Advanced Topics in Reinforcement Learning

Lecture 8: On-Policy Temporal Difference Learning

Josiah Hanna
University of Wisconsin — Madison

Announcements

- Homework 2 due next Thursday @ 9:29 AM
- Project proposals due tonight @ midnight central time!
- Start reading chapter 8 for next week (Models and Planning).

Today

- Finishing Prediction
  - Convergence of TD / Monte Carlo.
  - TD(λ)
- On-Policy SARSA for control.
- Q-learning for control.
- Off-Policy SARSA.

Review

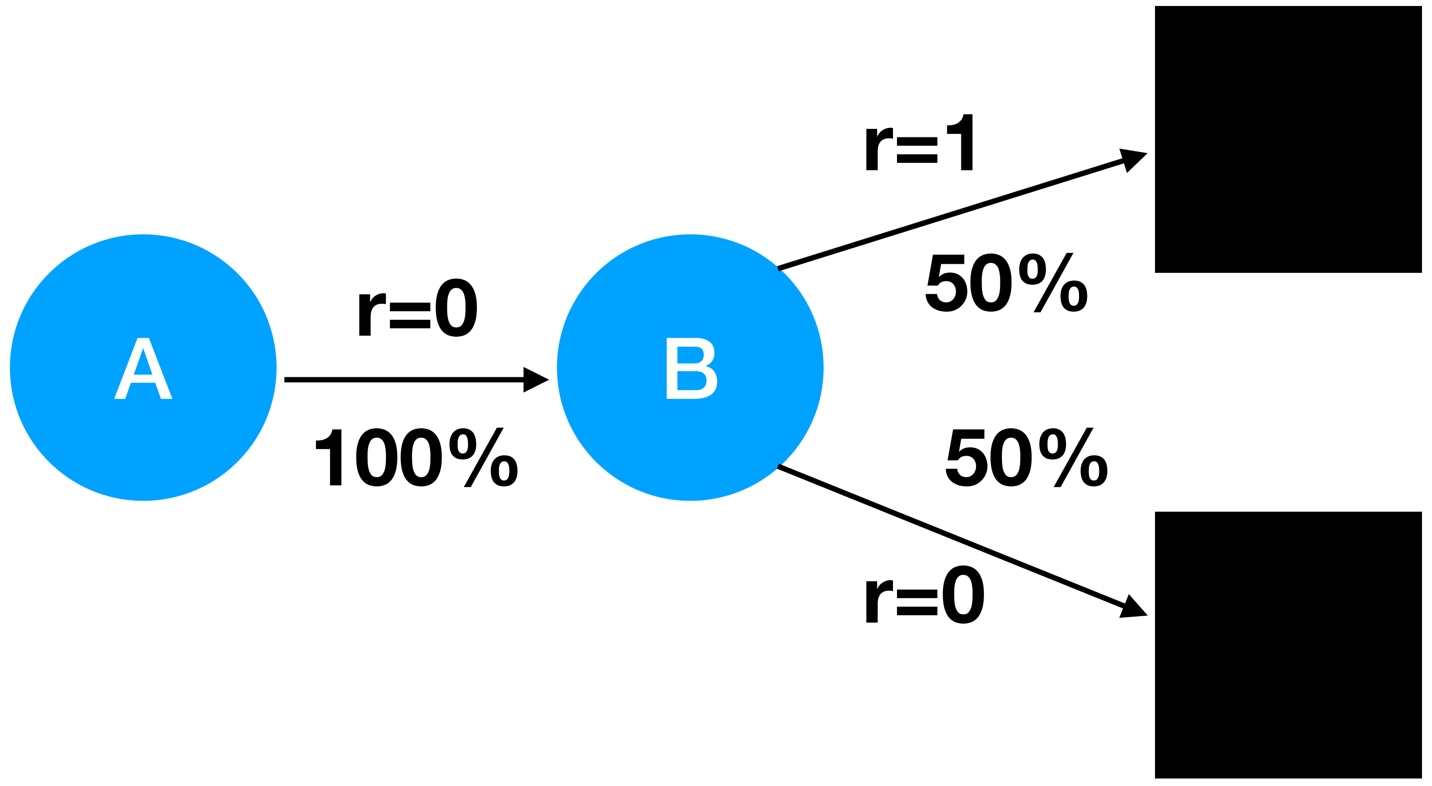
- Temporal difference (TD) learning learns from experience and bootstraps
    - Allows immediate learning without a model of the environment.
- In a batch setting, TD-learning converges to the certainty-equivalence estimate.
    - Highlights the connection between TD-learning and dynamic programming.
- TD-learning and Monte Carlo methods sit at either end of a spectrum of n-step return methods.

