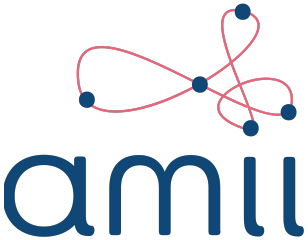
Work-in-progress: Comparing Feedback Distributions in Limited Teacher-Student Settings

Calarina Muslimani, Kerrick Johnstonbaugh, Matthew E. Taylor



Teacher-Student framework

- Teacher agent provides **guidance** to a student agent during the student's training process

- Teacher provides **evaluative feedback** to student agent

What can go wrong with including a teacher?

Problem	Possible Solution
Communication constraints	**Feedback budget!**

1. Fatigue, attention span
2. Cost, ability to teach multiple students

Goal of this work!

Given a limited feedback **budget**, ***when*** should teachers provide feedback to students in order to **maximize student performance**?

Possible feedback strategies?

- Adopted three teaching heuristics originally used in the action advising setting

Early advising

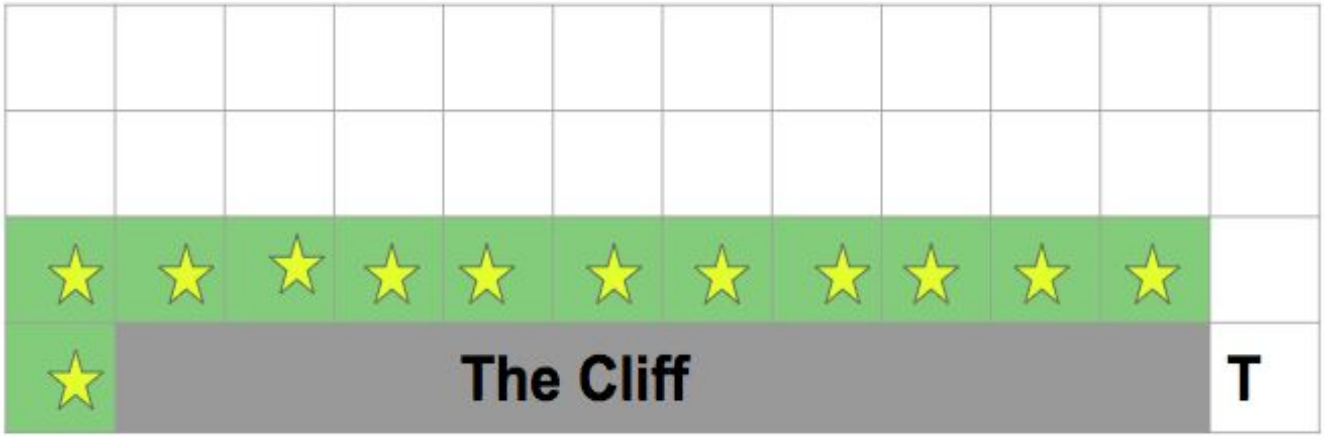
Torrey & Taylor ‘13

Alternating advising

➢ Provide feedback every ***u*** time steps

Importance advising

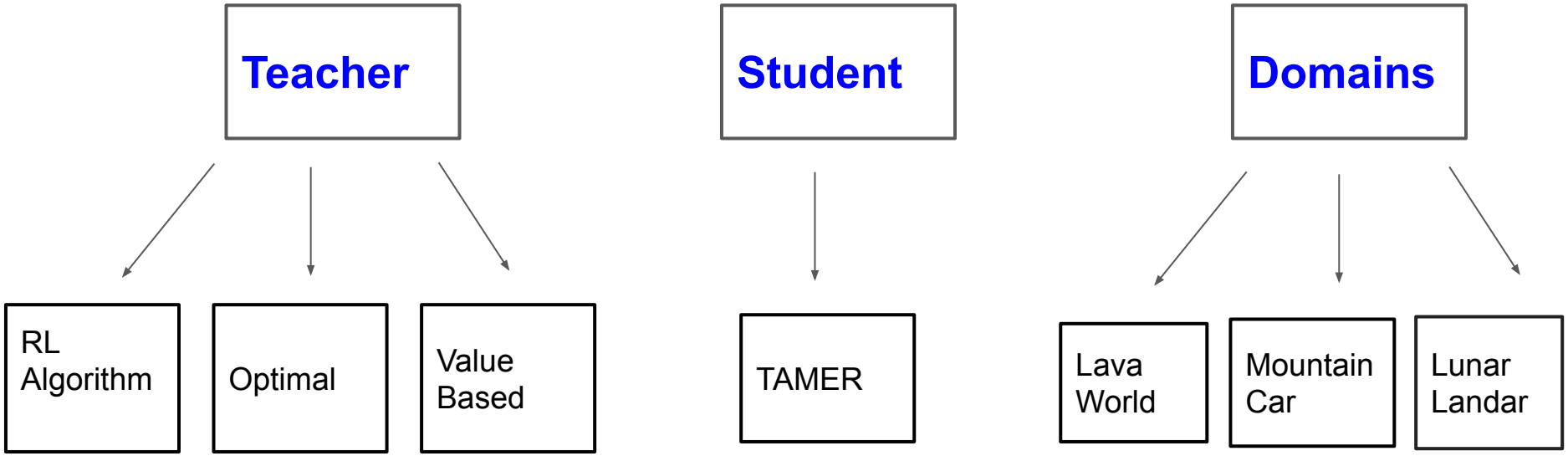
- Provide feedback at states where **difference between action values is large**



