Homework 1 Template

Use this template to record your answers for Homework 1. Add your answers using LaTeXand then save your document as a PDF to upload to Gradescope. You are required to use this template to submit your answers. **You should not alter this template in any way** other than to insert your solutions. You must submit all 15 pages of this template to Gradescope. Do not remove the instructions page(s). Altering this template or including your solutions outside of the provided boxes can result in your assignment being graded incorrectly.

You should also export your code as a .py file and upload it to the **separate** Gradescope coding assignment. Remember to mark all teammates on **both** assignment uploads through Gradescope.

Instructions for Specific Problem Types

On this homework, you must fill in blanks for each problem. Please make sure your final answer is fully included in the given space. **Do not change the size of the box provided.** For short answer questions you should **not** include your work in your solution. Only provide an explanation or proof if specifically asked.

Fill in the blank: What is the course number?

10-703

Problem 0: Collaborators

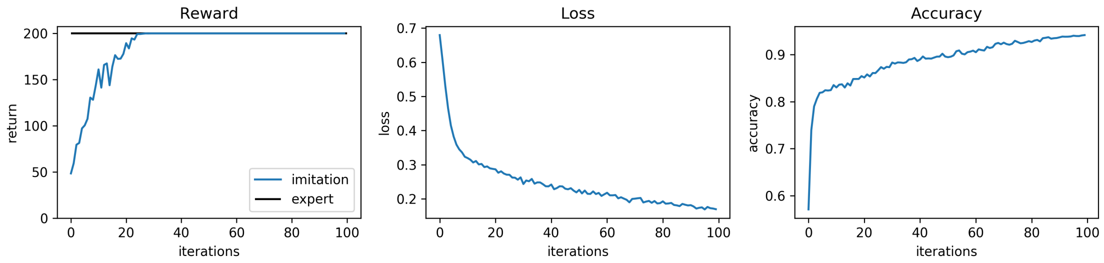
Enter your team members' names and Andrew IDs in the boxes below. If you worked in a team with fewer than three people, leave the extra boxes blank.

Name 1:	Eu Jing Chua	Andrew ID 1:	eujingc
Name 2:		Andrew ID 2:	
Name 3:		Andrew ID 3:	

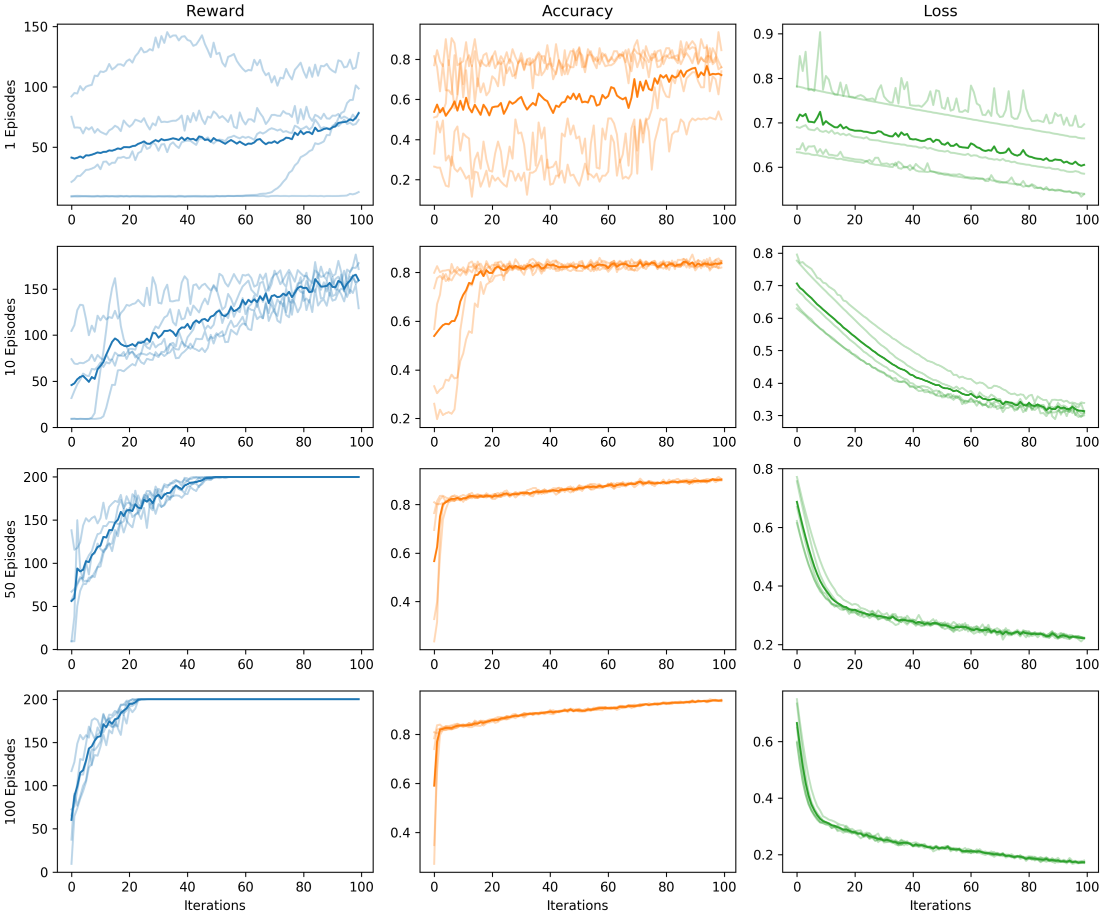
Problem 1: Behavior Cloning and DAGGER (50 pt)

