CS294 Deep RL Assignment 2: Policy Gradients

Mohamed Khodeir

December 25, 2018

Problem 1. State Dependent Baselines

(a)

As given in the question, we can use the chain rule to deconstuct $P_\theta(\tau)$ as:

$$P_\theta(\tau) = P_\theta(s_t, a_t) P_\theta(\tau / s_t, a_t | s_t, a_t)$$

We can then use the law of iterated expectations to express the expectation over τ as:

$$E_{\tau/s_t,a_t \sim P_\theta(\tau/s_t,a_t|s_t,a_t)}[E_{(s_t,a_t)\sim P_\theta(s_t,a_t)}[\nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)]]$$

Looking just at the inner expectation, we see:

$$E_{(s_t,a_t)\sim P_\theta(s_t,a_t)}[\nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)]$$

Expanding the expectation, we can rewrite that as a nested integral, the first over s_t and the second over a_t. We can also substitute the full form of $P_\theta(s_t, a_t)$ as a product of the policy and the state marginal.

$$\int_{s_t} \int_{a_t} P_\theta(s_t) \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)$$

Taking terms in common

$$\int_{s_t} P_\theta(s_t) b(s_t) \int_{a_t} \pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t)$$

Using the identity $\pi_\theta(a_t|s_t) \nabla_\theta \log \pi_\theta(a_t|s_t) = \nabla_\theta \pi_\theta(a_t|s_t)$, we get:

$$\int_{s_t} P_\theta(s_t) b(s_t) \int_{a_t} \nabla_\theta \pi_\theta(a_t|s_t)$$

becomes by linearity of differentiation:

$$\int_{s_t} P_\theta(s_t) b(s_t) \nabla_\theta \int_{a_t} \pi_\theta(a_t|s_t)$$

The inner integral, being an integral over a propper probility distribution just sums to 1.

$$\int_{s_t} P_\theta(s_t) b(s_t) \nabla_\theta (1)$$

becomes

$$\int_{s_t} P_\theta(s_t) b(s_t)(0) = 0$$

. So going back to the full equation in (12):

$$\sum_{t=1}^{T} E_{\tau \sim P_\theta}[\nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)] =$$

$$\sum_{t=1}^{T} E_{\tau/s_t, a_t \sim P_\theta(\tau/s_t, a_t|s_t, a_t)}[E_{(s_t, a_t)\sim P_\theta(s_t, a_t)}[\nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)]] =$$

$$\sum_{t=1}^{T} E_{\tau/s_t, a_t \sim P_\theta(\tau/s_t, a_t|s_t, a_t)}[0] =$$

$$0$$

(b)

(a)

Let's consider $P_\theta(s_{t+1:T}, a_{t:T}|s_{1:t}, a_{1:t-1})$, the probability of the "rest" of the trajerctory after $(s_1, a_1, s_2, a_2, ...a_{t-1}, s_t)$.

Because an MDP satisfies the Markov property, we know that given s_t and a_t, the probability of s_{t+1} is independent of previous states and actions.

Therefore $P_\theta(s_{t+1:T}, a_{t:T}|s_{1:t}, a_{1:t-1})$ should exactly equal $P_\theta(s_{t+1:T}, a_{t:T}|s_t)$

We can show this using Bayes rule and by substituting the full form of $P_\theta(\tau)$, where τ is the whole trajectory $(s_1, a_1, s_2, a_2, ...a_{T-1}, s_T)$. See Appendix.

