Balancing Test Accuracy and Security in Computerized Adaptive Testing

Wanyong Feng, Aritra Ghosh, Stephen Sireci, and Andrew S. Lan

University of Massachusetts Amherst
Contact Email: `wanyongfeng@umass.edu`

Abstract. Computerized adaptive testing (CAT) is a form of personalized testing that accurately measures students' knowledge levels while reducing test length. Bilevel optimization-based CAT (BOBCAT) is a recent framework that learns a data-driven question selection algorithm to effectively reduce test length and improve test accuracy. However, it suffers from high question exposure and test overlap rates, which potentially affects test security. This paper introduces a constrained version of BOBCAT to address these problems by changing its optimization setup and enabling us to trade off test accuracy for question exposure and test overlap rates. We show that C-BOBCAT is effective through extensive experiments on two real-world adult testing datasets.

1   Introduction

Compared to conventional testing/assessment mechanisms, computerized adaptive testing (CAT) is a form of personalized testing that adaptively selects the next question based on students' answers to previous questions to reduce the test length effectively [6,15,14]. A CAT system is generally comprised of the following components: a knowledge level estimator that estimates a student's current knowledge level given the answers to the previous questions, a response model that predicts how likely a student answers the question correctly given the knowledge level estimate and question features, and a question selection algorithm that selects the next most informative question given the output of response model [2]. Despite the fact that CAT has been widely applied to real-world tests, an important limitation is that most of the existing question selection algorithms are static, which means that they cannot be improved over time as more and more students go through tests. Recently, researchers proposed several conceptual ideas for learning a data-driven question selection algorithm [8].

Bilevel Optimization-Based Computerized Adaptive Testing (BOBCAT) [1] (and extensions [16]) learns a data-driven question selection algorithm while supporting different response models. BOBCAT learns the question selection algorithm by solving the following bilevel optimization problem: the inner level optimization updates the student's knowledge level estimate using questions selected by the question selection algorithm, while the outer level optimization updates the question selection algorithm and response model by evaluating

the latest knowledge level estimate on held-out data, i.e., student responses to questions not seen by the inner optimization. BOBCAT has several advantages over conventional CAT: i) BOBCAT can achieve the same knowledge level estimate accuracy with fewer questions selected, and ii) the question selection algorithm is agnostic of the response model, which enables the test administrators to explore the best combination that meets their needs.

Although BOBCAT has several advantages over conventional CAT, the learned question selection algorithm suffers from high question exposure and test overlap rates. The high question exposure rate means some questions are constantly being selected and the high test overlap rate means different students may receive similar test contents. This limitation could potentially affect the validity and fairness of the test by nefarious test preparation activities that focus on harvesting items administered to previous test takers [5]. Existing methods that address this issue are designed for traditional CAT settings [13], by either introducing randomness in question selection [4] or using maximum clique algorithms [12] and cannot be applied to data-driven CAT settings.

Contributions. In this paper, we propose a constrained version of BOBCAT (C-BOBCAT) that can trade off test accuracy for question exposure and test overlap rates. C-BOBCAT i) uses a stochastic question selection algorithm instead of a deterministic one and ii) adds a penalty term to BOBCAT's optimization objective to promote the learned question selection algorithm to select diverse questions across different students. Therefore, we can trade off test accuracy for question exposure and test overlap rates via a temperature hyper-parameter. We validate the effectiveness of C-BOBCAT through experiments on two real-world adult test datasets and our implementation will be publicly available at: https://github.com/umass-ml4ed/C-BOBCAT.

2   Methodology

We first review BOBCAT and then detail the C-BOBCAT framework.

