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

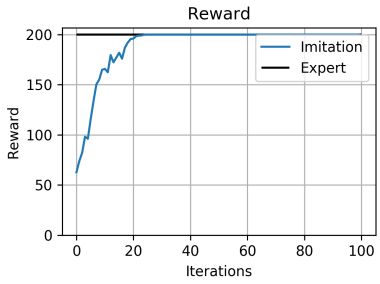
Enter your team members' names and Andrew IDs in the boxes below. If you worked in a team with fewer than three people, leave the extra boxes blank.

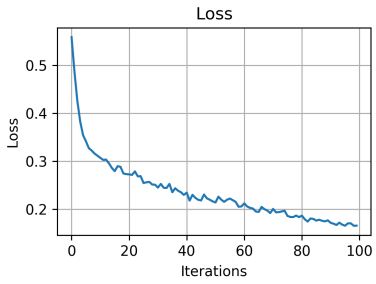
Name 1:	Karmesh Yadav	Andrew ID 1:	karmeshy
Name 2:	Heethesh Vhavle	Andrew ID 2:	hvhavlen
Name 3:	Aaditya Saraiya	Andrew ID 3:	asaraiya

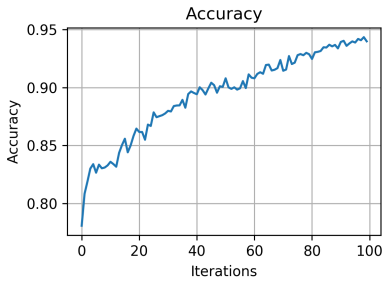
Problem 1: Behavior Cloning and DAGGER (50 pt)

1.1 Behavior Cloning (25 pt)

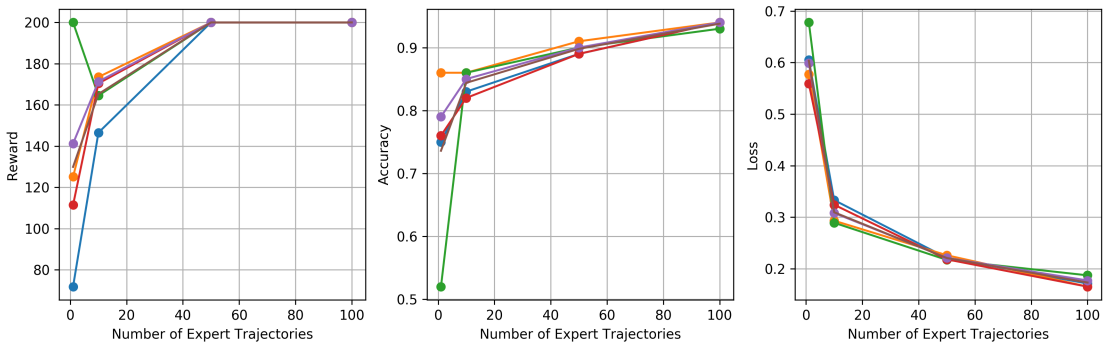
1.1.1 Plot Behavior Cloning (15 pt)



