

Faster Confidence Intervals for Item Response Theory via an Approximate Likelihood Profile

Benjamin Paaßen
German Research Center for
Artificial Intelligence
Berlin, Germany
benjamin.paassen@dfki.de

Christina Göpfert
Faculty of Technology
Bielefeld University
Bielefeld, Germany

Niels Pinkwart
Institute of Computer Science
Humboldt-University of Berlin
Berlin, Germany

ABSTRACT

Item response theory models the probability of correct student responses based on two interacting parameters: student ability and item difficulty. Whenever we estimate student ability, students have a legitimate interest in knowing how certain the estimate is. Confidence intervals are a natural measure of uncertainty. Unfortunately, computing confidence intervals can be computationally demanding. In this paper, we show that confidence intervals can be expressed as the solution to a feature relevance optimization problem. We use this insight to develop a novel solver for confidence intervals and thus achieve speedups by 4-50x while retaining near-indistinguishable results to the state-of-the-art approach.

Keywords

item response theory, confidence intervals, relevance intervals, approximation

1. INTRODUCTION

Item response theory (IRT) is a well-established method to model the responses of students on a test [1, 5, 8]. The basic version models the probability of student i to answer correctly on item j as $p_{i,j} = 1/(1 + \exp[b_j - \theta_i])$, where θ_i is a parameter representing the ability of student i , and b_j is a parameter representing the difficulty of item j . If a student's ability exceeds the item's difficulty, the success probability is larger than 0.5, and vice versa. Numerous extensions have been developed over the years, such as parameters for item discrimination (two-parameter IRT), base success chance due to random guessing (three-parameter IRT) [1, 5], IRT for multiple skills [10], performance factor analysis, which describes the increased success chance for repeated attempts [12], or combinations with machine learning [11, 13].

Whenever we use IRT to estimate student ability, students have a legitimate interest in making sure that our model does

B. Paaßen, C. Göpfert, and N. Pinkwart. Faster confidence intervals for item response theory via an approximate likelihood profile. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 555–559, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

© 2022 Copyright is held by the author(s). This work is distributed under the Creative Commons Attribution NonCommercial NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license.
<https://doi.org/10.5281/zenodo.6852950>

not underestimate them (in line with the European Union's concept of a right to explanation [6, 9]). Accordingly, it is important that we not only estimate each student's ability, but also our uncertainty. We can quantify uncertainty in IRT models via confidence intervals [3, 4]. Roughly speaking, an α -confidence interval describes a range of possible parameter values such that the 'true' value is outside the range with probability at most α . For example, Wald's method assumes that parameters are Gaussian-distributed and uses the Gaussian's standard deviation to estimate confidence intervals [3]. While this is efficient, the Gaussian approximation assumes symmetry of the distribution and is only valid for large numbers of items in the test [4]. This is often unrealistic, which is why Wald's method tends to provide inaccurate confidence intervals in practice [4]. Doebler et al. have reacted by providing more exact methods under the assumption of constant item parameters but the resulting confidence intervals are disconnected regions [4], which are challenging to interpret.

More recently, Chalmers et al. [3] suggested the likelihood profile method, which is based on a likelihood ratio test. In particular, we look for a parameter range in which the log-likelihood is at least the optimal log-likelihood minus half the $1 - \alpha$ -quantile of the χ^2 -distribution with one degree of freedom. Because this method does not assume symmetry and works on the 'true' likelihood, it improves the accuracy of confidence intervals considerably. However, the computation requires a nested optimization scheme for every single parameter, which is computationally demanding, especially for large numbers of students.

In this paper, we provide a new perspective by showing that likelihood profile confidence intervals are equivalent to feature relevance intervals, which can be found via an optimization problem [7]. Second, we utilize this theoretical insight to develop a novel solver for confidence intervals which is considerably faster (by factors of 4-50x), while the resulting confidence intervals are almost indistinguishable from direct computation. We evaluate our proposed solver on a range of synthetic experimental conditions from 30-500 students and tests with 10 or 20 items. The experimental code can be found at https://github.com/bpaassen/ability_bounds.

