Assignment 2: Policy Gradients

Due September 25, 11:59 pm

4 Policy Gradients

- Create two graphs:
  - In the first graph, compare the learning curves (average return vs. number of environment steps) for the experiments prefixed with `cartpole`. (The small batch experiments.)
  - In the second graph, compare the learning curves for the experiments prefixed with `cartpole_lb`. (The large batch experiments.)

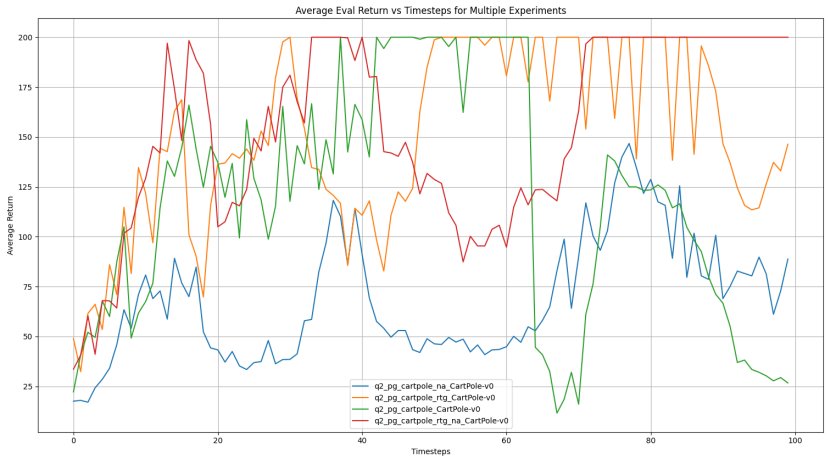


Figure 1: Learning curves for small batch experiments prefixed with `cartpole`.

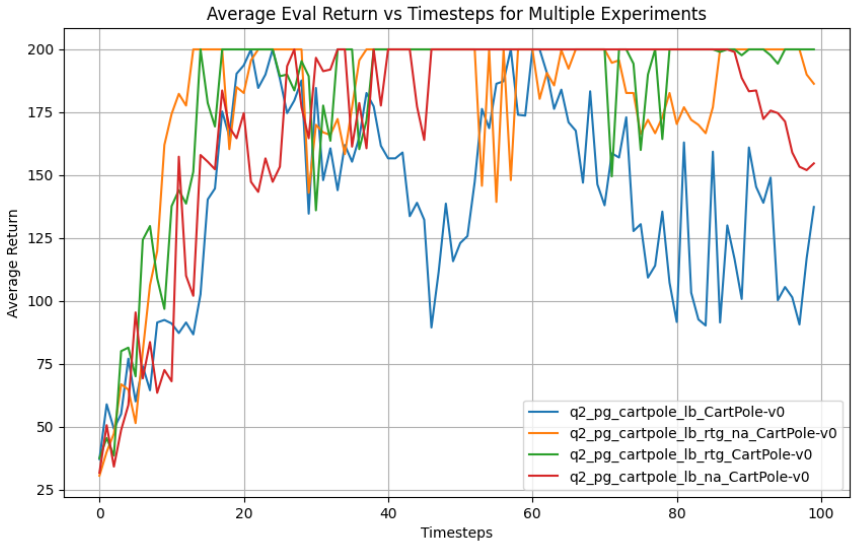


Figure 2: Learning curves for large batch experiments prefixed with `cartpole_lb`.

- Answer the following questions briefly:
  - Which value estimator has better performance without advantage normalization: the trajectory-centric one, or the one using reward-to-go?
    Answer: The value estimator using reward-to-go generally has better performance without advantage normalization because it provides a more accurate estimate of future rewards, leading to more effective policy updates. The overall variance is reduced when we consider only the rewards-to-go, which increases performance.
  - Did advantage normalization help?
    Answer: Yes, advantage normalization helped by reducing the variance of the policy gradient estimates, leading to more stable and faster learning. We see from the plots that the reward hits the maximum much quicker than in the case without the normalization.
  - Did the batch size make an impact?
    Answer: Yes, the batch size made a significant impact. Larger batch sizes tend to reduce the variance of the gradient estimates, leading to more stable learning, though they also increase the computational cost per update. We see from the plots that the reward reaches 200 much quicker, and the average returns are also higher when we use larger batches.
- Provide the exact command line configurations (or `#@params` settings in Colab) you used to run your experiments, including any parameters changed from their defaults.

