A Maximum Likelihood Based Offline Estimation of Student Capabilities and Question Difficulties

Shana Moothedath, Prasanna Chaporkar and Madhu N. Belur

***Abstract*— In recent years Computerized Adaptive Test (CAT) has gained popularity over conventional exams in evaluating student capabilities with desired accuracy. However, the key limitation of CAT is that it requires a large pool of pre-calibrated questions. In absence of such a calibrated question bank, offline exam with uncalibrated questions has to be conducted. Even today many important exams are offline, e.g., Graduated Aptitude Test in Engineering (GATE) and Joint Entrance Examination (JEE) that are conducted in India. In offline exams, typically, normalized marks are used as an estimate of the students' capabilities. In this work, our key contribution is to verify whether marks obtained are indeed a good measure of students' capabilities. To this end, we propose an evaluation methodology that mimics evaluation process of CAT. In our approach, based on the marks scored by students in various questions, we iteratively estimate question parameters like difficulty and discrimination, and student parameters like capability. Our algorithm uses alternating maximization to maximize the log likelihood estimate for the questions' and students' parameters given the marks. We prove that the alternating maximization process converges. We compare our approach with marks based evaluation using simulations. The simulation results show that our approach outperforms marks based evaluation.**

Keywords: **Logistic ogive, alternating maximization, concave function, offline exam.**

1. INTRODUCTION

Exam based assessment is the long-established method to filter out the most deserving candidates from the enormous number of applicants. The idea of conducting a subjective exam is not an appreciable method as it is time consuming and evaluation procedure could be non-uniform. Objective exam is a universally accepted alternative for this. All competitive exams with a large number of examinees follow the objective pattern where a correct response to a question carries a certain credit and an incorrect response carries either no credit or a negative credit. However, in this case guessing is a major concern, which led practitioners to think about a new way of answering in the competitive exams. The idea of entering the final answer value in the response sheet is the recently adopted refined way of answering subjective exams.

A special case of computerized exam conduction and evaluation is Computerized Adaptive Testing (CAT). The outstanding feature of an online adaptive test over conventional subjective exams and offline exams, is its ability to adaptively modify the set of questions in such a way that each examinee attempts a self tailored set of questions from a precalibrated pool, where a question that an examinee is going to answer next is determined by his previous responses [**?**]. The term 'offline' refers to the fact that scores of examinees not only depend on their own performance but also on the aggregate nature of the exam. More precisely, the evaluation results can be obtained only after all examinees complete their exam. Adaptive feature of CAT ensures that the number of easy question attempts by a highly capable student is very less, as easy questions gives very less information about the ability of a highly capable person. Similarly, in a CAT the number of difficult questions attempted by a less capable student is very less, as a difficult question response by a less capable student barely gives any information about his capability. This is because for a very difficult question, answer of a less capable is always biased by guessing, which thereby skews the whole assessment procedure. Thus by adaptively rendering questions of known difficulties, a bias free estimate of the student capability can be obtained through an optimal short exam with fewer questions.

Because of confidentiality issues and the difficulty involved in setting up pools of precalibrated questions for different disciplines, competitive exams like Joint Entrance Examination (JEE) and Graduate Aptitude Test in Engineering (GATE), conducted in India are offline. Besides, for competitive exams with validity more than a year, scores should be comparable across years and in certain circumstances across disciplines as well. Comparison of scores across disciplines is crucial when students with scores in different disciplines apply for a common course, like a student with score in computer science applying for electrical engineering. Now the relevant question is "How efficient is the existing assessment in ranking students based on their exam scores?". If the total marks is a good estimate of the ability, then the whole assessment procedure is easy and straight forward. The student who gets the maximum marks is considered as the most capable student and that particular question which has fewer number of correct responses is considered as the most difficult question. But then, we need to address the following question, when is the total marks indeed a good estimate of capability? The key point here is one or two unintended mistakes can cause significant lowering of the student's rank when marks based assessment is used. This is because, typically many examinees can get the same total score.

Further, since the question difficulties are not known a priori, assessing students with a small number of questions may result in an erroneous assessment. In this regard, we

The authors are in the Department of Electrical Engineering, Indian Institute of Technology Bombay, India. Corresponding author email: shana@ee.iitb.ac.in. This work was supported in part by SERB (DST) and BRNS, India.

are forced to conduct a long duration exam with a large number of questions. But then, factors like attention span of a student, incidental guessing due to lack of time, fatigue and more, also come into play which may result in a skewed response.

Additionally, computerized offline exams are conducted on slot basis, where question papers for different slots are different, even for the same discipline. However, all the candidates who take exams across all slots for the same discipline are often required to be ranked as one group. Since the questions are not pre-calibrated and students in different slots appear for different question sets, marks based ranking may not be a good option. In this situation it is imperative to find out an optimum method for offline assessment of examinees' abilities. That is, given the responses of all the exam takers for all the questions, we need to estimate their capabilities in the most accurate way, thereby ensuring that the assessment process sieves out the most deserving candidates.

