\mathbb{E}\left[Y^2\right] + \mathbb{E}\left\{[\nabla\mathcal{L}_{MLE}(\theta_0)]^2\right\}\right. \\ &\qquad \left. - 2\lambda\theta_0 I(\theta_0)\mathbb{E}[Y] + 2\lambda\mathbb{E}\left[Y\nabla\mathcal{L}_{MLE}(\theta_0)\right]\right] \\ &\approx \frac{\theta_0^2 I(\theta_0)^2 + \lambda^2\theta_0^2 - 2\lambda\theta_0^2 I(\theta_0)}{(I(\theta_0) - \lambda)^2} + \frac{\lambda^2 + I(\theta_0)^2 - 2\lambda I(\theta_0)}{(I(\theta_0) - \lambda)^2 tI(\theta_0)} \\ &= \theta_0^2 + \frac{1}{tI(\theta_0)}.\end{aligned}$$

Therefore the variance of estimator θ^*:

$$\mathbb{V}\text{ar}\left[\theta^*\right] = \mathbb{E}\left[(\theta^*)^2\right] - \mathbb{E}^2\left[\theta^*\right] \approx \theta_0^2 + \frac{1}{tI(\theta_0)} - \theta_0^2 = \frac{1}{tI(\theta_0)}.$$

The asymptotic variance $\frac{1}{tI(\theta_0)}$ equal to the Cramér–Rao lower bound, which proves our method is asymptotically efficient.