2.1   BOBCAT Background

The goal of BOBCAT is to solve the following bilevel optimization problem:

$$\underset{\boldsymbol{\gamma},\boldsymbol{\phi}}{\text{minimize}}\ \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\Gamma_i}\ell\Big(Y_{i,j}, g(j;\boldsymbol{\theta}_i^*)\Big) \tag{1}$$

$$\text{s.t.}\ \boldsymbol{\theta}_i^* = \underset{\boldsymbol{\theta}_i}{\text{argmin}}\sum_{t=1}^{T}\ell\Big(Y_{i,j_i^{(t)}}, g(j_i^{(t)};\boldsymbol{\theta}_i)\Big)+\mathcal{R}(\boldsymbol{\gamma},\boldsymbol{\theta}_i) \tag{2}$$

$$\text{where}\ j_i^{(t)} \sim \Pi(Y_{i,j_i^{(1)}},\ldots,Y_{i,j_i^{(t-1)}};\boldsymbol{\phi}), \tag{3}$$

where $\Pi(\cdot)$ represents the question selection algorithm with parameters $\boldsymbol{\phi}$. The inputs are the responses of a particular student i to the previously selected questions $(Y_{i,j_i^{(1)}},\ldots,Y_{i,j_i^{(t-1)}})$, and the output is the next selected question index

$j_i^{(t)}$. $g(\cdot)$ represents the global response model with parameters $\boldsymbol{\gamma}$. $\boldsymbol{\gamma}$ contains both the prior mean of student knowledge level estimate and question difficulties. The inputs are the student's knowledge level estimate $\boldsymbol{\theta}_i$ and question id j, and the output is the student's likelihood of responding to the question correctly. The global response model can either be IRT-based or neural-network-based. Within a batch of N students, BOBCAT splits each student i's response questions into training part Ω_i and held-out meta part Γ_i. For the inner-level optimization problem in (2), the question selection algorithm selects the next informative question from Ω_i. The current student knowledge level estimate $\boldsymbol{\theta}_i^*$, which is the local adaption to the prior mean of student knowledge level estimate, is calculated as a function of $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ by minimizing the summation of binary cross-entropy loss $\ell(.)$ on the selected questions and penalty function $\mathcal{R}(\boldsymbol{\gamma}, \boldsymbol{\theta}_i)$ that prevents large deviation. For the outer level optimization problem in (1), global response model and question selection algorithm are updated by minimizing the binary cross entropy loss $\ell(.)$ on the held-out meta data with the current student knowledge level estimate $\boldsymbol{\theta}_i^*$ as input. For further details, refer to [1].

2.2   C-BOBCAT

We now detail the modifications to BOBCAT. We turn BOBCAT's underlying question selection algorithm from deterministic to stochastic, which injects some randomness into the questions selected for each student. Specifically, we transform the original categorical question selection distribution to the Gumbel-Softmax distribution [3,7] with a fixed temperature hyperparameter. Moreover, since the entropy of a distribution increases as the distribution approaches the uniform distribution, maximizing the entropy of the categorical question selection distribution can further encourage the learned question selection algorithm to select a diverse set of questions for each student. Specifically, we add the negative summation of the entropy of categorical question selection distributions of all selected questions for each student in (5) to the outer level optimization function in (1) to create the new outer level function in (4).

$$\underset{\boldsymbol{\gamma},\boldsymbol{\phi}}{\text{minimize}} \ \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\Gamma_i}\ell\Big(Y_{i,j}, g(j;\boldsymbol{\theta}_i^*)\Big) - \lambda H_i(\boldsymbol{\phi}) \tag{4}$$

$$\text{where } H_i(\boldsymbol{\phi}) = \sum_{t=1}^{T}H\Big(\Pi(Y_{i,j_i^{(1)}}, \ldots, Y_{i,j_i^{(t-1)}}; \boldsymbol{\phi})\Big) \tag{5}$$

During the training process, the model needs to maximize the combination of both prediction accuracy on the held-out meta data and the uncertainty of the question selection algorithm, which is reflected in the entropy regularization term. This balance is controlled by the value of the hyperparameter λ: when $\lambda = 0$, the problem reduces to the original BOBCAT bi-level optimization problem; when $\lambda = \infty$, the solution to the problem is a question selection algorithm that selects each question with equal probability, i.e., the entropy is maximized when the question selection distribution is uniform distribution [10]. In practice, test administrators can explore different values of λ and select a suitable value, to

achieve a desirable balance between maximizing the accuracy of the test and maintaining acceptable question exposure and test overlap rates, according to the requirement of each testing scenario.