2. METHOD

Our goal is to develop a faster solver for confidence bounds for ability parameters of IRT models via the likelihood profile technique [3]. For an illustration of the technique, con-

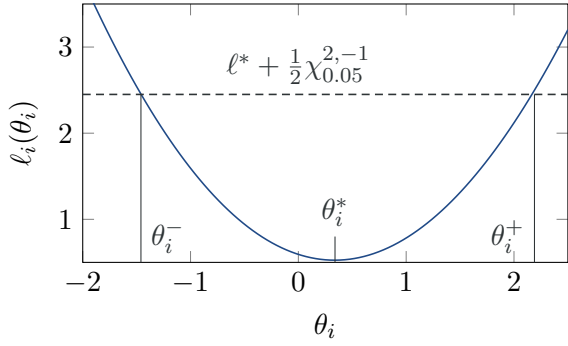


Figure 1: Illustration of the optimal NLL we can achieve for a certain value of θ_i in a single-student, single-item setup with $C = 1$ and $y_{1,1} = 1$. The dashed line shows the loss bound for $\alpha = .05$, the solid lines mark the corresponding confidence interval

sider Fig. 1. The blue curve illustrates the best negative log-likelihood (NLL) ℓ we can achieve when fixing the ability parameter θ_i to the value on the x axis. The likelihood profile technique now asks for the values θ_i^- and θ_i^+ , such that the NLL is exactly $\ell^* + \frac{1}{2}\chi_\alpha^{2,-1}$, where ℓ^* is the overall minimum NLL and $\chi_\alpha^{2,-1}$ is the $1 - \alpha$ quantile of the χ^2 distribution with one degree of freedom. Our α confidence interval is, then, given as $[\theta_i^-, \theta_i^+]$.

More precisely, the NLL of an IRT model on a dataset with m students and n items is given as

$$\begin{aligned} \ell(\vec{\theta}, \vec{b}) &= \sum_{i=1}^m \sum_{j=1}^n -y_{i,j} \cdot \log[p_{i,j}] - (1 - y_{i,j}) \cdot \log[1 - p_{i,j}] \\ &\quad + \frac{1}{2C} \cdot (\|\vec{\theta}\|^2 + \|\vec{b}\|^2), \end{aligned} \quad (1)$$

where $y_{i,j}$ is 1 if student i answered correctly on item j and is 0, otherwise, $p_{i,j} = 1/(1 + \exp[b_j - \theta_i])$, and C is the variance of a Gaussian prior on the ability parameters $\vec{\theta}$ and item parameters \vec{b} [1] (chapter 7).

Now, let $\vec{\theta}_{-i}$ denote the vector $\vec{\theta}$ without its i th element and let $\ell_i(\theta_i) := \min_{\vec{\theta}_{-i}, \vec{b}} \ell(\vec{\theta}, \vec{b})$. In other words, $\ell_i(\theta_i)$ is the minimum NLL value we can achieve if we hold θ_i fixed but optimize all other parameters. Then, the likelihood profile method solves the equation $\ell_i(\theta_i) = \ell^* + \frac{1}{2}\chi_\alpha^{2,-1}$, where $\ell^* = \min_{\vec{\theta}, \vec{b}} \ell(\vec{\theta}, \vec{b})$. Because ℓ is convex in θ_i , this equation has exactly two solutions, which correspond to our interval bounds. The drawback of the likelihood profile method is its computational demand. For each student i , we need to solve a nested optimization problem, where the outer optimization searches for a solution to the equation $\ell_i(\theta_i) = \ell^* + \frac{1}{2}\chi_\alpha^{2,-1}$, and the inner optimization computes $\ell_i(\theta_i)$ for each value of θ_i probed by the outer optimization. We look for a way to speed up this computation.

Our key inspiration is the concept of feature relevance intervals (FRI) proposed by Göpfert et al. [7]. The FRI for some parameter θ is defined as the interval between the minimum and maximum value that still retains a loss of at most $(1 - \delta)$ times the optimal loss, where δ is a user-defined hy-

perparameter. More precisely, given some loss function ℓ of some parameter vector $\vec{\theta}$ and some specific parameter θ_i , the FRI for θ_i is computed by solving the following minimization/maximization problem:

$$\min_{\vec{\theta}} / \max_{\vec{\theta}} \theta_i \quad \text{s.t.} \quad \ell(\vec{\theta}) \leq \ell^* \cdot (1 + \delta). \quad (2)$$

Note that this concept is more general than confidence intervals and is mostly intended for classifiers in machine learning [7]. However, for IRT, confidence intervals via the likelihood profile method and FRIs happen to be equivalent.

