Dakota Walker
CSC449
December 2, 2022
Programming Assignment 3

SARSA(λ) with Fourier Basis Linear Function Approximation

The Setup

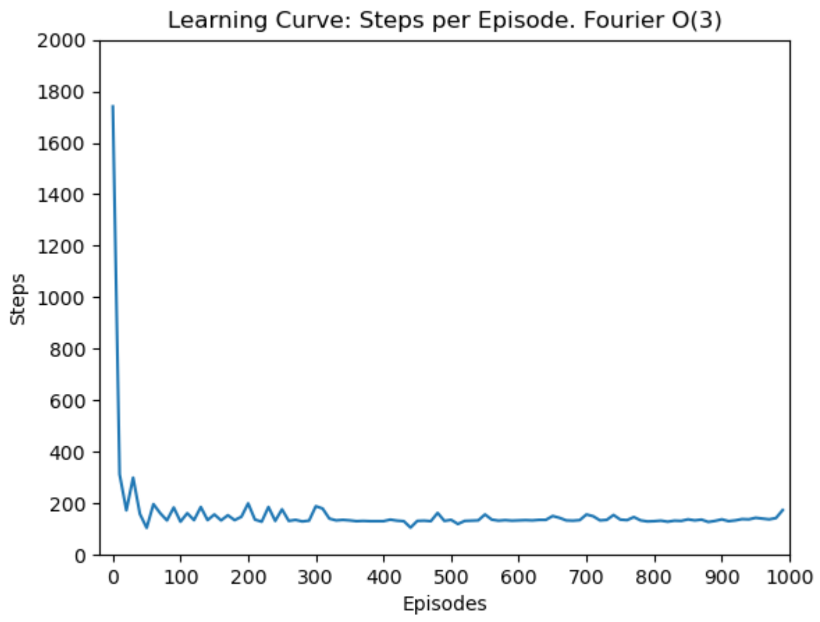
OpenAI's gym environment was used for my implementation and testing. I disabled the time step limit normally implemented, and let the car go until it found the goal. The position state spanned from -1.2 to 0.6, with >= 0.6 being the goal and episode termination state. The three possible actions available are accelerate to the left, no acceleration, or accelerate to the right. Every trial was a run of 1000 episodes. The learning curve graphs were sampled every 10 episodes to help reduce some of the noise. The value function surface graphs were created with a function that returns the calculated weight of the best action, and used as "steps to go."

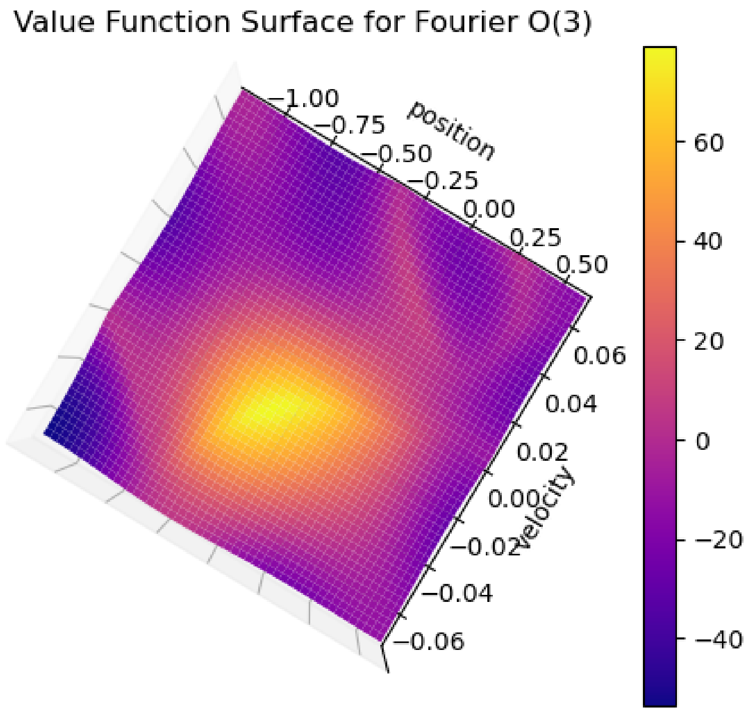
Hyperparameters used for all tests:

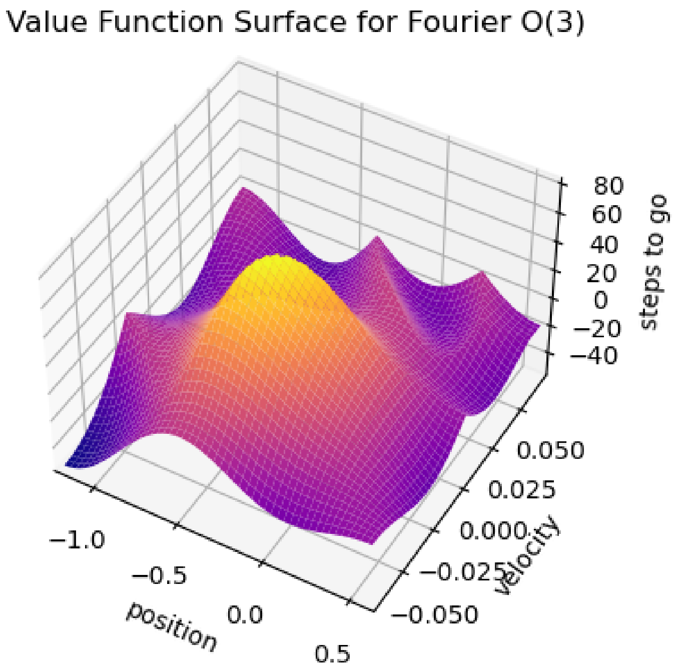
Alpha = 0.001 Learning Rate	Epsilon = 0 Random Action %	Gamma = 1 Trace Decay	Lambda = 0.9 Discount Factor

Order 3 Basis

Testing started with an order of 3 for the Fourier basis. During trials of 1000 episodes, the learning curves and value function surfaces remained rather consistent. The corners would occasionally change depending on if a great policy was found early.