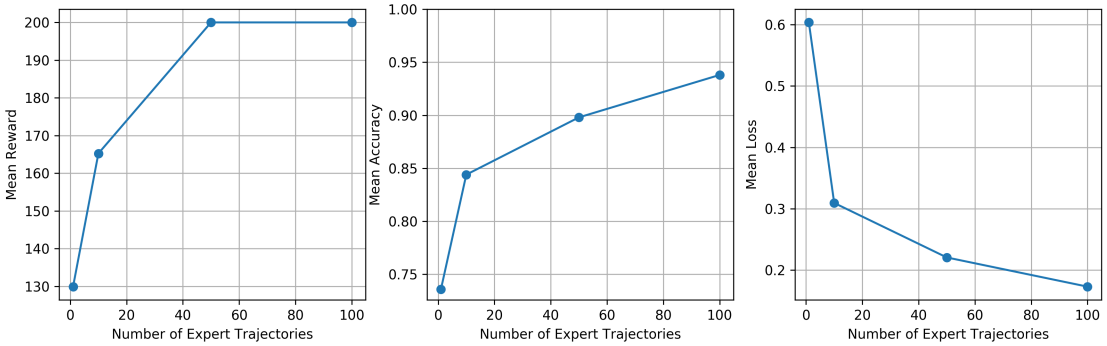




1.1.2 Plot Behavior Cloning with Varying Expert Episodes (10 pt)



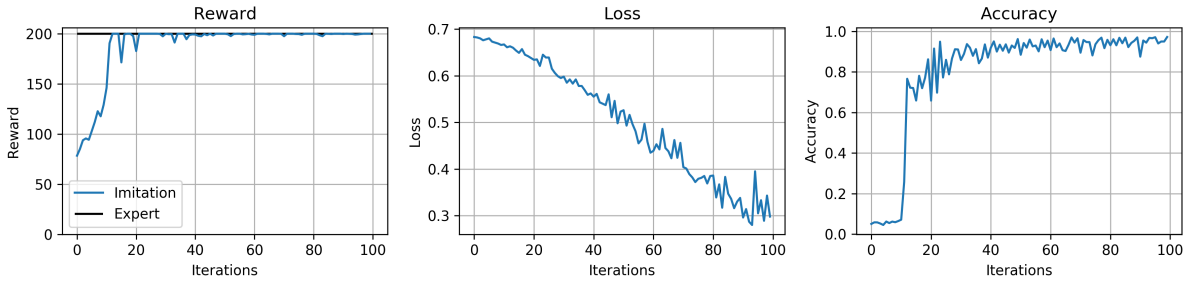
Each of the lines above represent the results of the trained student policy for a given random seed.



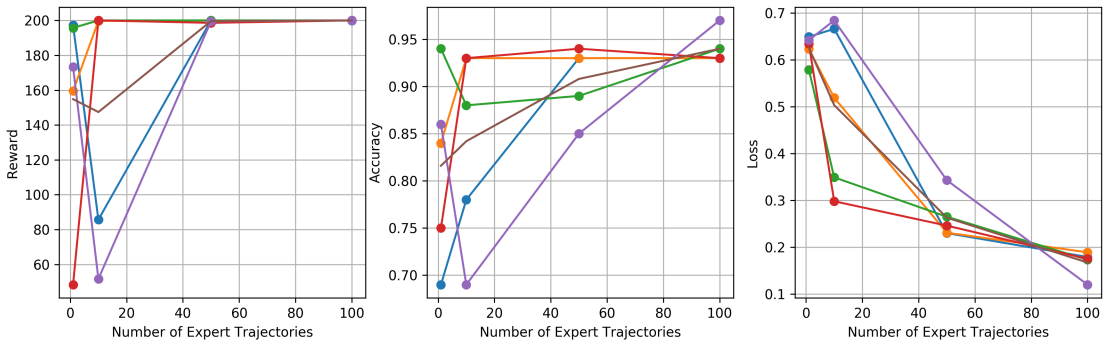
Each point in the above plot represents the mean result of the student policy across 5 different random seeds.

1.2 DAGGER (25 pt)

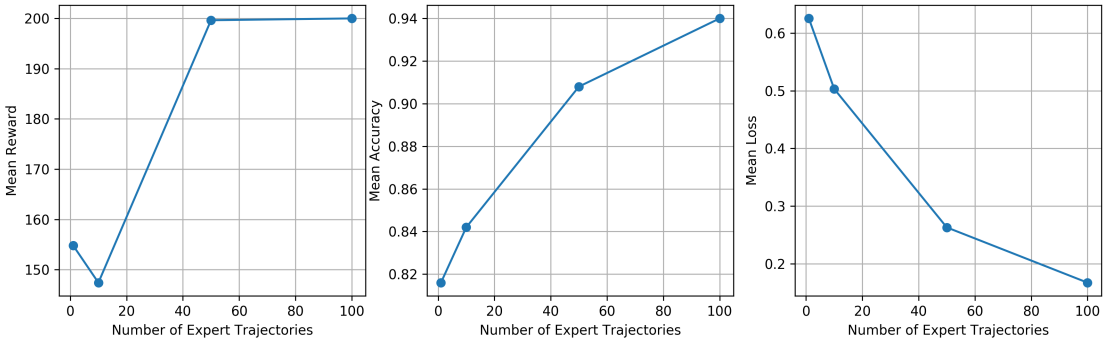
1.2.1 Plot DAGGER (10 pt)



1.2.2 Plot DAGGER with Varying Expert Episodes (10 pt)



Each of the lines above represent the results of the trained student policy for a given random seed.



