A Bounded Ability Estimation for Computerized Adaptive Testing

NeurIPS 2023

Yan Zhuang, Qi Liu, GuanHao Zhao, Zhenya Huang, Weizhe Huang
Zachary A. Pardos, Enhong Chen, Jinze Wu, Xin Li

1: Anhui Province Key Laboratory of Big Data Analysis and Application,
University of Science and Technology of China
2: State Key Laboratory of Cognitive Intelligence
3: University of California, Berkeley
4: iFLYTEK Co., Ltd

Computerized Adaptive Testing (CAT):

How to **efficiently** measure student's ability?

Traditional Paper-and-pencil Test

Computerized Adaptive Testing

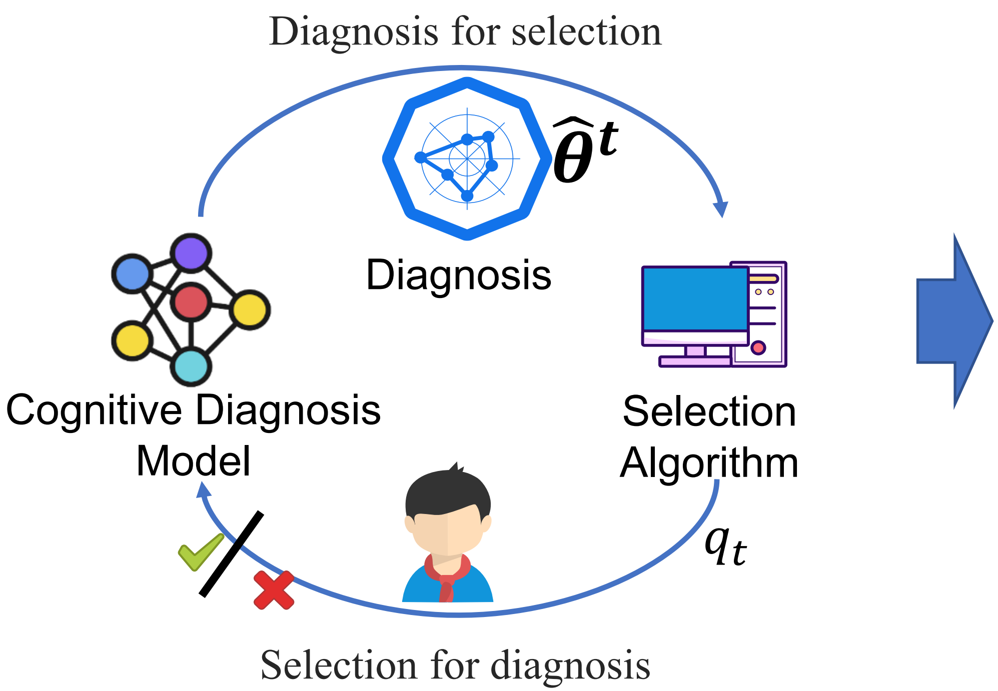
VS

Too many questions:
low-efficiency and heavy burden
for students

Providing students **the least number of questions** to accurately measure their ability (e.g. GRE)

Computerized Adaptive Testing (CAT):

- CAT's Goal: **Accurately assess student true ability $\boldsymbol{\theta_0}$ with as few questions as possible** **(accuracy and efficiency)**
  - (1) *Cognitive Diagnosis Model:* estimates the student's current ability θ^t based on previous t responses.
  - (2) *The Selection Algorithm:* find the next valuable or best-fitting question from the bank, guided by his/her θ^t.



$$\min_{|S|=T} \|\theta^T - \theta_0\|.$$

Our Work

- Problem:

$$\min_{|S|=T} \|\theta^T - \theta_0\|.$$

The exact **true ability $\boldsymbol{\theta_0}$ of student is unknown**, thus it is impossible to find such ground truth in datasets to directly optimize/design the selection algorithm. (Existing methods are implicit)

- Solution:
  - 1. Find $\theta_0's$ theoretical approximation θ^* as the new target ($\theta^* \approx \theta_0$)
  - 2. Design a selection algorithm: select questions set S such that the estimate can best approximates this new target.

$$\min_{|S|=T} \|\theta^T - \theta^*\| \Rightarrow \min_{|S|=T} \max_{\theta \in \Theta} \|\sum_{j \in S} \gamma_j \nabla l_j(\theta) - \sum_{i \in Q} \nabla l_i(\theta)\|.$$

