Optimizing Human Learning using Reinforcement Learning

Samuel Girard
Inria Saclay, SODA team
samuel.girard@inria.fr

Jill-Jênn Vie
Inria Saclay, SODA team
jill-jenn.vie@inria.fr

Françoise Tort
Pix
francoise.tort@pix.fr

Amel Bouzeghoub
Télécom SudParis
amel.bouzeghoub@telecom-sudparis.eu

ABSTRACT

Education is a field greatly impacted by the digital revolution. Online courses and MOOCs give access to education to most parts of the world, and many assessments are made online as they are easier to evaluate. This creates an important collection of learning analytics that can be used to provide and generate personalized content, which is essential to keep learners engaged and to have increased learning gains. The purpose of this thesis is to see how machine learning algorithms can be used to learn better knowledge representations of learners, and consequently to recommend learning tasks (exercises or courses) tailored to a student's needs. We are learning instructional policies from student data so that we can understand how students learn and which lessons/exercises in a course have a strong impact on learning for which students.

Keywords

Reinforcement Learning, Intelligent Tutoring Systems, Partially Observable Markov Decision Processes

1. INTRODUCTION

Reinforcement learning (RL) is a convenient setting for personalizing instruction: it consists in learning new instructional policies to assist teacher decision making. However, RL usually requires many samples, i.e., it is not sample-efficient, and collecting extra data can be expensive or time-consuming. In any case, we cannot ask too many questions to a given student, nor collect invasive data. While learning analytics may be various and abundant for a given student and can increase the efficiency of student models, it is still a challenge to combine various modalities of data or to assess external factors like engagement, motivation, etc.

Another significant problem in student modeling is to properly define and assess the validity of the knowledge state of a student. How can we be sure that a notion is learned or that a student has understood a course? We need to be robust to changes in data distribution or to people gaming the system. In this thesis, we intend to build a general RL framework to find new teaching policies from learning analytics. Notably, we need to estimate a broader context of how the student has interacted with an intelligent tutoring system (ITS) to understand with greater precision their learning process. This is essential for proposing personalized content and possibly generate content tailored to their specific needs. This is in continuity with the generative AI revolution where models are fine-tuned with user feedback [8].

In particular, should the RL approach be model-free or follow heuristics from cognitive student models (model-based)? Both methods come with advantages and drawbacks. Like in any machine learning application, it is also challenging to frame a reward objective while ensuring it does not discriminate one part of the population from another.

This paper is organized as follows. In Section 2, we present our research questions. Then, we present the related work in Section 3, we present our preliminary work on offline RL in Section 4, and we give a conclusion and expected contributions in Section 5.

2. RESEARCH QUESTIONS

To effectively use machine learning in education, it's important to rely on the research literature of psychometrics, cognitive science, and statistics and machine learning. A successful model should mimic the way we learn. Therefore, the research questions in our thesis cover these various fields.

2.1 RQ1: How to enrich student knowledge states using cognitive models?

To determine the knowledge state of students, several methods are focused mainly on the assessment of knowledge components. Typically, these student models are trained on existing data to predict future student performance, a process known as *knowledge tracing*. But when we rely only on knowledge components and attempted exercises, we lack other aspects of cognitive processes: it is important to assess the concentration or motivation of a student through temporal data such as response time, number of items attempted, etc. A work so far inspired by cognitive models has been focused mainly on memorization of items with spaced repetition techniques [14], but this can and should be encompassed in a more general framework for knowledge tracing. Using cognitive models to improve knowledge tracing models has also the advantage to boost their interpretability, which

S. Girard, J.-J. Vie, F. Tort, and A. Bouzeghoub. Optimizing human learning using reinforcement learning. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 974–977, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12730017

is essential for practitioners [20].

We should differentiate between students who attempt an exercise randomly and those who give it some thought while their outcomes appear the same. Cognitive models will be studied to track the level of engagement perceived through the data. The willingness to learn may be intrinsically linked to our dopamine system: in neuroscience, the state of full engagement with no self-referential thinking is called the *flow state* [15]. Several studies have proved that areas related to the brain's dopaminergic reward system are more active during flow. It has been established that, in order to experience flow, a key dimension is the match between a person's skills and the challenge from the task. Presenting students with exercises that are too straightforward can lead to a lack of commitment, i.e., the temporal difference between the expected reward signal and the real reward is small – similarly for too difficult exercises. This is related to the so-called *zone of proximal development* [16].

