The objective of this homework assignment is for you to get a chance to implement batch RL on the sepsis management task. You may work with others on this assignment, but your code and write-up must be your own. Include the names of collaborators at the top of your write-up. Include your code at the end of the assignment. You should submit your assignment on Canvas. This assignment may take you some time. Start early!

Problem 1: Off-Policy Evaluation. Implement the importance-sampling and doubly-robust policy evaluation schemes described in https://arxiv.org/abs/1604.00923. Test them out on your simple gridworld from Homework 1. Discuss your results for the “test maze” and “cliffworld” grids, but you may test other grids to gain intuition for how the estimators perform. As behavior policies, test several values of $\epsilon$ for $\epsilon$-greedy policies. Include $\epsilon = 0$ and $\epsilon = 1$, as well as two intermediate values for $\epsilon$. Set the action-error probability to be 0.2 (Bonus: test how your results depend on the action-error probability). Plot how the variances of the estimators change with the amount of data in the batch for evaluating a uniformly random policy. Next perform the same analysis for the optimal policy.

Problem 2: Basic Sepsis MDP. Use an off-the-shelf clustering algorithm to cluster the observations in the clinical data. Include ventilation, sedation, and RRT as part of the observation space, and only include intravenous fluids and vasopressors at actions. Discretize each action dimension into 5 bins, resulting in a total of 25 possible actions. Next, treating each observation cluster as a state, and use the states to learn an MDP. Solve the MDP to derive a policy.

For policy evaluation, plot the state value functions of different policies under the MDP. Compare the value functions of (a) the clinician policy (that is, the actions that were taken by the clinicians), (b) a policy that chooses among the 25 actions uniformly at random, (c) a policy that always chooses the “no intervention” action (the most common action), and (d) the optimal policy with respect to the MDP. How do they compare? Are your results sensitive to your choice of model parameters (for example choice of rewards and $\gamma$) and model estimation methods (for example how you estimated the transition probability of a state-action pair you observed only a few times or did not observe at all)?

In addition, plot the mortality rate of patients vs. deviation from the dose recommended by the policy (see Figure 4 in https://arxiv.org/pdf/1705.08422.pdf). Bonus: Try evaluating policies (a-d) using the off-policy doubly-robust method. Normally, this method is considered more reliable than the model-based method we asked you to use in this section. However, depending on your implementation method, your results may have high variability. Discuss your results and any problems you can identify when applying the doubly-robust estimator to the data.

Problem 3: Basic Sepsis Kernel-based Policy. The MDP approach above assumes that the patient’s history is Markov in their current set of observations, which may not be
true. Explore a simple way of direct policy learning via a kernel: First, create a function defining the similarity between two patient’s histories. Second, for any patient at any time, choose the action that had the best survival rate for similar patients. Plot the performance of this policy to those above, and discuss how it compares.

Finally, describe the choices you had to make—how did you define patient similarity? How did you choose which patients were similar to your current patient? How did you compute survival rates? We realize that there are a lot of choices to be made here. For this assignment, make reasonable kernel choices and do some parameter exploration; you will be evaluated primarily based on a successful implementation of the core idea and your discussion. Exploring how to optimize the kernel is an option for the semester project.