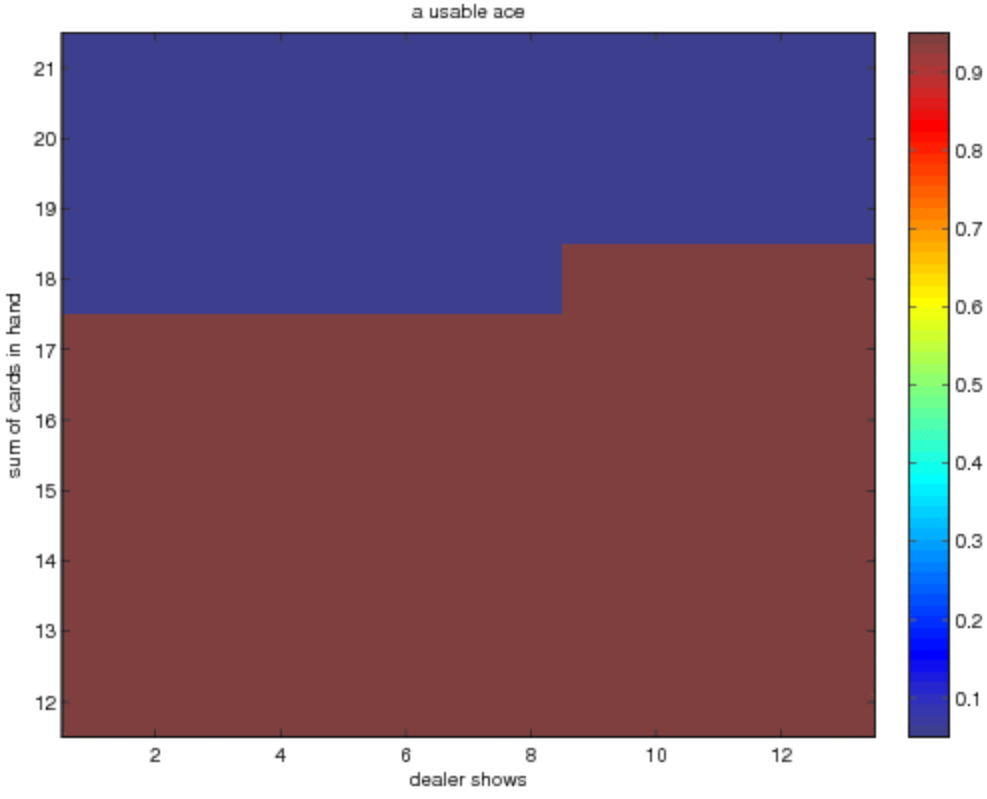
I ran various experiments to find the optimal epsilon soft policy via Monte Carlo simulation for blackjack. I computed the optimal policy under this class of policies. Since my state space for the dealer card consists of several individual cards 11, 12, 13 (the face cards) while the book collapses all of these states into one I need many more Monte Carlo trials to properly compute the optimal action value function and the corresponding policy. In the following plots I simulated 44 million games of black jack and computed the policies under a couple different variations

- Save all rewards that follow each state and a count of the number of times each state is visited. Directly compute the action value by dividing the two or the direct (non-recursive) average formulation.
- Recursively update our action value estimates
- Use the fixed step size learning algorithm.
- Set the initial values take for the action value function to be relatively large (taken to be +5) to encourage exploration.

