CS294 Deep RL Assignment 5: Exploration Strategies

Mohamed Khodeir

February 9, 2019

Problem 1

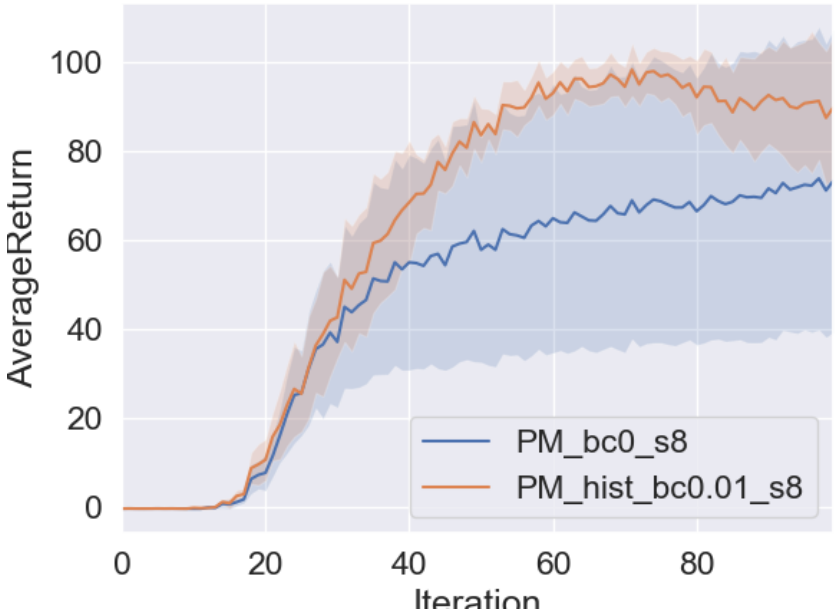

Figure 1: Comparing an agent with histogrambased exploration and an agent with no exploration.

Problem 2

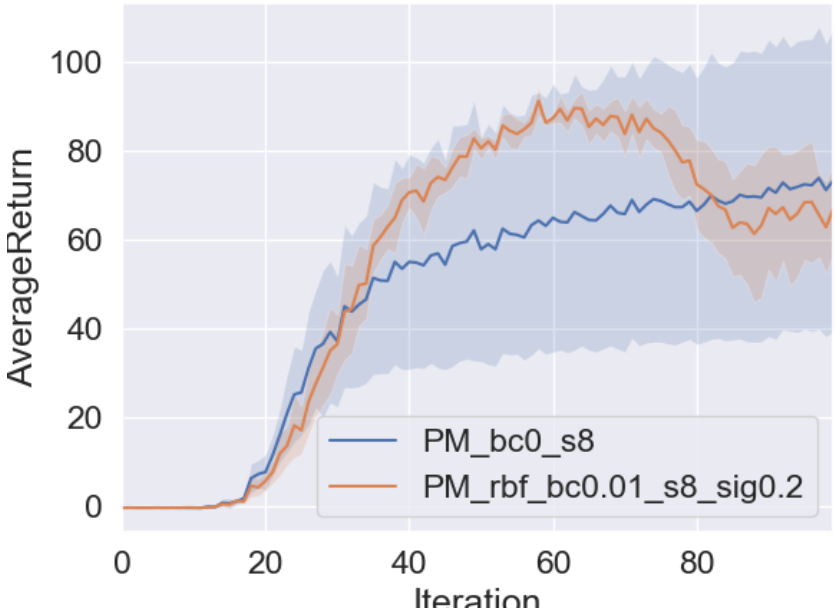

Figure 2: Comparing an agent with KDEbased exploration and an agent with no exploration.