SNO	Experiment Name	Iterations	Batch Size	RTG	Normalize Advantages
1	cartpole	100	1000	No	No
2	cartpole_rtg	100	1000	Yes	No
3	cartpole_na	100	1000	No	Yes
4	cartpole_rtg_na	100	1000	Yes	Yes
5	cartpole_lb	100	4000	No	No
6	cartpole_lb_rtg	100	4000	Yes	No
7	cartpole_lb_na	100	4000	No	Yes
8	cartpole_lb_rtg_na	100	4000	Yes	Yes

Table 1: Experiment configurations and parameters.

In addition, the following parameters were fixed across all experiments:
- Discount Rate: 1
- Learning Rate: 0.005
- Number of Layers: 2
- Layer Size: 64

5 Neural Network Baseline

- Plot a learning curve for the baseline loss.

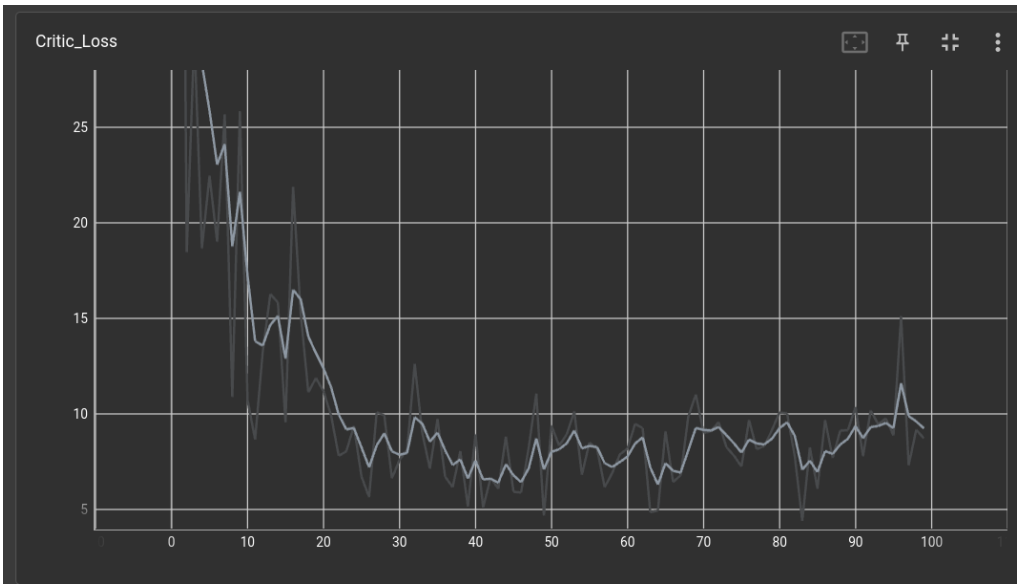


Figure 3: Learning curve for the baseline loss.

- Plot a learning curve for the eval return. You should expect to achieve an average return over 300 for the baselined version.

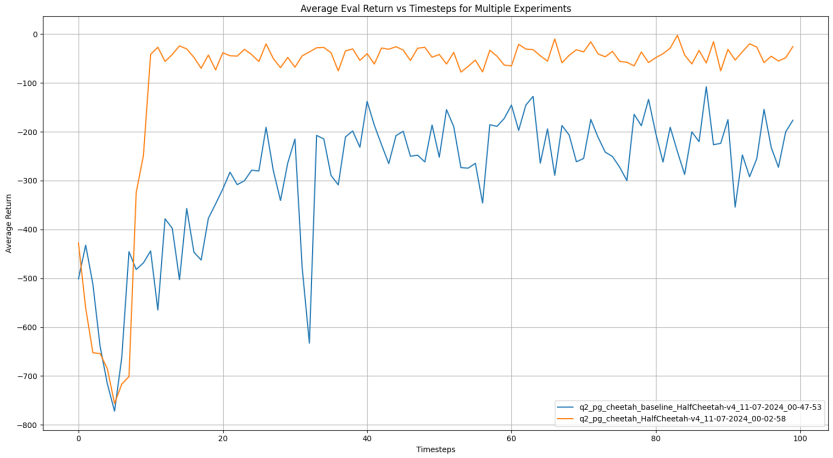


Figure 4: Learning curve for the eval return with baseline.