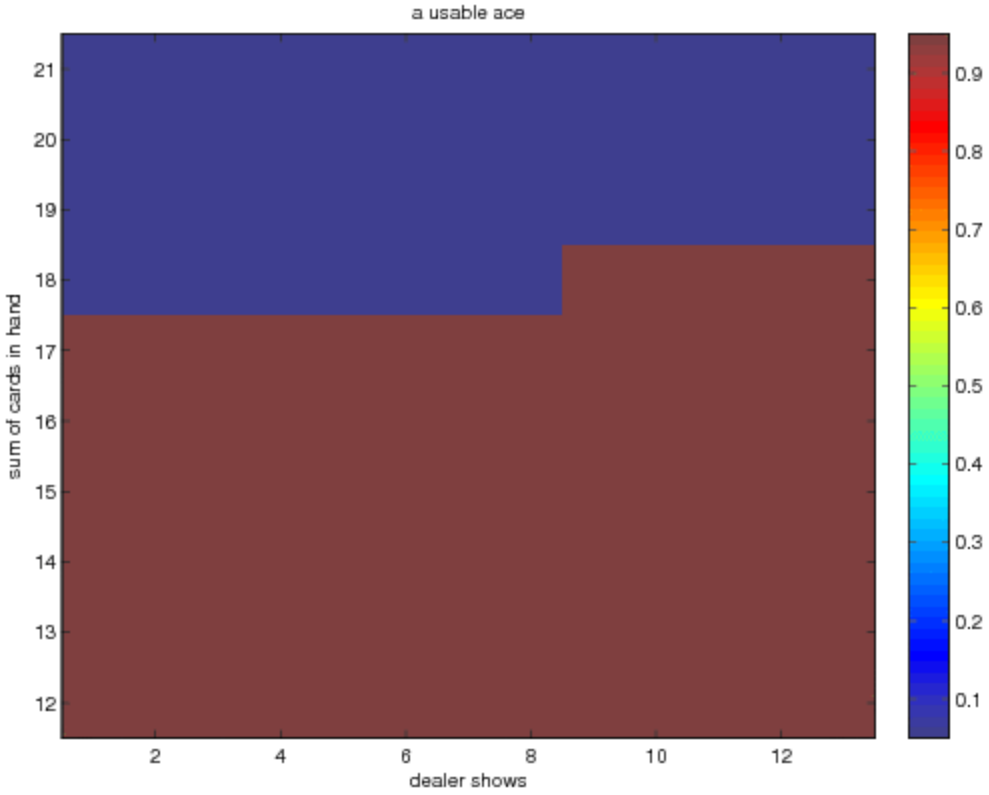
In general, the fixed step size learning algorithm (with alpha = 0.1) performed quite poorly, with resulting policies that were not continuous and had many "holes". These results are not presented. They might improve with a smaller value of alpha however.

The results of from using the direct averages are all very compatible. Some minor differences seem to exist in the computed policies however. The minor differences have to do with the sampling of the 11,12,13 cards (the face cards) and action to take when the dealer shows an ace and the player does not posses an usable ace. These differences maybe caused by sampling issues from under-sampling some of the states (dividing up our state space into the individual cards 11,12,13 is not as efficient as holding on to a single state since each state then would get one third the number of samples.

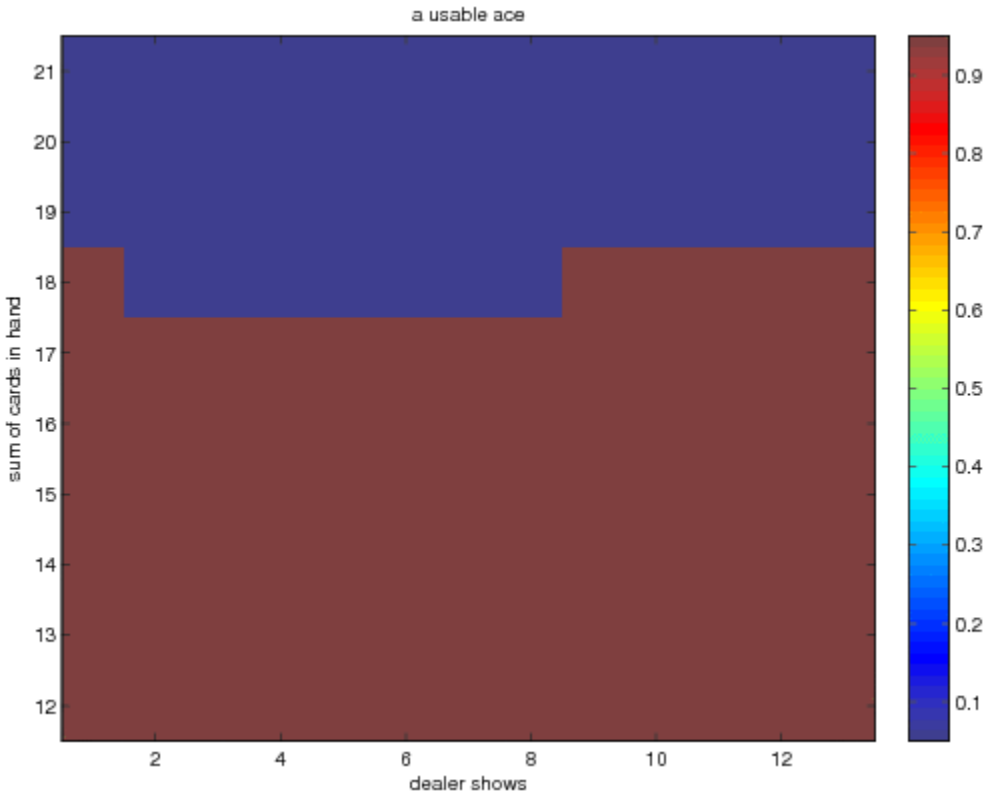
The optimal policies when the player has a usable ace for the direct average technique