A. Background

The first theory of psychometric test analysis, called classical test theory (CTT) arose in the year 1906 through the seminal work of Charles Spearman. However the basic constructs of the theory was reformulated using modern mathematical statistical approach by Lord and Novick in the year 1966 [**?**]. CTT, also known as the true score theory, relies on the accurate estimation of *true score* from the *observed score*, where the *true score* is the number of correct responses of a student to infinite number of independent items. The major shortcoming of CTT is that instead of separating examinee and item characteristics, it can only interpret nature of one in terms of other. This limitation of CTT has led to the emergence of item based test theory, called Item Response Theory (IRT). Start of the item based test theory was marked in the year 1916 by the work of Binet-Simon [**?**]. IRT attracted more attention of practitioners in the year 1943, when Lawley through his well appreciated paper [**?**], showed how to obtain the maximum likelihood estimates of parameters of the item characteristic curve.

Most of the previous work in this area use iterative Newton-Raphson, Marginal Maximum Likelihood Estimation (MMLE) and Bayesian parameter estimation techniques for the estimation of exam parameters. Through many practical examples it has been shown that Newton-Raphson method is robust and Newton-Raphson/Fisher equations always converges nearly [**?**]. The convergence of MMLE and Bayesian parameter estimation methods are data set dependent [**?**]. Existence of a global maxima for the likelihood function is stated by Baker by plotting for different values of parameters [**?**]. However, there are no formal proofs for the existence of a unique maxima of the likelihood function. Also there are no known algorithms that guarantee to give the unique optimizing value of the parameters always.

B. Summary of Contribution

In this paper, we propose a maximum likelihood estimation based algorithm for offline exam assessment using alternating maximization. Given the responses of all the students as input, our algorithm outputs the maximum likelihood estimate of student capability and question difficulty and discrimination. The main feature of our algorithm is that it allows us to conduct a short offline exam with fewer number of questions, and even then gives accurate results like the online adaptive exam. Convergence of the alternating maximization of the log-likelihood function is guaranteed by our result that proves the concave nature of the function in all variables individually when others are fixed.

C. Organization of the Paper

The organization of the paper is as follows: Section 2 details the probabilistic model and the different parameters of the model used in the paper. In Section 3, we summarize Maximum Likelihood Estimation (MLE) and formulate log-likelihood function for the problem at hand. Section 4 explains the proposed algorithm with a pseudo-code. Results showing the concave nature of the log-likelihood function and the convergence of the proposed algorithm is also given in this section. For comparing the proposed method with the existing marks based method, we define certain comparison metrics and variables. Description of these are included in Section 5. The comparison results of the proposed algorithm with marks based ranking scheme is depicted through figures and tables in section 6. Section 7 contains concluding remarks and proofs of the results are given in Appendix.

2. Definitions of Parameters

This section details the different test parameters that affects the assessment. The only student parameter is capability c_j, and the two item parameters are difficulty d_i and discrimination a_i. Henceforth, j is a student index and i is a question index. ICC is the functional relationship between the probability of correct response to an item and a criterion variable, where the criterion variable is some unobserved hypothetical variable like intelligence, ability, difficulty and so on. This relationship is characterized by the location of the item on the criterion variable scale and its discriminating power. In this paper the probabilistic model that is used for analysis is the logistic ogive model, which is a well studied model [**?**] both theoretically and verified by fitting various sample data. The cumulative form of the logistic function, often called the logistic law, is given by,

$$P_i(c_j) = P(d_i, a_i, c_j) = \Psi(Z) = \frac{e^{a_i(c_j - d_i)}}{1 + e^{a_i(c_j - d_i)}}.$$

Figure **??** shows the variation of the probability of correct response with respect to c_j, where $-\infty \leqslant c_j \leqslant +\infty$ with $d_i = 0.5$ and $a_i = 4.255$, with $P(d_i, a_i, c_j)$ defined as 1 and 0 for $c_j = +\infty$ and $-\infty$ respectively. Here the criterion variable is capability and the location variable is difficulty. As capability increases, the probability to answer correctly also increases and finally saturates to 1 beyond a certain value of c_j. Thus given a question of a fixed difficulty and discrimination, the probability of answering correctly is higher for a more capable student. The difficulty of a

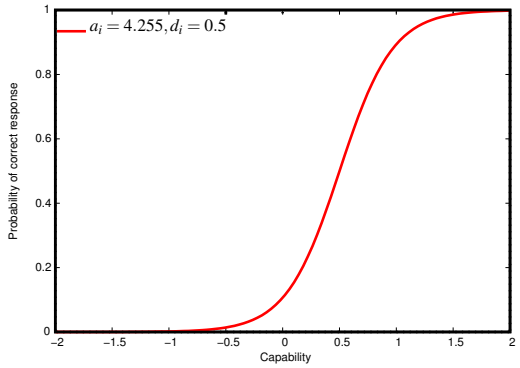

Fig. 1: ICC for correct response with $d_i = 0.5$ and $a_i = 4.255$

question is defined as the capability at which the probability of answering correct is 0.5.

Figure **??** shows the variation of the probability of incorrect response with respect to d_i, where $-\infty \leqslant d_i \leqslant +\infty$ and $c_j = 0.5$ and $a_i = 4.255$, with $1 - P(d_i, a_i, c_j)$ defined as 1 and 0 for $d_i = +\infty$ and $-\infty$ respectively. Here the criterion variable is difficulty and the location variable is capability. From figure it is clear that as difficulty increases, the probability to answer incorrect also increases and finally saturates to 1. Thus, a question that is answered correctly by very few students, is considered as a difficult question and a question that is correctly answered by many candidates is considered as a less difficult question. Note that given a student of known capability c_j, he/she is more likely to correctly answer a question of difficulty $d_i < c_j$, but more likely to make questions of higher difficulties wrong. The capability of a student is defined as the difficulty at which the probability of answering incorrect is 0.5.