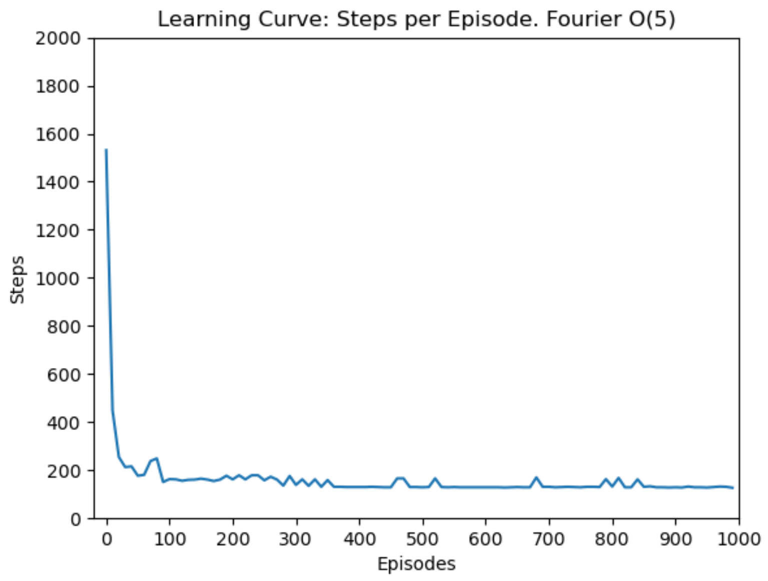


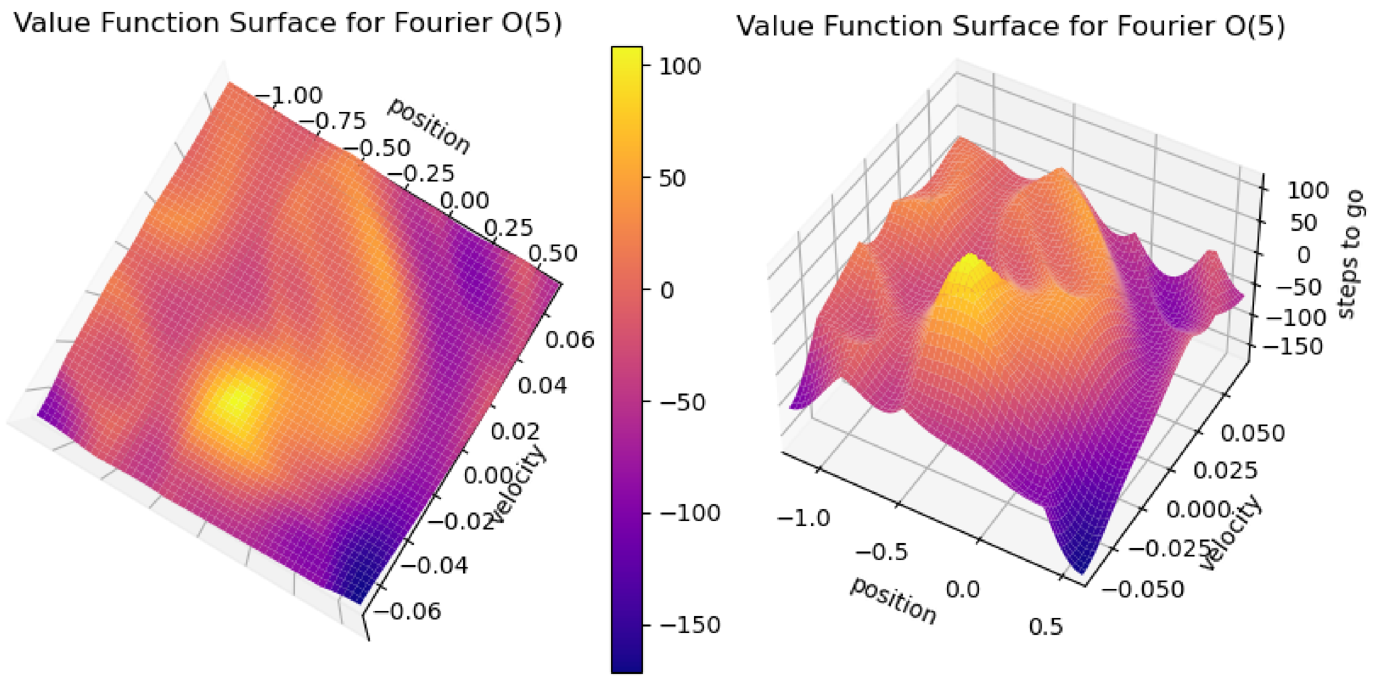




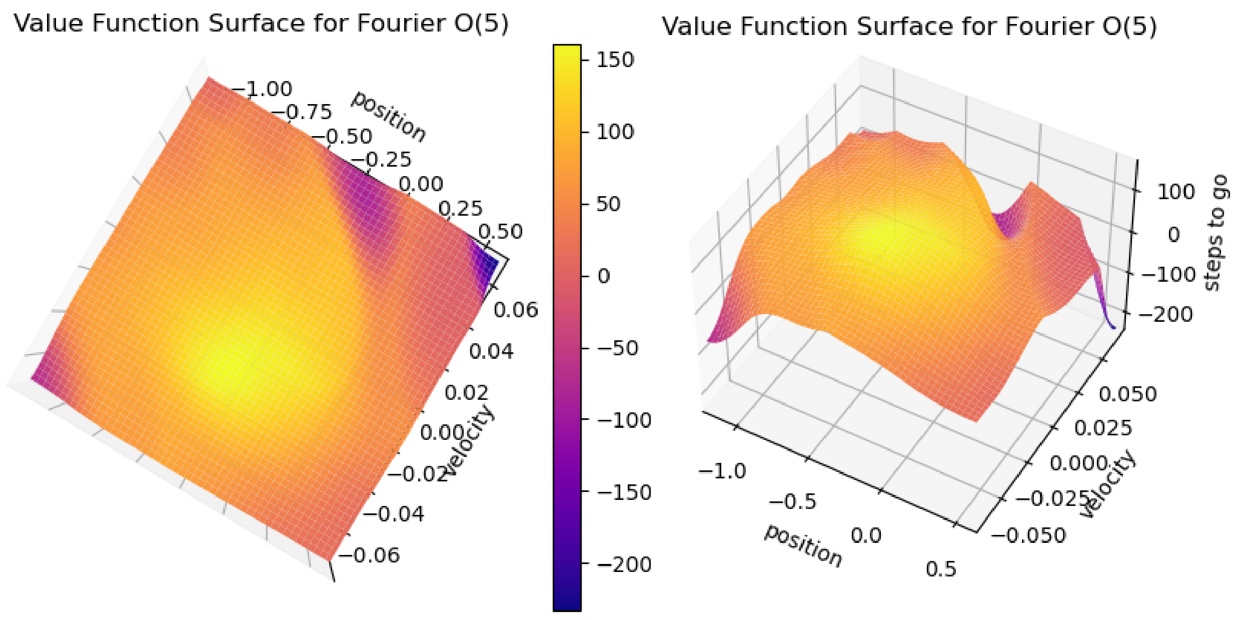
Order 5 Basis

With an order 5 Fourier basis, trial results started to become inconsistent. While some features of the value function remained consistent, there was variety in the surface maps created. Rather defined features would stand out in some runs, but then look like a giant blob with minimal features for others. When looking at the learning curve differences, the surfaces with more defined features appeared to have found a more optimal policy earlier on which helped shape the overall map from then on. The blob maps never really settled and appeared to continue to try rather similar policies. An example image is included below.