Gagan’s Presentation

Slides

Certainty Equivalence Updating

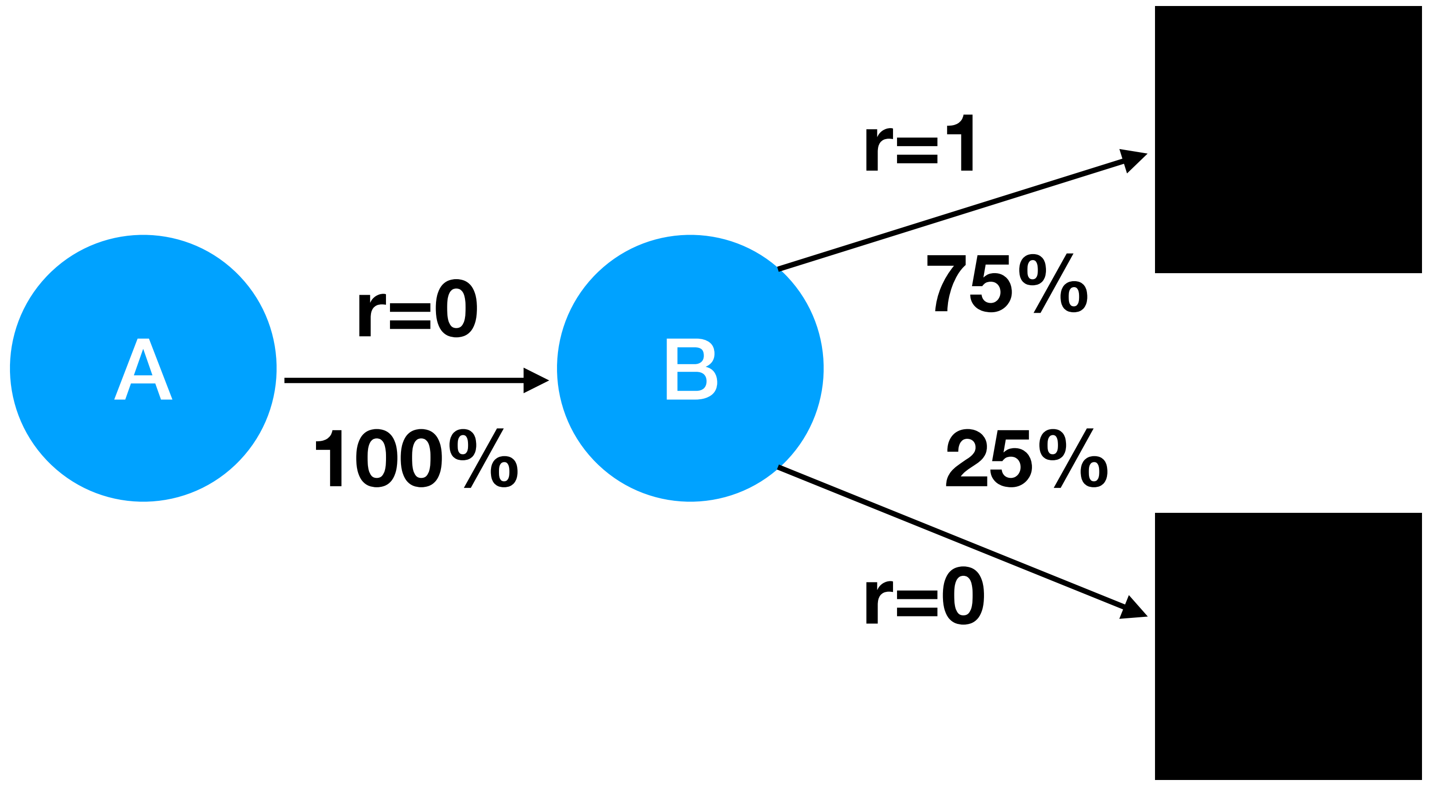
- Certainty Equivalence Learning: use data to estimate Markov process and then compute value function for the estimated process.