3   Experiments

We now detail experiments we conducted on two real-world adult test datasets to validate C-BOBCAT's effectiveness.

3.1   Data, Experimental Setup, Baseline, and Evaluation metrics

We adopt two new publicly available datasets collected from the Massachusetts Adult Proficiency Test (MAPT), the reading comprehension test (MAPT-Read), and the math test (MAPT-Math). In both datasets, there are a total of more than 90K students, 1.7K questions, and 4M question responses [11]. We use these datasets since they are collected in real tests while datasets used in the original BOBCAT paper are collected during longer periods of student learning. As a result, the MAPT datasets are closer to the actual real-world situation addressed by CAT, i.e., a test taker's ability roughly remains unchanged during a short test. We choose the gradient-calculation-based method from BOBCAT to learn the question selection algorithm with C-BOBCAT framework (**C-BINN-Approx** and **C-BIIRT-Approx**). We split the datasets over students into train, validation, and test sets with the ratio of $6 : 2 : 2$ respectively. For each student i, we partition the questions they answer (and their responses) into the training part (Ω_i, 80%) and the held-out meta part (Γ_i, 20%). We choose the IRT-based active learning method (**IRT-Active**), and the IRT-based random question selection method (**IRT-Random**) as the baseline models. We use the area under the receiver operating characteristics curve (**AUC**) as metrics to evaluate the predictive performance on the held-out meta data Γ_i. We use the Scaled Chi-square Statistics of the question exposure rate (**EXPOSE-CHI**) and the average value of test overlap rate between every two students (**OVERLAP-MU**) to measure question exposure and test overlap rates [9].

3.2   Results and Discussion

In Figure 1, we use a series of plots to demonstrate how C-BOBCAT trades off test accuracy (AUC on held-out meta data) versus test security (EXPOSE-PHI and OVERLAP-MU). For both C-BINN-Approx and C-BIIRT-Approx, each point on the curve corresponds to a specific value of the hyperparameter λ. The baselines correspond to single points since they do not support this tradeoff. We observe that AUC and EXPOSE-PHI are positively correlated in the first column of the graph. Both metrics reach the minimum as λ approaches infinity, i.e., when the categorical question selection distribution's entropy is maximized. In this scenario, the question selection algorithm selects each question with equal probability, which decreases both the question exposure rate and test
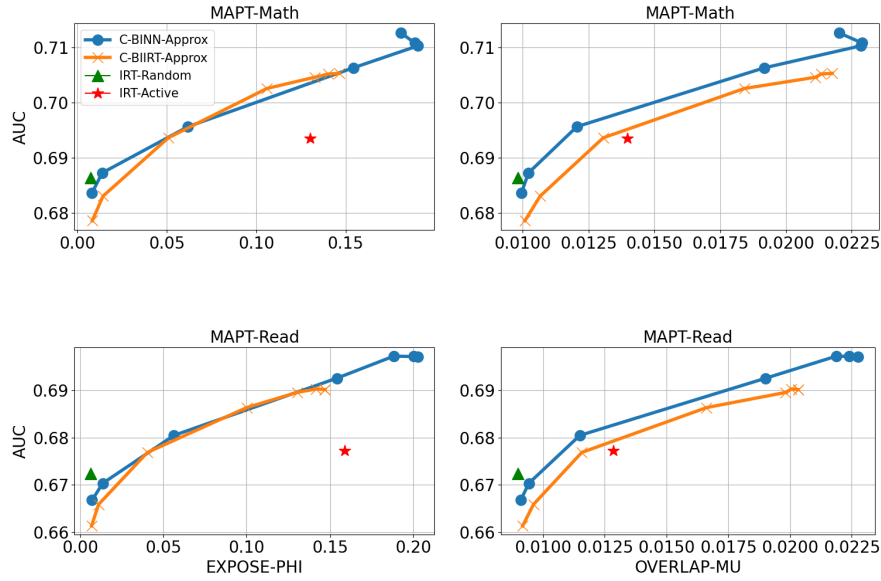
Fig. 1: C-BOBCAT can effectively trade off test accuracy (y-axis, AUC on held-out meta data) for test security (x-axes, EXPOSE-PHI and OVERLAP-MU) on both datasets.