1.1 Behavior Cloning (25 pt)

1.1.1 Plot Behavior Cloning (15 pt)

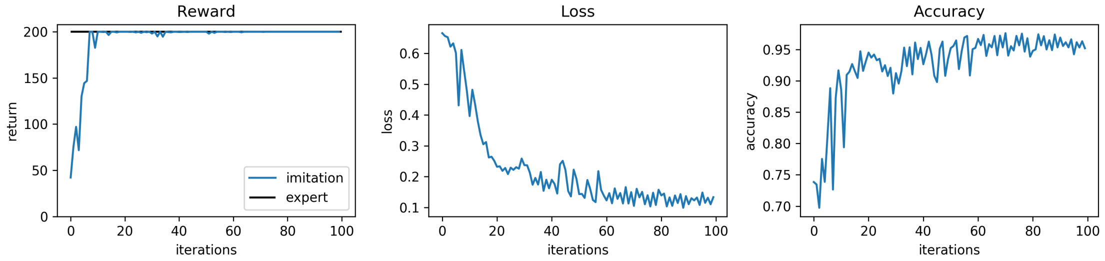


1.1.2 Plot Behavior Cloning with Varying Expert Episodes (10 pt)

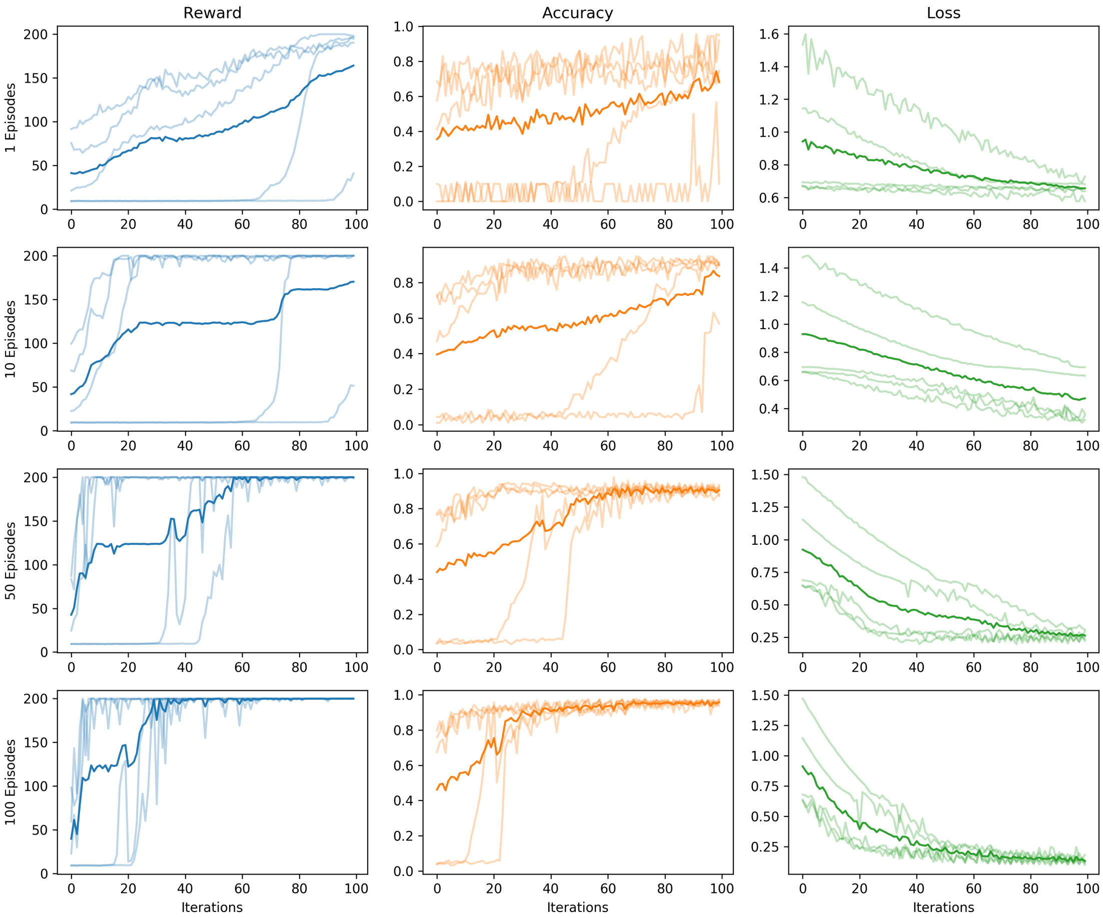


1.2 DAGGER (25 pt)

1.2.1 Plot DAGGER (10 pt)



1.2.2 Plot DAGGER with Varying Expert Episodes (10 pt)