True Markov Process

Data:

A, 0, B, 0
B, 1
B, 1
B, 1
B, 1
B, 1
B, 1
B, 0



Estimated Markov Process

S A R S A

- Same generalized policy iteration scheme from past two weeks.
  - Evaluate π_k.
  - Make π_{k+1} greedy with respect to q_k.
- Now, use TD(0) to learn action-values:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

- Is this update on- or off-policy?
- What does generalized policy iteration with TD action-values and ϵ-greedy exploration converge to?

Q-Learning

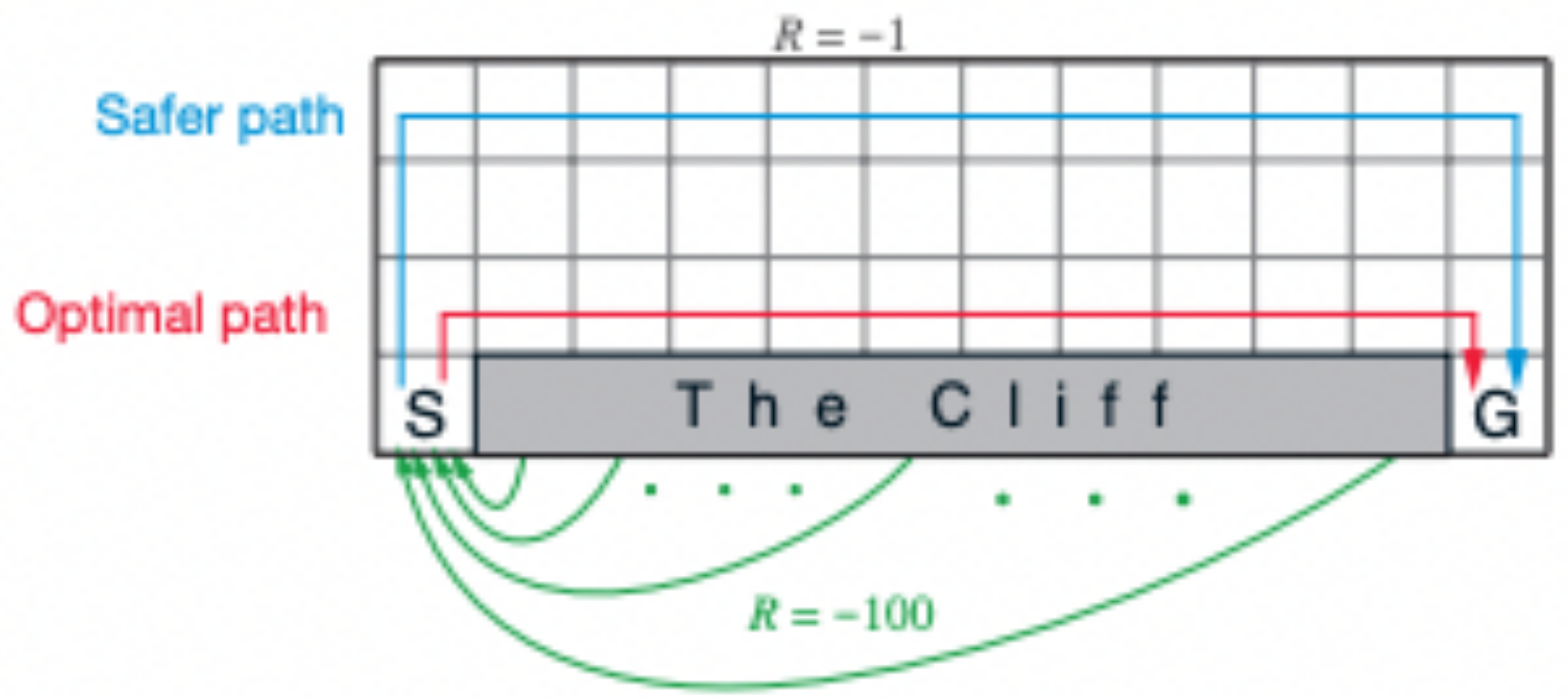
- SARSA is essentially policy iteration.
- Can we also use value iteration without a model of the environment?
- Q-learning update:

