Towards approximately optimal assessments — A tractable methodology

Androniki Sapountzi[1] **Sandjai Bhulai**[2], **Jaap Storm**[3], **Martijn Meeter**[1]

[1] Faculty of Behavioral and Movement Sciences, Vrije Universiteit Amsterdam, De Boelelaan 1111,
1081 HV Amsterdam, The Netherlands.
[2] Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1111,
1081 HV Amsterdam, The Netherlands.
[3] Department of Mathematics and Computer Science, Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands.

Abstract

An important challenge in education is estimating whether an individual has attained competency based on multiple aspects simultaneously, e.g., response accuracy and response time data. It is not trivial to obtain reliable estimates with few samples and provide decisions in an online and personalized manner. We propose an approach to solve that problem. The approach is grounded in stochastic control theory combined with the statistical power of Reinforcement Learning (RL) and machine learning. It leads to the development of an offline RL method that can efficiently leverage student data sequences of accuracy and response time. We develop a novel method that combines three components: (1) the learner model combines the data in a reward function effectively and efficiently; (2) a reinforcement learning-based prediction technique estimates the parameters of the model, and (3) a machine learning model generalizes the quality of decisions to sequential problems. We provide support that our methodology has a merit enabling robust, computationally efficient and tractable planning of personalized competency assessments.

Introduction

The topic of this paper is the real-time adaptive assessment of a student's level of competence while providing the student with exercises. This setting appears, for instance, in online educational environments, where it has to be determined (e.g., by an algorithm) whether a student has mastered the current topic of study and can proceed to study the more advanced topics. This decision has to be made based on a sequence of the student's responses to exercises that contain: whether the student solved the exercise correctly and what the response time of the student was. The challenge is that the assessment has to be updated after every response (in real-time), without overwhelming the student with lengthy assessments (efficient), and with an adequate level of certainty (reliable). In this paper, we address this challenge by providing a tractable method to assess a student's level that satisfies these three criteria.

The problem described above fits the setting of a sequential learning problem where a measure of the student's level of competence has to be learned via a sequence of paired accuracy and response-time data. There have been many attempts in the literature for dealing with this problem, found in psychometrics and student modeling (Kyllonen and Zu 2016; De Boeck and Jeon 2019; Pelánek 2017; Klinkenberg, Straatemeier, and van der Maas 2011; Käser, Klingler, and Gross 2016). However, these solutions either suffer from incorporating only the accuracy of student's responses, leading to a less informative measure of student's competence (Pelánek 2017), or they assume a homogeneous student population such that individual student assessment is no longer possible (Lee and Brunskill 2012).

The methods to tackle sequential learning problems can be divided into two categories: data-driven and model-based approaches. Data-driven approaches rely on methods from, for instance, statistics or machine learning. Although many of these methods are highly flexible, making them attractive solutions in many cases, their drawback is that they usually require large quantities of data. Since our problem requires an efficient solution, data-driven methods are unsuitable. Moreover, in our problem, it is not evident how one should translate a student's responses into a measure of competency; as the data is not labeled, the competency measure has to be modeled using the available data.

Model-based approaches to sequential learning problems are more challenging than data-driving methods, since they require the user to model their understanding of the problem and its relation to the learning objective. However, once this step is taken, they are much more efficient in using data and are transparent in their inner workings.

In this paper, we present a model-based approach that efficiently uses a sequence of student response data to assess their competency. To be precise, we model the student's competency assessment as a Markov reward process in which the accuracy and response time observations are translated into rewards using the speed-accuracy metric introduced in (Klinkenberg, Straatemeier, and van der Maas 2011). The rewards then reflect how good the student's answers are. In turn, the reward function defines the value function associated with the Markov reward process, which can be interpreted as the student's future performance on exercises, which we take as a measure of competency.

The advantage of modeling the problem as a Markov reward process is that evaluation of the student's competence is reduced to knowing the value associated with the process' state that results from the student's responses. Deriving this so-called value function is a complex task, since the model's state space is continuous and the time horizon is infinite. We provide a detailed description of how to overcome this challenge. Our solution involves modern methods of reinforcement learning and machine learning, customized to our setting. The value function can be used in a Markov decision process where the decision is to stop or to continue the assessment. An important feature of the whole procedure is that it is based on a tractable model rather than black-box methods, giving the results a high degree of interpretability.