accuracy. Both metrics reach the maximum as λ approaches zero, i.e., when the categorical question selection distribution's entropy is minimized. In this scenario, the question selection algorithm greedily selects the question that leads to the highest test accuracy, which also results in a high question exposure rate. A similar relationship can also be found between OVERLAP-MU and AUC in the second column of the graph. More importantly, with the same EXPOSE-PHI and OVERLAP-MU values, both C-BINN-Approx and C-BIIRT-Approx achieve higher AUC values than the IRT-Active baseline. This observation implies C-BOBCAT is agnostic of the underlying response model and always achieves a better balance between test accuracy and question exposure and test overlap rates than the IRT-Active baseline.

4   Conclusions and Future Work

In this paper, we proposed C-BOBCAT, a framework that attempts to strike a balance between test accuracy and security in CAT settings, which we demonstrated to be effective via experiments on two real-world datasets. Avenues of future work include i) investigating the effect of combining questions' info with the responses from students as the new inputs to the question selection algorithm, ii) for multiple choice questions, investigating the effect of predicting the exact option that students select instead of correctness in both inner and outer level optimization, and iii) deploying C-BOBCAT in real-world tests.

5   Acknowledgements

References

1. Ghosh, A., Lan, A.: Bobcat: Bilevel optimization-based computerized adaptive testing. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI). p. 2410–2417 (2021)
2. Han, K.C.T.: Components of the item selection algorithm in computerized adaptive testing. Journal of Educational Evaluation for Health Professions **15** (2018)
3. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. ICLR (2017)
4. Kingsbury, C.G., Zara, A.R.: A comparison of procedures for content-sensitive item selection in computerized adaptive tests. Applied measurement in education **4**(3), 241–261 (1991)
5. Lee, H., Dodd, B.G.: Comparison of exposure controls, item pool characteristics, and population distributions for cat using the partial credit model. Educational and Psychological Measurement **72**(1), 159–175 (2012)
6. Linden, W.J., van der Linden, W.J., Glas, C.A.: Computerized adaptive testing: Theory and practice. Springer (2000)
7. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. ICLR (2017)
8. Nurakhmetov, D.: Reinforcement learning applied to adaptive classification testing. In: Theoretical and Practical Advances in Computer-based Educational Measurement, pp. 325–336. Springer, Cham (2019)
9. Ozturk, N.B., Dogan, N.: Investigating item exposure control methods in computerized adaptive testing (2015)
10. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)
11. Sireci, S.G., Baldwin, P., Martone, A., Zenisky, A.L., Kaira, L., Lam, W., Shea, C.L., Han, K.T., Deng, N., Delton, J., et al.: Massachusetts adult proficiency tests technical manual, version 2. Center for Educational Assessment Research Report No **677** (2008)
12. Ueno, M., Miyazawa, Y.: Uniform adaptive testing using maximum clique algorithm. In: Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20. pp. 482–493. Springer (2019)
13. Ueno, M., Miyazawa, Y.: Two-stage uniform adaptive testing to balance measurement accuracy and item exposure. In: Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part I. pp. 626–632. Springer (2022)
14. Vats, D., Lan, A.S., Studer, C., Baraniuk, R.G.: Optimal ranking of test items using the rasch model. In: 2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton). pp. 467–473. IEEE (2016)
15. Vats, D., Studer, C., Lan, A.S., Carin, L., Baraniuk, R.: Test-size reduction for concept estimation. In: Educational Data Mining 2013 (2013)
16. Zhuang, Y., Liu, Q., Huang, Z., Li, Z., Shen, S., Ma, H.: Fully adaptive framework: Neural computerized adaptive testing for online education (2022)