THEOREM 1. *Let $\ell^* = \min_{\vec{\theta}, \vec{b}} \ell(\vec{\theta}, \vec{b})$ for the NLL (1) and some $C > 0$. Then, for any $\alpha \in (0, 1)$ and any $i \in \{1, \dots, m\}$, it holds: Problem (3) is convex with a global optimum $\vec{\theta}^\pm, \vec{b}^\pm$, such that $\ell(\vec{\theta}^\pm, \vec{b}^\pm) = \ell_i(\theta_i^\pm) = \ell^* + \frac{1}{2}\chi_\alpha^{2,-1}$.*

$$\min_{\vec{\theta}, \vec{b}} / \max_{\vec{\theta}, \vec{b}} \theta_i \quad \text{s.t.} \quad \ell(\vec{\theta}, \vec{b}) \leq \ell^* + \frac{1}{2}\chi_\alpha^{2,-1}. \quad (3)$$

PROOF. As a notational shorthand, let $L_\alpha := \ell^* + \frac{1}{2}\chi_\alpha^{2,-1}$. Note that the objective function of (3) is linear and hence convex. Now, consider the feasible set $\mathcal{X} = \{(\vec{\theta}, \vec{b}) | \ell(\vec{\theta}, \vec{b}) \leq L_\alpha\}$. Let $x, y \in \mathcal{X}$ and let $z_\gamma = \gamma \cdot x + (1 - \gamma) \cdot y$ for some $\gamma \in [0, 1]$, that is, z_γ is on the connecting line between x and y . Then, it must hold: $\ell(z_\gamma) \leq \gamma \cdot \ell(x) + (1 - \gamma) \cdot \ell(y)$, because ℓ is convex with respect to $\vec{\theta}$ and \vec{b} . Further, because $x, y \in \mathcal{X}$, $\ell(x) \leq L_\alpha$ and $\ell(y) \leq L_\alpha$. Therefore, $\ell(z_\gamma) \leq L_\alpha$, implying that $z_\gamma \in \mathcal{X}$. Hence, \mathcal{X} is convex and (3) is convex.

Next, note that, for any $\alpha \in (0, 1)$, $\chi_\alpha^{2,-1}$ is strictly larger zero. Therefore, the feasible set \mathcal{X} is not empty (it contains at least the minimizer of ℓ). In turn, problem (3) must have a optimum $(\vec{\theta}^\pm, \vec{b}^\pm)$ with $\ell(\vec{\theta}^\pm, \vec{b}^\pm) = L_\alpha$ because the objective function is strictly increasing.

Finally, we need to show that $\ell_i(\theta_i^\pm) = L_\alpha$. Assume this is not the case. If $\ell_i(\theta_i^\pm) > L_\alpha$, then $\min_{\vec{\theta}_{-i}, \vec{b}} \ell(\vec{\theta}, \vec{b}) > \ell(\vec{\theta}^\pm, \vec{b}^\pm)$, which is a contradiction. If $\ell_i(\theta_i^\pm) < L_\alpha$, we need to inspect the behavior of ℓ_i in more detail. Let (θ_i^*, b^*) be the solution to the unconstrained problem $\min_{\vec{\theta}_{-i}, \vec{b}} \ell(\vec{\theta}, \vec{b})$. Then, we know that ℓ_i attains its minimum at θ_i^* with $\ell_i(\theta_i^*) = \ell^*$. Further, because ℓ_i is defined as the component-wise minimum of a convex function, it is convex itself [2] (p. 87). In turn, we know that for any $\theta_i > \theta_i^*$, ℓ_i is increasing and for any $\theta_i < \theta_i^*$, ℓ_i is decreasing. Because $C > 0$, it is also guaranteed that $\ell_i(\theta_i)$ exceeds L_α for sufficiently large/small θ_i , e.g., $\ell_i(\pm\sqrt{2 \cdot C \cdot L_\alpha}) > \frac{1}{2C} \sqrt{2 \cdot C \cdot L_\alpha}^2 = L_\alpha$. Now, consider the minimization version of (3). In that case, we certainly have $\theta_i^- \leq \theta_i^*$, otherwise θ_i^- would not be minimal. Now, because $\ell_i(\theta_i^-) < L_\alpha$, because ℓ_i is decreasing for all values $\theta_i < \theta_i^*$, and because ℓ_i is continuous, there must exist some value $\theta_i < \theta_i^-$ with $\ell_i(\theta_i^-) < \ell_i(\theta_i) \leq L_\alpha$. Therefore, $(\vec{\theta}^-, \vec{b}^-)$ is not a solution for the minimization version of (3), which is a contradiction. The argument for the maximization version is analogous. \square