- Run another experiment with a decreased number of baseline gradient steps (`-bgs`) and/or baseline learning rate (`-blr`). How does this affect (a) the baseline learning curve and (b) the performance of the policy?
  One thing that seemed problematic with this experiment was that it did not converge. I tried implementing various tricks such as gradient clipping, xavier initialization of the neural networks, different learning rates, optimizers and also varying network sizes. I also tried normalizing advantages. It initially

seemed to be a problem with all continuous environments. Upon introducing an entropy component to the policy loss, I was able to fix it for the simpler continous environment such as Inverted Pendulum. However, this one did not get resolved. I have attached the plots of the best performing experiments here nonetheless.

- **Optional:** Add `-na` back to see how much it improves things. Also, set `video_log_freq 10`, then open TensorBoard and go to the "Images" tab to see some videos of your HalfCheetah walking along!

6 Generalized Advantage Estimation

- Provide a single plot with the learning curves for the `LunarLander-v2` experiments that you tried. Describe in words how λ affected task performance. The run with the best performance should achieve an average score close to 200 (180+).

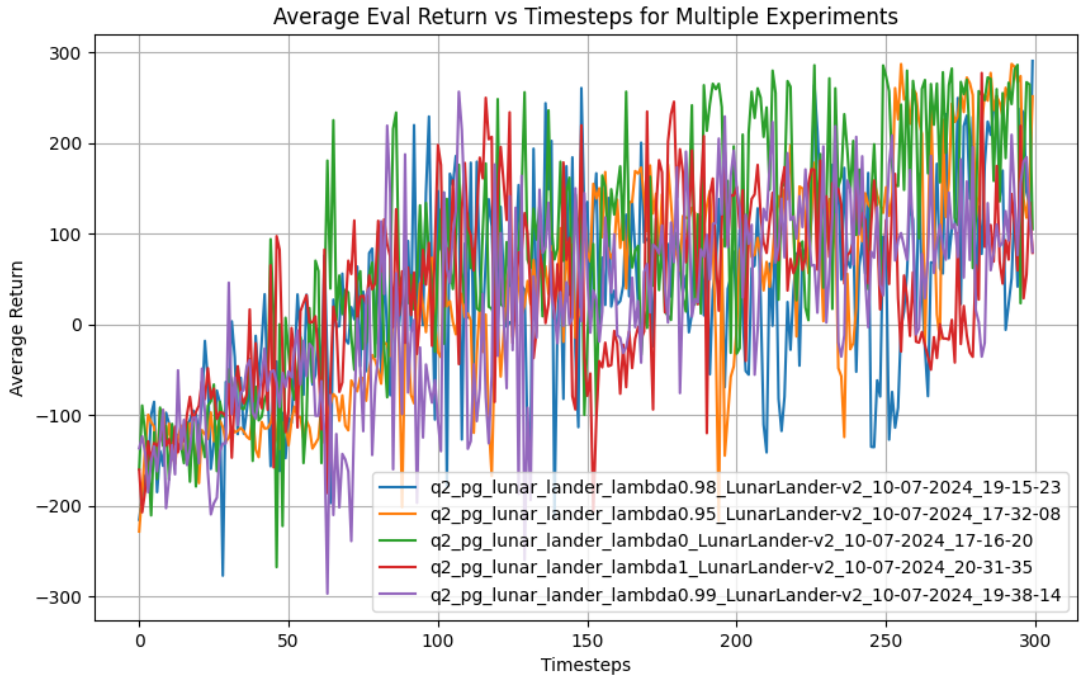


Figure 5: Learning curves for `LunarLander-v2` with different λ values.

