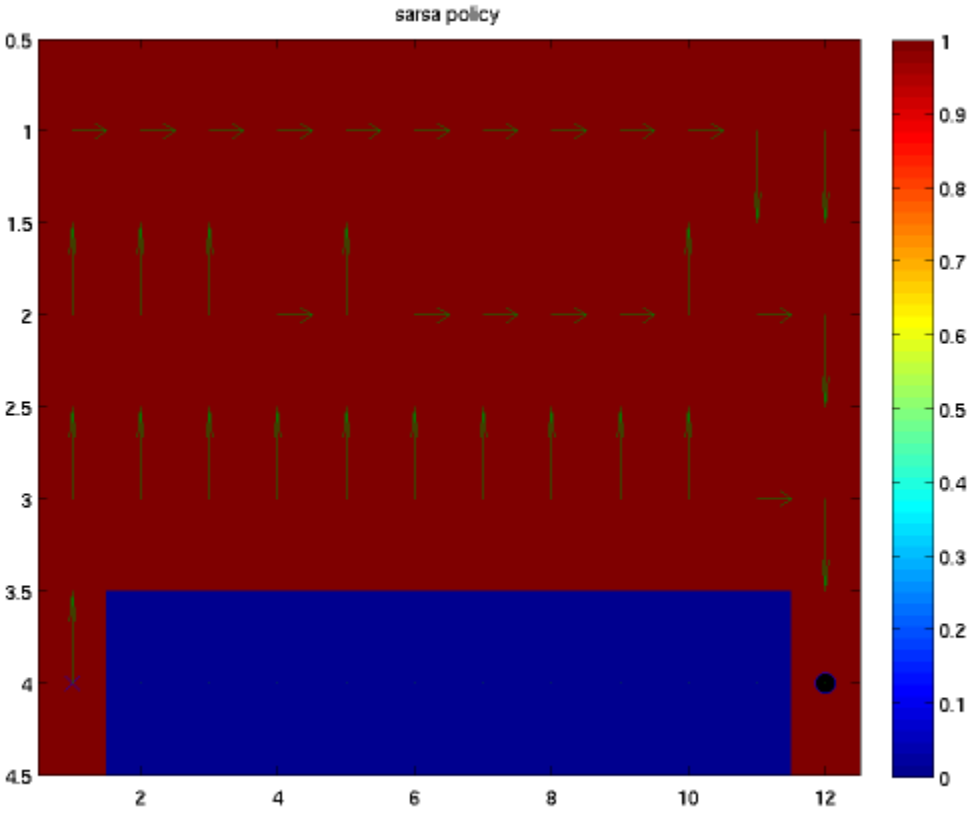
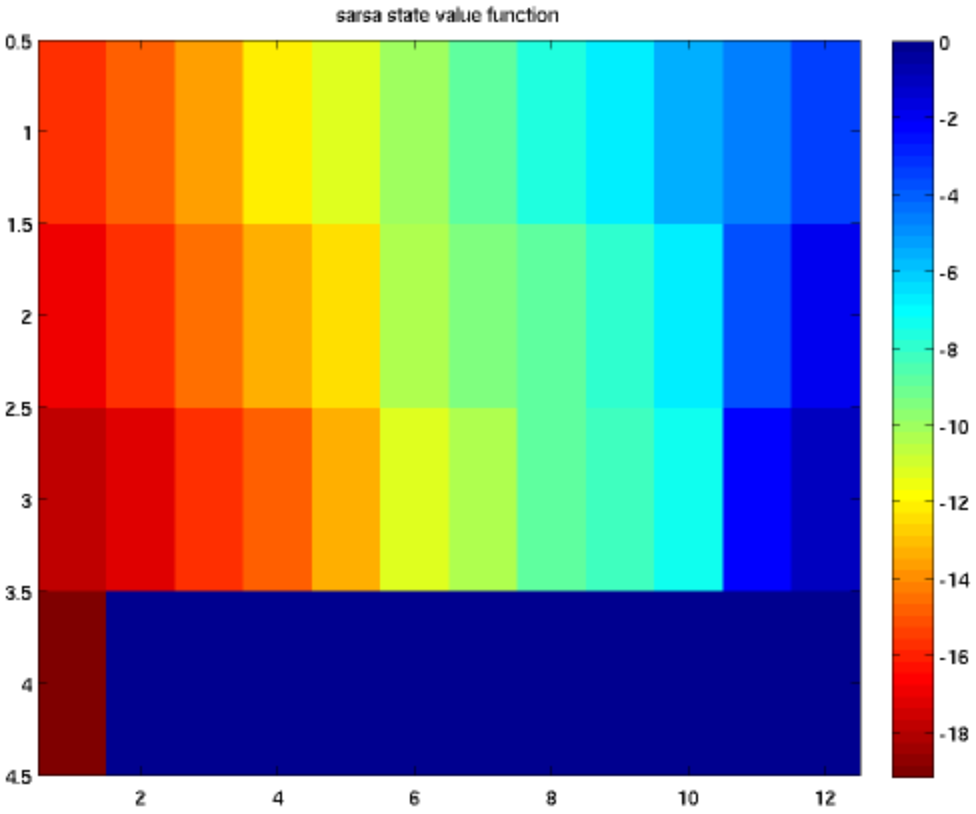
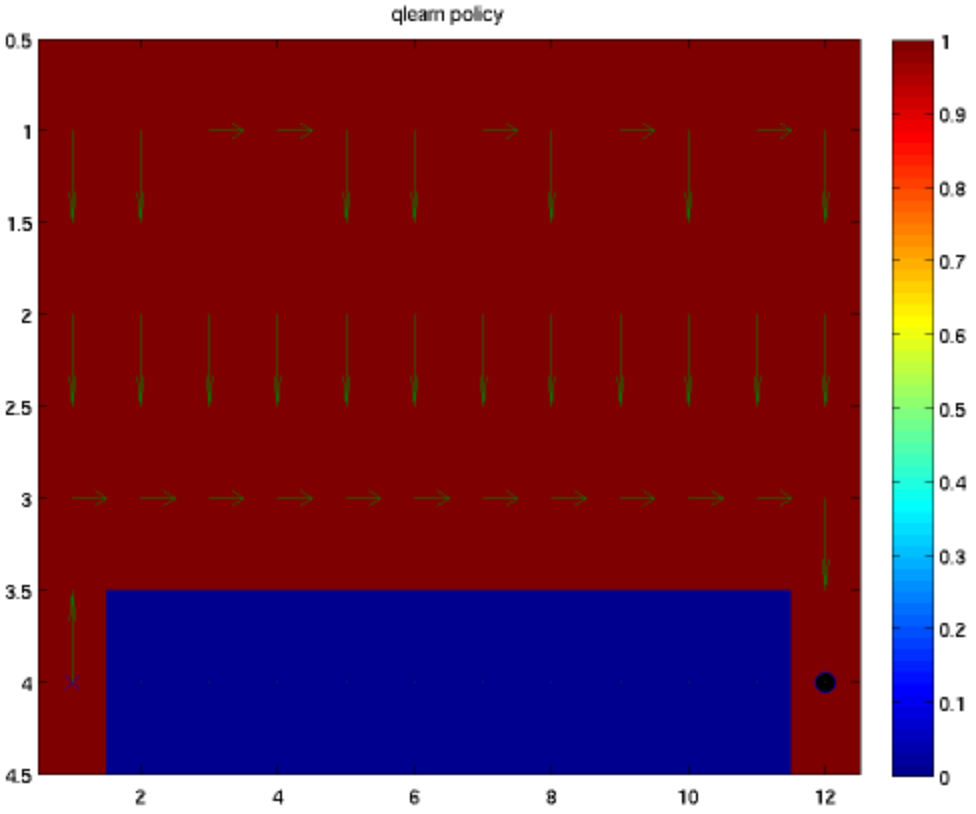
Here you will find experiments and results obtained when performing SARSA and Q learning algorithms on the cliff walk problem from the book. The next plot is the learned policy when using the SARSA on policy algorithm for this problem and ten thousand episodes. Here arrows represent the greedy policy direction to be taken at each point.