(b)

I will start by rewriting the expression for the probability of the "rest" of the trajectory using bayes rule:

$$P_\theta(s_{t+1:T}, a_{t:T}|s_{1:t}, a_{1:t-1}) = P_\theta(s_{t+1:T}, a_{t+1:T}|s_{1:t}, a_{1:t}) P_\theta(a_t|s_{1:t}, a_{1:t})$$

This allows us to write:

$$P_\theta(\tau) = P_\theta(s_{1:t}, a_{1:t-1}) P_\theta(a_t|s_{1:t}, a_{1:t}) P_\theta(s_{t+1:T}, a_{t+1:T}|s_{1:t}, a_{1:t})$$

Note that, in our case $P_\theta(a_t|s_{1:t}, a_{1:t}) = \pi_\theta(a_t|s_t)$.

$$E_{\tau \sim P_\theta}\Big[\nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)\Big] =$$

$$E_{(s_{1:t}, a_{1:t-1})\sim P_\theta}\Big[E_{a_t \sim P_\theta}\big[E_{s_{t+1:T}, a_{t+1:T}\sim P_\theta}\big[\nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)\big]\big]\Big]$$

$$\int_{(s_{1:t}, a_{1:t-1})} P_\theta(s_{1:t}, a_{1:t-1}) \Big[\int_{a_t} \pi_\theta(a_t|s_t) \big[\int_{s_{t+1:T}, a_{t+1:T}} P_\theta(s_{t+1:T}, a_{t+1:T}|s_{1:t}, a_{1:t}) \nabla_\theta \log \pi_\theta(a_t|s_t) b(s_t)\big]\Big]$$

$$\int_{(s_{1:t},a_{1:t-1})} P_\theta(s_{1:t},a_{1:t-1})\Big[\int_{a_t}\pi_\theta(a_t|s_t)\nabla_\theta\log\pi_\theta(a_t|s_t)b(s_t)\big[\int_{s_{t+1:T},a_{t+1:T}} P_\theta(s_{t+1:T},a_{t+1:T}|s_{1:t},a_{1:t})\big]\Big]$$

$$\int_{(s_{1:t},a_{1:t-1})} P_\theta(s_{1:t},a_{1:t-1})\Big[\int_{a_t}\pi_\theta(a_t|s_t)\nabla_\theta\log\pi_\theta(a_t|s_t)b(s_t)\big[Const\big]\Big]$$

Making use of that useful identity again, and moving constants out of the inner integral:

$$\int_{(s_{1:t},a_{1:t-1})} P_\theta(s_{1:t},a_{1:t-1})b(s_t)\big[Const\big]\Big[\nabla_\theta\int_{a_t}\pi_\theta(a_t|s_t)\Big]$$

$$\int_{(s_{1:t},a_{1:t-1})} P_\theta(s_{1:t},a_{1:t-1})b(s_t)\big[Const\big]\Big[\nabla_\theta Const\Big]$$

$$\int_{(s_{1:t},a_{1:t-1})} P_\theta(s_{1:t},a_{1:t-1})b(s_t)\big[Const\big]\Big[0\Big] = 0$$

As we've shown that

$$E_{\tau\sim P_\theta}\Big[\nabla_\theta\log\pi_\theta(a_t|s_t)b(s_t)\Big] = 0$$

it follows that

$$\sum_{t=1}^{T} E_{\tau\sim P_\theta}\Big[\nabla_\theta\log\pi_\theta(a_t|s_t)b(s_t)\Big] = 0$$

