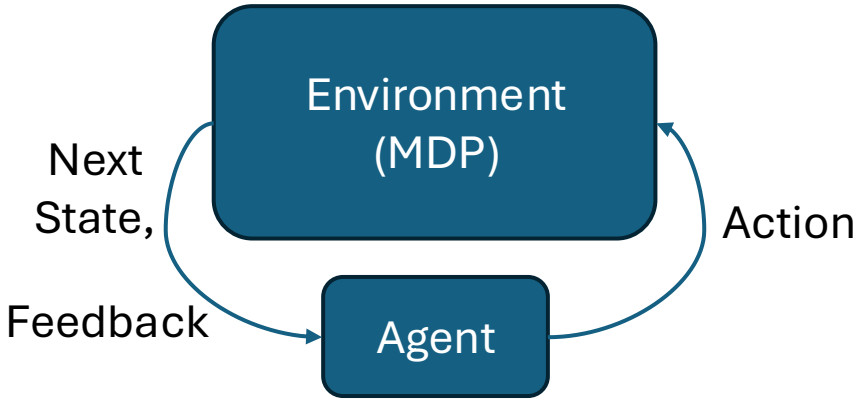
Paper

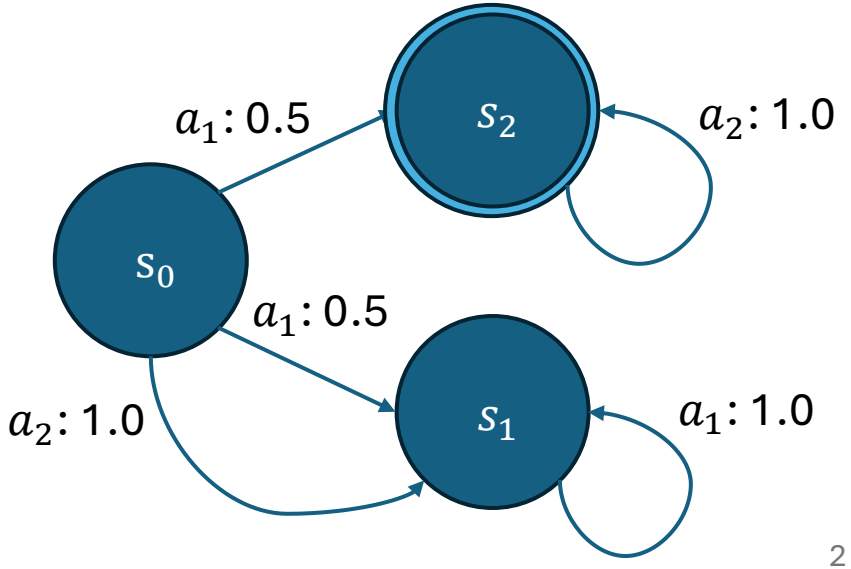
Reinforcement Learning for Reachability: Guaranteeing Asymptotic Optimality

Authors: Amogh Palasamudram Jakub Svoboda, Krishnendu Chatterjee, Suguman Bansal

Background: Reinforcement Learning (RL)

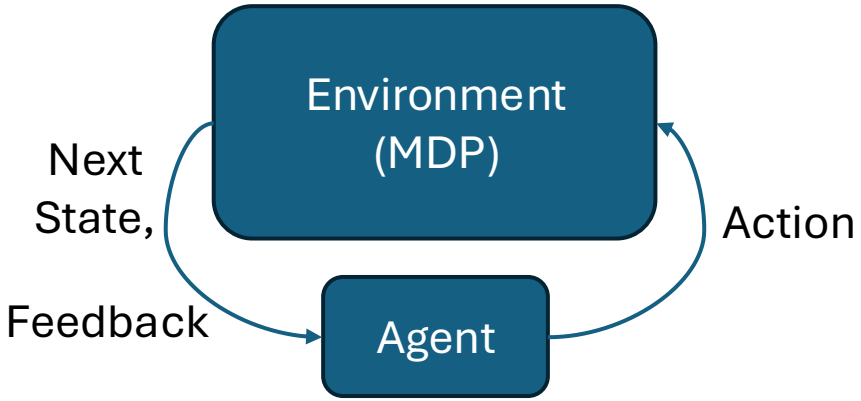
- **Environment:** an MDP that represents the 'world'
  - States, actions, transitions
- **Agent:** explores and learns the environment
- **Policy (π):** dictates which actions to take at each state