1.2.3 Compare Behavior Cloning and DAGGER (5 pt)

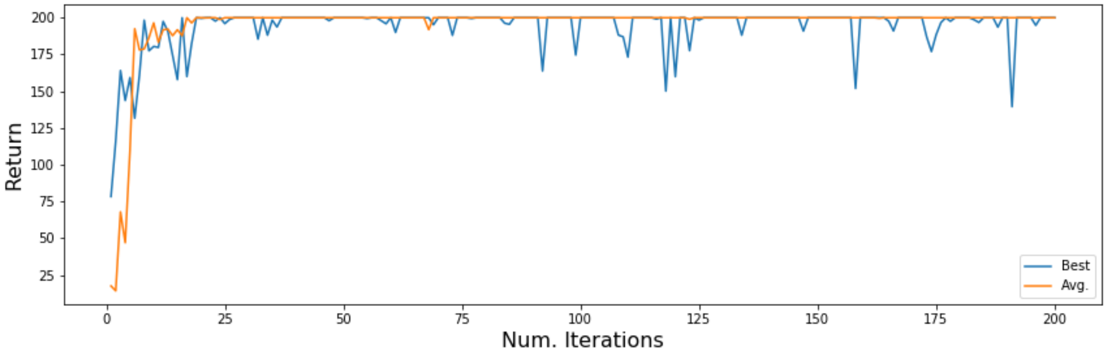
My DAGGER implementation does not outperform my behaviour cloning implementation as it has high variance in reaching convergence and usually at a later iteration than Behavior Cloning, across multiple runs with different seeds.