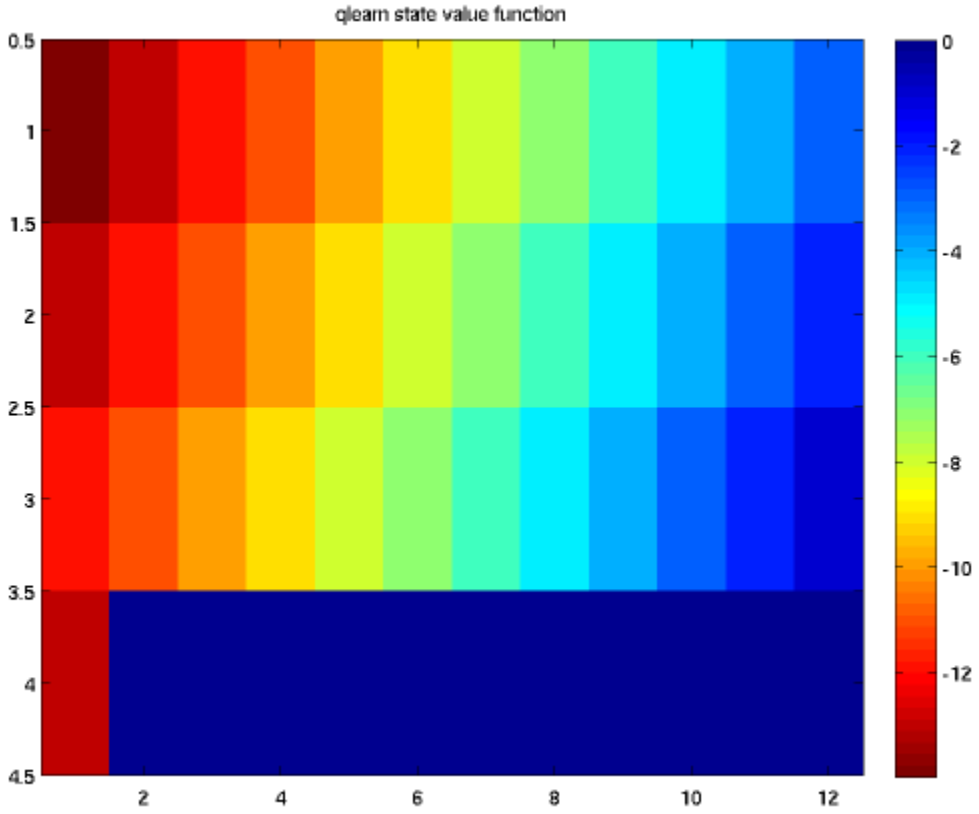
Note that this algorithm finds the "safe" path or the one that walks as far from the cliff as possible. The corresponding state value function is given by