The plot in Figure 5 shows the learning curves for the `LunarLander-v2` environment with different values of λ. From the plot, we observe that while λ values of 0.99 and 0.98 achieve the best performance with average scores close to 200, $\lambda = 1$ performs worse. This suggests that using a slightly lower value than 1 helps stabilize learning and improve performance by reducing the variance introduced by using the full trajectory of rewards.

- Consider the parameter λ. What does $\lambda = 0$ correspond to? What about $\lambda = 1$? Relate this to the task performance in `LunarLander-v2` in one or two sentences.

  The parameter λ in Generalized Advantage Estimation (GAE) controls the bias-variance trade-off in the advantage estimates. A value of $\lambda = 0$ corresponds to using only the immediate rewards (high bias, low variance), while $\lambda = 1$ corresponds to using the full trajectory of rewards (low bias, high variance). In the `LunarLander-v2` environment, values of λ close to but slightly less than 1 (such as 0.99 and 0.98) provide the best task performance by balancing the trade-off, whereas $\lambda = 1$ introduces too much variance, resulting in worse performance.

7 Hyperparameter Tuning

1. Provide a set of hyperparameters that achieve high return on `InvertedPendulum-v4` in as few environment steps as possible.
2. Show learning curves for the average returns with your hyperparameters and with the default settings, with environment steps on the x-axis. Returns should be averaged over 5 seeds.

- **Environment Name:** InvertedPendulum-v4
- **Number of Iterations:** 100
- **Experiment Name:** pendulum_default_s{seed}
- **Use Reward to Go:** True
- **Normalize Advantages:** True
- **Batch Size:** 5000
- **GAE Lambda:** 0.98
- **Learning Rate:** 0.005

The above configuration was able to attain a reward of 1000 within 48 iterations.

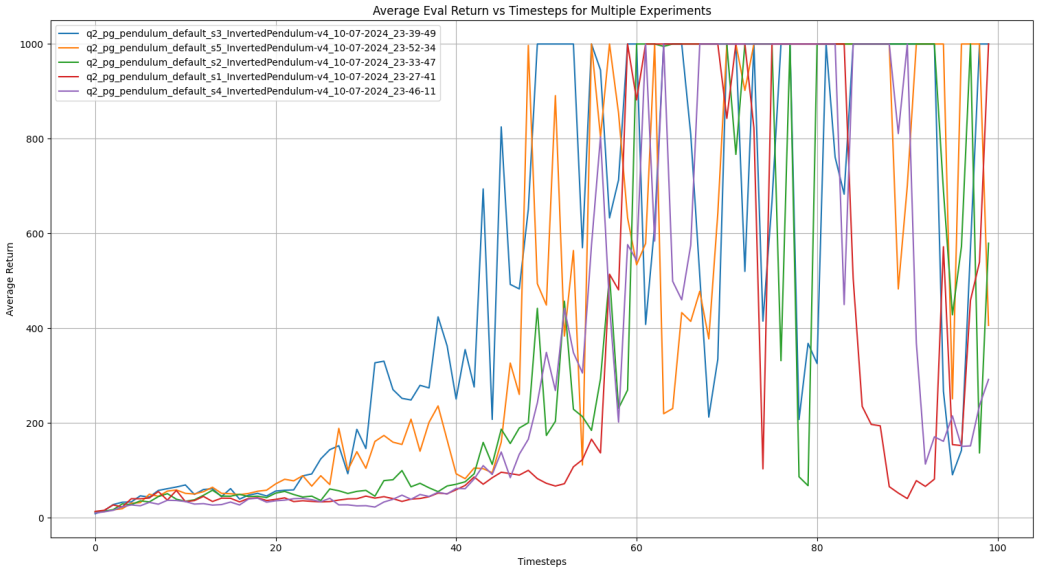


Figure 6: Learning curve with default hyperparameters.