Each point in the above plot represents the mean result of the student policy across 5 different random seeds.

1.2.3 Compare Behavior Cloning and DAGGER (5 pt)

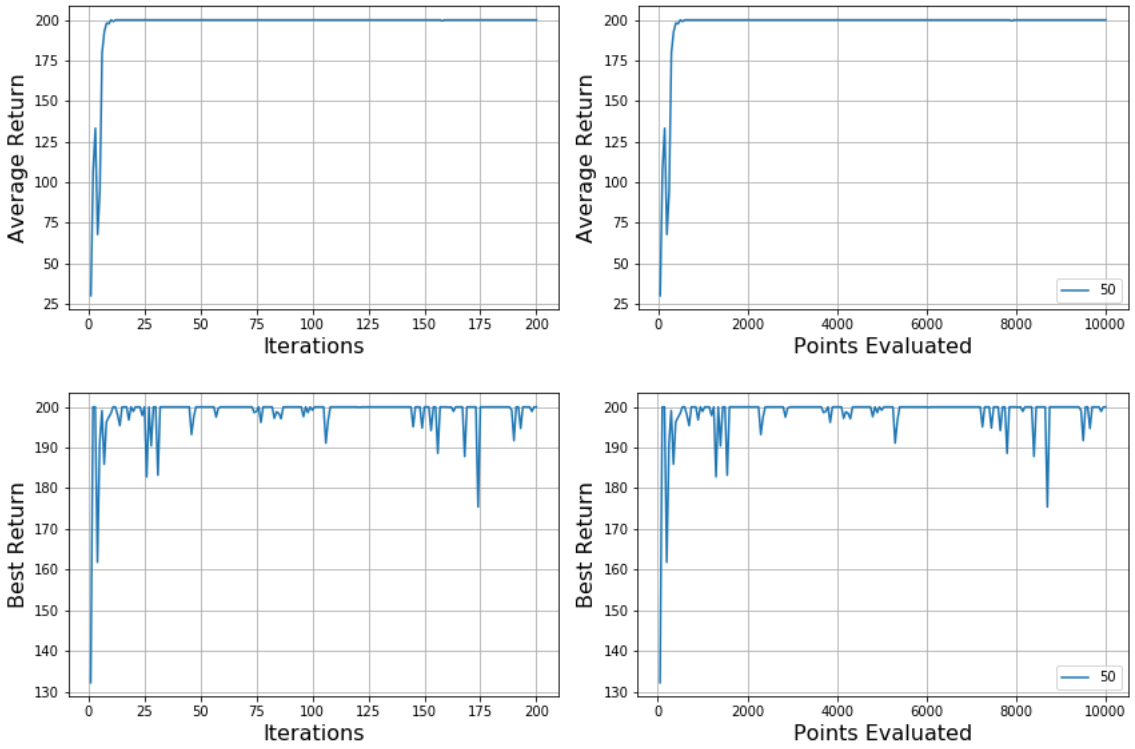
The loss obtained in DAgger is higher than that obtained in behaviour cloning because the samples that the network trains on are much more varied. But since the DAgger network has seen more states, it is able to generalise better in the environment and get a higher reward.

To support this hypothesis, let us assume the example of a robotic manipulator to pick a given object. If we used behaviour cloning, our student policy would have been trained only on ideal state/action pairs shown by the expert. The dataset would not have seen extreme/special states and the policy might not be able to recover or perform in such states. Policy trained with DAgger would work on states not seen by the expert because we would have queried those states later on during training.