Fig. 1 provides a graphical intuition for the proof: We can find the two solutions to $\ell_i(\theta_i) = \ell^* + \frac{1}{2}\chi_\alpha^{2,-1}$ by starting at the minimum θ_i^* of ℓ and then decreasing/increasing θ_i

as much as possible while maintaining a loss that is at most $\ell^* + \frac{1}{2}\chi_\alpha^{2,-1}$. To do so, we automatically need to adjust all other parameters to allow for as much slack as possible to increase θ_i .

Alternating optimization solver. The key insight for our solver is that problem (3) becomes much more efficient if we *only* optimize over θ_i and not over any remaining parameters. In particular, we can re-write the NLL (1) as:

$$\begin{aligned} \tilde{\ell}_i(\theta_i) &= \sum_{j=1}^n -y_{i,j} \cdot \log[p_{i,j}] - (1 - y_{i,j}) \cdot \log[1 - p_{i,j}] \\ &\quad + \frac{1}{2C} \cdot \theta_i^2 + \ell_{-i}, \end{aligned} \quad (4)$$

where ℓ_{-i} is a constant term that does not depend on θ_i . Computing $\tilde{\ell}_i$ only requires $\mathcal{O}(n)$ computations, whereas the full NLL (1) requires $\mathcal{O}(m \cdot n)$. Overall, we obtain the simplified problem:

$$\min_{\theta_i} \pm\theta_i \quad \text{s.t.} \quad \tilde{\ell}_i(\theta_i) \leq \ell^* + \frac{1}{2}\chi_\alpha^{2,-1}. \quad (5)$$

By extension of Theorem 1, solving (5) is equivalent to solving the equation $\tilde{\ell}_i(\theta_i) = \ell^* + \frac{1}{2}\chi_\alpha^{2,-1}$, which we can do efficiently via standard nonlinear equation solvers.

Importantly, our solution of (5) will be sub-optimal according to the original problem (3) because $\tilde{\ell}_i$ only approximates ℓ_i . Therefore, we update $\tilde{\ell}_i$ by keeping θ_i fixed and minimizing over all other parameters $\vec{\theta}_{-i}$ and \vec{b} . Then, we can solve (5) again, and so forth, until convergence. This is our proposed alternating optimization (AO) solver.

Figure 2 shows an illustration of the algorithm. To infer the upper bound θ_i^+ , we start at $\theta_{0,i}^+ = \theta_i^*$, i.e. the optimal value according to the NLL. We obtain the next estimate $\theta_{1,i}^+$ by solving the $-\theta_i$ version of (5) with respect to the current surrogate NLL $\tilde{\ell}_{0,i}^+$. Then, we update all other parameters $\vec{\theta}_{-i}$ and \vec{b} by minimizing the NLL (1), yielding a new surrogate $\tilde{\ell}_{1,i}^+$. Solving (5) again yields the next estimate $\theta_{2,i}^+$. In this example, $\theta_{2,i}^+$ would already be indistinguishable from the optimal θ_i^+ according to (3) because $\tilde{\ell}_{1,i}^+$ and ℓ_i strongly overlap. In general, we can prove that $\theta_{t,i}^+$ converges to θ_i^+ .

THEOREM 2. *Let $\theta_{0,i}^\pm = \theta_i^*$, where $(\vec{\theta}^*, \vec{b}^*)$ minimizes the NLL $\ell(\vec{\theta}, \vec{b})$. Let $\tilde{\ell}_{t,i}^\pm$ be the surrogate NLL (4) for parameters $\min_{\vec{\theta}_{-i}, \vec{b}} \ell(\vec{\theta}, \vec{b})$ for fixed $\theta_i = \theta_{t,i}^\pm$, and let $\theta_{t+1,i}^\pm$ be the solutions of the \mp versions of (5) for $\tilde{\ell}_i = \tilde{\ell}_{t,i}^\pm$. Then, for $C > 0$ and $t \rightarrow \infty$, $\theta_{t,i}^\pm$ converge to solutions of the minimization/maximization version of (3).*