To summarize our contributions, we provide a novel tractable methodology to solve the adaptive mastery assessment problem. The basis of our approach is a Markov reward model that overcomes the limited observability of the student's knowledge. Because of the model's complex nature, we provide an in-depth explanation of how to derive the value function of this model using reinforcement and machine learning, supported with preliminary results. The associated value function can be evaluated offline, and be used in sequential decision problems for personalized assessment.

Related work

While there exist recent papers modeling educational problems as a stochastic optimal control problem (Rafferty et al. 2016; Whitehill and Movellan 2017; Bassen et al. 2020; Doroudi, Aleven, and Brunskill 2019; Singla et al. 2021), to our knowledge our work is the first to personalize adaptive assessment stopping policies within an MDP model. Bayesian methods in connection with stopping criteria have been used in the past for adaptive assessments (Lewis and Sheehan 1990; Vos 2000; Weiss and Kingsbury 1984; Choi, Grady, and Dodd 2011; Pelánek 2017; Käser, Klingler, and Gross 2016). These works assume a discrete model of competency, a homogeneous population of students, or they are based on accuracy data only. Statistical models of response times (Maris and Van der Maas 2012; Kyllonen and Zu 2016; De Boeck and Jeon 2019; Klinkenberg, Straatemeier, and van der Maas 2011) face similar limitations, and they do not directly translate to policies, let alone individualized policies (Lee and Brunskill 2012). The main difference in the context of learning is that the goal is to increase knowledge during practice by choosing the optimal content to offer the student. On the side of assessments, the objective is to estimate the static student knowledge reliably at the time of the assessment by selecting the optimal number of questions to show the student (Pelánek 2017; Pardos 2017). One of the major similarities of these two contexts is that they share the same educational challenge: the partial observability of the student's state and the need for explainable results (Singla et al. 2021).

In our prior work (Sapountzi et al. 2021), we developed a novel MDP framework with a reward function that combines the aspect of speed-accuracy trade-off in a similar manner as defined in (Maris and Van der Maas 2012; Klinkenberg, Straatemeier, and van der Maas 2011). We made an initial investigation of the model's properties. We established that it accurately retrieves the competency distribution with simulated student performances. However, we did not address the computational challenges related to evaluating the value function that can lead to using the model in practice inducing optimal policies. Our ideas towards that are presented in this paper.

Model and objective function

In this section, we introduce the Markov reward process $X(\cdot) := \{X(n)\colon n \geq 1\}$ that models the student assessment. We do so by providing the states, transition kernel, and reward function of the process.

The state of $X(n)$ is given by a tuple $(\alpha, \beta, n, \gamma)$, where $\alpha, \beta, n \in \mathbb{Z}_+$ and $\gamma \in [0, \infty)$; note that n is both the time-index of the process and in the state, but by writing $n = k + n_0$, where n_0 is the initial state and k is time, the process remains Markov.

To define the transitions of the process, consider a state $s = (\alpha, \beta, n, \gamma)$. Given s, we draw independent realizations c_n and t_n from, respectively, the random variables $C(s)$ and $T(s)$, where $C(s)$ has a Bernoulli distribution with parameter θ that has a Beta(α, β) distribution and $T(s)$ has an exponential distribution with parameter λ that has a Gamma(n, γ) distribution. We then define $X(n+1)$ to be $(\alpha + c_n, \beta + (1 - c_n), n + 1, \gamma + t_n)$. The idea behind the transitions is that the random parameters of the distribution model the uncertainty about the student's skill. In particular, α and β represent the number of correct and incorrect responses, respectively, and n and γ can be interpreted as having a total response time of γ in n observations. See Section 3 of (Sapountzi et al. 2021) for more intuition behind this way of modeling and the construction of this model from a POMRP.

The reward at state $s = (\alpha, \beta, n, \gamma)$ is defined as $Z_n = c_n(1 - t_n/d)^+$, where $(x)^+ := \max\{x, 0\}$ for $x \in \mathbb{R}$ and $d > 0$. The interpretation of the rewards is that students receive zero contribution to their competency if they answer a question incorrectly or if they exceed some response time threshold d. Otherwise, the reward decreases linearly between 1 and 0 in the length of the response time. This metric combines both speed and accuracy of a student in dealing with the exercises and was introduced in (Klinkenberg, Straatemeier, and van der Maas 2011).

Optimal stopping policies

Our objective is to determine the minimum number of required questions, after which one can decide whether a student has mastered a certain topic. We explain how our model can be used for this purpose. From the sequence of student responses and a starting state $X(1)$, we can sequentially update the state $X(n)$ with the response data. This gives us a sequence of values of states that are estimates for the student's mastery level. Once these values level off, i.e., the differences between consecutive estimates in n become small enough, we know that the student's mastery level has been learned, and the assessment can be stopped. At that point, one can compare if the final estimate is above a preset threshold that indicates mastery; if the student has a value above this threshold, he or she has mastered the topic.