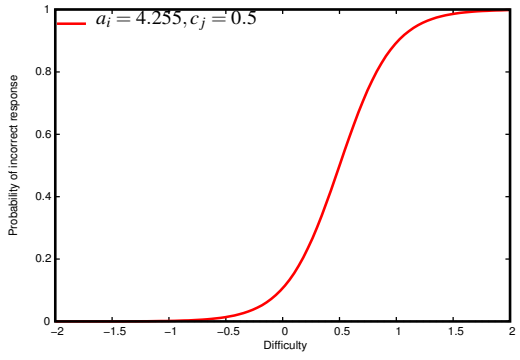

Fig. 2: ICC for incorrect response with $a_i = 4.255$ and $c_j = 0.5$.

$P_i(c_j)$ also depends on another item parameter called discrimination, which gives an index of the steepness/slope of the item characteristic curve at d_i. It is the inverse of the standard deviation of the item characteristic curve, i.e. $a_i = 1/\sigma_i$, where σ_i is the standard deviation of item i. Higher the value of a_i steeper is the ICC near d_i and the spread of the normal distribution is very less. For smaller value of a_i the ICC is not very steep near d_i and normal distribution is flat. Even though the level of a_i varies in the range $-\infty \leqslant a_i \leqslant +\infty$, it is always taken to be non-negative, with $P(d_i, a_i, c_j)$ defined as 1 and 0 for $a_i = +\infty$ and $-\infty$ respectively. Moreover, the value of a_i usually taken in practice is $\leqslant 2.5$. Figure **??** shows the variation of the probability to correct answer for different values of discrimination, with d_i and c_j fixed to 0.5. As the value of a_i increases, the slope of ICC also increases. For large a_i ICC ≈ 0, if $c_j < d_i$ and ICC ≈ 1 if $c_j > d_i$. Thus ICC can tell if capability is above or below difficulty with great precision. In short a question with a large value of a_i is capable of differentiating students of different capabilities better and hence is called a question with high discrimination. The above discussions thus points to the fact that c_j, d_i and a_i are interrelated and their independent estimation is not possible.

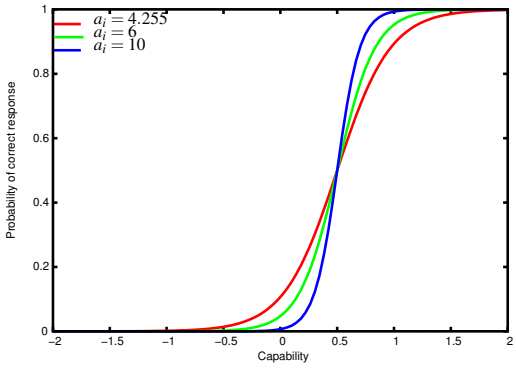

Fig. 3: ICC for correct response with $c_j = 0.5$ and $d_i = 0.5$.

Normal ogive and logistic ogive are the two commonly used models for the item characteristic curve. The cumulative normal distribution function is

$$P_i(c_j) = P(d_i, a_i, c_j) = \Phi(Z) = \int_{-Z_{ij}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}Z^2} dZ,$$

where $Z_{ij} = a_i(c_j - d_i)$ is the normal deviate, a_i is the item discrimination, c_j is the student capability and d_i is the item difficulty. The cumulative form of the logistic function, often called the logistic law, is given by

$$P_i(c_j) = P(d_i, a_i, c_j) = \Psi(Z) = \frac{e^{a_i^*(c_j - d_i)}}{1 + e^{a_i^*(c_j - d_i)}}.$$

By plotting both the normal and logistic ogive models it was found that d_i values are the same for both the models, but their slope at that point i.e. a_i is different.

Haley [**?**] has shown that if the logistic deviate $a_i(c_j - d_i)$ is multiplied by a constant factor 1.702 the absolute difference between $P_i(c_j)$ of the normal ogive and the logistic ogive is less than 0.01 over the full range of c,

$$|(\Phi(Z) - \Psi(Z))| < 0.01.$$

Thus to make both the models in agreement with each other 1.702 is multiplied to the numerical value of a_i in normal ogive metric, to obtain the discrimination parameter of the logistic ogive model, i.e. $a_i^* = 1.702 a_i$. Henceforth the

notation a_i denotes a_i^* as logistic ogive is the model used in this paper.

3. Maximum Likelihood Estimation

Likelihood of a set of parameter values, θ, given outcomes X, is equal to the probability of those observed outcomes given those parameter values, that is

$$\mathscr{L}(\theta|X) := P(X|\theta). \tag{1}$$

Given the data, maximum likelihood estimation is a widely accepted method for estimating parameters of a statistical model [**?**]. This method evaluates the values of the parameters in such a way that the likelihood function is maximized, hence the name maximum likelihood estimation. The objective of this paper is to estimate the capability vector $C = [c_1 \cdots c_{n_S}]$, the difficulty vector $D = [d_1 \cdots d_{n_Q}]$ and the discrimination vector $A = [a_1 \cdots a_{n_Q}]$, where n_S is the number of students and n_Q is the number of questions in the exam.