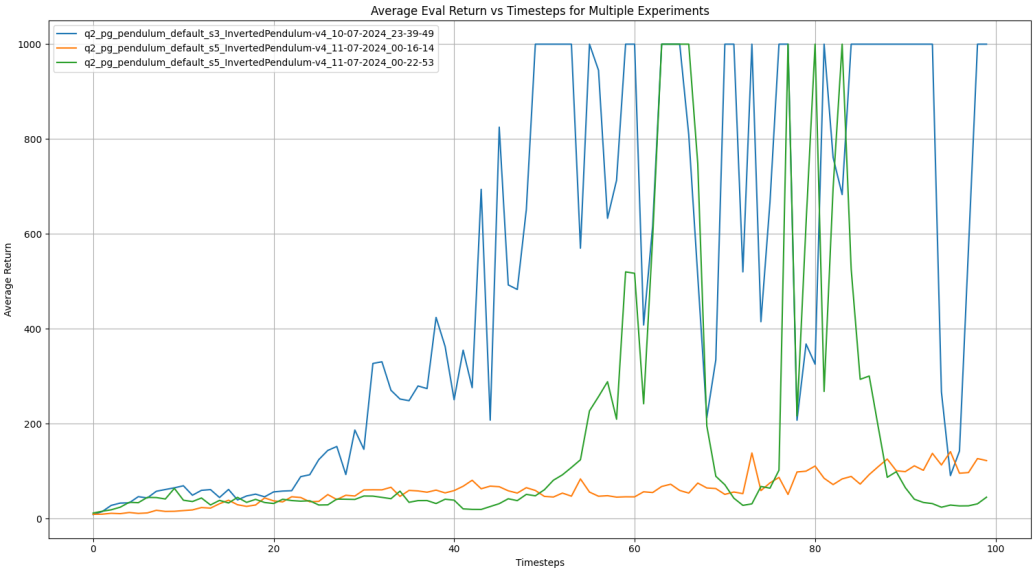


Figure 7: Learning curve with tuned hyperparameters.

8 (Extra Credit) Humanoid

1. Plot a learning curve for the Humanoid-v4 environment. You should expect to achieve an average return of at least 600 by the end of training. Discuss what changes, if any, you made to complete this problem (for example: optimizations to the original code, hyperparameter changes, algorithmic changes).

9 Analysis

Consider the following infinite-horizon MDP:

$$a_1 \circlearrowright s_1 \xrightarrow{a_2} s_F$$

At each step, the agent stays in state s_1 and receives reward 1 if it takes action a_1, and receives reward 0 and terminates the episode otherwise. Parametrize the policy as stationary (not dependent on time) with a single parameter:

$$\pi_\theta(a_1|s_1) = \theta, \pi_\theta(a_2|s_1) = 1 - \theta$$

1. **Applying policy gradients**

   (a) Use policy gradients to compute the gradient of the expected return $J(\theta) = \mathbb{E}_{\pi_\theta}[R(\tau)]$ with respect to the parameter θ. **Do not use discounting.**

   Hint: to compute $\sum_{k=1}^{\infty} k\alpha^{k-1}$, you can write:

   $$\sum_{k=1}^{\infty} k\alpha^{k-1} = \sum_{k=1}^{\infty} \frac{d}{d\alpha}\alpha^k = \frac{d}{d\alpha}\sum_{k=1}^{\infty}\alpha^k$$

   Solution:

   The policy gradient theorem states:

   $$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\tau) R(\tau)\right]$$

   Here, $\pi_\theta(\tau)$ is the probability of the trajectory τ, and $R(\tau)$ is the return of the trajectory.

   Let's break this down step by step.

   1. **Compute $\pi_\theta(\tau)$**: The trajectory τ consists of a sequence of k actions a_1 followed by a_2. The probability of such a trajectory is $\pi_\theta(\tau) = \theta^k(1-\theta)$.

   2. **Compute $\log \pi_\theta(\tau)$**: The log probability of the trajectory is:

   $$\log \pi_\theta(\tau) = \log(\theta^k(1-\theta)) = k\log\theta + \log(1-\theta)$$

   3. **Compute $R(\tau)$**: The return $R(\tau)$ for this trajectory is k, as the agent receives a reward of 1 for each step until it terminates.