PROOF. *As a notational shorthand, let $L_\alpha := \ell^* + \frac{1}{2}\chi_\alpha^{2,-1}$. For simplicity, we only consider $\theta_{t,i}^+$ here; the proof for $\theta_{t,i}^-$ is analogous. First, observe that, for all t , we have $\tilde{\ell}_{t,i}^+(\theta_{t,i}^+) = \ell_i(\theta_{t,i}^+)$ by construction, and we have $\tilde{\ell}_{t,i}^+(\theta_{t+1,i}^+) = L_\alpha$, otherwise $\theta_{t+1,i}^+$ would not be a solution to (5). Further, by definition, we have $\tilde{\ell}_{t,i}^+(\theta_i) \geq \ell_i(\theta_i)$ for all t and all $\theta_i \in \mathbb{R}$. Therefore, we obtain $\tilde{\ell}_{t,i}^+(\theta_{t,i}^+) = \ell_i(\theta_{t,i}^+) \leq \tilde{\ell}_{t-1,i}^+(\theta_{t,i}^+) = L_\alpha$.*

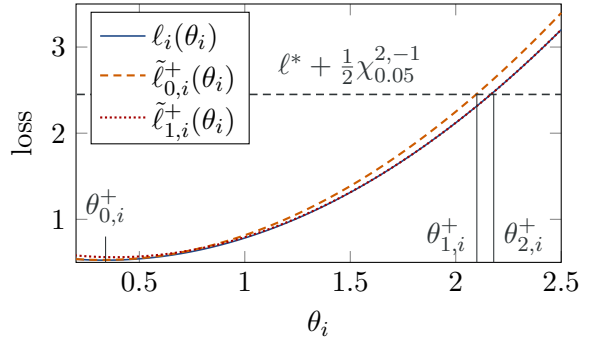


Figure 2: Illustration of the alternating optimization algorithm. We start with the optimum value for $\theta_{0,i}^+ = \theta_i^*$, then maximize θ_i via (5), yielding $\theta_{1,i}^+$, then update $\tilde{\ell}_i$, then maximize θ_i via (5) again, and so forth.

Accordingly, whenever we solve (5) in step $t + 1$, $\theta_{t,i}^+$ is a feasible initial point. Therefore, $\theta_{t+1,i}^+ \geq \theta_{t,i}^+$.

Whenever $\theta_{t+1,i}^+ = \theta_{t,i}^+$, we have $\tilde{\ell}_{t+1,i}^+ = \tilde{\ell}_{t,i}^+$ and, thus, $\theta_{t,i}^+ = \theta_{t+1,i}^+ = \theta_{t+2,i}^+ = \dots$, that is, we have a fixed point. Further, we have $\ell_i(\theta_{t+1,i}^+) = \tilde{\ell}_{t+1,i}^+(\theta_{t+1,i}^+) = \tilde{\ell}_{t+1,i}^+(\theta_{t,i}^+) = L_\alpha$, which is a solution to the maximization version of (3). Before the fixed point, the sequence $\theta_{0,i}^+, \theta_{1,i}^+, \dots$ is strictly increasing. Since ℓ_i is convex with a minimum at $\theta_{0,i}^+$, the sequence $\ell_i(\theta_{0,i}^+), \ell_i(\theta_{1,i}^+), \dots$ is also increasing. For $C > 0$, it would eventually grow beyond bounds due to the regularization term, but since it is upper-bounded by L_α , it must converge to L_α and, hence, $\theta_{t,i}^+$ must converge to θ_i^+ . \square

While Theorem 2 only holds in the limit, very few steps t suffice for a close-to-optimal solution in practice (e.g., Figure 2). We investigate this issue in more detail in our experiments.

3. EXPERIMENTS

In our experiments, we simulate synthetic data from a ground-truth IRT model with ability and difficulties sampled from a standard normal distribution. We varied the number of students m in the range $\{30, 50, 100, 500\}$ and the number of items n in the range $\{10, 20\}$ (similar to the protocol of [3]). For each model, we repeated the sampling 10 times to assess variation. In each repeat, we computed confidence intervals for θ_i at $\alpha = 0.05$ via Wald’s method, the likelihood profile-method via solving $\ell_i(\theta_i) = \ell^* + \frac{1}{2}\chi_\alpha^{2,-1}$ for each i (LP), a log-barrier solver (barrier) for (3), and our proposed alternating optimization scheme (Theorem 2) for $t = 1$ (AO(1)), $t = 2$ (AO(2)), and $t = 3$ (AO(3)) steps. All experimental code is available at https://github.com/bpaassen/ability_bounds.

