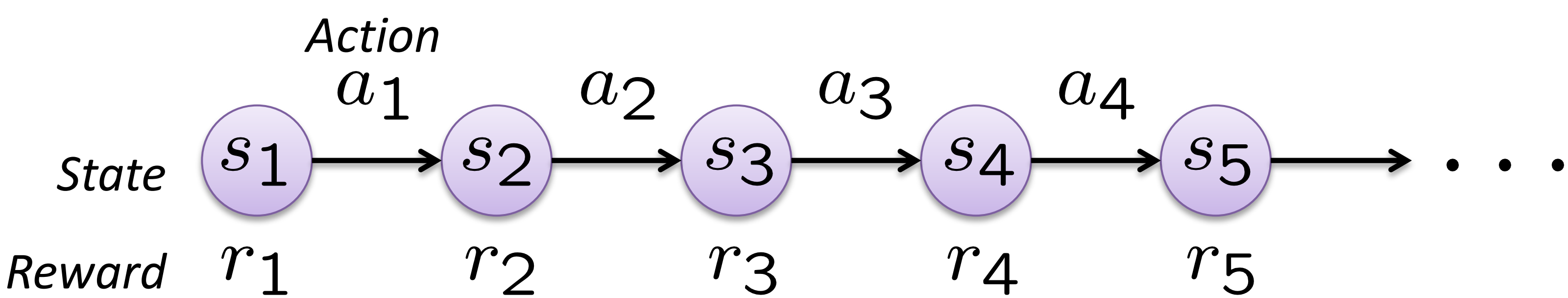
The Fixed Points of Off-Policy TD

J. Zico Kolter
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology

Poster T6

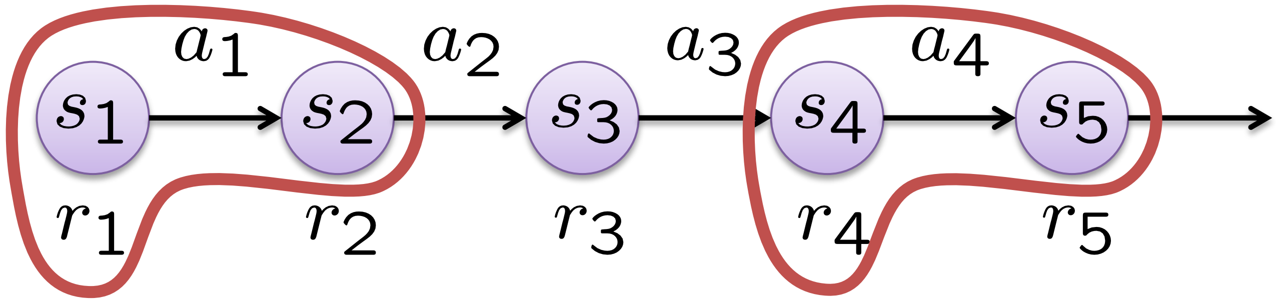
- **Given sequence of experience:**



- **For an *arbitrary* policy, determine value (expected sum of discounted rewards) for acting under that policy**

- **Can be solved, in principle, by Temporal Difference learning:**

$$\pi : S \to A$$



Repeat: $\hat{V}(s_i) \leftarrow \hat{V}(s_i) + \alpha\left(r_i + \gamma \hat{V}(s_{i+1}) - \hat{V}(s_i)\right)$

- **Works when values are represented explicitly, but might not work with value function approximation**

$$\hat{V}(s) = \theta^T \phi(s)$$

On-Policy	**Off-Policy**
TD converges [Tsitsiklis and Van Roy, 1997]	TD can fail to converge [Boyan, 1994] … fixed! [Sutton et al., 2008]
TD solution close to true value function [Tsitsiklis and Van Roy, 1997]	TD solution can be arbitrarily poor [example in paper]

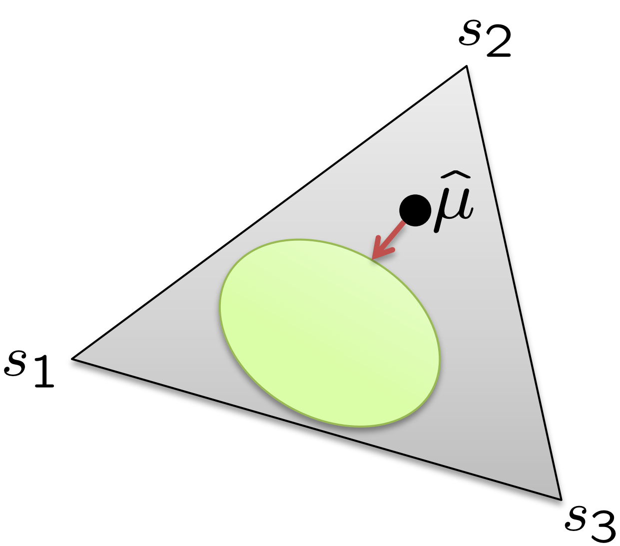
- **This work is about fixing off-policy TD**

  Basic idea: reweight samples so that TD solution has quality guarantees (and so that TD converges)

- **Technical idea**

Off-policy sampling → $s_i \not\sim \mu(\pi)$ → "TD operator" not a contraction

"filtered" states   stationary distribution of policy



Key contribution: we can find set of *all* distributions for which TD operator *is* a contraction (represented efficiently via a linear matrix inequality)

Algorithm: Project empirical distribution on to this set

- **Guarantees on resulting solution quality**

$$\boxed{\|\hat{V} - V^{\pi}\| \leq K \|V^{\star} - V^{\pi}\|}$$

Modified off-policy solution

Best possible approximation in function class

- **Efficient projection via low-rank optimization of dual problem**

- **Provides much better solutions in practice**