   Putting it all together, we have:

   $$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta(k\log\theta + \log(1-\theta)) \cdot k\right]$$

   The gradient of $\log \pi_\theta(\tau)$ with respect to θ is:

   $$\nabla_\theta \log \pi_\theta(\tau) = \nabla_\theta(k\log\theta + \log(1-\theta)) = \frac{k}{\theta} - \frac{1}{1-\theta}$$

   Therefore, the policy gradient becomes:

   $$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\left(\frac{k}{\theta} - \frac{1}{1-\theta}\right)k\right]$$

   Now, we need to sum this over all possible trajectories:

   $$\nabla_\theta J(\theta) = \sum_{k=0}^{\infty} \theta^k(1-\theta)\left(\frac{k^2}{\theta} - \frac{k}{1-\theta}\right)$$

Simplifying, we get:

$$\nabla_\theta J(\theta) = \sum_{k=0}^{\infty} \left(\theta^{k-1}(1-\theta)k^2 - \theta^k k\right)$$

Breaking this sum into two parts:

$$\nabla_\theta J(\theta) = (1-\theta)\sum_{k=0}^{\infty} \theta^{k-1}k^2 - \sum_{k=0}^{\infty} \theta^k k$$

Using the known series sum identities:

$$\sum_{k=0}^{\infty} k\theta^k = \frac{\theta}{(1-\theta)^2}$$

$$\sum_{k=0}^{\infty} k^2\theta^k = \frac{\theta(1+\theta)}{(1-\theta)^3}$$

Plugging these into our gradient expression:

$$\nabla_\theta J(\theta) = (1-\theta)\cdot\frac{1+\theta}{(1-\theta)^3} - \frac{\theta}{(1-\theta)^2}$$

Simplifying:

$$\nabla_\theta J(\theta) = \frac{1+\theta}{(1-\theta)^2} - \frac{\theta}{(1-\theta)^2}$$

Simplifying the fractions:

$$\nabla_\theta J(\theta) = \frac{1}{(1-\theta)^2}$$

(b) Compute the expected return of the policy $\mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$ directly. Compute the gradient of this expression with respect to θ and verify that this matches the policy gradient.

Solution:

As computed above, the expected return $J(\theta)$ is:

$$J(\theta) = \frac{\theta}{1-\theta}$$

The gradient of this expected return with respect to θ is:

$$\frac{d}{d\theta}J(\theta) = \frac{1}{(1-\theta)^2}$$

This matches the policy gradient computed in part (a).

2. Compute the variance of the policy gradient in closed form and describe the properties of the variance with respect to θ. For what value(s) of θ is variance minimal? Maximal?

Solution:

The variance of the policy gradient can be computed using the expression for the gradient:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\left(\frac{k}{\theta} - \frac{1}{1-\theta}\right)k\right]$$

To compute the variance, we need to compute the expectation of the square of the gradient and subtract the square of the expectation:

$$(\nabla_\theta J(\theta)) = \mathbb{E}\left[\left(\left(\frac{k}{\theta} - \frac{1}{1-\theta}\right)k\right)^2\right] - \left(\mathbb{E}\left[\left(\frac{k}{\theta} - \frac{1}{1-\theta}\right)k\right]\right)^2$$

Let's break it down step by step:

$$\mathbb{E}\left[\left(\left(\frac{k}{\theta} - \frac{1}{1-\theta}\right)k\right)^2\right] = \sum_{k=0}^{\infty} \theta^k(1-\theta)\left(\left(\frac{k^2}{\theta} - \frac{k}{1-\theta}\right)^2\right)$$

After some algebraic manipulation, we find:

$$(\nabla_\theta J(\theta)) = \sum_{k=0}^{\infty} \theta^k(1-\theta)\left(\frac{k^4}{\theta^2} - 2\frac{k^3}{\theta(1-\theta)} + \frac{k^2}{(1-\theta)^2}\right) - \left(\frac{1}{(1-\theta)^2}\right)^2$$

Simplifying and using known series sums, we get:

$$(\nabla_\theta J(\theta)) = \frac{\theta(1+4\theta+\theta^2)}{(1-\theta)^5}$$

The variance is minimal when θ is close to 0 or 1, and maximal at some intermediate value of θ.

