**CS282r – Reinforcement Learning in HealthCare**

**Checkpoint 1**

Donghun Lee, Srivatsan Srinivasan, Linying Zhang

* **Problem Statement**

A challenge in optimizing treatment for sepsis using reinforcement learning is the delay between actions and resulting rewards. By introducing intermediate reward (IR) in patient trajectory, we can better utilize the intermediate transitions in the data to potentially derive a better policy, a policy that can take the right action earlier during a patient’s ICU stay to save patients’ life. Our goal in this project is to find the best choice of intermediate reward that allow us to obtain higher predicted state value.

* **Outline of the approach**

1) proof-of-concept

To first check out the assumption of intermediate reward on policy improvement, we will try some simple form of intermediate rewards, such as

1) patient-uniform intermediate reward

*intermediate reward for a given patient = the end-point reward / the number of time stamps*

2) discounted intermediate reward

Another way is to have a discounted intermediate reward back-tracking from the end stage for each patient.

*intermediate reward t-steps away from the last timestamp= the end-point reward \**

Either way, we expect to see an increase of some state value functions compared to the optimal policy from using terminal reward.

2) Intermediate reward based on SOFA score

Among all the observational features, we found a clear positive association between SOFA score and mortality, which could be the starting point of associating intermediate reward with state. We can use the mortality vs SOFA plot from homework 2, and perform a linear transformation of mortality to intermediate reward, for example, 0% mortality is +10 reward, 50% mortality is 0 reward, and 100% mortality is -10 reward. Other scaling from mortality to intermediate reward can also be experimented here.

3) Intermediate reward based on factors strongly associated with mortality

Once 2) established, we can move onto including other features strongly associated with mortality besides SOFA to fine-tune our intermediate reward. We can perform any state-of-art supervised learning method, such as tree-based method or SVM to find the top significant features, and using the same mortality-IR transformation we use in 2) for this part.

4) Intermediate reward from 3) with temporal discount factor

By far, we should establish a good mapping of mortality-related features to intermediate reward. However, this mapping does not take into account the temporal character of states, namely when does the state appear during a patient’s ICU stay. We would like physicians to make a right decision at the earliest time possible, so giving a higher weight to state-action pairs that appear early in the patient trajectory may make intuitive sense here. Need to be discussed further.

5) Moving from discrete to continuous state space

If the intermediate reward we establish improve the policy on discrete state space, there is no reason not to try a continuous state space as suggested by Raghu *et al* 2017.

* **Evaluation Methodology**

Since our approach is based on MDP, comparing state value functions under different policy is a good method for comparing the performance of policies. Doubly robust estimator seems to outperform IS based on HW3, so DR can be the estimator for variance assessment. What else?

* **Deliverables for checkpoint 2**

We would like to get 1) and 2) done for next checkpoint and start the feature selection for 3).

* References

1. Raghu *et al*. Continuous State-Space Models for Optimal Sepsis Treatment - a Deep Reinforcement Learning Approach.

2. More about intermediate reward in RL in general?

My questions:

1. would intermediate reward and state have strong correlation if both are a representation of partially same features? In other words, if the factors we select for intermediate reward are the most significant in predicting mortality, wouldn’t those also be significant in defining state?