Problem 4. CartPole

Learning Curves for Small/Large Batch Sizes

Figure 1: Small Batch

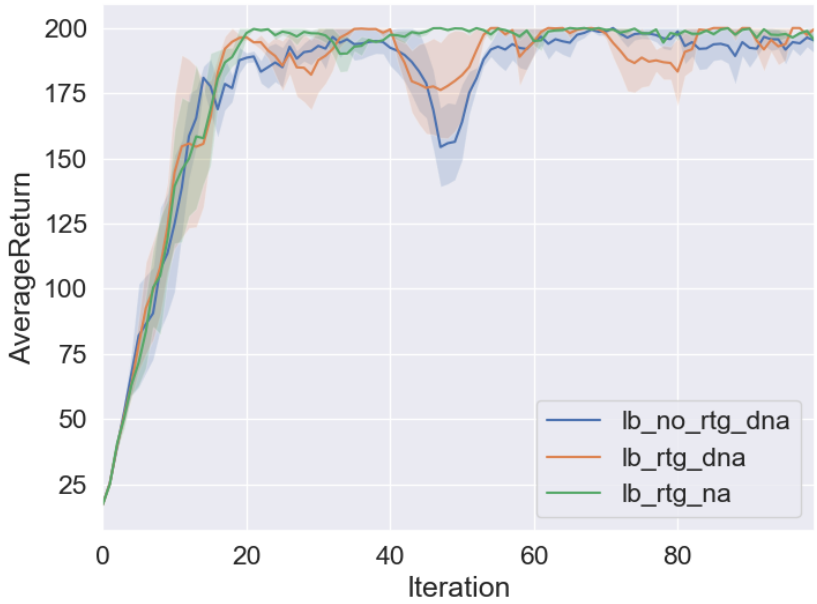

Figure 2: Large Batch

Analysis questions

Trajectory-Centric vs Reward-To-Go w/out Advantage Centering

We can see that the reward to go estimator displays higher performance, though the effect seems to be significantly less pronounced with larger batch sizes.

Advantage Centering

Advantage centering certainly seems to have helped by reducing the variance of the estimator, which we can see in the more stable learning curve for both small and larger batch sizes.

Batch Size

The batch size also seems to be very effective at reducing the variance of the gradient estimators both in the reward-to-go estimator as well as the trajectory-centric one. It also converges more quickly in the number of iterations in all cases.

Problem 5. Inverted Pendulum

Learning Curve for Smallest Batch Size and Largest Learning Rate

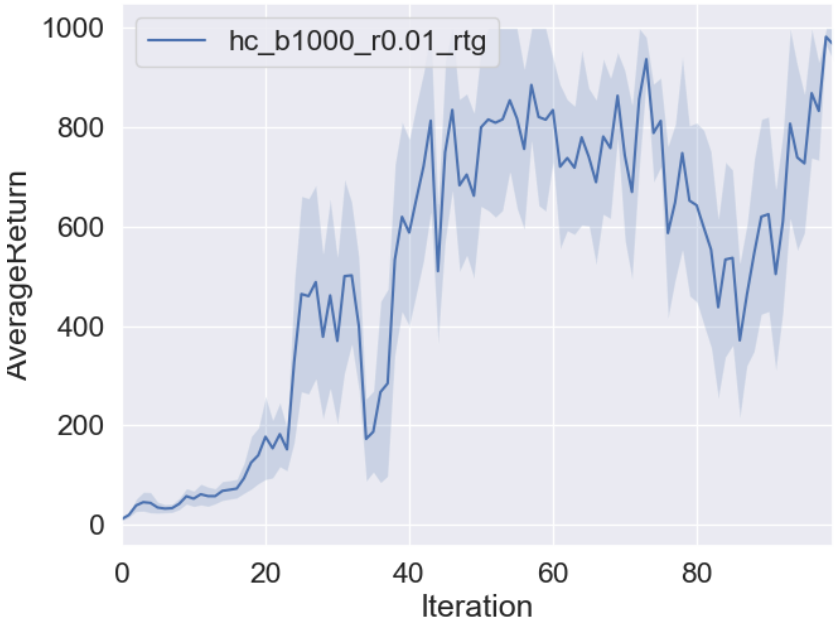

Figure 3: The learning rate used is 0.01, and the batch size is 1000.