2.2 RQ2: What reward functions do we want to optimize?

An important goal is also to define the reward. Do we expect students to score better at a future test [6], to be more engaged in the platform, or to be knowledgeable on a wider range of skills [19]? Ideally, a little of everything, but this objective is solely reached by the reward signal, defined arbitrarily. Different articles have shown different rewards adapted to the end goal [5].

The challenge lies in finding a reward signal that would provide maximum information while requiring minimal estimation. In an educational setting, observations are usually simplified to binary outcomes (exercise was correct or incorrect) as it is hard to model nuance. More data, such as response time, should be taken into account, but this multimodal data leads to more complex models.

Rewards can be short-term, i.e., correct answer, or long-term, e.g., the score at an exam [2], but this highlights another open problem in the RL literature: the *temporal credit assignment problem* [13, 7]. What would be the most impactful action from a sequence on the long-term reward? Should we optimize several rewards at the same time, making it a multi-objective RL problem?

2.3 RQ3: How to assess the validity of RL in educational settings?

How to make sure that the findings in simulated experiments are actually validated in the real world with real students? Usually, we do cross-validation. One goal of our research project is to elaborate a robust reinforcement learning framework. In order to find robust policies, i.e., that resist changes in data distribution or variance in the data, it is important that our contexts (the information retained by a student) are defined in a compact manner while retaining maximum information. Having interpretable contexts is a plus.

Sometimes, it is expensive to conduct user studies. Therefore, it is good to conduct offline experiments to make sure that a learned policy works no matter the student model [4], this is called *offline reinforcement learning*. This implies we can first evaluate teaching policies on logged data without performing more online experiments on new students.

Another line of work would be to validate inferred knowledge states: How can we be confident in our assessment? We could trick students with adversarial items. For example, presenting an MCQ with distractors, with all wrong but plausible answers, and then assessing the level of confidence of the students. This is sometimes done in a two-step process, inviting students to change their answer after they see that a classmate answered differently [1].

2.4 RQ4: How to properly define the structures of items to generate?

An important aspect to consider in reinforcement learning is the action space; an exercise has a particular structure we should take advantage of. By defining the structure of an item through expert knowledge or unsupervised learning, we could infer much more than its knowledge components or difficulty. Each question has a purpose; sometimes, the difficulty lies in the understanding more than the answering part, and most exercises follow a path of reasoning that might not always be adapted to students. By structuring properly an item into different components, we could refine the action space into a continuous multi-dimensional space: such a space would not only contain the corpus of existing exercises but also fully generated exercises, as variants of existing ones. A method to consider is to use LLMs to retrieve the structure of exercises in the corpus into an embedding and possibly generate new questions that would be more adapted to a given student. Some items rely on much more than their textual content, therefore the use of LLMs for item generation is essential but not sufficient: additional constraints are necessary to improve the mathematical validity of item generation such as arithmetic math word problems [17]. Another approach could be seen as an extension of the work proposed by [3] by choosing an action where each parameter is selected by a bandit algorithm.

3. RELATED WORK

Reinforcement Learning in education has been studied extensively using POMDPs [10, 18]: we do not observe the knowledge of a student, only their interactions with items, hence the name "partially observable", and those interactions allow us to update the actual belief state of the student knowledge. POMDPs allow us not to make the assumption that learners' understanding can be directly observed or approximated by a large number of features. A major inconvenience with POMDPs is that their learning is usually intractable unless under certain conditions, as it requires a lot of data, something hard to counter given the scarcity of educational data.

Consequently, the majority of the literature focuses on the idea of assuming a model of student behavior and then determining the best sequence of learning exercises or policies to achieve specific educational objectives according to this model. Various techniques are used, such as HOT-DINA [12], a hybrid of Bayesian Knowledge Tracing (BKT) and Item Response Theory (IRT); and deep knowledge tracing [9], among others.

However, relying solely on model-based techniques has its limitations, as these models are not perfect and can lead to overfitting, making the learned policies less effective in real-world educational settings (cf. our research question RQ3). To address this issue, [4] suggests using off-policy estimation with multiple student models to find simpler, more robust policies that can perform well regardless of the chosen student model. Alternatively, when no student model is explicitly assumed (model-free), the policy learned is learned solely through the reward function.

A lot of work has been done on automatic item generation (AIG), we are able to generate exercises of better quality and even explanations if the student makes a mistake [11] but we are still lacking on adapting generated items to other aspects than proficiency.