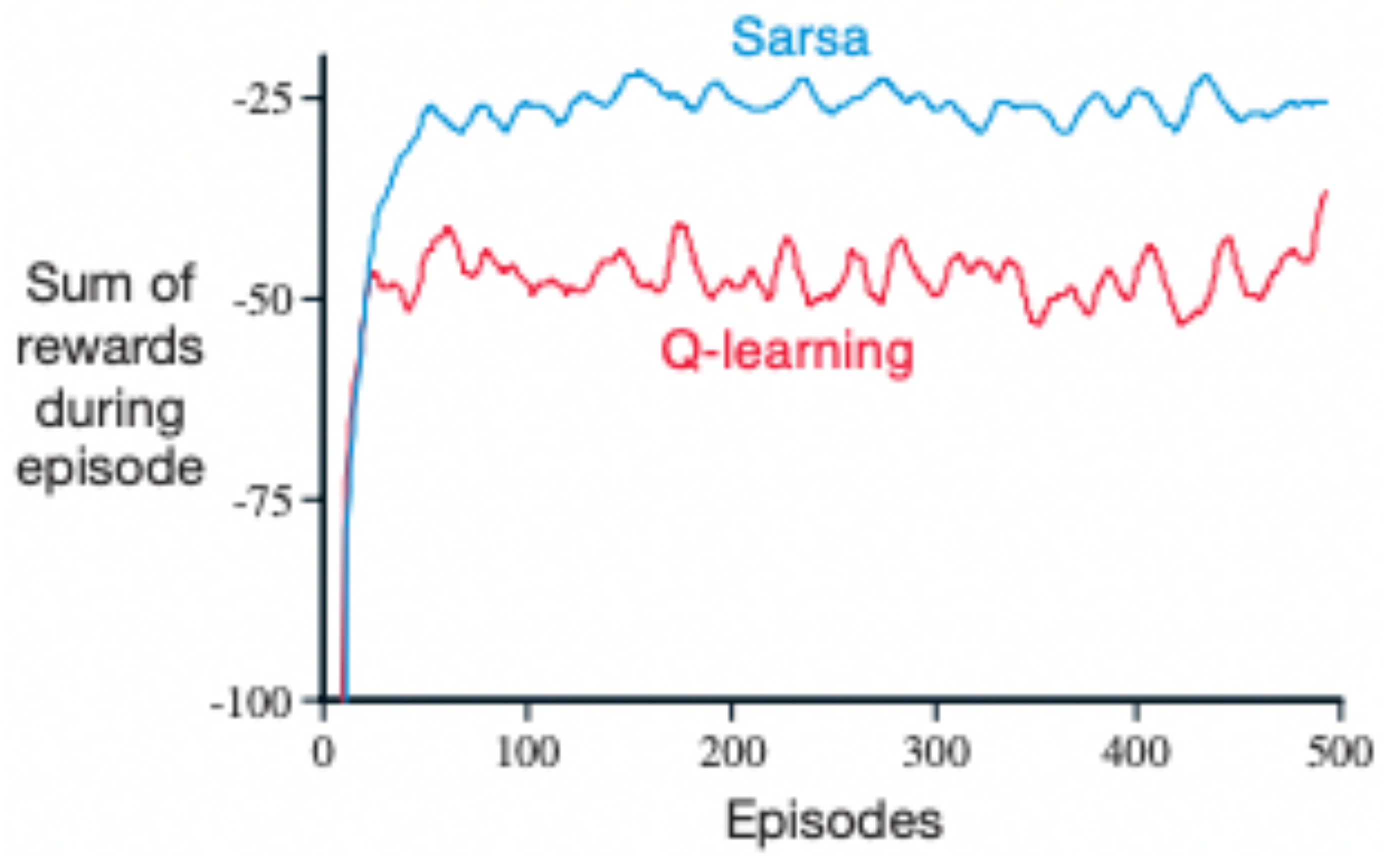
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t)]$$