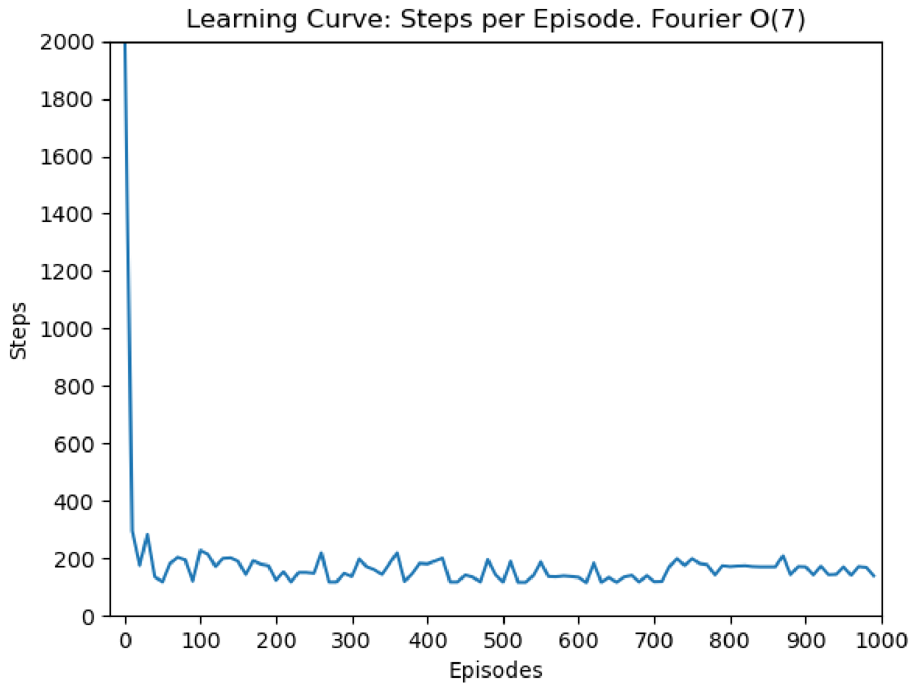
Order 5 Value Function with Definition

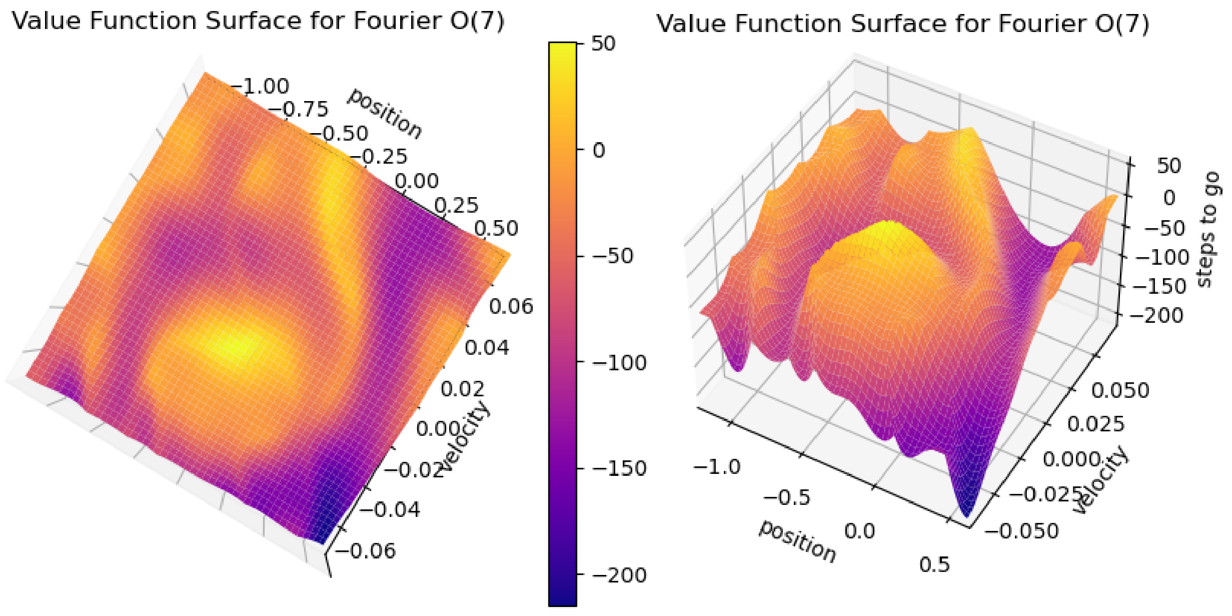


Extreme example of an Order 5 Value Function with minimal Definition (Blob)

Order 7 Basis

The higher the order became, the more features started to stand out. Compared to the order 3 basis surface, the order 7 basis multiple distinct peaks and ridges along the value function surface. Some of the shapes that started to appear in the order 5 surface now have even more prominence. The order 7 learning curves did have more motion to them though as more features were checked for importance. Like the previous basis though, some surfaces would occasionally vary from the norm, with either blob like shapes or rarely a sharp peak in an odd spot.





Answers to #3

Gamma (γ) is used as a trace decay variable. When gamma = 1, traces decay only based on the lambda (λ) future discount value. With a gamma of less than 1, the traces start to decay at a faster rate. The result of this quicker decay is a shorter “chain” of active traces being updated as the episode continues. If the optimal solution to a problem requires a large number of steps, less active traces will hamper the program’s ability to learn the optimal policy. The convergence time will grow.

In mountain car, each step has a negative reward, except for the goal which offers a reward of zero. If this were flipped where each step was zero and then the goal offered a value, the program would not be able to compare two different solutions to the problem. Let us say the goal offered a reward of +1. The best possible solution, π^*, will have a total reward of +1. Now a very suboptimal solution that takes 100x longer than the optimal policy will still offer a final total reward of +1 after reaching the goal. It will also take a *long* time for the program to converge on any one policy, optimal or not, even with gamma = 1. Now with gamma < 1, the episode count required to converge will sky rocket as it will take even longer for the path to the goal to propagate back to the starting state.