Reinforcement Learning Formulation

To formulate the AI Teacher task as a reinforcement learning problem, we first assess the *Markov assumption* by assuming that the student's current mastery levels fully summarize their future learning potential, independently of their past performance history. Under this assumption, the system can be modeled as Markovian. Given this assumption, the core components of the RL formulation are defined as follows:

- **States**: A continuous state vector $S \in [0,1]^2$, representing the student's mastery in algebra and arithmetic.
- **Actions**: The type of assignment provided by the teacher, namely algebra-only, arithmetic-only, or mixed.
- **Rewards**: Either the sum of the student's improvements or the reaching of a specific threshold.
- **Transition dynamics**: The student's skills evolve probabilistically based on their current mastery levels and the assigned task[1].

The specific type of RL problem can be classified based on the knowledge and observability of the environment. From the teacher's perspective, the student's learning dynamics are unknown and cannot be explicitly modeled, so the problem is best addressed using *model-free reinforcement learning*, rather than approaches that rely on known transition dynamics.

Regarding observability, the student's true mastery levels are not directly observable and must be inferred from assignment outcomes, making the environment *partially observable*. However, for the scope of this project, we assume that the AI Teacher generates high-quality assignments that accurately assess student mastery, allowing us to approximate the problem as a fully observable Markov Decision Process (MDP).

Simulation Environment

The student learning process is modeled as an MDP with a continuous state space. We assume the state transition follows:

$$S' = S + \Delta(S, a, d), \tag{1}$$

where $d \sim \mathcal{U}([1, 1.5]^2)$ represents the student's innate disposition. The learning update is defined as:

$$\Delta(S,a,d) \sim \mathcal{N}\left(-\frac{S \odot (S-1)}{5} \odot d \odot A(a), \sigma^2 diag(A(a))\right), \text{ with } \sigma = 0.01, A(a) = \begin{cases} (1,0) & \text{if } a \text{ is algebra-only} \\ (0,1) & \text{if } a \text{ is arithmetic-only} \\ \left(\frac{1}{2}, \frac{1}{2}\right) & \text{if } a \text{ is mixed.} \end{cases} \tag{2}$$

The learning update term $-\frac{S \odot (S-1)}{5}$ is chosen to be parabolic, with zeros at mastery levels [0,0] and [1,1] and a maximum update of [0.05,0.05], in order to model the increased difficulty of making progress at the early stages of learning a new topic as well as the challenge of achieving complete mastery.

Implementation[1]

Q-learning was chosen because the problem is model-free; compared to on-policy methods such as SARSA, Q-learning typically converges faster, which is relevant in this setting where minimizing the time to reach high mastery is a primary objective. Since Q-learning is formulated for discrete state spaces, the continuous mastery states were discretized into 100 bins to make the problem tractable.

To account for stochasticity in the student's learning process, the model was evaluated by repeating training over multiple independent runs and averaging the results. Two reward formulations were considered, each tested under both *greedy* and ϵ-*greedy* Q-learning policies. In the *dense-reward setting*, the reward was defined as the instantaneous learning gain, computed as the sum of the mastery updates across both skills. In the *sparse-reward setting*, a single terminal reward was provided when the student reached at least 90% mastery in both skills[2].

Results

The results shown[3] indicate that all tested configurations exhibit comparable learning efficiency. Across runs, the median number of assignments required to reach [0.9,0.9] mastery is approximately 44 for all settings. The sparse-reward, ϵ-greedy configuration reaches mastery slightly earlier, with an average improvement of about one assignment relative to the other configurations.

Outlook

Adaptive education systems can be beneficial in certain contexts, but they should function primarily as a supplement to classroom teaching or as tools for self-directed improvement. Schools play a fundamental role in the development of children and adolescents, not only by facilitating knowledge transfer and enhancing learning abilities, but also by fostering social skills through interaction with peers and teachers. Human interaction is a crucial component of education, as evidenced by the increased levels of anxiety and depression observed among students who spent a significant portion of their developmental years in lockdown[2].

