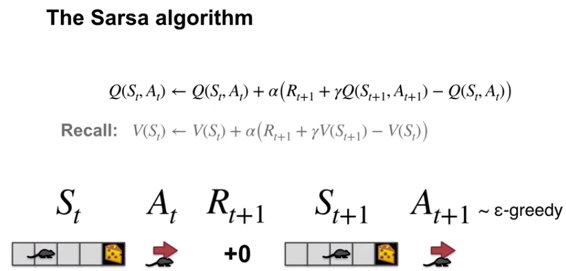
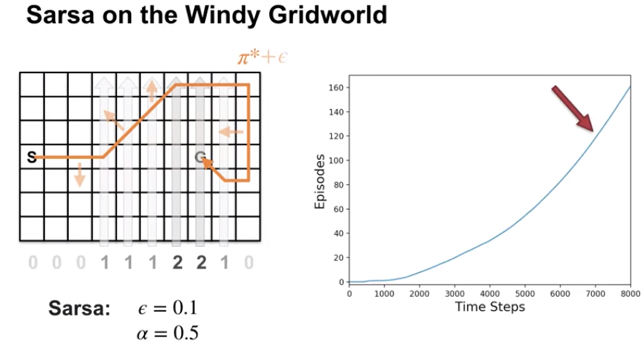
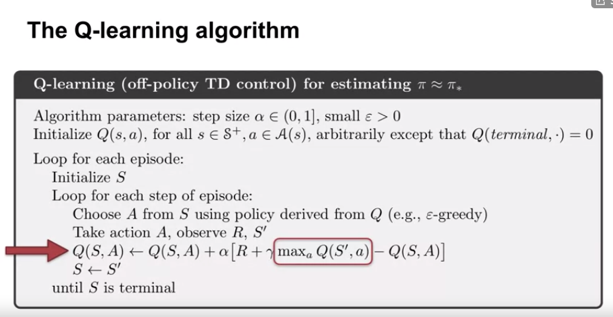
# Week 4 Notes

**Lesson 1: TD for Control**

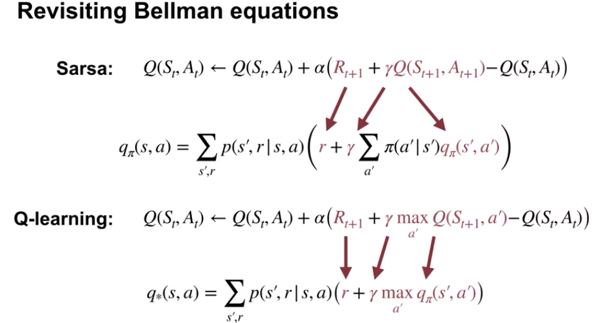
* Explain how generalized policy iteration (GPI) can be used with TD to find improved policies
  + Policy evaluation
  + Policy improvement
  + MC does eval and improvement after each episode.
  + TD improves policy after 1 evaluation step.
  + TD transitions S|A -> S|A
* Describe the Sarsa control algorithm
  + SARSA = State, Action, reward, next\_state, next\_action
  + 
* Understand how the Sarsa control algorithm operates in an example MDP
  + SARSA in windy grid world
  + 
* Analyze the performance of a learning algorithm

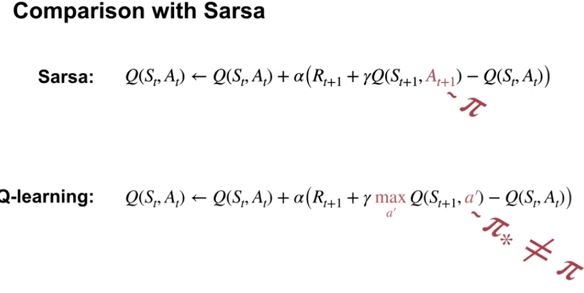
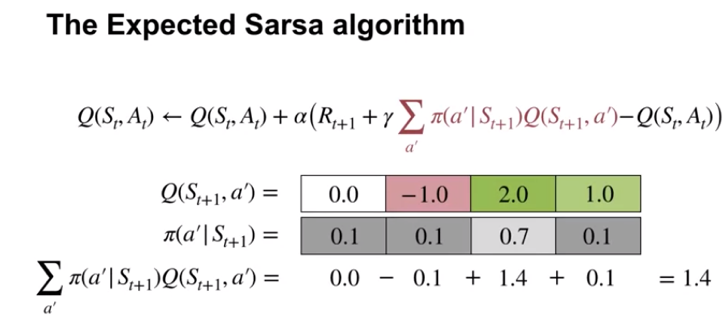
**Lesson 2: Off-policy TD Control: Q-learning**

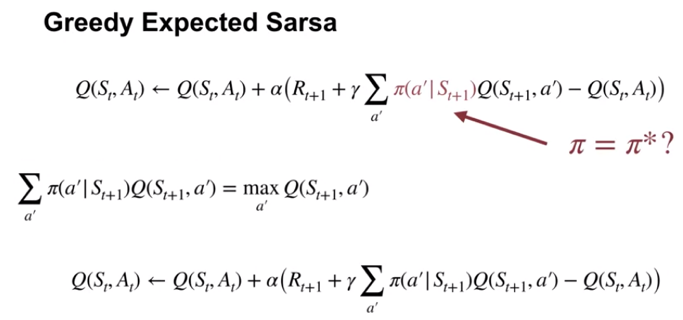
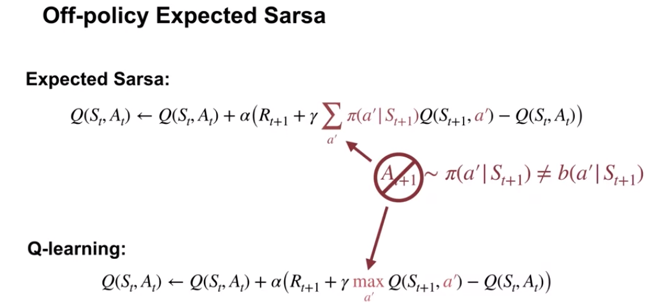
* Describe the Q-learning algorithm (11989)



* Explain the relationship between Q-learning and the Bellman optimality equations.



* + Sarsa is sample-based version of policy iteration which uses Bellman equations for action values, that each depend on a fixed policy
  + Q-learning is a sample-based version of value iteration
* Apply Q-learning to an MDP to find the optimal policy
* Understand how Q-learning performs in an example MDP
* Understand the differences between Q-learning and Sarsa
* Understand how Q-learning can be off-policy without using importance sampling
  + Target policy == behavior policy is on-policy
  + 
* Describe how the on-policy nature of Sarsa and the off-policy nature of Q-learning affect their relative performance
  + Sarsa reaches goal reliably and safely
  + Q learning actions are epsilon greedy so optimal but falls off the cliff with negative rewards and heads back to start.
* **Lesson 3: Expected Sarsa**
* Describe the Expected Sarsa algorithm
  + 
* Describe Expected Sarsa’s behaviour in an example MDP
* Understand how Expected Sarsa compares to Sarsa control
* Understand how Expected Sarsa can do off-policy learning without using importance sampling
* Explain how Expected Sarsa generalizes Q-learning
  + expected Sarsa and Q-Learning both use the expectation over their target policies in their update targets. This allows them to learn off-policy without importance sampling. Expected Sarsa with the target policy that's greedy with respect to its action values, is exactly Q-learning.



## 