when we use incremental averaging

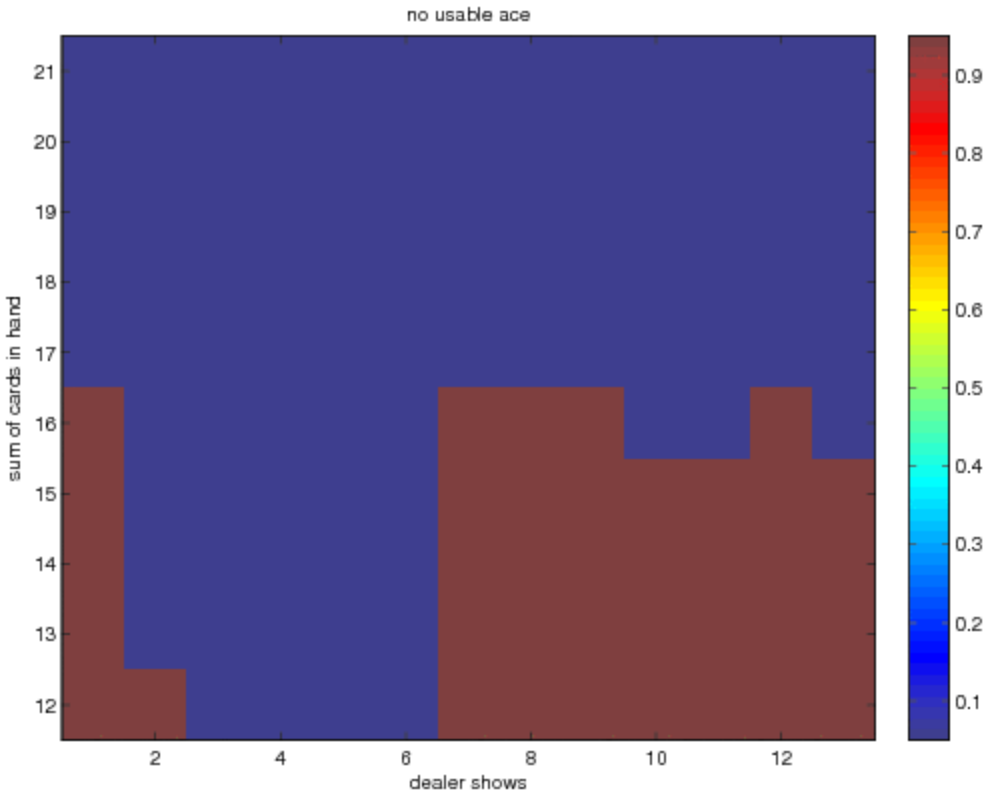


and finally when we use the method of exploring starts with incremental averaging

