We had a discrete, finite set of 4 actions where our observations were 8 dimensional continuous objects which required scaled-up, parametrized RL methods.Firstly we tried naïve policy gradient based optimization to update solely the actor.Here, the simple discounted future reward term was used in the unbiased gradient estimator which as expected, resulted in a lot of variance induced oscillations in the gradient steps and thus a poor convergence.Then, to adjust policy updates such that we correctly take into account only the actions which lead to high returns downstream and increase their probability, we replaced the discounted future returns of the trajectory with the rewards to go.Since this is merely obtained via subtracting a baseline depending on the previous steps of the episode, it didn’t disturb the unbiasedness and resulted in much more coherent gradient updates.Lastly, we used an actor critic method (GAE) where to policy gradient was estimated alternatively using an advantage function estimate which was estimated via discounted residuals of value function estimates with parameter gamma pre-tuned to find a sweetspot for the bias-variance tradeoff of the advantage estimator.Here, the critic network was updated using many iterations of TD updates via SGD using the past transitions we stored in the replay buffers.The actor network was updated using this refined policy gradient estimate for SGD which was once more approximated by MC sampling of trajectories at each epoch.