4. OUR PRELIMINARY WORK

So far, we have focused on research questions RQ2 and RQ3 in offline RL. The purpose of our methodology is to have an increase in the student's learning gains, i.e. we want them to solve harder problems correctly, hence we defined a short-term reward signal that is 0 if the student answers incorrectly, and that is positively correlated to the difficulty of the exercise if solved correctly. This is a short-term reward, and this setting is also related to contextual bandits: a machine learning model that selects actions based on context (input information) to maximize a reward, often used in recommendation systems and online advertising. Given a context x, that is some summarized student history, we seek to select an action a with a probability $\pi(a \mid x)$ that leads to a reward r. Our primary objective is to maximize the average reward, which can be calculated using this formula:

$$V(\pi) = \iiint r\, p(r \mid x, a)\pi(a \mid x)p(x)dx\, da\, dr.$$

The objective of off-policy evaluation is to assess a policy π_e using logged data collected from a distinct policy π_0. In this scenario, we are constrained to use existing logged data $\mathcal{D}_0$, represented as samples (θ_i, a_i, r_i) for $1 \leq i \leq n$, and conducting new experiments is impractical or costly. To perform offline evaluations, we first estimate a behavior policy $\pi_0(a \mid \theta)$ from the existing logged data $\mathcal{D}_0$. Then, we assess the potential outcomes r_i that would have occurred if a different question-asking policy, $\pi_e(a \mid \theta)$, had been employed.

Once we have selected an estimator $\widehat{V}$ for the average reward of new policies, we can even use it as an objective function for the purpose of learning new policies (off-policy learning), all while relying only on the existing logged data $\mathcal{D}_0$:

$$\widehat{\pi}^{\widehat{V}} = \underset{\pi}{\operatorname{argmax}}\, \widehat{V}(\pi).$$

There are several average reward estimators; some of them are model-based, and others are model-free (such as inverse probability weighting), leading to different variance-bias trade-offs. In model-based methods, the idea is to learn a model $\hat{r}$ that estimates missing rewards. It has the advantage of having a very low variance but a high bias due to the learned reward model that might be itself biased. On the other side, model-free methods can be unbiased but have subsequent variance and may require more data.

In our first experiment, we assumed the context for a student is a scalar that is characterizing the "proficiency" of a student. In this particular model, we can estimate the difficulties of each action, i.e., exercise, using an item response theory model such as the IRT-1PL model.

In one dimension, the proficiency of a given student is given by a single parameter $\theta \in \mathbb{R}$. In its simplest form, the one-parameter model (1PL), each item a is assigned a parameter d_a representing the difficulty of the item. In IRT-1PL, the probability that a student with a knowledge θ answers item a correctly is given by $\Pr(O = 1 \mid \theta, a) = \sigma(\theta - d_a)$ where d_a is the difficulty of item a and θ is the estimated ability of the student. Item difficulties d_a are estimated from prior data.

When we initially encounter a student, we lack any prior information, so we set the context as $\theta_0 = 0$. At each time step, we update the context of a student with IRT $\theta_{t+1} = \theta_t + K(o_t - p_t)$ where o_t is the binary outcome, i.e. 1 if the answer is correct, 0 otherwise, and p_t is the probability that they got the answer correct according to IRT, that is $p_t = \Pr(O = 1 \mid \theta_t, a_t) = \sigma(\theta_t - a_t)$. We can see that if students get a question correct, then θ increases; if they fail, θ decreases. If the initial estimation of IRT was far away, i.e., $|p_t - o_t|$ is high, then the context will change a lot.

As stated earlier, our objective is to ask harder questions while making sure that students will solve them. In order to do so, we define the reward for a context θ and action a as $R(\theta, a_t, o_t) = d_{a_t} \times o_t$. Therefore, a question answered incorrectly will have a reward of 0 no matter its difficulty (we assume here that difficulty values are shifted so that they are always positive). Our early results have shown that the learned policies had better average reward than the baseline, i.e., existing teaching strategies. As we expected, the model-free estimator has shown great variance due to the limited quantity of data, but it has also led to more interesting and visualizable policies. This work will be available as a preprint at the time of the conference.

5. CONTRIBUTIONS AND IMPACT

So far, we have been focusing on performing offline RL experiments (RQ2 & RQ3), but we also plan to possibly conduct experiments on real students. Studying cognitive models may consistently lead to better student models that can be used in the knowledge-tracing community (RQ1).

One aim of this thesis is to define a relevant RL framework in an educational setting. The choice of the reward (RQ2) or the state-action space structure will lead to the creation of different RL environments that can benefit both the RL community for benchmarking algorithms and the EDM community for understanding thought processes.