Given that the estimates have converged, comparing the final estimate to a threshold can formally be done by turning $X(\cdot)$ into a Markov decision process (MDP). We consider the case where we want to decide whether a student exceeds a mastery threshold $\xi/(1-\eta)$ for $\xi > 0, \eta \in (0,1)$. To define the MDP, we add a set of actions $a \in A = \{0,1\}$ for every state. When action 1 is chosen, the assessment is continued as was the case before. When and action 0 is chosen, the assessment is stopped. This can be modeled by giving reward ξ upon choosing action 0 and having the process stay in its current state with probability one.

Now recall that in an MDP, a state's value is the discounted sum of future rewards, where we use η as the discount factor. By characterizing optimality as minimizing future rewards, action 0 is chosen in an optimal policy if and only if the value of the process' state exceeds $\xi/(1-\eta)$. Therefore, the assessment is stopped when the student has attained a sufficient level of mastery as indicated by the value of ξ. The value of ξ can be calibrated offline by assessing students that are known to have achieved mastery.

With the MDP, the optimal policy is now characterized by the Bellman equation. Denote $V(\alpha, \beta, n, \gamma)$ for the value of the state $(\alpha, \beta, n, \gamma)$ associated to the optimal stopping policy. We have

$$\begin{aligned} V(\alpha,\beta,n,\gamma) = \min\Bigg\{ \int_0^\infty \Bigg(& \frac{\alpha}{\alpha+\beta}\Big[\Big(1-\frac{t}{d}\Big)^+ \\ & + \eta V(\alpha+1,\beta,n+1,\gamma+t)\Big] \\ & + \frac{\beta}{\alpha+\beta}\eta V(\alpha,\beta+1,n+1,\gamma+t)\Bigg) \\ & \cdot \frac{n}{\gamma+t}\Big(\frac{\gamma}{\gamma+t}\Big)^n \mathrm{d}t \,;\, \frac{\xi}{1-\eta}\Bigg\} \end{aligned} \tag{1}$$

The optimal policy is then given by choosing action 1 when the first term in the minimum of Equation (1) is minimal and choosing action 0 otherwise.

The key challenge now is to evaluate the value function of $X(\cdot)$. Computing exact solutions is an intractable problem (Papadimitriou and Tsitsiklis 1987), calling for approximate solution techniques (Lovejoy 1991). The difficulties are that the state space of $X(\cdot)$ is continuous, and the process has an infinite time horizon. In the next section, we provide a method with which the value function can be evaluated.

Methodology

The continuous nature of the state space stems from the continuous response times. To evaluate the value function, we consider a discrete state-space version of our model by discretizing the response time distribution. Moreover, we truncate the response times to some level $\tau > 0$, for which the probability of exceeding τ is small for states with positive value. Denote the resulting MDP by $\tilde{X}(\cdot)$. We evaluate the value function of this discrete model; our method is explained below. After learning the value function of this discrete model, we use a machine learning model to interpolate between the values of the states to continuous response times. Because the MDP is expected to have a relatively smooth value function, we expect that this interpolation should give an accurate approximation to the value function of $X(\cdot)$. The discretization of the response time distribution should be taken as fine-grained as possible for accuracy, but coarse-grained enough for tractability.

Evaluating the value function of $\tilde{X}(\cdot)$ is done with reinforcement learning. First, however, we have to overcome that the process is defined on an infinite time horizon. We do so by truncating the process to a finite time $N \geq 1$, limiting the number of exercises that can be provided in any assessment. This N cannot have a large value since otherwise, the number of states will increase beyond what can be handled with reinforcement learning. Typically, a value of $N = 21$ (20 questions) is expected to be feasible, while the characteristics of students are known within this time frame; the latter we demonstrate later.

The reinforcement learning agent is trained with data that we generate with simulation. Since the model is defined with explicit distributions, we can simulate the process from many initial states and thus gather data to feed to the reinforcement learning algorithm. This simulation is fast since we are essentially sampling from probability distributions.

To summarize, we simulate $\tilde{X}(\cdot)$ to generate data. This data is used in conjunction with a reinforcement learning algorithm to evaluate the value function of $\tilde{X}(\cdot)$. Finally, we train a machine learning model on the state-value pairs for $\tilde{X}(\cdot)$ to interpolate between discrete response times and approximate the value function of $X(\cdot)$. This value function then characterizes optimal stopping policies with Equation (1).