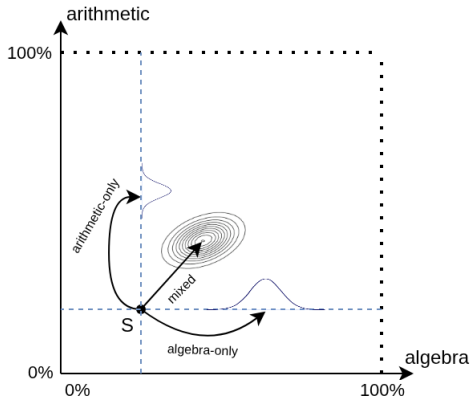

Figure 1: Transition dynamics representation

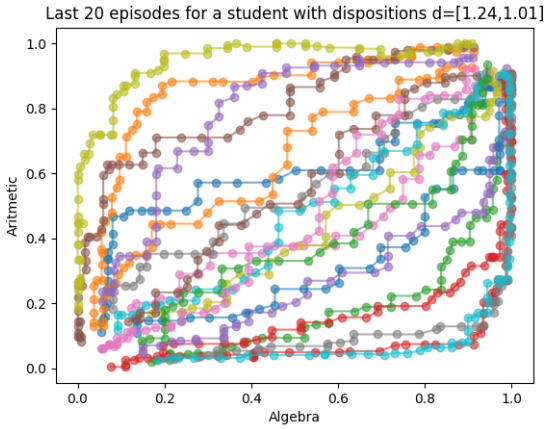

Figure 2: Traces from a run with sparse-reward and ϵ-greedy

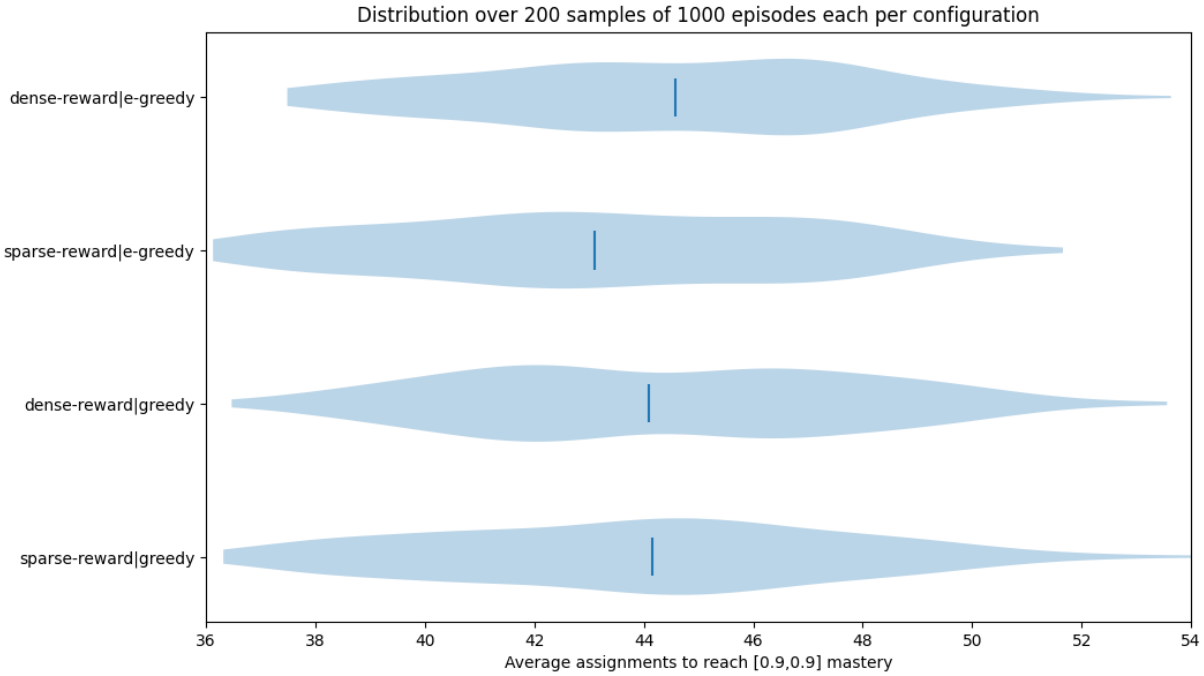

Figure 3: Violin plot of the learning efficiency

1. GitHub project repository
2. The impact of COVID-19 lockdown on child and adolescent mental health: systematic review