My hypothesis is that the state and action space of this experiment is small enough, and the dynamics of the system simple enough (with little stochasticity) such that there is not much exploration to be done and simply imitating an expert covers most of the situations that will be faced. The exploratory nature of DAGGER thus introduces higher variance in the results as seen.

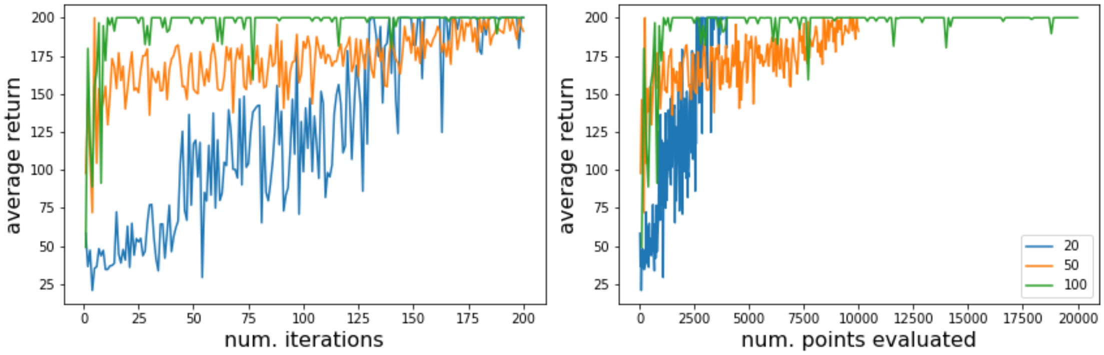
We can test this by testing on another environment with a larger state space where just following an expert would only cover a tiny proportion of the entire state space, such as a game of Tetris.

Problem 2: CMA-ES (25 pts)