Preliminary results

In the previous sections, we explained our method for efficiently assessing a student's mastery of a certain topic. We now provide support for these ideas with numerical experiments. For every step of our method, we present some preliminary findings that demonstrate the validity of the ideas behind this setup.

The first experiment demonstrates that for arbitrary prior, the student's skill is eventually learned by our procedure within a time frame of 20 questions, i.e., before $n = 21$. For this experiment, we fix two parameters $\theta \in (0,1)$ and $\lambda > 0$. We generate a sequence of student responses by drawing i.i.d.realizations from an Bernoulli(θ) and exponential(λ) distribution. These parameters thus fix the students 'true' mastery level. After each student response, we update the state $(\alpha, \beta, n, \gamma)$ based on these responses and evaluate the distribution of the rewards for that state. In Figure 1, we have plotted the mean reward as a function of n, for different values of θ and λ and for two different priors p_0. The figures demonstrate that the reward distribution converges in n, demonstrating that our model eventually learns the student's 'true' mastery level, i.e., the one associated with θ and λ.

With the second experiment, we demonstrate that that reinforcement learning is capable of evaluating the value function of $\tilde{X}(\cdot)$. Our setup is as follows. We use Temporal Difference Learning TD(0) as we experimentally found

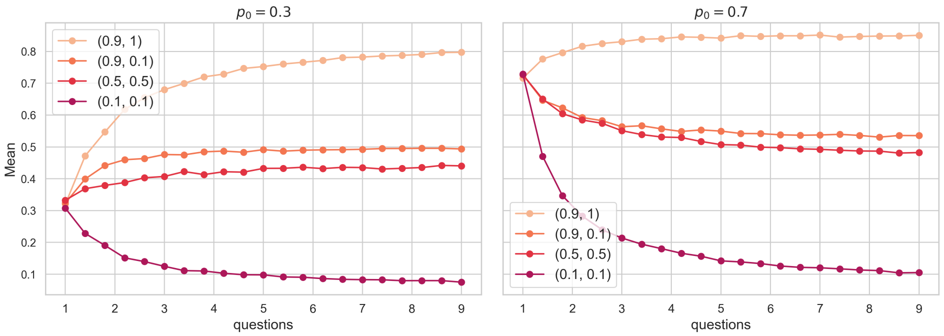

Figure 1: Convergence of the reward distribution for different values of θ and λ. Four student distributions are simulated for two prior competency beliefs $p_0 = 0.3$ and $p_0 = 0.7$.

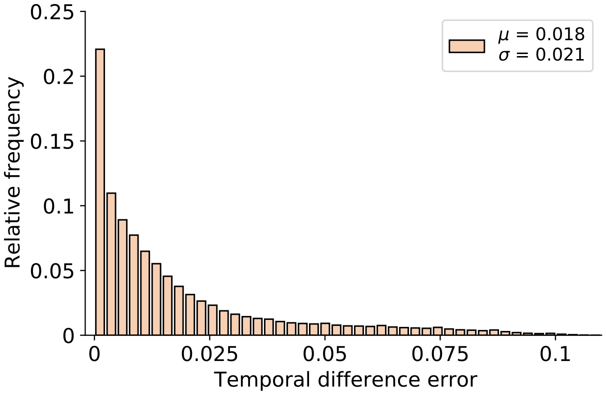

Figure 2: Distribution of the temporal difference errors for states that were explored at least once.

that it outperforms other algorithms (e.g., TD(λ) and Monte Carlo). We restrict our state space to states that allow us to model students with both high and low mastery levels. This is done by taking a set of 567 priors, consisting of states $(\alpha, \beta, 1, \gamma)$ where $\alpha, \beta \in \{1, \dots, 9\}$ and $\gamma \in \{2.5, 5, \dots, 17.5\}$. The state space is the set of states that can be reached from this set of priors in 21 steps. The response time distribution is discretized by mapping the continuous response times to the closest value in $\{2.5, 5, \dots, 100\}$; since the response-time distribution has heavy tails, we require large support. For every prior, we run 10^5 episodes of length 31 questions, where we exceed the bounds of $n = 21$ to avoid estimation issues at the boundary. The learning rate is $\alpha = 0.1$ and the discount factor is $\eta = 0.5$. The number of episodes is relatively small for this size of the state space, but we can easily scale this up later in our research; the computation time for this experiment was around 5 hours.