Our Work

- Problem transformation and optimization:

$$\min_{|S|=T} \|\theta^T - \theta^*\| \Rightarrow \min_{|S|=T} \max_{\theta \in \Theta} \|\sum_{j\in S} \gamma_j \nabla l_j(\theta) - \sum_{i \in Q} \nabla l_i(\theta)\|.$$

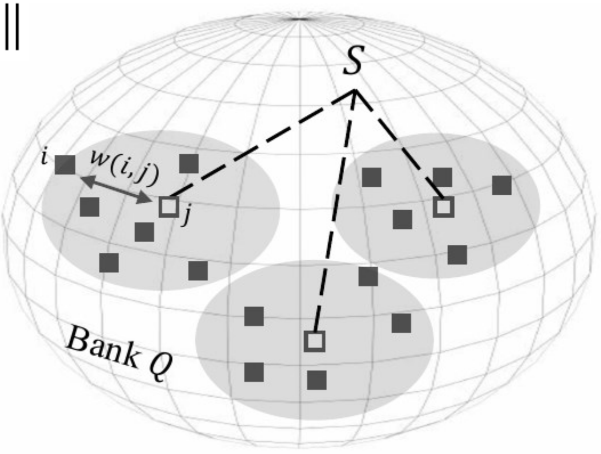
$$\Rightarrow \min_{|S|=T} \max_{\theta \in \Theta} \sum_{i \in Q} \min_{j\in S} \|\nabla l_i(\theta) - \nabla l_j(\theta)\|$$

$$\Rightarrow \max_{|S|=T} \sum_{i \in Q} \max_{j\in S} w(i,j),$$

(submodular facility location function)

where $w(i, j) = d - \max_{\theta \in \Theta} \|\nabla l_i(\theta) - \nabla l_j(\theta)\|$

This is equivalent to selecting the most **representative** items to form the subset S



Our Work

- The above algorithm is **impractical**: The responses to the bank Q are unavailable (gradient difference $\|\nabla l_i(\theta) - \nabla l_j(\theta)\|$ in $w(i,j)$ can not be calculated without labels)
- Solution: we propose an **expected gradient difference approximation method** to replace the original to calculate their similarity $w(i,j)$:

$$w(i,j) \triangleq d - \max_{\theta \in \Theta} \|\nabla l_i(\theta) - \nabla l_j(\theta)\|$$

$$\widetilde{w}(i,j) \triangleq d - \max_{\theta \in \Theta} \mathbb{E}_{y \sim p_{\theta^t}} \left[\|\nabla l_i(\theta) - \nabla l_j(\theta)\|\right],$$

Soft Pseudo-Labels

Our Work

- Theoretical guarantees for ability estimation: **Upper-bound of expected estimation error** (Theorem 1):

$$\mathbb{E}\left[\|\theta^{t+1}-\theta^*\|^2\right] \leq \frac{2\epsilon D\alpha+\sigma_l^2+2\sigma_f D\alpha H_p(\theta^t,\theta^*)}{\alpha^2}$$

- Experiments:

CDM	IRT			NeuralCDM		
Metric@Step	ACC/AUC@5	ACC/AUC@10	ACC/AUC@20	ACC/AUC@5	ACC/AUC@10	ACC/AUC@20
Random	66.45/69.05	68.23/71.66	70.23/74.82	67.19/69.32	68.44/71.56	70.57/74.99
FSI	67.70/70.60	69.62/73.62	71.03/76.24	–	–	–
KLI	67.09/69.79	69.27/73.30	70.42/75.73	–	–	–
MAAT	66.70/70.32	69.13/72.41	69.07/74.46	67.86/70.12	70.07/72.58	70.66/75.83
BOBCAT	**69.51/74.42**	70.94/75.73	71.73/76.58	71.13/76.00	72.52/77.87	73.47/79.00
NCAT	67.30/72.11	70.68/75.80	71.91/76.66	70.47/74.10	72.81/77.99	73.47/79.12
BECAT	66.98/73.15	**71.61/75.87**	**72.00/76.82**	**71.33/76.30**	**73.09/78.34**	**73.58/79.36**

It surpasses the data-driven methods, which requires training on large-scale data.

Thanks