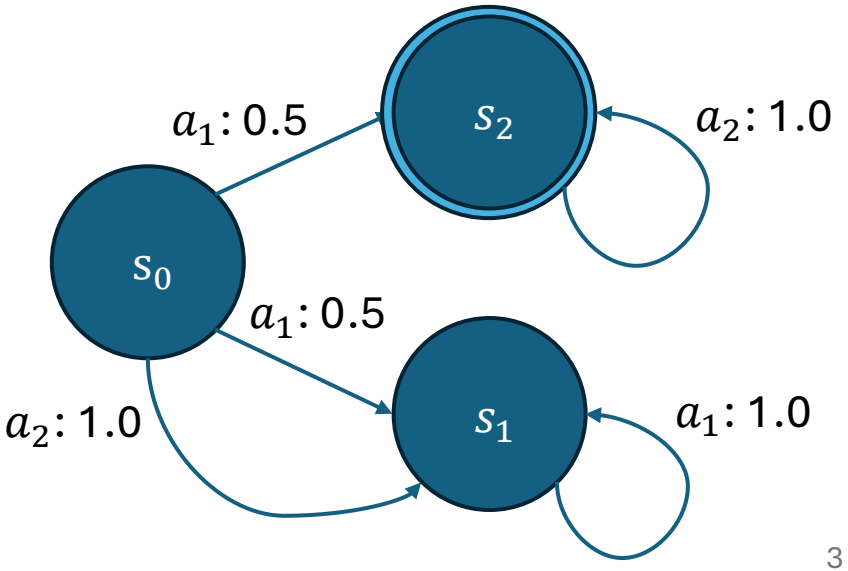




Background: RL for Reachability

- Given a reachability objective, find a policy that satisfies it
- **Optimal Policy (π_{opt}):**
  - Policy that maximizes the probability of reaching a goal state (*reachability value*)
- We know nothing about the MDP
  - (Black Box Environment)
- Perform actions to sample and learn the environment





Open Problem: RL for Reachability

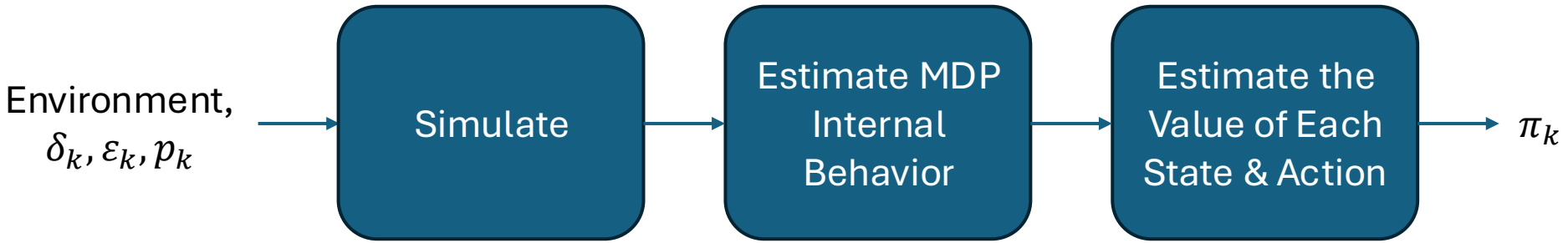
- Is there an RL Reachability algorithm that can asymptotically find the optimal policy for a black-box environment?
- Intuitively, this should be true
  - If we sample a lot, we should be able to accurately estimate the MDP and find the optimal policy
- There is no mathematical proof that such an algorithm can exist

Our Contributions

- We created an algorithm that asymptotically finds the optimal policy for a black-box environment
  - Current algorithms require some constraint on the MDP to work