After conducting an offline exam the input data available is the response matrix R, whose rows are indexed by students and columns are indexed by questions. In a dichotomous test the entries of the marks matrix are binaries, i.e. 1's and 0's [1]. An entry 1 in the marks matrix indicates that the student corresponding to that particular row got a correct response for that particular question induced by that column. Similarly, an entry 0 indicates that the student corresponding to that particular row made a wrong response for that particular question induced by that column. An example of a typical marks matrix is given below.

$$\begin{array}{c} \\ \text{Student Capabilities} \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} \text{Question difficulties} & & & \\ d_1 & d_2 & \dots & d_{n_Q} \end{array} \\ \begin{array}{c} c_1 \\ c_2 \\ \vdots \\ c_{n_S} \end{array} \begin{bmatrix} 1 & 1 & \cdots & 0 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{n_S} \end{bmatrix} \text{Row-wise totals} \\ \begin{bmatrix} q_1 & q_2 & \cdots & q_{n_Q} \end{bmatrix} \\ \text{Column-wise totals} \end{array}$$

Thus the likelihood function is

$$\mathscr{L}(C,D,A) := \text{Prob}(R|\ C,D,A), \tag{2}$$

and the entire problem condenses to estimating the vectors C,D and A such that (**??**) is maximized. Given the complete response vector of each student and every question, the likelihood function can be calculated. For a student, the likelihood function depends on his/her response to all the questions. While for a question, the likelihood function depends on the responses to that particular question by all the students. The likelihood function of the entire test is the product of the likelihood functions of each student, under the assumption that all examinees are independent. For the logistic ogive model, the logistic function is given by

$$P_{ij} := P(d_i,a_i,c_j) := \frac{e^{a_i(c_j-d_i)}}{1+e^{a_i(c_j-d_i)}}, \tag{3}$$

$$Q_{ij} := 1-P(d_i,a_i,c_j). \tag{4}$$

The global likelihood function of the exam is formulated as

$$\text{Prob}(R|\ C,D,A) := \Pi_{j=1}^{n_S}\Pi_{i=1}^{n_Q} P_{ij}^{m_{ij}}(1-P_{ij})^{1-m_{ij}}, \tag{5}$$

where n_S is the number of students, n_Q is the number of questions and m_{ij} is the indicator for student j correctly answering question i. That is, $m_{ij} = 1$ if student j addresses question i correctly and 0 otherwise. Taking natural logarithm of the likelihood function we get

$$\begin{aligned} L(C,D,A) &:= \log\ \text{Prob}(R|\ C,D,A) \\ &= \sum_{j=1}^{n_S}\sum_{i=1}^{n_Q} m_{ij}\log\left(\frac{e^{a_i(c_j-d_i)}}{1+e^{a_i(c_j-d_i)}}\right) \\ &- \sum_{j=1}^{n_S}\sum_{i=1}^{n_Q}(1-m_{ij})\log\left(\frac{1}{1+e^{a_i(c_j-d_i)}}\right) \\ &= \sum_{j=1}^{n_S}\sum_{i=1}^{n_Q}\left[m_{ij}a_i(c_j-d_i) - \log\left(1+e^{a_i(c_j-d_i)}\right)\right], \end{aligned} \tag{6}$$

where c_j is the capability level of the j^{th} student, d_i is the difficulty level of the i^{th} question. Since maxima of the likelihood and the log-likelihood functions occurs at the same point, maximizing (**??**) is equivalent to maximizing (**??**). Thus L in equation (**??**) serves as the objective function to be maximized and it depends on the parameters: difficulty vector D, discrimination vector A and capability vector C. In the next section we detail a maximum likelihood based estimation algorithm for estimating C,D and A by maximizing (**??**).

4. Algorithm and Main Results

In this section, we discuss the proposed algorithm for capability assessment and the affirmative results for the convergence of the proposed method. Given the response matrix of the exam we need to estimate C,D and A in such a way that (**??**) is maximized over all [C, D, A] $\in \mathbb{R}^{n_S} \times \mathbb{R}^{n_Q} \times \mathbb{R}^{n_Q}$. Since (**??**) is a function in three vectors of variables and all these parameters are interrelated, their individual optimization is not possible. Hence we have adopted an alternating optimization technique, also called as nonlinear Gauss-Siedel algorithm [**?**]. In alternating optimization the set of parameters are divided in to blocks and optimization is done over blocks sequentially. That is, a function $F(X)$, $X \in \mathbb{R}^n$ can be represented as $F(X) = F(x_1,x_2,\dots,x_m)$ where $x_i \in \mathbb{R}^{n_i}$ and $\Sigma_{i=1}^m n_i = n$. The algorithm proceeds by updating all the block components of X, one at a time, in every step. For the log-likelihood function given in (**??**), the steps for alternating maximization is of the form: given $\mathscr{C},\mathscr{D},\mathscr{A}$ and a function $F : \mathscr{C}\times\mathscr{D}\times\mathscr{A} \to \mathbb{R}$, maximize F over $\mathscr{C}\times\mathscr{D}\times\mathscr{A}$. That is,

$$\max_{C,D,A\in\mathscr{C}\times\mathscr{D}\times\mathscr{A}} F(C,D,A). \tag{7}$$

[1] Negative marks for incorrect responses is not considered in this paper

Start with an initial guess C_0, D_0, A_0 and iteratively find for $n \geqslant 1$,

$$C_n := \arg \max_{C \in \mathscr{C}} F(C, D_{n-1}, A_{n-1}) \quad (8)$$

$$D_n := \arg \max_{D \in \mathscr{D}} F(C_n, D, A_{n-1}) \quad (9)$$

$$A_n := \arg \max_{A \in \mathscr{A}} F(C_n, D_n, A) \quad (10)$$

Alternating optimization is an intuitive idea, but it need not converge. If it converges, it may converge to one of many local maxima depending on the initial condition. In order to prove the convergence of the alternating optimization algorithm when maximizing $L(C,D,A)$, we prove the following two results, Lemma **??** and Lemma **??**.

