1 Problem 3

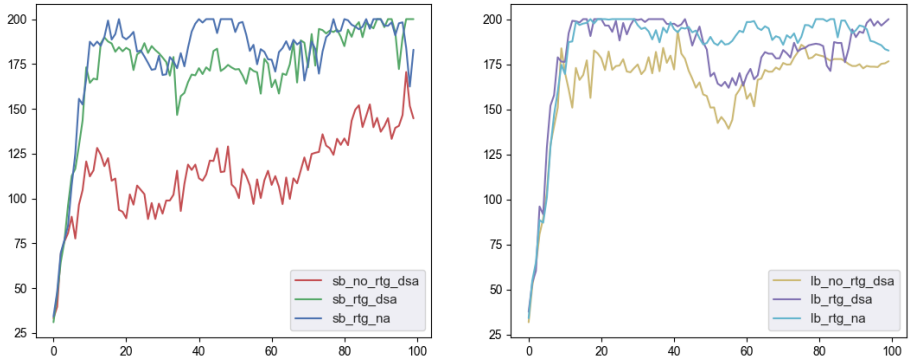


Figure 1: Problem 3. Results are averaged over 5 seeds.

Please see Figure 1.
Questions:

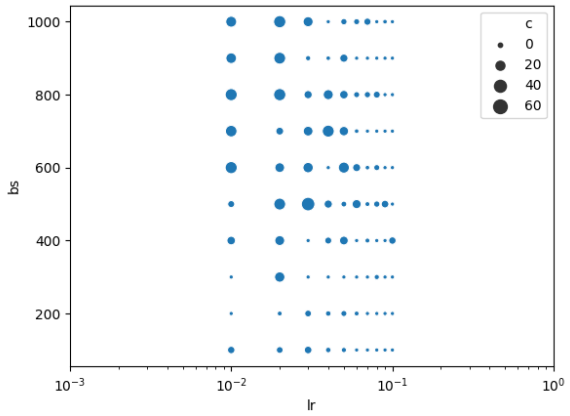
- Which value estimator has better performance without advantage-standardization: the trajectory-centric one, or the one using reward-to-go?
  - Using reward-to-go is always better.
- Did advantage standardization help?
  - In these configurations, there is no evidence that it helps
- Did the batch size make an impact?
  - Yes. The model converges significantly faster with larger batch size.

2 Problem 4

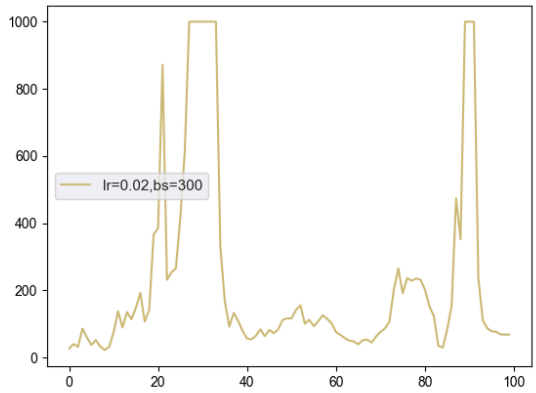
The hyper-parameter search in Figure 2a shows that the best configuration is when batch size is 300 and learning rate is $2e-2$. So I am using this. See Figure 2b for the result. However even though this configuration reaches 1000 for the seed that we choose, it is actually very unstable.

3 Problem 6

See Figure 3 for the result.



(a) Hyper-parameter search. Size indicates number of steps where the score is larger than 900.



(b) With batch size 300, learning rate 0.02.

Figure 2: Problem 4.

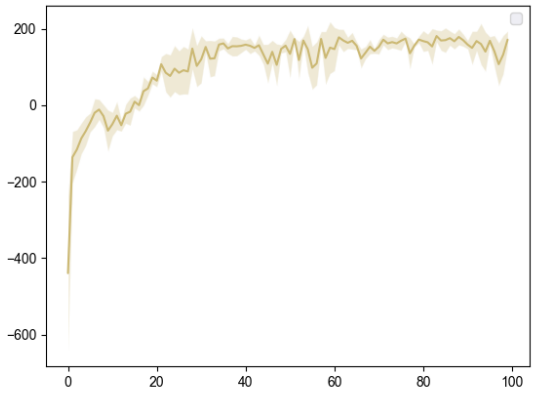
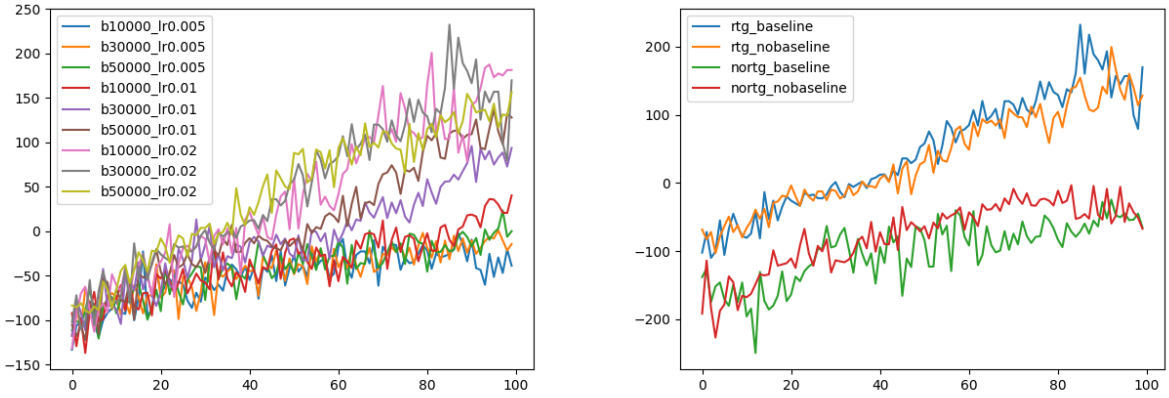


Figure 3: Lunar Lander



(a) Hyper-parameter search.

(b) With batch size 30k, learning rate 0.02.

Figure 4: Problem 7.

4 Problem 7

See Figure 4a and Figure 4b. In general, using large learning leads to faster convergence, and using larger batch size also has this effect, but not obvious in some cases.