

The Monte Carlo method is similar to the on-policy Monte-Carlo control, TD learning method as defined by n= T. It is not much difference on the pseudocode. Therefore, it is not reasonable to give pseudocode.

The performance of Monte Carlo method for the task of Mountain Car could be costly as it would not start to learn until the first episode is completed. And also, for each step that does not reach the goal state would receive a negative reward, which make the cost higher and longer to converge for learning compared to TD or SARSA method.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text, letter

Description automatically generated

Text

Description automatically generated

I design to use different weight vectors for each aggregate states. The weight vector has the same length with aggregations. By using one-hot coding method, the gradient of weight is correlated to the specific feature that needs to be used.

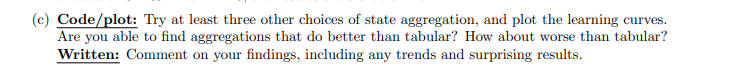
Text

Description automatically generated

Chart

Description automatically generated

Figure 1. Tabular Equivalence



Choice 1: row aggregation method by wrapping each row together

Choice 2: room aggregation method by wrapping into four different rooms

Choice 3: 3x3 aggregation method by wrapping the grid world into small groups of size 3 by 3

Choice 4: 2x2 aggregation method by wrapping the grid world into small groups of size 2 by 2

Chart, line chart

Description automatically generatedChart

Description automatically generated

Chart

Description automatically generatedChart, histogram

Description automatically generated

The smaller wrapping fromat leads to the graph more similar to the tabular one. Larger aggregation decrease earlier than smaller ones. The result is not surprising since less quantities will update more efficient than more quanties(tabular one). Therefore, the tabular converges faster than state aggregation wrapping. But the running speed is slower.

Text

Description automatically generated with medium confidence

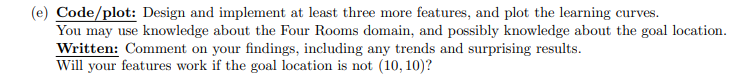
The constant is necessary as it work as the bias term in the normal regression method. Without a bias term, the function would always across (0, 0) no matter what update that made. It would cause large error.

Chart, line chart

Description automatically generated

Figure 6. Feature x, y, 1

Compared to the result of state aggregation, the feature 1, x, y is not as good as the state aggregation one. It is close to a line feature and the learning result is bad and a huge difference. It is because that the number of features is not enough to estimate.



Choice e\_1: room features

Choice e\_2: room number features

Choice e\_3: goal distance features (1, x, y, xy)

Chart, histogram

Description automatically generated Chart, line chart

Description automatically generated

Chart, histogram

Description automatically generated

Both the feature of room and goal distance is showed with a sharp decrease in the steps, but the feature of room number is closed to a linear format. It is surprising to find the feature with room number is not perform as expected.

The feature will still be working if the goal location is not (10, 10) but the performance might be varied based on the different goal states.

Text

Description automatically generated

Text

Description automatically generated

Chart, surface chart

Description automatically generated

Figure 10. Mountain Car with Cost to Go Function

Histogram

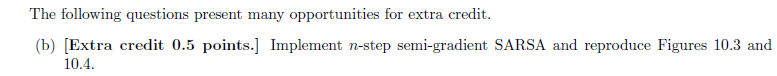
Description automatically generated

Figure 11. Mountain Car learning curves (log)

Chart, line chart

Description automatically generated

Figure 12. Mountain Car learning curves



Chart, histogram

Description automatically generated

Figure 13. Mountain Car Performance of one-step vs 8-step

Chart, line chart

Description automatically generated

Figure 14. Mountain Car Performance Effect of and

Text

Description automatically generated with medium confidence



By replacing , the TD error will be minimized from the objective function as followed:

It is not a good idea to involve expectation over S’ as the loss function depends on next state. This will lead to a bias that changes the converge state and value.

Text

Description automatically generated



In the approaching method above, the problem is that the formulation involves production of two expectations. Both independent samples of the next state are required to get unbiased sample. But for it is not promised that two examples are equal except they are independent to each other. In the real world with some normal interaction circumstances, only one expectation is needed. Therefore, this algorithm is biased and will be converged on a wrong value. For circumstances like deterministic environment, where the transition to next state is determined, the two samples will be the same and therefore the algorithms should work properly. Or in simulated environment that independent sample obtain to work for different expectations, then the algorithms could converge to the minimize value.