A function $f(X,Y) : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, $X \in \mathbb{R}^m$, $Y \in \mathbb{R}^n$, is said to be concave in X, if **dom** f is a convex set, and for all x_1, $x_2 \in X$ and all $y \in Y$, and θ with $0 \leqslant \theta \leqslant 1$,

$$f(\theta x_1 + (1-\theta)x_2, y) \geqslant \theta f(x_1, y) + (1-\theta) f(x_2, y). \quad (11)$$

The following lemma is about the concave nature of the log-likelihood function $L(C,D,A)$ in $C \in \mathbb{R}^{n_S}$, $D \in \mathbb{R}^{n_Q}$ and $A \in \mathbb{R}^{n_Q}$.

Lemma 4.1. *The log-likelihood function $L(C,D,A)$, see* (**??**)*, is strictly concave separately in C, D and A.*

Proof. Given in Appendix. □

Lemma **??** proves the concave nature of the log-likelihood function jointly in $C \in \mathbb{R}^{n_S}$ and $D \in \mathbb{R}^{n_Q}$ for a fixed $A \in \mathbb{R}^{n_Q}$.

Lemma 4.2. *The log-likelihood function $L(C,D,A)$, see* (**??**)*, is jointly concave in C and D for a fixed A.*

Proof. Given in Appendix. □

We propose an iterative algorithm for the accurate estimation of test parameters C, D and A using log-likelihood function. Pseudo-code of the proposed algorithm is presented below. Convergence of this algorithm is guaranteed by the block-concave nature of the log-likelihood function. Alternating maximization over the three blocks of the log-likelihood function is done as follows: For each student we optimize $L(\bar{C},D,A)$, as shown in step **??**. The function $L(\bar{C},D,A)$ is separable for each student and so optimizing for each c_j separately is equally good as optimizing for the full vector C. The optimum point that we get, (C^*,D,A) is the global maxima of the function $L(\bar{C},D,A)$, since the function is concave in C, for fixed D and A. Similarly, for each question we optimize the functions, $L(C,\bar{D},A)$ and $L(C,D,\bar{A})$ as shown in steps **??** and **??** respectively. These functions are also separable for each question separately, and so optimizing for each d_i and a_i separately is equally good as optimizing for the full vector D and A respectively. The optimum point that we get for these two optimization problems are (C,D^*,A) and (C,D,A^*) respectively. They are the global maxima of the respective functions, since the log-likelihood function is concave in D, for fixed C and A and also concave in A, for fixed C and D. Finally by step **??** the algorithm converges to a point (C^*,D^*,A^*).

Algorithm 4.1 Pseudo-code of the ML based algorithm

Input: Raw marks matrix.
Output: Student capability vector C, question difficulty vector D and question discrimination vector A.

1: Initialize $D, A \in \mathbb{R}^{n_Q}$
2: **while** Error norms of estimated levels in previous iteration $\geqslant$ tolerance value **do**
3: **for** each student j **do**
4: Using D and A find $c \in [0,1]$ such that L is maximum.
5: $c_j := \arg \max_c L(\bar{C},D,A)$
6: **end for**
7: **for** each question i **do**
8: Using C and A find $d \in [0,1]$ such that L is maximum.
9: $d_i := \arg \max_d L(C,\bar{D},A)$
10: **end for**
11: **for** Each question i **do**
12: Using D and C find $a \in [0,6]$ such that L is maximum.
13: $a_i := \arg \max_a L(C,D,\bar{A})$
14: **end for**
15: Go to Step **??**
16: **end while**

Theorem 4.3. *The ML based alternating optimization algorithm* **??** *always converges to a stationary point of the log-likelihood function $L(C,D,A)$. Also, the converging point is either the global maxima of $L(C,D,A)$ or a stationary point on a flat surface of the function.*

Proof. It is proved in Lemma **??** that, $L(C,D,A)$, is a strict concave function individually in vectors C, D and A. Thus every step in alternating maximization algorithm gives a better estimate of C, D, A such that $L(C,D,A)$ value in the current step is higher than the previous step. Thus the converging point of the algorithm will not be a local maxima. This is because, the function will always improve at a local maxima, and so the algorithm will give updated values of C,D,A if at all it reaches one. Thus the converging point of the algorithm will be either a stationary point on a flat surface or the global maxima of the log-likelihood function. Thus the convergence of the proposed algorithm to a stationary point is guaranteed. □

5. Comparison Metrics and Variables

In order to compare the performance of the proposed method with the conventional marks based scheme, we conducted few experiments. Analyses of these experiments are made by monitoring the following parameters:

- Number of candidates qualified
- Number of 'false-positives'
- Number of desired candidates qualified

In any hypothesis testing, there are four categories, true-positive, true-negative, false-positive, and false-negative. These four categories get defined based on the hypothesis. In

our case, we hypothesize that if a student is capable, and our test indeed detects him/her to be capable, it is true-positive. If the test detects a truly not capable student as not capable, then it is true-negative. On the other side there can exist cases where a not capable student is identified as capable by the testing method. This is false-positive. Lastly the case where the test identifies a capable student as not capable is false-negative.