Problem 2: CMA-ES (25 pts)

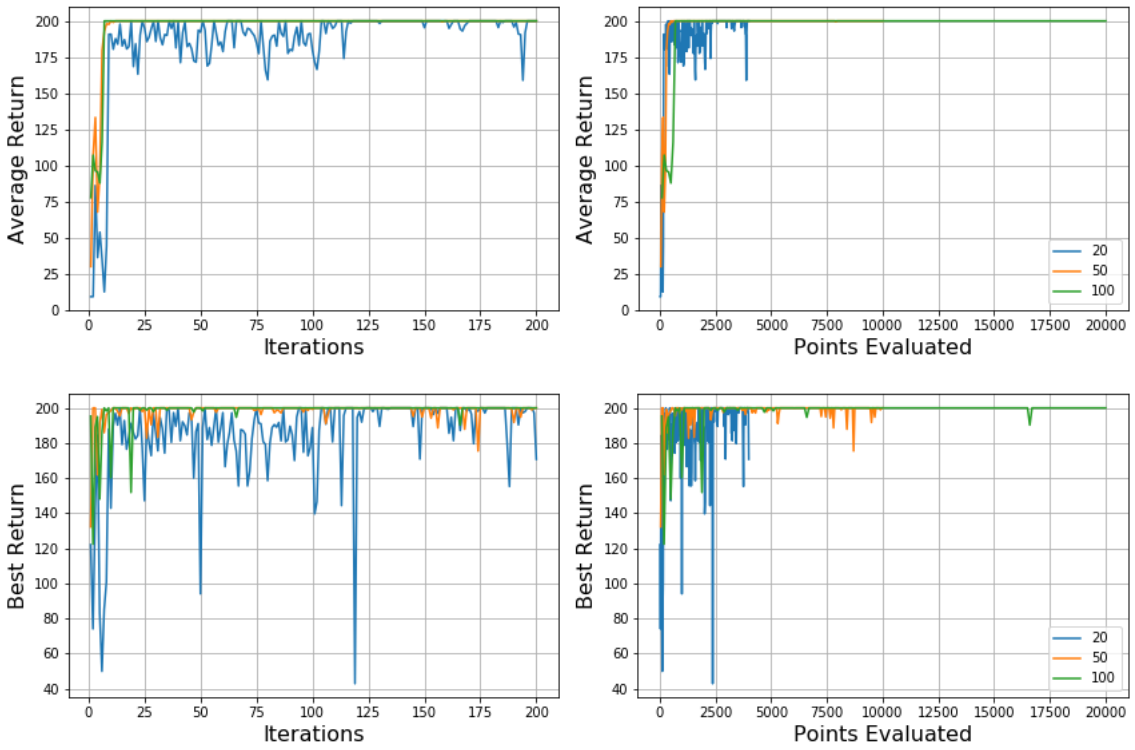
2.1 Plot CMA-ES (15 pts)

The following are the plots showing the average and best rewards over 200 iterations for a population of size 50.



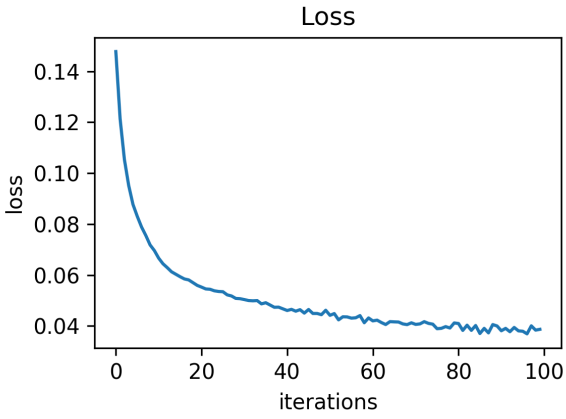
2.2 Plot CMA-ES with Varying Populations (10 pts)

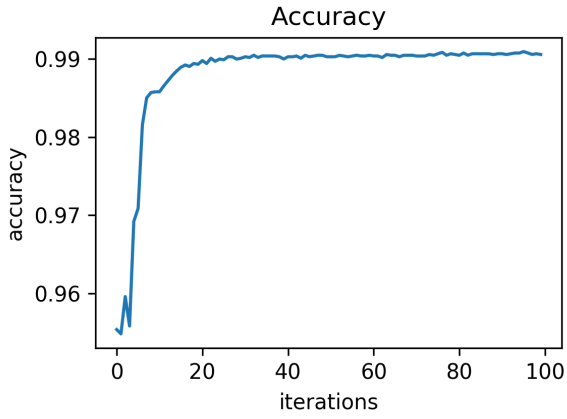
The following are the plots showing the average and best rewards over 200 iterations for population of sizes 20, 50, and 100.