Problem 7. Lunar Landing

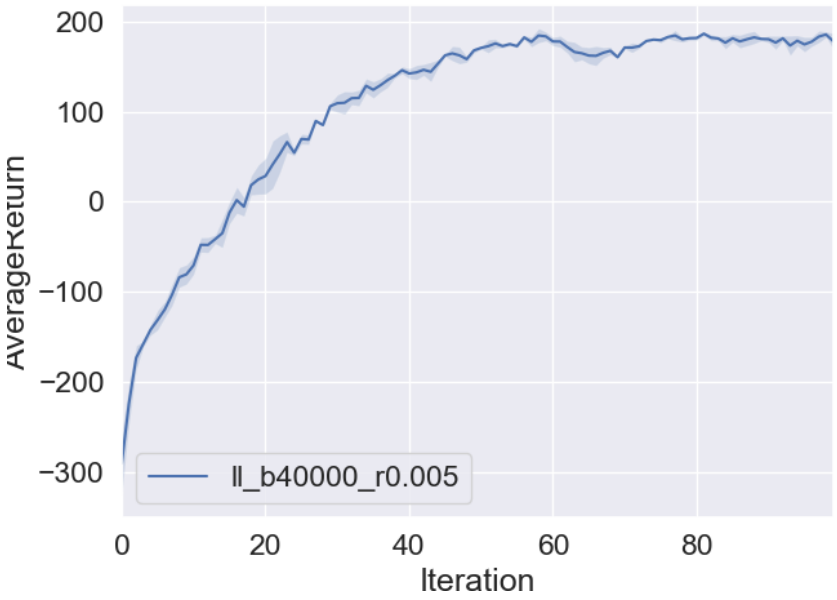

Figure 4: S

Problem 8. HalfCheetah

batch size and learning rate

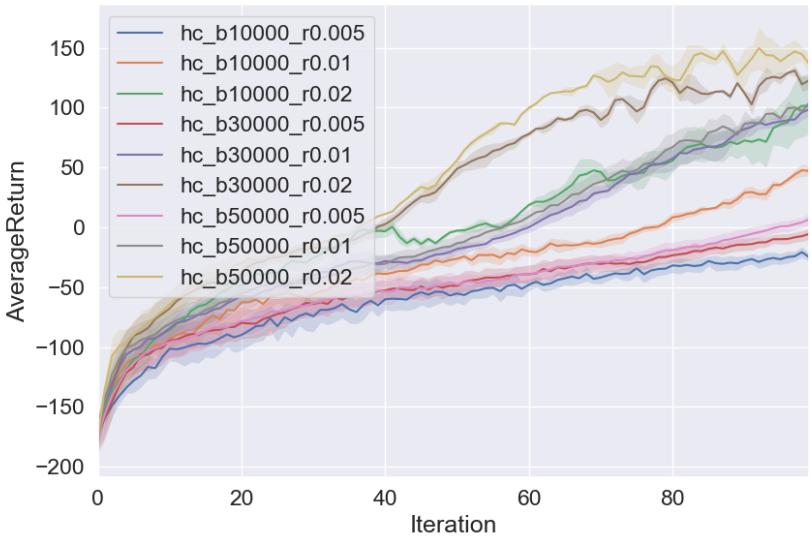

Figure 5: S

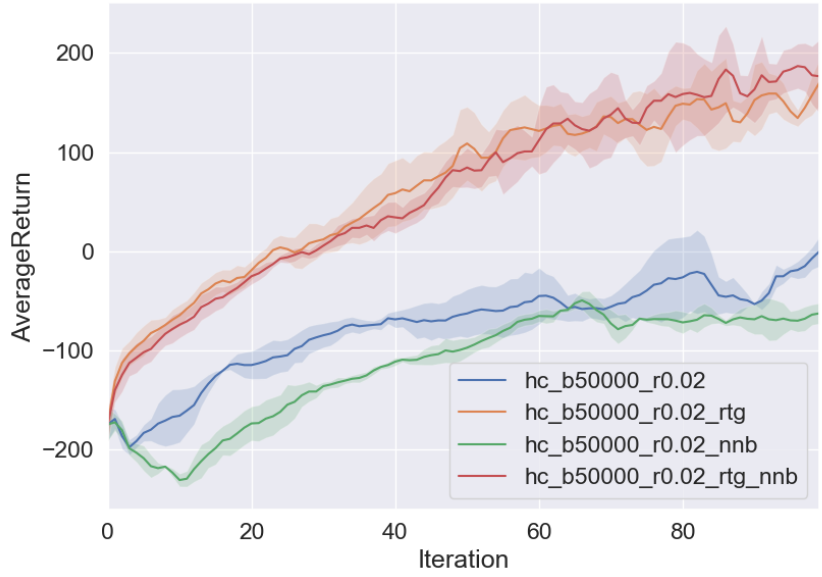

Figure 6: S

Appendix

Problem 1 (b)

$$P_\theta(s_{t+1:T}, a_{t:T}|s_{1:t}, a_{1:t-1}) = \frac{P_\theta(\tau)}{P(s_{1:t}, a_{1:t-1})}$$