- Is this update on- or off-policy?
    - Off-policy: can follow any policy (e.g., ϵ-greedy) while learning $q_\star$.
    - "Follow a policy derived from Q" — still off-policy!
- What does the Q-learning update converge to?
    - $q_\star$

Q-Learning or SARSA?

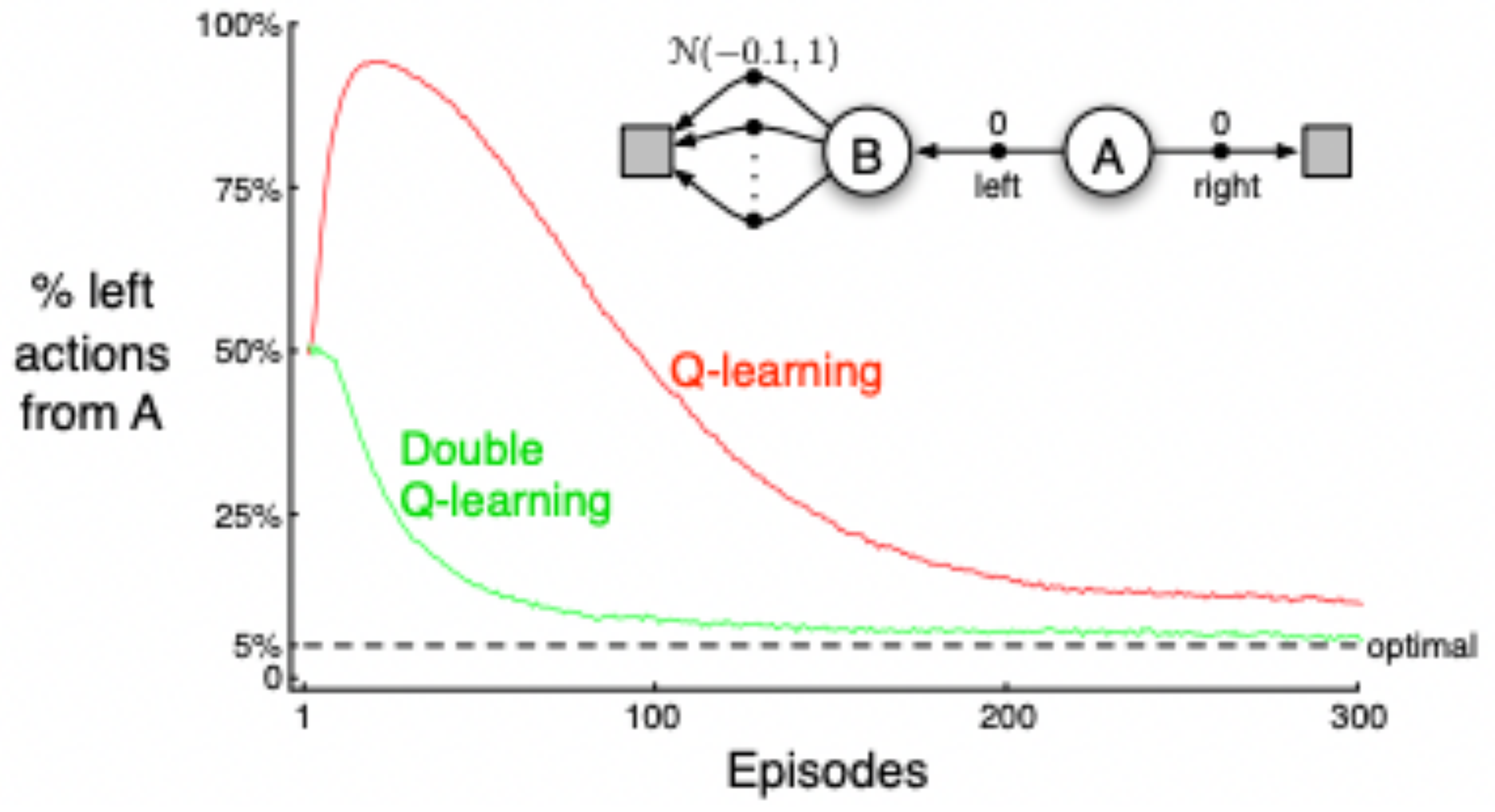
- Q-learning is off-policy; SARSA is on-policy.
  - Q-learning follows an exploration policy and learns $q_\star$.
  - SARSA follows an exploration policy, π, and learns q_π.
- What if exploration policy is greedy?