Finally, personalized exercise generation tailored to students has been a key goal in the EDM community. One focus of this thesis is to reason at the item template level and manipulate structures of items in order to generate multiple variants of items without relying on too large sources of educational data.

The challenges associated with the application of RL in edu-

cational settings serve to address several open issues within the RL community. These include concerns regarding sample efficiency, managing continuous time dynamics (due to the irregular intervals at which students respond to exercises), and navigating within a continuous action space of items. These issues will also form a focal point of investigation in my thesis, with the aim of offering theoretical contributions to the field.

6. REFERENCES

[1] R. Andriamiseza, F. Silvestre, J.-F. Parmentier, and J. Broisin. Recommendations for orchestration of formative assessment sequences: a data-driven approach. In *Technology-Enhanced Learning for a Free, Safe, and Sustainable World: 16th European Conference on Technology Enhanced Learning, EC-TEL 2021, Bolzano, Italy, September 20-24, 2021, Proceedings 16*, pages 245–259. Springer, 2021.

[2] J. Bassen, B. Balaji, M. Schaarschmidt, C. Thille, J. Painter, D. Zimmaro, A. Games, E. Fast, and J. C. Mitchell. Reinforcement learning for the adaptive scheduling of educational activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[3] B. Clément, D. Roy, P.-Y. Oudeyer, and M. Lopes. Multi-armed bandits for intelligent tutoring systems. *Journal of Educational Data Mining*, 7(2):20–48, 2015.

[4] S. Doroudi, V. Aleven, and E. Brunskill. Robust evaluation matrix: Towards a more principled offline exploration of instructional policies. In *Proceedings of the fourth ACM conference on learning @ scale, I@s 2017, cambridge, MA, USA, april 20-21, 2017*, pages 3–12. ACM, 2017.

[5] S. Doroudi, V. Aleven, and E. Brunskill. Where's the reward? a review of reinforcement learning for instructional sequencing. *International Journal of Artificial Intelligence in Education*, 29(4):568–620, 2019.

[6] A. S. Lan and R. Baraniuk. A contextual bandits framework for personalized learning action selection. In *Educational Data Mining*, pages 424–429, 2016.

[7] T. Mesnard, T. Weber, F. Viola, S. Thakoor, A. Saade, A. Harutyunyan, W. Dabney, T. S. Stepleton, N. Heess, A. Guez, et al. Counterfactual credit assignment in model-free reinforcement learning. In *International Conference on Machine Learning*, pages 7654–7664, 2021.

[8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

[9] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep knowledge tracing. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 505–513. Curran Associates, Inc., 2015.

[10] A. Rafferty, E. Brunskill, T. Griffiths, and P. Shafto. Faster teaching by pomdp planning. In *Advances in Neural Information Processing Systems*, volume 24, pages 280–287, 2011.

[11] S. Sarsa, P. Denny, A. Hellas, and J. Leinonen. Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 1*, ICER '22, page 27–43, New York, NY, USA, 2022. Association for Computing Machinery.

[12] J. Subramanian and J. Mostow. Using deep reinforcement learning to train and evaluate instructional sequencing policies for an intelligent tutoring system. Spotlight at RL4ED - EDM'21 Workshop on Reinforcement Learning for Education, 2021.

[13] R. S. Sutton. Temporal credit assignment in reinforcement learning. *University of Massachusetts Amherst*, 1984.

[14] B. Tabibian, U. Upadhyay, A. De, A. Zarezade, B. Schölkopf, and M. Gomez-Rodriguez. Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences*, 116(10):3988–3993, 2019.

[15] M. Ulrich, J. Keller, K. Hoenig, C. Waller, and G. Grön. Neural correlates of experimentally induced flow experiences. *Neuroimage*, 86:194–202, 2014.

[16] L. Vygotsky. Interaction between learning and development. *Readings on the Development of Children*, 23(3):34–41, 1978.

[17] Z. Wang, A. S. Lan, and R. G. Baraniuk. Math word problem generation with mathematical consistency and problem context constraints, 2021.

[18] J. Whitehill and J. Movellan. Approximately optimal teaching of approximately optimal learners. *IEEE Transactions on Learning Technologies*, 11(2):152–164, 2017.

[19] A. Yessad. Personalizing the sequencing of learning activities by using the q-learning and the bayesian knowledge tracing. In *17th European Conference on Technology Enhanced Learning*, pages 638–644, September 2022.

[20] H. Zhou, R. Bamler, C. M. Wu, and Álvaro Tejero-Cantero. Predictive, scalable and interpretable knowledge tracing on structured domains, 2024.