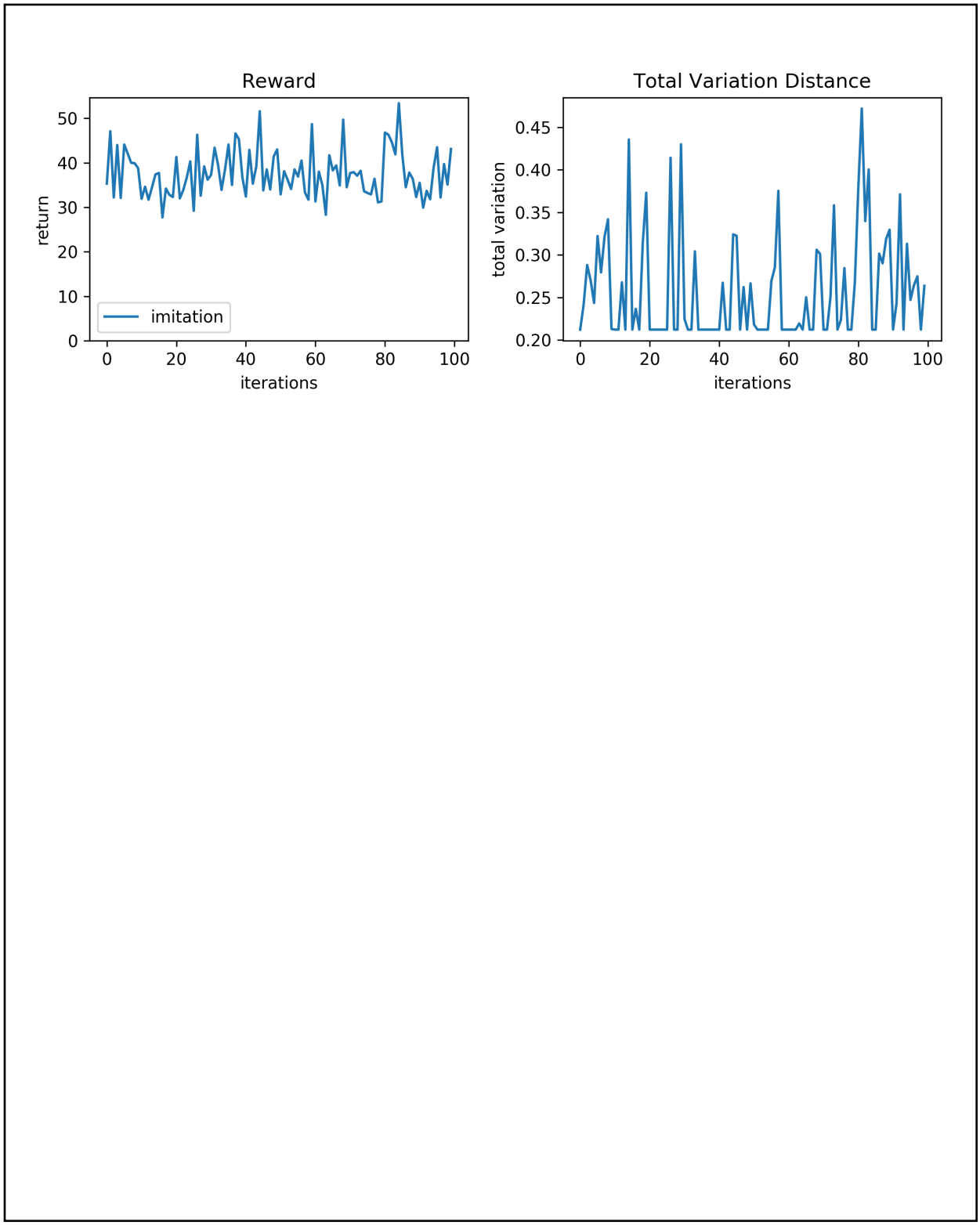
Problem 3: GAIL (25 pts)