Double Q-Learning

- Q-learning may suffer from maximization bias?
    - What is it?
- Double Q-learning **mitigates this bias** by learning two action-value functions:
  $$Q_1(S_t, A_t) \leftarrow Q_1(S_t, A_t) + \alpha[R_{t+1} + \gamma Q_2(S_{t+1}, \arg\max_{a'} Q_1(S_{t+1}, a')) - Q_1(S_t, A_t)]$$
- Is this on- or off-policy?
- What does double Q-learning converge to?

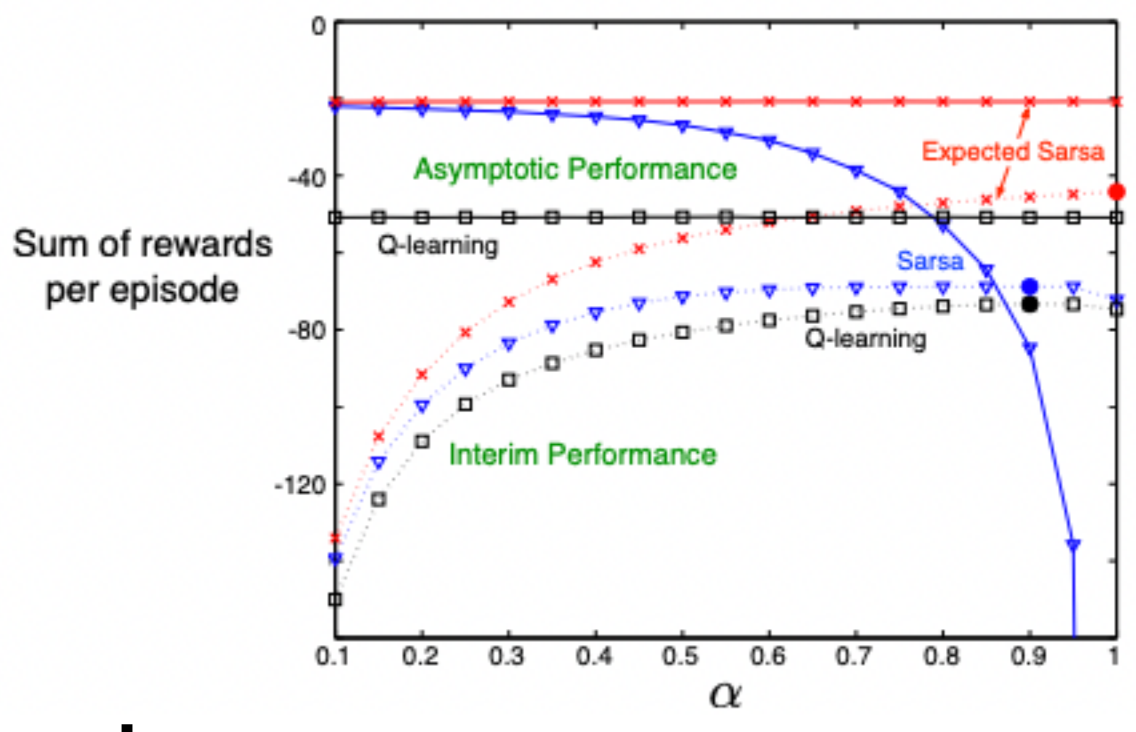


Off-Policy SARSA

- Can SARSA learn off-policy?
- Yes, with importance sampling!
- $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\rho_t[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$
- Where $\rho_t := \dfrac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$.
- Note that we only have a single factor in the importance weight.
    - What advantage would this have compared to off-policy Monte Carlo?
    - What is the off-policy variant of n-step returns?