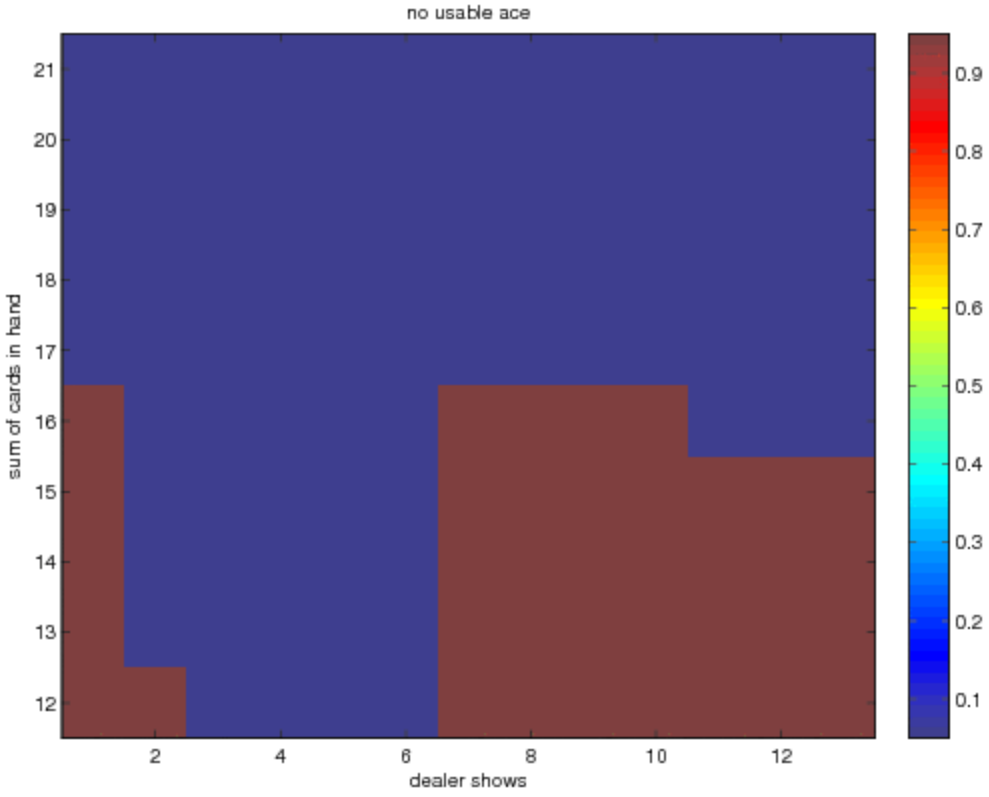


Note that these policies disagree on the action to take when the dealer shows an ace.

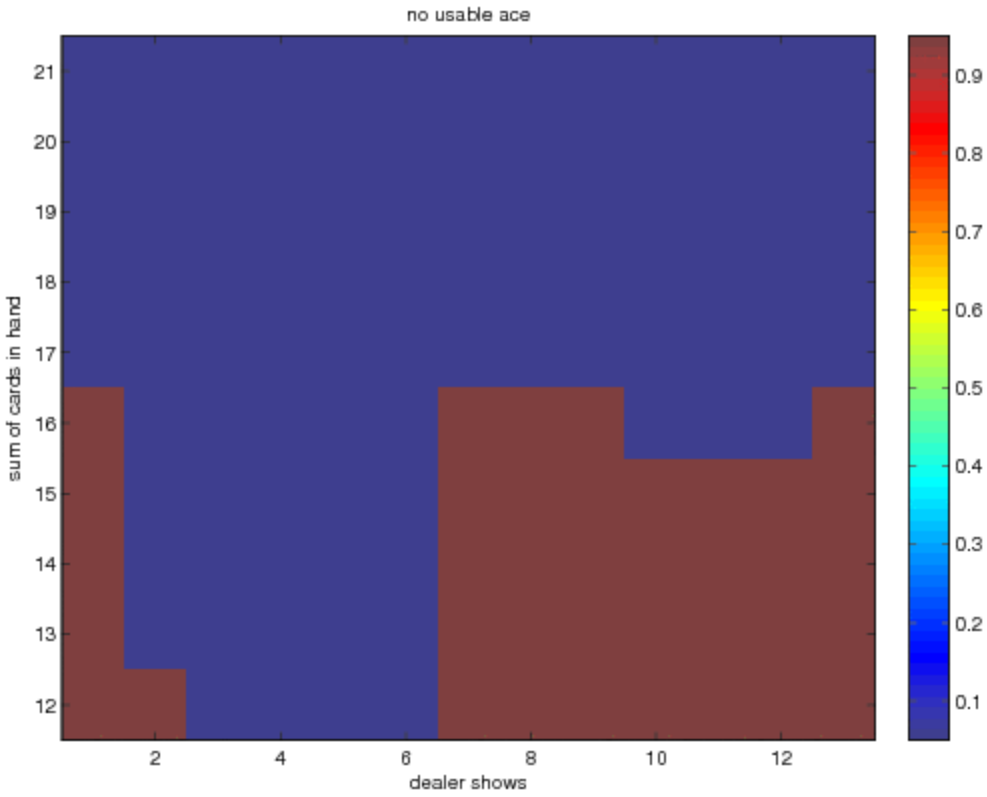
I find that the optimal policies when the player does not have a usable ace becomes, for the direct average technique



when we use incremental averaging



finally when we use the method of exploring starts with incremental averaging



In general the policies are very similar.

John Weatherwax