Novel teaching strategy: **Visitation Advising**

- Provide student feedback only at states that were **visited often by the teacher**

Experimental Design

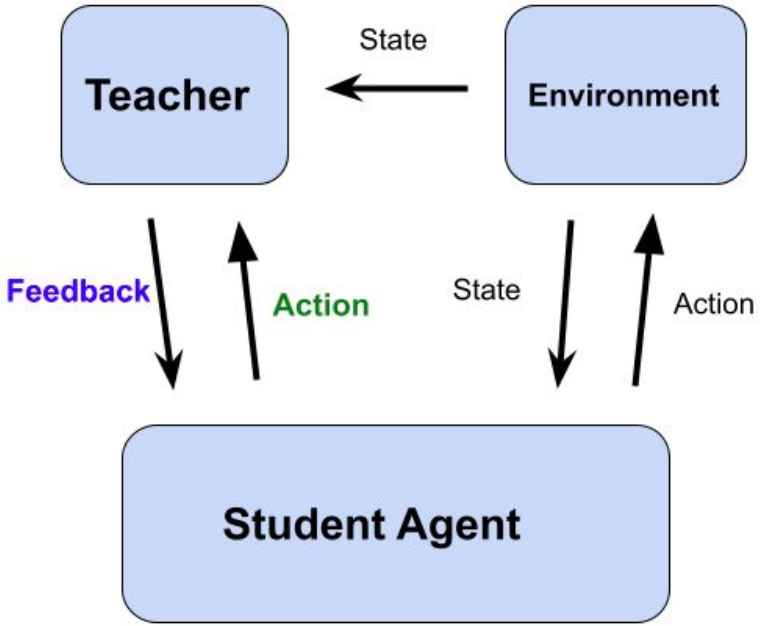


Student Agent: TAMER

- Supervised learning algorithm
- Student learns **teacher's reinforcement function**

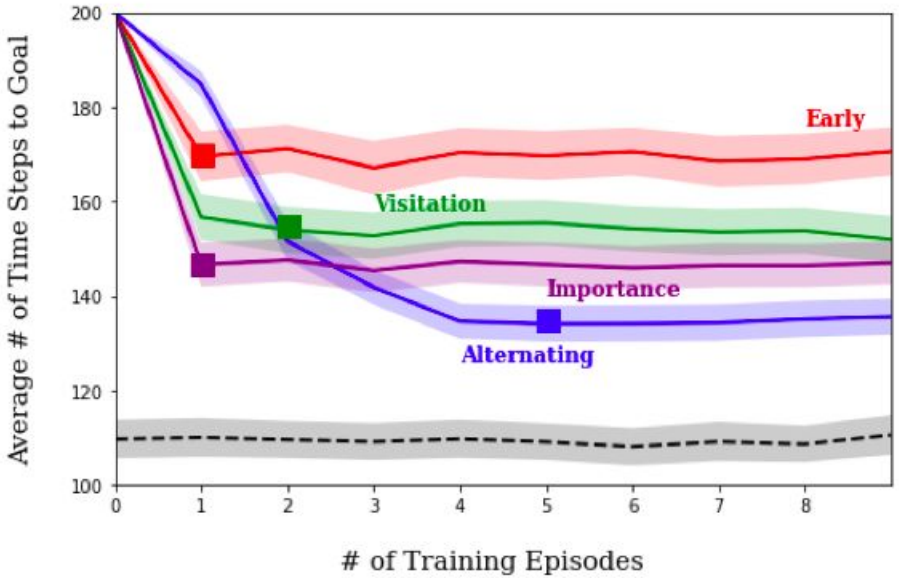
$$H : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$$