These experiments are done with top 10% as the cut-off bound. There are three different ranklists which are compared using the above metrics:

- based on their Actual Capability (AC), AC ranklist,
- based on the Raw Marks (RM), RM ranklist,
- based on the Maximum Likelihood Estimate of Capability (MLC), ML ranklist.

Cut-off MLC (cut-off RM) is the MLC (RM) of the last candidate in the top 10% of the ML ranklist (RM ranklist). Qualified people in ML ranklist (RM ranklist) are those candidates whose MLC (RM) is greater than or equal to the cut-off MLC (cut-off RM).

False-positives in ML ranklist (RM ranklist) are those examinees who actually are not within the cut-off bound in AC ranklist, but make it within the cut-off bound in ML ranklist (RM ranklist). These candidates are considered as the non-deserving section which gets qualified after exam. A good ranking scheme should have a small fraction of false-positives in every top percentile bracket.

Number of desired candidates qualified in ML ranklist (RM ranklist) refers to the number of people who are present in the AC ranklist as well as in the ML ranklist (RM ranklist). More the number of desired candidates better is the ranking scheme, as it is more in accordance with the AC ranklist.

Exams for these analyses are randomly generated and averaged over 50 runs for the same set of students and questions. The proposed algorithm always outperforms the conventional marks based scheme, by giving a significantly lesser number of false-positives in the estimated capability ranklist as compared to the marks based ranklist.

6. Results and Discussions

Tables **??**, **??** and **??** are the comparison results of the conventional marks based scheme and the proposed maximum likelihood based scheme for different experiments conducted. Labels ML and RM in the comparison tables represents proposed Maximum Likelihood (ML) based scheme and conventional Raw Marks (RM) based scheme respectively. Simulation results **??** and **??** presents the band of number of false-positives in ML ranklist for 90% of exams over which each experiment is done. This is done to clarify the fact that better results of the proposed scheme is not because of some favourable cases in the 50 exams over which they were averaged. Experiments conducted for making these analyses are: (i) fixed number of students and number of questions varied, (ii) fixed number of questions and number of students varied and (iii) normalized marks based exam used for multiple sessions where students take exams in batches answering different question papers for the same discipline, to enter in a single ranklist.

Because of multiple students getting same score, number of qualified people in RM scheme is much more than the cut-off percentile. However, the estimated capabilities of students are less likely to be same. It is evident from tables that the number of qualified people beyond the cut-off limit is at most one in the proposed method for the experiments conducted, whereas a lot in the case of RM scheme. Large number of people with same score results in ties in the RM ranklist unlike in the proposed method.

From tables it is clear that, number of false-positives in all the experiments is always less for the proposed ML scheme. However, the number of desired candidates qualified is more for the RM scheme. This can be justified by the fact that the number of qualified people in the RM scheme is more than the cut-off percentage, and this difference is much more than the number of extra desired candidates qualified when comparing with our method. It is also observed that if the proposed algorithm is qualifying the same number of candidates that gets qualified through RM scheme, even then the number of false-positives in the proposed method is less.

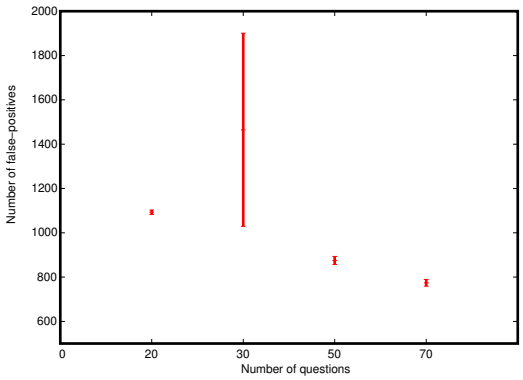

Fig. 4: Demonstration of 90% band of false-positives for $n_S = 20,000$ and different n_Q

Figure **??** shows 90% band of number of false-positives in ML ranklist for $n_S = 20,000$ and n_Q varied as 20, 30, 50 and 70. The plot validates that in case of 20, 50 and 70 the number of false-positives lies within a narrow band for 90% of the exams over which it is averaged. However, for $n_Q = 30$ the band is wide, which means the number of false-positives is spread over a wide band for 90% of the exams. Thus $n_Q = 30$ is considered as the worst case and the rest of the analyses are made with $n_Q = 30$.

Figure **??** shows 90% band of number of false-positives in ML ranklist for $n_Q = 30$ and n_S varied as 2000, 5000, 10,000 and 20,000. The 90% band of number of false-positives is in agreement with the better performance of the proposed algorithm. The number of false-positives for 90% of the exams is within acceptable band.