3. **Apply return-to-go as an advantage estimator.**

   (a) Write the modified policy gradient and confirm that it is unbiased.

   Solution:

   The return-to-go $R_t(\tau)$ is the sum of rewards from time step t to the end of the episode. For this specific problem, since each reward is 1 until termination, $R_t(\tau) = 1$ for each step until the termination action a_2 is taken.

   The modified policy gradient using return-to-go is:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) R_t(\tau)\right]$$

   Since $R_t(\tau) = 1$, this reduces to:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\right]$$

   This estimator is unbiased because:

$$\mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\right] = \nabla_\theta J(\theta)$$

   This equality holds because the expectation of the return-to-go equals the expectation of the total return.

(b) Compute the variance of the return-to-go policy gradient and plot it on $[0, 1]$ alongside the variance of the original estimator.

Solution:

The variance of the return-to-go policy gradient can be computed similarly to the original policy gradient but considering that $R_t(\tau) = 1$.

Let's compute the variance of the return-to-go policy gradient:

$$(\nabla_\theta J(\theta)_{\text{RTG}}) = \mathbb{E}\left[\left(k\left(\frac{1}{\theta} - \frac{1}{1-\theta}\right)\right)^2\right] - \left(\mathbb{E}\left[k\left(\frac{1}{\theta} - \frac{1}{1-\theta}\right)\right]\right)^2$$

Since $R_t(\tau) = 1$, the variance computation simplifies to considering the distribution of k:

$$(\nabla_\theta J(\theta)_{\text{RTG}}) = \sum_{k=0}^{\infty} \theta^k (1-\theta)\left(\frac{k^2}{\theta^2} - 2\frac{k}{\theta(1-\theta)} + \frac{1}{(1-\theta)^2}\right) - \left(\frac{1}{(1-\theta)^2}\right)^2$$

Using the known series sums, we get:

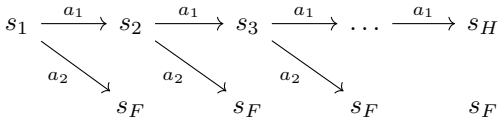
$$\mathbb{E}\left[k^2\right] = \frac{\theta(1+\theta)}{(1-\theta)^3}$$

Therefore, the variance of the return-to-go policy gradient is:

$$(\nabla_\theta J(\theta)_{\text{RTG}}) = \frac{1+\theta}{\theta(1-\theta)^3} - \frac{1}{(1-\theta)^4}$$

This variance is different from the original estimator. The variance is minimal when θ is close to 0 or 1, and maximal at some intermediate value of θ.

4. **Consider a finite-horizon H-step MDP with sparse reward:**



The agent receives reward $R_{\max}$ if it arrives at s_H and reward 0 if it arrives at s_F (a terminal state). In other words, the return for a trajectory τ is given by:

$$R(\tau) = \begin{cases} 1 & \tau \text{ ends at } s_H \\ 0 & \tau \text{ ends at } s_F \end{cases}$$

Using the same policy parametrization as above, consider off-policy policy gradients via importance sampling. Assume we want to compute policy gradients for a policy π_θ with samples drawn from $\pi_{\theta'}$.

(a) Write the policy gradient with importance sampling.

Solution:

The policy gradient with importance sampling is:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta'}}\left[\left(\frac{\pi_\theta(\tau)}{\pi_{\theta'}(\tau)}\right)\nabla_\theta \log \pi_\theta(\tau) R(\tau)\right]$$

In this case, the probability of reaching s_H under policy π_θ is:

$$\pi_\theta(\tau) = \theta^H$$

For behavior policy $\pi_{\theta'}$:

$$\pi_{\theta'}(\tau) = \theta'^H$$

Therefore, the importance sampling ratio is:

$$\frac{\pi_\theta(\tau)}{\pi_{\theta'}(\tau)} = \left(\frac{\theta}{\theta'}\right)^H$$