3.1 Plot Training Accuracy (5 pts)

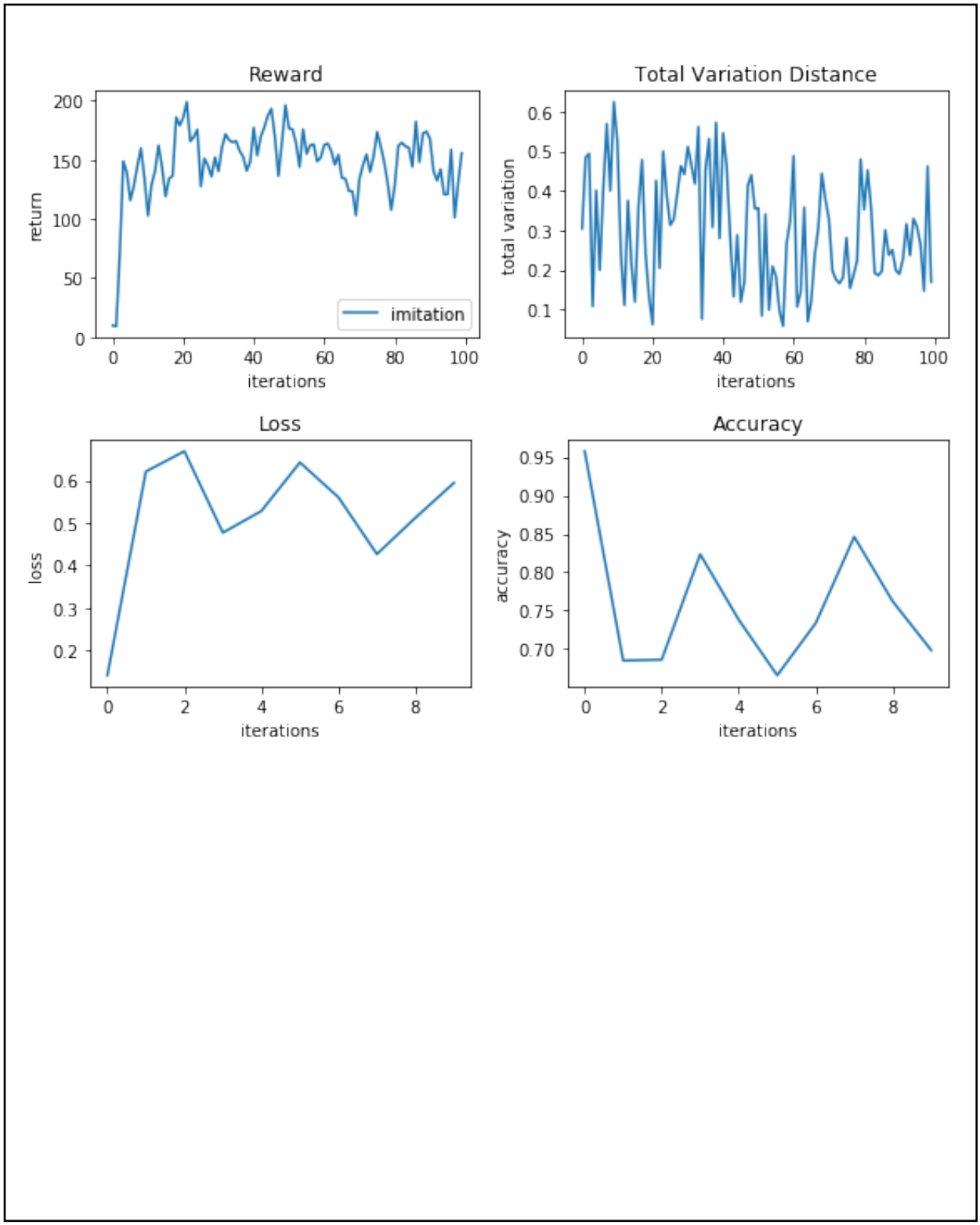




3.2 Plot CMA-ES Task Reward and TV Distance (5 pts)



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



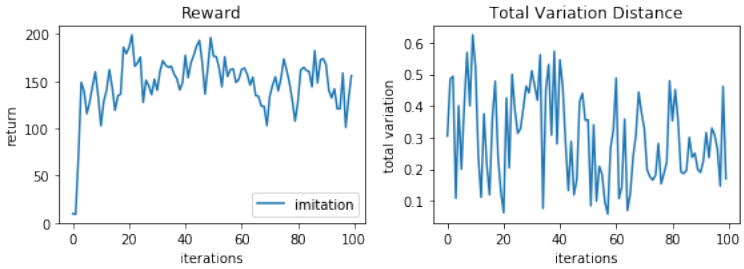
3.4 Vary Frequency(5 pts)

Describe your findings (3-5 sentences):

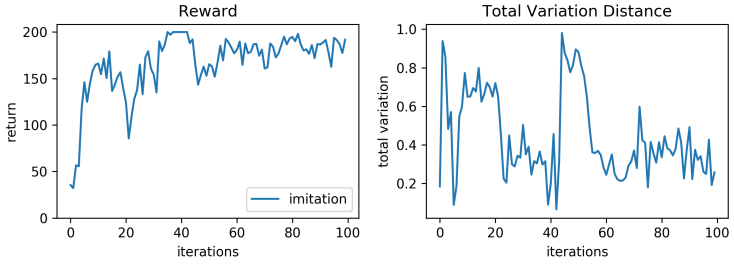
GAIL is overall a little bit difficult to train because of the structure of the problem. The discriminator needs to be trained carefully so that it doesn't overfit. An overfit discriminator will starts giving really low rewards to the student and therefore the student won't learn anything. As we can see in the plots below, when the discriminator update period is 30, this student is able to reach a reward of 200 while it fluctuates a lot with an update frequency of 3.

Plot(s):

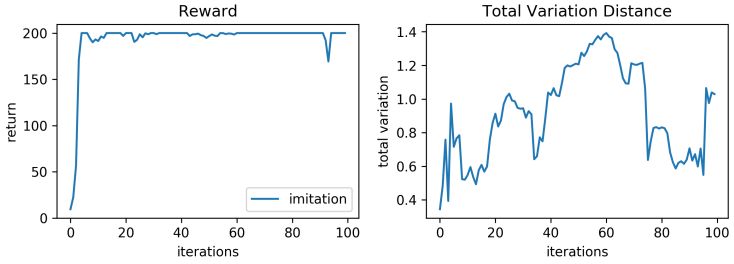
With discriminator_update_period = 3



With discriminator_update_period = 10



With discriminator_update_period = 30



3.5 Overall Findings (5pts)

1) DAgger worked the best out of the three methods. This was because supervised learning is a easier task to solve than RL and if we augment the dataset with states which were not seen in the expert dataset the policy will generalize better.

2) We would use imitation learning when we don't have access to the agent for queries like DAgger.

3) GAIL can be used in situations when we don't have access to even the expert actions. In such cases, GAIL can be used to do occupancy matching of the agent states with the expert states.

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 Cart-Pole problem was a fairly simple problem for the RL algorithms to solve which got very high rewards in just few iterations of training. This was not very helpful to make intuitive comparisons and understand the practical aspects of the different RL algorithms covered in this assignment. There were a lot of mistakes in the codes in the beginning. But the FAQ section in Piazza was really helpful during the course of the homework. It was a little difficult to work with Colab notebooks as we lost data and progress a few times initially and when the notebook got disconnected with the server. Version control and collaboration was another issue as well. We would prefer working with independent Python files instead of notebooks in the future.

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

Alone	12
With teammates	2
With other classmates	0
At office hours	0