2.1 Plot CMA-ES (15 pts)

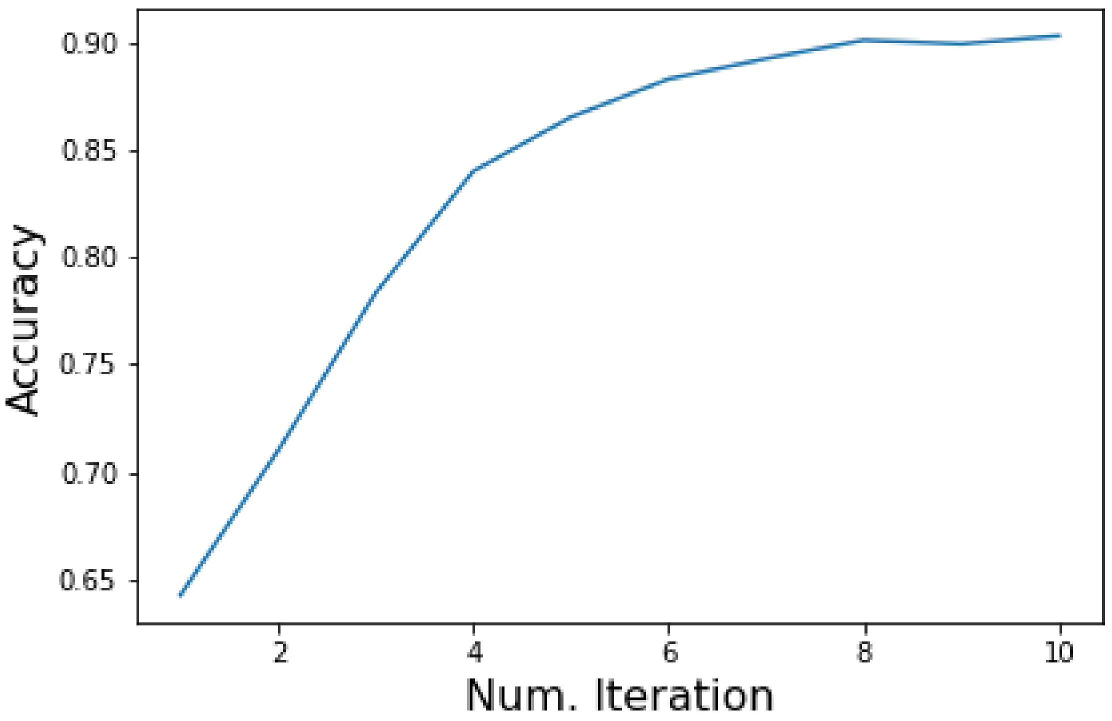


2.2 Plot CMA-ES with Varying Populations (10 pts)

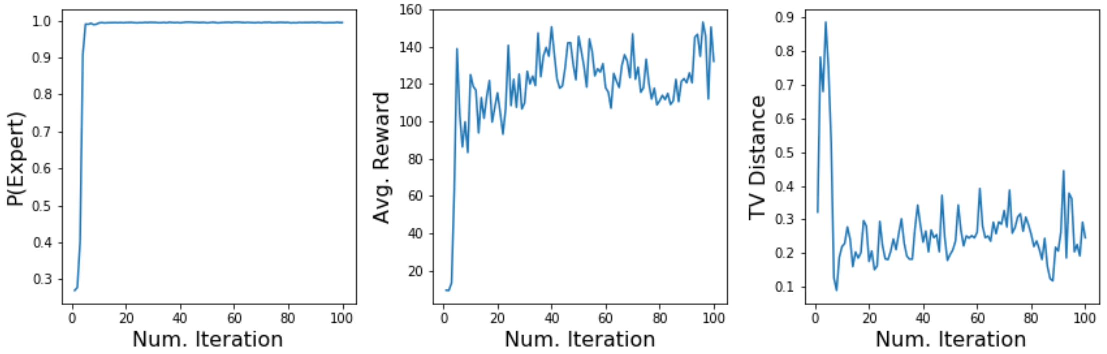


Problem 3: GAIL (25 pts)

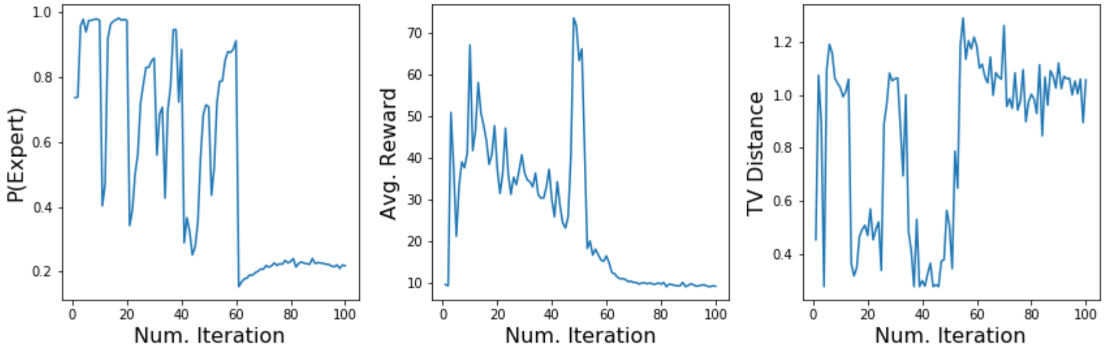
3.1 Plot Training Accuracy (5 pts)



3.2 Plot CMA-ES Task Reward and TV Distance (5 pts)



3.3 Plot GAIL Task Reward and TV Distance (5 pts)



It is interesting to see the pattern in the P(Expert) graph, where the probability increases until the iteration is a multiple of 10, where the discriminator updates. Right after, the probability drops sharply and the cycle continues.

This makes sense according to the GAIL algorithm, as the generator is trying to maximize its task reward, the probability, and does so across each set of 10 iterations, while the discriminator is trying to decrease it and does so every 10^{th} iteration.