- Student goal is to **maximize immediate feedback**
- Student only **updates its model when the teacher provides feedback**



Knox & Stone '09

Results for Mountain Car

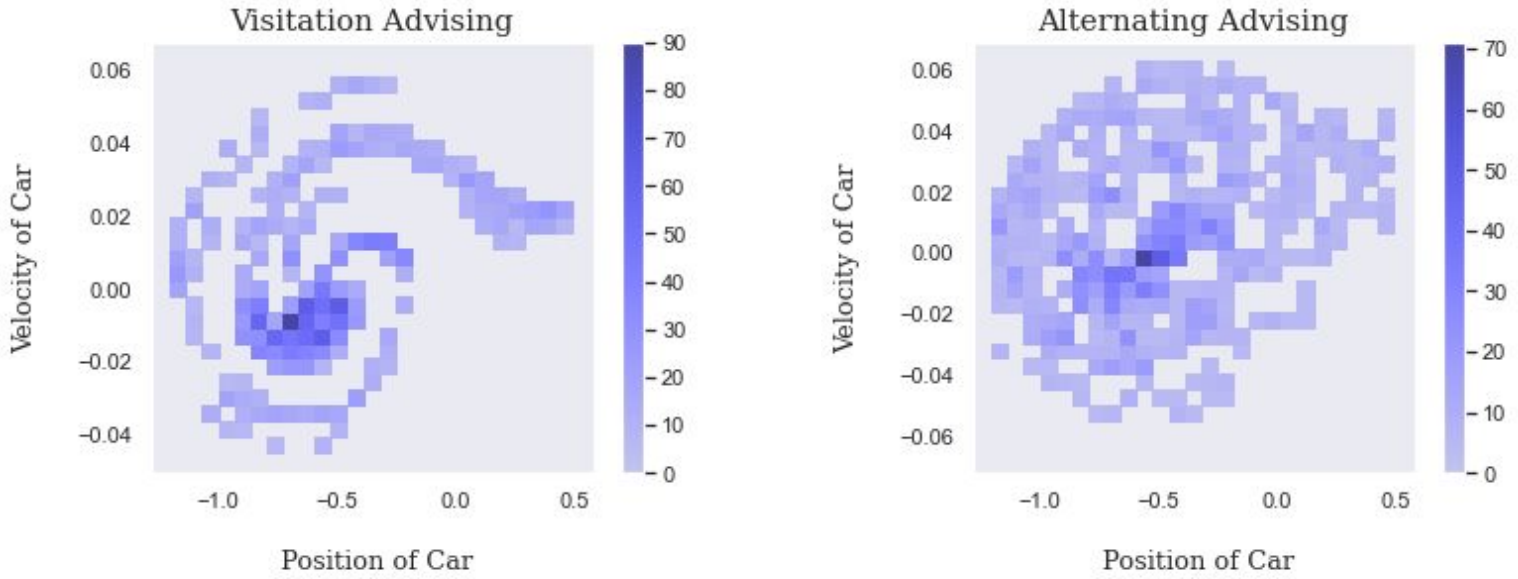


Performance of the student in Mountain Car in terms of average number of time-steps to reach the goal state. Black curve represents the teacher's performance. Blocks indicate the training episode where feedback budget was complete.

- SARSA teacher
- Budget = 75 feedbacks

- Student achieved **best performance** being taught with **alternating advising** but unable to reach optimal behavior

Where did learning occur in Mountain Car?



States the student was advised at with visitation advising (Left) and alternating advising (Right). Darker blue indicates more learning updates occurred at that state. Start state region has position [-.6, -.4] and velocity of 0.

And the winner is ….

Best performing teaching strategy

Lava World: Visitation Advising

Mountain Car: Alternating Advising

Lunar Lander: Importance + Alternating Advising

- **No clear winning feedback strategy**
  - Efficacy of strategy is environment dependent
- Distribution of feedback does impact student learning
  - Learning efficiency
  - Quality of policy

What comes next?

- TAMER's inability to propagate feedback to past states and actions was a limiting factor
  - Other teacher-student algorithms?

- Visitation advising is constrained by its inability to directly associate visit counts with states in non-linear function approximation settings
  - Define a visitation metric over neural networks?

Calarina Muslimani
MSc Student, Computing Science
https://www.linkedin.com/in/calarina-muslimani-696a13138/