In our experiment, we save for each state the absolute difference between the previous value estimate and the new one (i.e., the temporal difference error) for the last time it was updated. About 73% of the state space was explored in this short run. In Figure 2, we show the distribution of temporal difference errors for these states after running the experiment. The distribution has the bulk of its mass close to zero, indicating that the temporal difference errors tend to become uniformly small even for this small-scale experiment. Hence, our method is capable of learning the value function.

With our third experiment, we demonstrate that machine learning can predict the values for the continuous state space of $X(\cdot)$ from the discrete state space values of $\tilde{X}(\cdot)$ as estimated in the second experiment. We fitted a simple linear regression model to test this, where we added polynomial features of α, β, n, and γ up to degree 5, i.e., $\alpha^2, \alpha\beta, \alpha^2\beta^2 n$, etc. We split the data set into a training and test set with a 0.75-0.25 ratio and used five-fold cross-validation. We found that fitting a linear model to this data gave a coefficient of determination $R^2 = 0.908$ and a mean squared error of 0.0054 on both the training and test set.

Since these value estimates are far from perfect, we expect they should only improve after running the reinforcement learning algorithm for a longer time. These results show that a simple machine learning method can predict the value function of $\tilde{X}(\cdot)$ from the states, suggesting a high degree of regularity for the value function. Thus, interpolating with the machine learning model to continuous values of γ should give us an accurate approximation of the value function of $X(\cdot)$.

Discussion

This paper presented a tractable model-based methodology for personalizing adaptive assessments in real-time. The policy can efficiently leverage data of students, offer reliable estimates, and provide decisions in real-time. We adopt an MDP framework with Bayesian updates to map students' responses to a knowledge state to overcome this issue. The response effectively combines the information of model accuracy and response time. The evaluation of the value function, the proxy for students' mastery level, is based on established methods from reinforcement learning and machine learning to maintain the computational tractability of the control process. We support each part of our idea with numerical experiments, showing that we need only a few data points to provide accurate estimates.

References

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

Choi, S. W.; Grady, M. W.; and Dodd, B. G. 2011. A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, 71(1): 37–53.

De Boeck, P.; and Jeon, M. 2019. An overview of models for response times and processes in cognitive tests. *Frontiers in psychology*, 10: 102.

Doroudi, S.; Aleven, V.; and Brunskill, E. 2019. Where's the reward? *International Journal of Artificial Intelligence in Education*, 29(4): 568–620.

Käser, T.; Klingler, S.; and Gross, M. 2016. When to stop?: towards universal instructional policies. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 289–298. ACM.

Klinkenberg, S.; Straatemeier, M.; and van der Maas, H. L. 2011. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, 57(2): 1813–1824.

Kyllonen, P. C.; and Zu, J. 2016. Use of response time for measuring cognitive ability. *Journal of Intelligence*, 4(4): 14.

Lee, J. I.; and Brunskill, E. 2012. The Impact on Individualizing Student Models on Necessary Practice Opportunities. *International educational data mining society*.

Lewis, C.; and Sheehan, K. 1990. Using Bayesian decision theory to design a computerized mastery test. *ETS Research Report Series*, 1990(2): i–48.

Lovejoy, W. S. 1991. A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research*, 28(1): 47–65.

Maris, G.; and Van der Maas, H. 2012. Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, 77(4): 615–633.

Papadimitriou, C. H.; and Tsitsiklis, J. N. 1987. The complexity of Markov decision processes. *Mathematics of operations research*, 12(3): 441–450.

Pardos, Z. A. 2017. Big data in education and the models that love them. *Current opinion in behavioral sciences*, 18: 107–113.

Pelánek, R. 2017. Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, 27(3-5): 313–350.

Rafferty, A. N.; Brunskill, E.; Griffiths, T. L.; and Shafto, P. 2016. Faster teaching via pomdp planning. *Cognitive science*, 40(6): 1290–1332.

Sapountzi, A.; Bhulai, S.; Cornelisz, I.; and van Klaveren, C. 2021. Personalized Stopping Rules in Bayesian Adaptive Mastery Assessment. *arXiv preprint arXiv:2103.03766*.

Singla, A.; Rafferty, A. N.; Radanovic, G.; and Heffernan, N. T. 2021. Reinforcement Learning for Education: Opportunities and Challenges. *arXiv preprint arXiv:2107.08828*.

Vos, H. J. 2000. A Bayesian procedure in the context of sequential mastery testing. *Psicológica*, 21(1): 191–211.

Weiss, D. J.; and Kingsbury, G. G. 1984. Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4): 361–375.

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