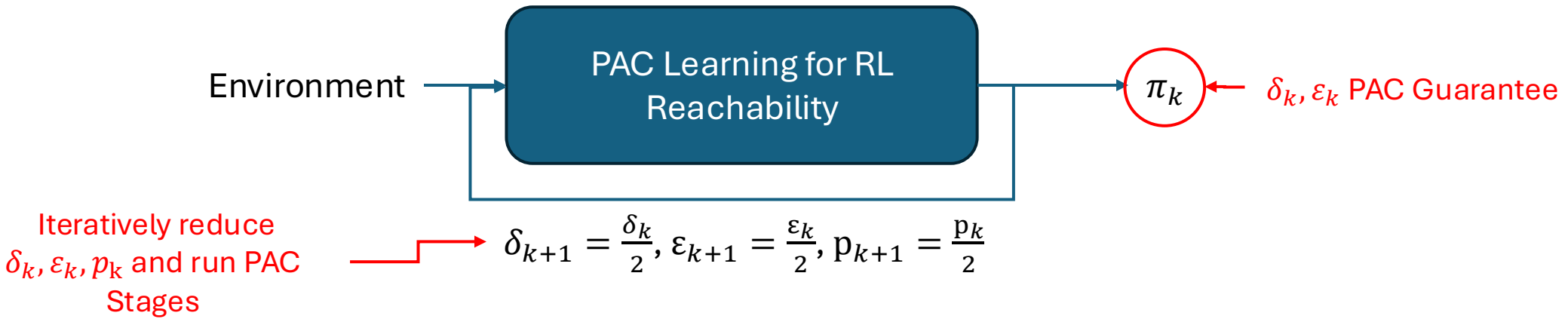
- We provide theoretical and empirical insights regarding the convergence dynamics of our algorithm

Algorithm: PAC Stage



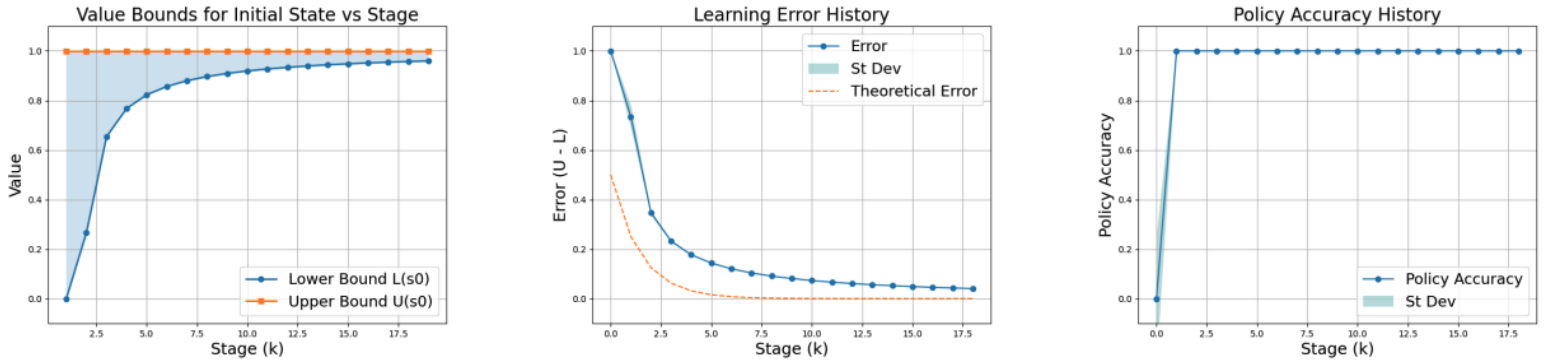
Asymptotic Guarantees using PAC Stages

- Intuition:
  - Iteratively run the PAC algorithm, but exponentially decrease δ, ε, p
  - $\delta, \varepsilon, \mathrm{p} \to 0$, so eventually we will get the optimal policy with probability 1



Empirical Results

- We used the PRISM MDP Benchmark dataset to test our algorithm
- We relaxed some of the theoretical requirements for the runs:
  - E.g. number of simulations to run



(a) Value Bounds vs. Stage k *(b)* Error vs. Stage k *(c)* Policy Accuracy vs. Stage k

Figure 3. Performance on Dining Philosophers benchmark. The values plotted are the median values over 10 trials.

Conclusion

- We have solved the open problem:
  - Create an algorithm that asymptotically guarantees finding the optimal policy with probability 1 in the black-box MDP setting.
- We provided an intuitive algorithm to solve this problem
- We provided a theoretical analysis and proof as to why this works

Thank You