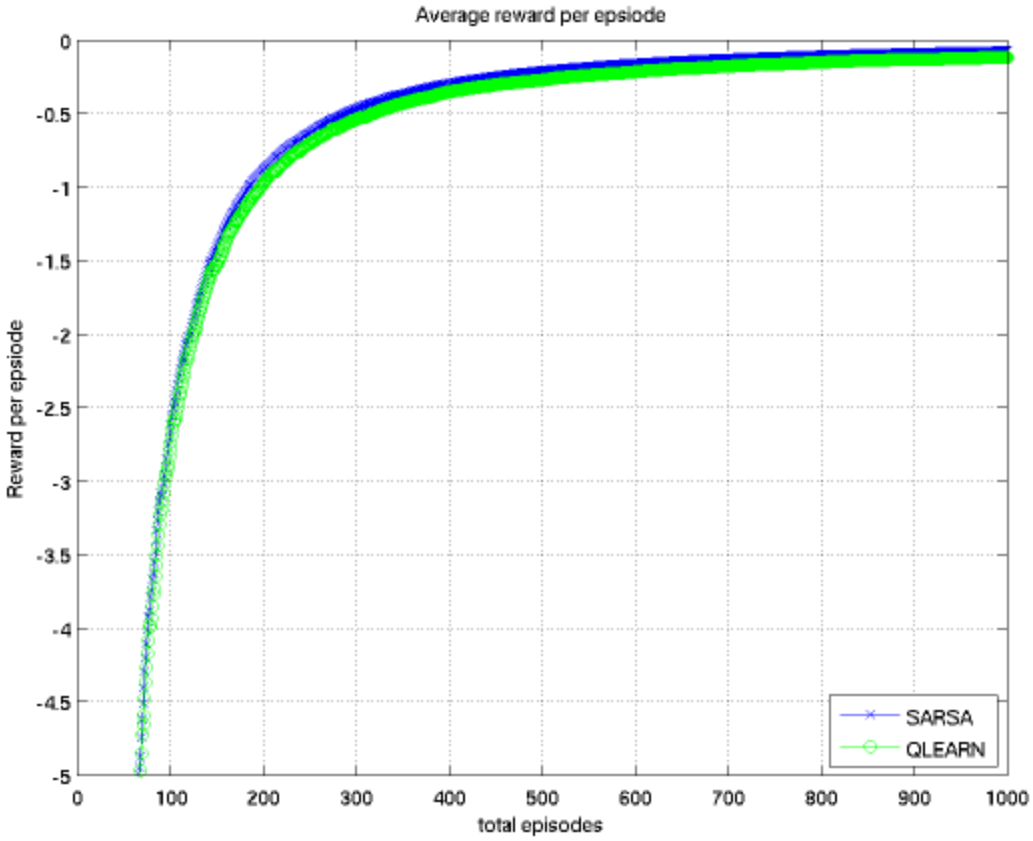
The next plot is the learned policy when using the Q-learning algorithm off-policy algorithm for this problem and ten million episodes. Again arrows represent the greedy policy direction to be taken at each point.



Note that this algorithm finds the quickest path or the one that walks as close to the cliff as possible. The corresponding state value function is given by



These match quite well with the similar results presented in the book. It is interesting to note that the Q-learn policy can look somewhat strange in regions where we don't visit very often. There the statistics is quite poor and can result in correspondingly poor action estimates. Since during our episodes we don't actually visit these states very often the fact that we have poor action estimates is of no consequence. If we compute the average reward per episode under each algorithm for many episodes we get the following plots.



John Weatherwax
Last modified: Sun May 15 08:46:34 EDT 2005