Coverage. Table 1 shows the coverage rates of all methods, that is, the rate at which the ground truth θ_i value was included in the confidence interval $[\theta_i^-, \theta_i^+]$ [3]. For $\alpha = 0.05$, the coverage rate should be as close as possible to 0.95. We observe that Wald’s method selects too large confidence intervals, yielding rates close to 100%. The barrier method

Table 1: Coverage rates of all methods across experimental conditions

m	n	Wald	LP	barrier	AO(1)	AO(2)	AO(3)
30	10	0.983 ± 0.022	0.943 ± 0.054	0.910 ± 0.056	0.943 ± 0.054	0.943 ± 0.054	0.943 ± 0.054
30	20	1.000 ± 0.000	0.943 ± 0.021	0.083 ± 0.050	0.940 ± 0.025	0.943 ± 0.021	0.943 ± 0.021
50	10	0.998 ± 0.006	0.926 ± 0.041	0.884 ± 0.047	0.920 ± 0.041	0.926 ± 0.041	0.926 ± 0.041
50	20	0.998 ± 0.006	0.942 ± 0.039	0.098 ± 0.039	0.938 ± 0.040	0.942 ± 0.039	0.942 ± 0.039
100	10	0.998 ± 0.004	0.934 ± 0.028	0.892 ± 0.029	0.933 ± 0.027	0.934 ± 0.028	0.934 ± 0.028
100	20	0.999 ± 0.003	0.937 ± 0.030	0.059 ± 0.012	0.933 ± 0.031	0.937 ± 0.030	0.937 ± 0.030
500	10	0.998 ± 0.002	0.940 ± 0.011	0.898 ± 0.015	0.939 ± 0.010	0.940 ± 0.011	0.940 ± 0.011
500	20	1.000 ± 0.001	0.949 ± 0.008	0.097 ± 0.016	0.948 ± 0.008	0.949 ± 0.008	0.949 ± 0.008

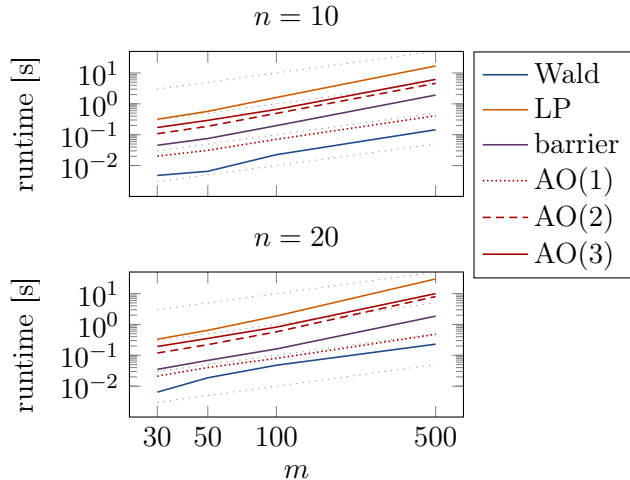


Figure 3: Log-Log plots for the runtime in seconds versus m for $n = 10$ (top) and $n = 20$ (bottom). Dotted, gray lines show linear functions at .1, 1, 10, and 100ms per student.

selects smaller confidence intervals, yielding rates around 90% for $n = 10$ items. If we choose $n = 20$ items, the barrier method becomes numerically unstable and the coverage rates degrade (smaller 10%). The likelihood profile (LP) method consistently achieves rates between 92% and 95% and thus is closest to the desired value of 95%. Alternating optimization achieves indistinguishable results to LP at 3 significant digits for $t \geq 2$ (AO(2) and AO(3)).

Runtime. Figure 3 displays the time it took to compute confidence intervals for all students in every experimental settings in log-log plots. Dotted, gray lines are linear reference functions at .1, 1, 10, and 100ms per student. Using linear fits, we find that Wald’s method is fastest (at about .3ms per student), followed by AO(1) (about .8ms per student), barrier (about 4ms per student), AO(2) (about 9ms per student), AO(3) (about 12ms per student), and finally LP (about 37ms per student). Accordingly, AO(1) is roughly 50x faster than LP, and AO(2) is roughly 4x faster.