Recall that the numerator, $P_\theta(\tau)$ is:

$$P_\theta(\tau) = P(s_1)\prod_{i=2}^{i=T} P_\theta(s_i|a_{i-1}, s_{i-1})P_\theta(a_{i-1}|s_{i-1})$$

We can equivalently represent that as:

$$P_\theta(\tau) = \left(P(s_1)\prod_{i=2}^{i=t} P_\theta(s_i|a_{i-1}, s_{i-1})P_\theta(a_{i-1}|s_{i-1})\right)\prod_{i=t+1}^{i=T} P_\theta(s_i|a_{i-1}, s_{i-1})P_\theta(a_{i-1}|s_{i-1})$$

The denomenator $P(s_{1:t}, a_{1:t-1})$ simply marginalizes P_θ over all possible assignments of the remaining states and actions. i.e.

$$P(s_{1:t}, a_{1:t-1}) = \sum_{a_{t:T}}\sum_{s_{t+1:T}} P_\theta(\tau) = \sum_{a_{t:T}}\sum_{s_{t+1:T}} P(s_1)\prod_{i=2}^{i=T} P_\theta(s_i|a_{i-1}, s_{i-1})P_\theta(a_{i-1}|s_{i-1}) =$$

Factoring out the terms that dont depend on the summation domains:

$$\left(P(s_1)\prod_{i=2}^{i=t}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})\right)\sum_{a_{t:T}}\sum_{s_{t+1:T}}\prod_{i=t+1}^{i=T}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})$$

Now, substituting this for our numerator and denomenator we get:

$$P_\theta(s_{t+1:T},a_{t:T}|s_{1:t},a_{1:t-1})=\frac{\prod_{i=t+1}^{i=T}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})}{\sum_{a_t}^{T}\sum_{s_{t+1:T}}\prod_{i=t+1}^{i=T}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})}$$

We can follow a similar procedure starting from $P_\theta(s_{t+1:T},a_{t:T}|s_t)$ to show that they reduce to the same expression.

$$P_\theta(s_{t+1:T},a_{t:T}|s_t)=\frac{\sum_{a_1:t-1}\sum_{s_1:t-1}P_\theta(\tau)}{\sum_{a_1:t-1}\sum_{s_1:t}\sum_{a_t:T-1}\sum_{s_t+1:T}P_\theta(\tau)}$$

Substituting our factored form for P_θ, looking only at numerator:

$$\sum_{a_1:t-1}\sum_{s_1:t-1}\left(P(s_1)\prod_{i=2}^{i=t}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})\right)\prod_{i=t+1}^{i=T}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})=$$

$$\prod_{i=t+1}^{i=T}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})\sum_{a_1:t-1}\sum_{s_1:t-1}\left(P(s_1)\prod_{i=2}^{i=t}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})\right)$$

Now denomenator:

$$\sum_{a_1:t-1}\sum_{s_1:t-1}\sum_{a_t:T-1}\sum_{s_t+1:T}\left(P(s_1)\prod_{i=2}^{i=t}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})\right)\prod_{i=t+1}^{i=T}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})=$$

$$\sum_{a_{1:t-1}}\sum_{s_{1:t-1}}\left(P(s_1)\prod_{i=2}^{i=t}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})\right)\sum_{a_t:T-1}\sum_{s_t+1:T}\prod_{i=t+1}^{i=T}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})=$$

Putting it all together:

$$P_\theta(s_{t+1:T},a_{t:T}|s_t)=\frac{\prod_{i=t+1}^{i=T}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})}{\sum_{a_t}^{T}\sum_{s_{t+1:T}}\prod_{i=t+1}^{i=T}P_\theta(s_i|a_{i-1},s_{i-1})P_\theta(a_{i-1}|s_{i-1})}$$