Finally the multiple session case where students take exam in batches for the same discipline answering different question papers is considered. In this case the marks based ranklist is formed by normalizing all the sessions using

TABLE I: Comparison for $n_S = 20,000$ and $n_Q = 20, 30, 50$ and 70 and cut-off bound = 10%

Metrics and Parameters	$n_S = 20000$							
	$n_Q = 20$		$n_Q = 30$		$n_Q = 50$		$n_Q = 70$	
	ML	RM	ML	RM	ML	RM	ML	RM
Num. of candidates qualified	2001	2205	2001	2699	2001	2555	2001	2197
Num. of false-positives	1093	1288	1465	2068	875	1307	774	965
Num. of deserving candidates qualified	908	916	536	631	1126	1248	1227	1232

TABLE II: Comparison for $n_Q = 30$ and $n_S = 2000, 5000, 10000$ and 20000 and cut-off bound = 10%

Metrics and Parameters	$n_Q = 30$							
	$n_S = 2000$		$n_S = 5000$		$n_S = 10000$		$n_S = 20000$	
	ML	RM	ML	RM	ML	RM	ML	RM
Num. of candidates qualified	201	225	501	691	1001	1203	2001	2699
Num. of false-positives	115	137	272	423	513	692	1465	2068
Num. of deserving candidates qualified	86	88	229	268	489	511	536	631

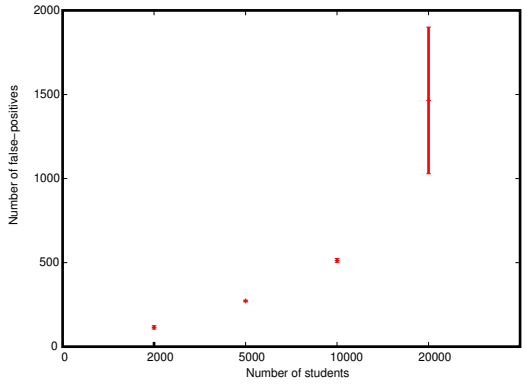

Fig. 5: Demonstration of 90% band of false-positives for $n_Q = 30$ and different n_S

the following relation and then combining them. The normalizing equation used is,

$$\bar{m_{ij}} = \frac{(m_t^g - m_q^g)}{m_{ti} - m_{iq}}(m_{ij} - m_{iq}) + m_q^g, \tag{12}$$

where m_{ij} is the actual marks obtained by the j^{th} candidate in the i^{th} session, m_t^g is the average marks of the toppers in all sessions, m_q^g is the mean of marks of all students in all sessions, m_{ti} is the top marks of the i^{th} session, m_{iq} is the average marks of all the students in the i^{th} session.

The combined normalized marks based ranklist is compared with the combined ML ranklist. Simulation results affirms that our method out-performs the normalized marks based scheme.

7. Conclusion

We propose a maximum likelihood based alternating maximization algorithm for estimating student capabilities and question difficulties of an offline exam. The model employed in this paper is the logistic ogive model, which is a well studied item response model. The individual concavity of the log-likelihood function on student capability vector C, question difficulty vector D and question discrimination vector A is being proved. We also proved that the log-likelihood function is jointly concave in C and D. Using these two results, convergence of the proposed method is also proved.

Experimental tests confirms the improved performance of the proposed scheme over the conventional marks based scheme. Student capabilities and question difficulties were estimated and maximum likelihood estimated capability based ranklist (MLC ranklist) is compared with the raw marks based ranklist (RM ranklist) for the following experiments: (i) fixed number of students with different number of questions (ii) fixed number of questions with different number of students (iii) different question paper distributions and (iv) finally multiple session exam where students take exams in batches answering different question papers for the same discipline. The number of false-positives in the top 10% are compared for both the ranklists with the actual capability based ranklist (AC ranklist) and found that the number of false-positives in the ML based method is less for all the experiments.

References

[1] F. B. Baker. *The Basics of Item Response Theory*. Heinemann, 2 edition, 1985.

[2] F. B. Baker. The item log-likelihood surface for two- and three-parameter item characteristic curve models. *Applied Psychological Measurement*, 12:387–395, 1985.

[3] A. Binet and T. H. Simon. *The Development of Intelligence in Young Children*. Vineland, NJ: The training school, 1916.

[4] G. H. Golub and C. F. V. Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 4 edition, 2013.

[5] D. C. Haley. Estimation of the dosage mortality relationship when the dose is subject to error. *CA: Stanford University, Applied Mathematics and Statistics Laboratory*, Technical Report 15, 7, 1952.

[6] D. N. Lawley. On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, Mathematical and Physical Sciences*, 61:273–287, 1943.

[7] F. M. Lord and M. R. Novick. *Statistical Theories of Mental Test Scores*. MA: Addison-Wesley, 1968.

[8] I. J. Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100, 2003.

TABLE III: Comparison for multiple session exam for $n_S = 20000, n_Q = 30$ done in four sessions with cut-off bound = 10%

Metrics and Parameters	$n_S = 5000, n_Q = 30$							
	session 1		session 2		session 3		session 4	
	ML	RM	ML	RM	ML	RM	ML	RM
Num. of candidates qualified	501	689	501	706	501	562	501	553
Num. of false-positives	280	429	239	389	286	351	274	321
Num. of deserving candidates qualified	221	260	249	311	215	211	227	232

[9] U. Niesen, D. Shah, and G. W. Wornell. Adaptive alternating minimization algorithms. *IEEE Transactions on Information Theory*, 55:1423–1429, 2009.

[10] W. D. Way, L. L. Davis, and S. Fitzpatrick. Practical questions in introducing computerized adaptive testing for k-12 assessments. pages 1–20, 2006.