4. DISCUSSION AND CONCLUSION

In this paper, we introduced a new solver for confidence intervals on item response theory parameters via the likelihood profile method. In particular, we found that our

alternating optimization solver was 4-50 times faster than the existing solver while achieving almost indistinguishable results. For anyone who seeks to compute confidence intervals, this should provide massive speedups in practice. More generally, though, we hope that our new solver can help the community to respond to an emerging need in educational data mining: more and more, policy makers and the general public expect machine learning models to provide explanations for their decisions. This is exemplified by recent policy initiatives in the European Union for a right to explanation and a risk-based approach to regulate artificial intelligence. Grading student ability—like in item response theory—will likely be under increasing scrutiny in the years to come. As such, we believe that it is crucial to quantify a model’s uncertainty precisely, which our solver can help to do.

We also provide a key theoretical insight in our paper: Traditionally, confidence intervals express the range of parameter values which is likely to contain the ‘true’ value. We showed that the likelihood profile method for confidence intervals can also be interpreted as an optimization problem: We try to find the minimum/maximum ability value for a student which is still consistent with a high likelihood of the data. This interpretation provides a new way to explain an ability estimate: The upper bound of our confidence interval is the highest possible grade we can give to a student while still being consistent with the responses they provided.

Beyond this paper, there remains ample room for future work. Our evaluation has only covered synthetic data to validate the runtime advantage and the closeness to existing methods. Future work could investigate how large confidence intervals tend to be in practical scenarios. Further, our experiments indicated that the size of our confidence is still larger than it would need to be to cover the ‘true’ ability value. Future work could try to find new methods to compute confidence intervals which are more precise. In particular, it may be promising to investigate combinations with recently proposed models for variational item response theory [14].

Acknowledgements

Funding by the German Ministry for Education and Research (BMBF) under grant number 21INVI1403 (project KIPerWeb) is gratefully acknowledged.

5. REFERENCES

[1] F. Baker and S.-H. Kim. *Item Response Theory: Parameter Estimation Techniques*. CRC Press, Boca

Raton, FL, USA, 2 edition, 2004.

- [2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2009.
- [3] R. P. Chalmers, J. Pek, and Y. Liu. Profile-likelihood confidence intervals in item response theory models. *Multivariate Behavioral Research*, 52(5):533–550, 2017.
- [4] A. Doebler, P. Doebler, and H. Holling. Optimal and most exact confidence intervals for person parameters in item response theory models. *Psychometrika*, 78(1):98–115, 2013.
- [5] S. Embretson and S. Reise. *Item response theory for psychologists*. Psychology Press, New York, NY, USA, 2000.
- [6] European Commission. Laying down harmonised rules on artificial intelligence and amending certain union legislative acts, 2021.
- [7] C. Göpfert, L. Pfannschmidt, J. P. Göpfert, and B. Hammer. Interpretation of linear classifiers by means of feature relevance bounds. *Neurocomputing*, 298:69–79, 2018.
- [8] R. Hambleton and H. Swaminathan. *Item response theory: Principles and applications*. Springer Science+Business Media, New York, NY, USA, 1985.
- [9] M. E. Kaminski. The right to explanation, explained. *Berkeley Technology Law Journal*, 34:190–218, 2019.
- [10] R. P. McDonald. A basis for multidimensional item response theory. *Applied Psychological Measurement*, 24(2):99–114, 2000.
- [11] K. Niemeijer, R. Feskens, G. Kreml, J. Koops, and M. J. S. Brinkhuis. Constructing and predicting school advice for academic achievement: A comparison of item response theory and machine learning techniques. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (LAK)*, page 462–471, New York, NY, USA, 2020. Association for Computing Machinery.
- [12] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance factors analysis –a new alternative to knowledge tracing. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, page 531–538, 2009.
- [13] K. Pliakos, S.-H. Joo, J. Y. Park, F. Cornillie, C. Vens, and W. Van den Noortgate. Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers & Education*, 137:91–103, 2019.
- [14] M. Wu, R. L. Davis, B. W. Domingue, C. Piech, and N. Goodman. Variational item response theory: Fast, accurate, and expressive. In A. Rafferty, J. Whitehill, V. Cavalli-Sforza, and C. Romero, editors, *Proceedings of the 13th International Conference on Educational Data Mining (EDM)*, pages 257–268, 2020.