Notes for Assignmnet -1(RL) – RL1

1)DAGGER:

Goal : To mimic policy from expert policy using supervised learning.

Procedure :

a) Trajectories are sampled using current policy and stored in rollouts variable as in Batch size. Here trajectory means sequence of (observation,action,new_state,reward) until some finite steps.
Also here our current policy is a Neural network which outputs mean of action space given observation and also variance (a parameter) will also be trained. (because of continuous action space.)

b) Now the actions are relabelled using expert policy given the observation from the rollouts.

c) Now these batches will be fed into our policy network to get loss, based on the loss the gradients will be calculated and weights will be updated of our training policy network.

1) How actions will be sampled using our policy?
Given the mean and var of the action space from our Network, we can build a Normal distribution out of it and sample a random action.(for better exploration.)

2) How is loss computed for our policy network?
Given an expert relabelled (obs,action) pair . Then loss = - logprob(Normal_dist(action))
Where Normal_dist parameters are guessed by our network.

Here the intuition is the mean and variance should be adjusted such that higher probability will be assigned to labeled action from expert.

Evaluation metrics:

a) Training batch Total reward.
b) Eval batch Total reward.

Here Total reward means sum of all rewards in all trajectories.