The policy gradient becomes:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim\pi_{\theta'}}\left[\left(\left(\frac{\theta}{\theta'}\right)^H\right)\nabla_\theta \log \pi_\theta(\tau) R(\tau)\right]$$

Given $R(\tau) = R_{\max}$ only when the agent reaches s_H:

$$\nabla_\theta J(\theta) = R_{\max}\left(\frac{\theta}{\theta'}\right)^H \nabla_\theta \log \theta^H$$

The log probability for θ is:

$$\log \pi_\theta(\tau) = \log \theta^H = H \log \theta$$

Therefore:

$$\nabla_\theta \log \pi_\theta(\tau) = \nabla_\theta (H \log \theta) = \frac{H}{\theta}$$

Putting it all together:

$$\nabla_\theta J(\theta) = R_{\max}\left(\frac{\theta}{\theta'}\right)^H \frac{H}{\theta}$$

(b) Compute its variance. How does it change when H becomes large?

Solution:

The variance of the importance-sampled policy gradient is:

$$(\nabla_\theta J(\theta)) = \mathbb{E}_{\tau\sim\pi_{\theta'}}\left[\left(\left(\frac{\pi_\theta(\tau)}{\pi_{\theta'}(\tau)}\right)\nabla_\theta \log \pi_\theta(\tau) R(\tau)\right)^2\right] - \left(\mathbb{E}_{\tau\sim\pi_{\theta'}}\left[\left(\frac{\pi_\theta(\tau)}{\pi_{\theta'}(\tau)}\right)\nabla_\theta \log \pi_\theta(\tau) R(\tau)\right]\right)^2$$

Since $R(\tau) = 0$ for all trajectories except those ending at s_H, we only need to consider those specific trajectories in our expectations. Given the expectation over $\pi_{\theta'}$:

$$\mathbb{E}_{\tau\sim\pi_{\theta'}}\left[R_{\max}^2\left(\frac{\theta}{\theta'}\right)^{2H}\frac{H^2}{\theta^2}\right] = R_{\max}^2\left(\frac{\theta}{\theta'}\right)^{2H}\frac{H^2}{\theta^2}\cdot\theta'^H$$

Simplifying, we get:

$$\mathbb{E}_{\tau\sim\pi_{\theta'}}\left[R_{\max}^2\left(\frac{\theta}{\theta'}\right)^{2H}\frac{H^2}{\theta^2}\cdot\theta'^H\right] = R_{\max}^2\frac{H^2}{\theta^2}\left(\frac{\theta^{2H}}{\theta'^{2H-H}}\right) = R_{\max}^2\frac{H^2}{\theta^2}\left(\frac{\theta^{2H}}{\theta'^H}\right)$$

For the second term:

$$\left(\mathbb{E}_{\tau\sim\pi_{\theta'}}\left[R_{\max}\left(\frac{\theta}{\theta'}\right)^H\frac{H}{\theta}\right]\right)^2=\left(R_{\max}\left(\frac{\theta}{\theta'}\right)^H\frac{H}{\theta}\cdot\theta'^H\right)^2=\left(R_{\max}\frac{H}{\theta}\cdot\theta^H\right)^2=R_{\max}^2\frac{H^2\theta^{2H}}{\theta^2\theta'^2H}=R_{\max}^2\frac{H^2\theta^{2H}}{\theta^2\theta'^2H}$$

Therefore, the variance simplifies to:

$$(\nabla_\theta J(\theta))=R_{\max}^2\frac{H^2}{\theta^2}\left(\frac{\theta^{2H}}{\theta'^H}\right)-\left(R_{\max}\left(\frac{\theta}{\theta'}\right)^H\frac{H}{\theta}\right)$$

As H becomes large, the importance sampling ratio $\left(\frac{\theta}{\theta'}\right)^H$ can become very large if θ and θ' are not close to each other. This can cause the variance to increase significantly, making the estimator less reliable.

If $\theta\approx\theta'$, the variance remains manageable, but as the difference between θ and θ' increases, the variance increases rapidly, especially for large H.