APPENDIX

Proof of Lemma ?? : The log-likelihood function of the exam, as derived in (**??**), is given by

$$L(C,D,A) = \sum_{j=1}^{n_S}\sum_{i=1}^{n_Q}[m_{ij}a_i(c_j - d_i) - \log(1+e^{a_i(c_j-d_i)})].$$

Taking partial derivatives of logistic function given in (**??**) and (**??**) w.r.t c_j, a_i and d_i we get,

$$\frac{\partial P_{ij}}{\partial c_j} = -\frac{\partial Q_{ij}}{\partial ci} = P_{ij}Q_{ij}a_i, \quad (13)$$

$$\frac{\partial P_{ij}}{\partial d_i} = -\frac{\partial Q_{ij}}{\partial d_j} = -P_{ij}Q_{ij}a_i, \quad (14)$$

$$\frac{\partial P_{ij}}{\partial a_i} = -\frac{\partial Q_{ij}}{\partial a_i} = P_{ij}Q_{ij}(c_j - d_i), \quad (15)$$

Taking partial derivatives of (**??**) with respect to c_j, d_i and a_i, and substituting (**??**), (**??**) and (**??**) we get

$$\frac{\partial L}{\partial c_j} = \sum_{i=1}^{n_Q}(m_{ij} - P_{ij})a_i, \quad (16)$$

$$\frac{\partial^2 L}{\partial c_j^2} = -\sum_{i=1}^{n_Q} a_i^2 P_{ij}Q_{ij}, \quad (17)$$

$$\frac{\partial L}{\partial d_i} = \sum_{j=1}^{n_S} -(m_{ij} - P_{ij})a_i, \quad (18)$$

$$\frac{\partial^2 L}{\partial d_i^2} = -\sum_{j=1}^{n_S} a_i^2 P_{ij}Q_{ij}, \quad (19)$$

$$\frac{\partial L}{\partial a_i} = \sum_{j=1}^{n_S}(m_{ij} - P_{ij})(c_j - d_i), \quad (20)$$

$$\frac{\partial^2 L}{\partial a_i^2} = -\sum_{j=1}^{n_S}(c_j - d_i)^2 P_{ij}Q_{ij}. \quad (21)$$

A function $f : \mathbb{R}^n \to \mathbb{R}$, which is twice differentiable, is jointly concave in two or more variables if its Hessian matrix is negative semi-definite, i.e. $\nabla^2 f(x) \preccurlyeq 0$. For a function $f : \mathbb{R} \to \mathbb{R}$ the condition reduces to $f'' \leqslant 0$. To prove the negative semi-definiteness of H, we show $-H$ is positive semi-definite. A matrix A is called positive semi-definite if it is symmetric and all its eigenvalues are non-negative.

The $(n_S \times n_S)$ Hessian of L w.r.t C, H_{11}, is a diagonal matrix with all negative entries along the diagonal.
$\Rightarrow H_{11} \prec 0$. Thus (**??**) is a strictly concave function in C.

Similarly the $(n_Q \times n_Q)$ Hessian of L w.r.t D, H_{22}, is also a diagonal matrix with all negative entries along the diagonal.
$\Rightarrow H_{22} \prec 0$. Thus (**??**) is a strictly concave function in D.

Also the $(n_Q \times n_Q)$ Hessian of L w.r.t a is a diagonal matrix with all negative entries.
$\Rightarrow H_{33} \prec 0$. Thus (**??**) is a strictly concave function in A. □

We use Gershgorin's Circle theorem [**?**, Theorem 7.2.1] for proving Lemma **??**. We state it here for completeness.

Proposition 1. *If $X^{-1}AX = D + F$ where $D =$ diag$(d_{11}, ..., d_{nn})$ and F has zero diagonal entries, then*

$$\lambda(A) \subseteq \cup_{i=1}^{n} D_i$$

$$\text{where,} \quad D_i = \{z \in \mathbb{C} : |z - d_{ii}| \leq \sum_{j=1}^{n} |f_{ij}|\}.$$

Proof of Lemma ?? : The log-likelihood function is given by,

$$L(C,D,A) = \sum_{j=1}^{n_S}\sum_{i=1}^{n_Q}[m_{ij}a_i(c_j - d_i) - \log(1+e^{a_i(c_j-d_i)})].$$

Taking partial derivative of (**??**) w.r.t d_i,

$$\frac{\partial^2 L}{\partial d_i \partial c_j} = a_i^2 P_{ij}Q_{ij}, \quad (22)$$

Using (**??**), (**??**) and (**??**), the $(n_S + n_Q) \times (n_S + n_Q)$ symmetric Hessian matrix of L can be represented as:

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 L}{\partial C^2} & \frac{\partial^2 L}{\partial C \partial D} \\ \frac{\partial^2 L}{\partial D \partial C} & \frac{\partial^2 L}{\partial D^2} \end{bmatrix} \quad (23)$$

where H_{11} is a $(n_S \times n_S)$ diagonal matrix, H_{12} is a $(n_S \times n_Q)$ matrix and H_{22} is a $(n_Q \times n_Q)$ diagonal matrix.

The Hessian H has all the diagonal entries as negative and the absolute value of the diagonal term in each row is equal to the sum of the absolute values of the off-diagonal terms of each row.
Since $-H$ is diagonally dominant and all diagonal entries are non-negative, by Gershgorin's circle theorem, all eigenvalues of $-H$ are non-negative.
$\Rightarrow H \preccurlyeq 0$. Thus the log-likelihood function is jointly concave in C and D. □