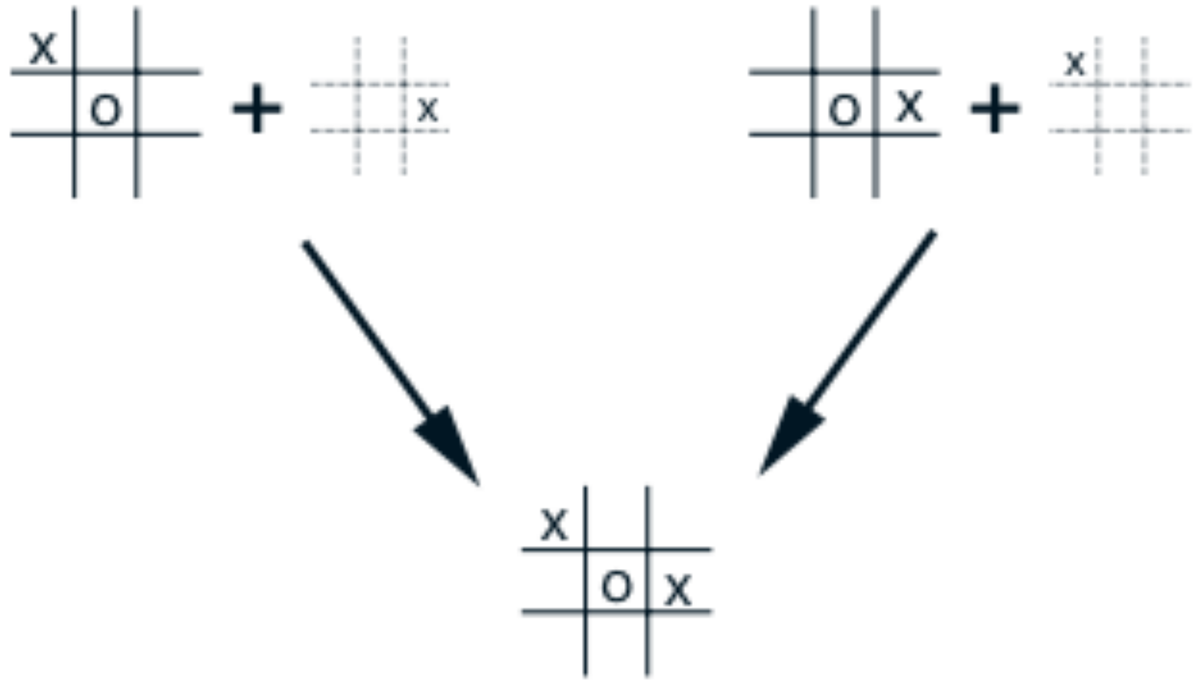
Expected SARSA

- SARSA samples the final acton A’. How could this be harmful?
    - We know π so we can compute the expected action-value exactly.
- $$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \sum_{a'} \pi(a' | S_{t+1}) Q(S_{t+1}, a') - Q(S_t, A_t)]$$
- How is this update useful? What are its limitations?
    - (+) Lower variance —> more data efficient learning.
    - (-) Computational cost for large or continuous action sets.



After-States

- In RL, the environment is usually a blackbox.
- But sometimes we have intermediate state changes that are available immediately after an action is taken.
- Such knowledge can be built into RL algorithms to help generalize learning.



Summary

- TD-learning can be integrated into generalized policy iteration in several ways.
  - SARSA uses on-policy TD-learning.
  - Q-learning learns $q_\star$ while acting off-policy.
  - Expected SARSA generalizes Q-learning and usually improves upon SARSA.
- These methods enable fully incremental, online, model-free learning.

Action Items

- Homework 2 due Thursday @ 9:29 am.
- Project proposal due midnight tonight.
- Begin reading Chapter 8.