However, we see the actual environmental reward decreasing over time, only temporarily sharply increasing around iteration 50 and then back down to around 10.

This is probably due to the nature of the task reward used in GAIL, which is the log probability which lies from $(-\infty, 0]$. The problem is that an episode length is variable, with shorter lengths corresponding to bad environmental reward and longer lengths corresponding to good environmental reward.

By summing up the task rewards in an episode, we actually see roughly two optimums:

1) A long sequence of high probability states or

2) A short sequence of lower probability states

My hypothesis here is that the algorithm has converged to the second optimum in this case, where the generator has learnt to fail as quickly as possible to minimize an episode's length.

3.4 Vary Frequency(5 pts)

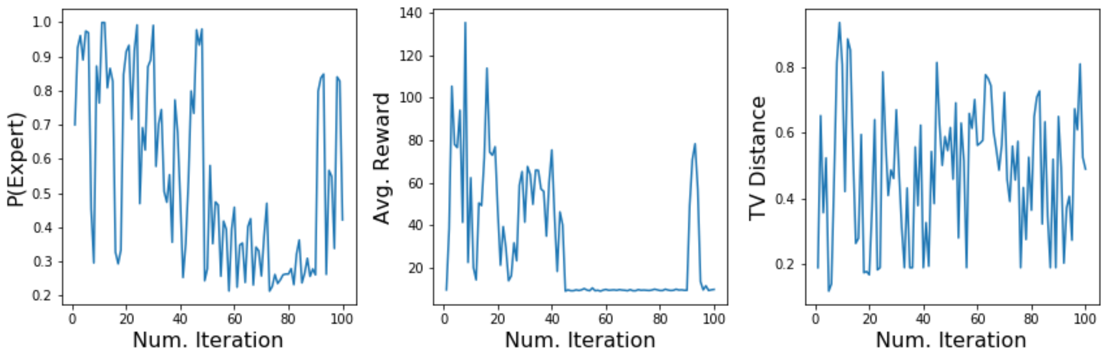
Describe your findings (3-5 sentences):

We test if the discriminator is not adapting fast enough with respect to the generator, so we try interleaving more discriminator updates (every 3 steps). This resulted in slightly better results as we do see more occurrences of high rewards above 100 now. Although the TV distance still varies very highly, its is still roughly lower than that of before, so we can conclude that for this discriminator a higher frequency of update was required for the TV distance to go down.

However, there is still the problem of the generator tending to learn to favor generating short sequences by failing fast, rather than long sequences that do not fail. As the score of each set of parameters in CMA-ES is actually the sum of log probabilities, which are individually negative, a short sequence of low probabilities might actually have a higher score than a long sequence of higher probabilities. Thus CMA-ES might tend to favor parameters that generate these short sequences that fail fast but actually have low task reward.

Perhaps a different reward function, such as $\log(P(\text{Expert}) + 1)$, which ranges from $[0, \infty)$ will push the optimizer towards the real optimum that corresponds to higher environmental reward.

Plot(s):



3.5 Overall Findings (5pts)

Behavior cloning actually performed the best out of all three algorithms.

I expect behavior cloning to work the best on problems with small state spaces where little exploration is required to generalize behavior, while DAGGER would work better in situations where more exploration is required, with larger state spaces.

However in the case where we have a very high-dimensional state space but there is some true causal structure in a lower manifold, using an algorithm like GAIL can reduce causal confusion when compared to the previous two algorithms.

Extra (2pt)

Feedback (1pt): You can help the course staff improve the course by providing feedback. You will receive a point if you provide actionable feedback. What was the most confusing part of this homework, and what would have made it less confusing?

The vagueness of the terms environmental reward and task reward.

Time Spent (1pt): How many hours did you spend working on this assignment? Your answer will not affect your grade.

Alone	15 Hours
With teammates	
With other classmates	
At office hours	