Problem 3

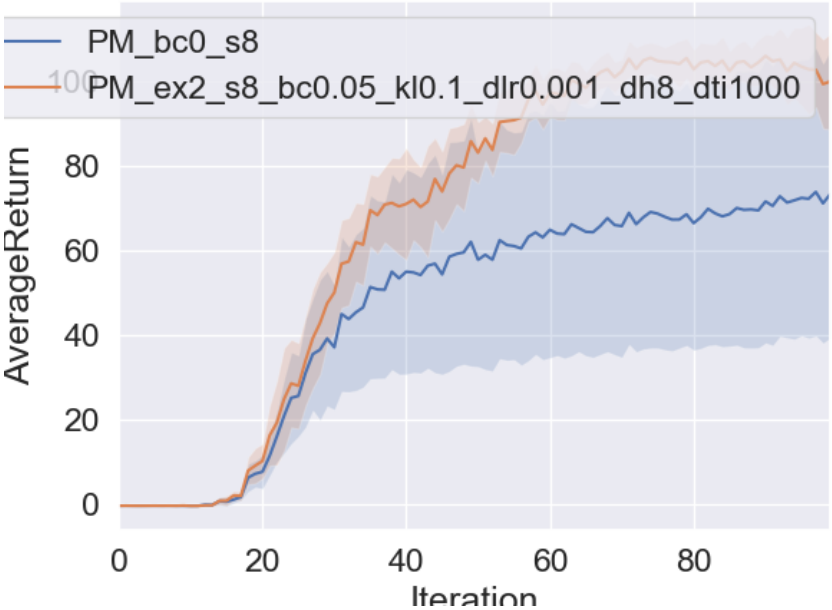

Figure 3: Comparing an agent with EX2-based exploration and an agent with no exploration.

Problem 4

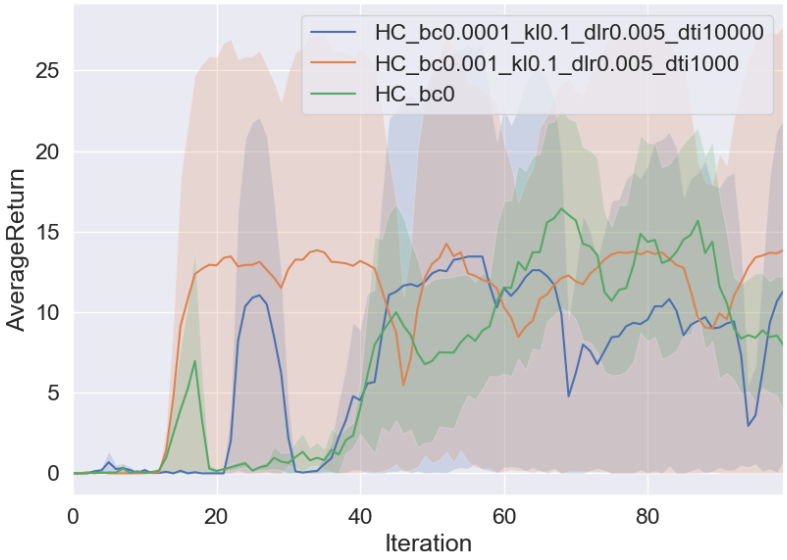

Figure 4: Comparing the agents with EX2-based exploration and an agent with no exploration.

In this experiment, we varied the bonus coefficient as well as the number of iterations on which the descriminator was trained. (1) With regard to the shape of the curves, one obvious thing to note is that the curve corresponding to the higher bonus coefficient seems to display significantly faster improvements in AverageReturn. However, both display faster improvement than the run without an exploration bonus. This could be attributed to faster exploration early on (as higher weight is given to novel states). This also explains to some extent, the higher variance (evident on the error bands) as more exploration is done vs exploitation of known good states. The difference in variance between the two ex2 runs, might also be explained to some extent by the difference in the number of iterations for which the discrimnator was trained. Intuitively, perhaps the discrimnator, having not been trained to convergence, displays higher variance in its state novelty estimation. (2) There is not a significant difference between the peak performance of the runs. Surprisingly, the run with a higher bonus coefficient seems to be more stable. I would have expected the opposite as more weight is given to exploration of unknown states should cause more fluctuations in average return.