CS 234: Assignment #3
Winter 2019

Li Quan Khoo (SCPD)

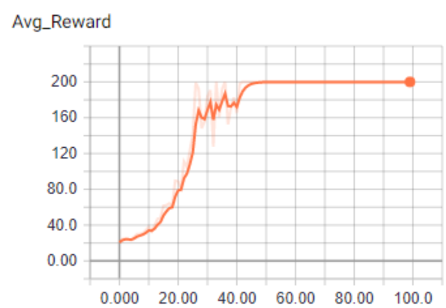
1 Policy Gradient Methods (50 pts + 15 pts writeup)

1.1 Coding Questions (50 pts)

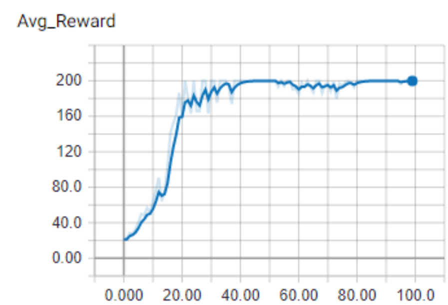
- https://github.com/lqkhoo/cs234-winter-2019/blob/master/assignment3/pg.py

1.2 Writeup Questions (15 pts)

(a) (4 pts) (CartPole-v0)



(a) CartPole-v0 with baseline. Average reward over 100 iterations.



(b) CartPole-v0 without baseline. Average reward over 100 iterations.

Where

$$r(\tau) = \sum_t r(s_t, a_t), \quad and \quad J(\theta) = \mathbb{E}_{\tau \sim P_\theta(\tau)}[r(\tau)],$$

We have derived the Monte-Carlo estimate of the gradient of $J(\theta)$ for policy gradient to be:

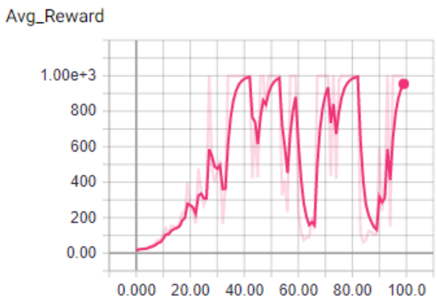
$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta log \pi_\theta(\tau)[r(\tau) - b]$$

Where b is some constant scalar bias of the reward function. Mathematically, the baseline is just another bias, and we already know that this gradient $\nabla_\theta J(\theta)$ is indifferent to the bias *in expectation*. However, the actual size and value of the bias still influences convergence (consider a bias close to +/- infinity) due to how the probability mass of the policy is shifted when we perform the gradient update, as well as the variance of the quantity $r(\tau) - b$. (Berkeley CS294, lecture 4)

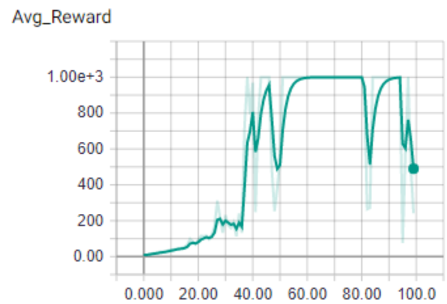
Here, the baseline we are taking is the state value function $V(s_t)$, which represents the average of returns of all actions from that state. Suppose we have a good estimate of $V(s_t)$. Mathematically, $V(s_t)$ already summarizes all future returns, and the advantage $A(s_t, a_t) = r(s_t, a_t) - V(s_t)$ becomes insulated from the variance of returns of all future time steps *in the current trajectory*. This lowers the variance of our estimated action values. Semantically, we are shifting the probability mass of the policy to make actions which are *better than average* (our chosen baseline) to be more likely.

Empirically, we can see from the plots above that using the baseline tends to give "better" convergence, although this is far from being guaranteed. Our observation here is that the plateau is more stable. The standard deviation of rewards (not shown here) tends to be lower as well.

(b) (4 pts) (InvertedPendulum-v1)



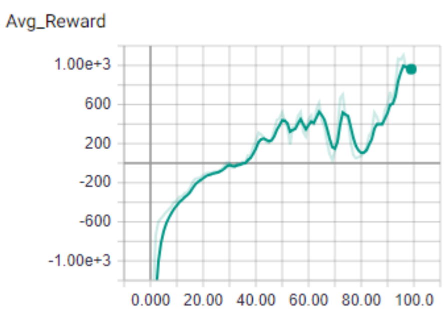
(a) Pendulum-v1 with baseline. Average reward over 100 iterations.



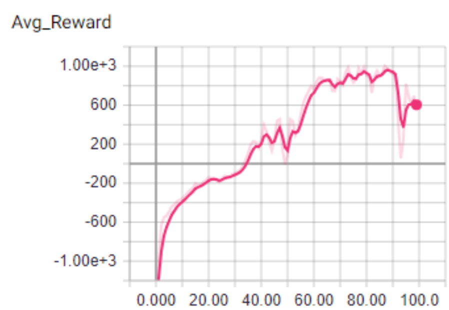
(b) Pendulum-v1 without baseline. Average reward over 100 iterations.

In this case, we notice that running with baseline tends to have faster convergence, but here, it doesn't seem to help very much with the instability of the algorithm. This instability is due to the high variance of REINFORCE (Monte-Carlo "vanilla" policy gradient), relative to comparable methods such as natural policy gradient and TRPO.

(c) (7 pts) (HalfCheetah-v1)



(a) HalfCheetah-v1 with baseline. Average reward over 100 iterations.



(b) HalfCheetah-v1 without baseline. Average reward over 100 iterations.

In this particular run, the baseline didn't seem to help, although on average it's supposed to have the same properties as explained in (a) by reducing the variance of the gradient.

2 Best Arm Identification in Multiarmed Bandit (35pts)

In this problem we focus on the Bandit setting with rewards bounded in $[0,1]$. A Bandit problem instance is defined as an MDP with just one state and action set $\mathcal{A}$. Since there is only one state, a "policy" consists of the choice of a single action: there are exactly $A = |\mathcal{A}|$ different deterministic policies. Your goal is to design a simple algorithm to identify a near-optimal arm with high probability.
Imagine we have n samples of a random variable x, $\{x_1, \ldots, x_n\}$. We recall Hoeffding's inequality below, where $\overline{x}$ is the expected value of a random variable x, $\widehat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean (under the assumption that the random variables are in the interval [0,1]), n is the number of samples and $\delta > 0$ is a scalar:

$$\Pr\left(|\widehat{x} - \overline{x}| > \sqrt{\frac{\log(2/\delta)}{2n}}\right) < \delta.$$

Assuming that the rewards are bounded in $[0,1]$, we propose this simple strategy: allocate an identical number of samples $n_1 = n_2 = ... = n_A = n_{des}$ to every action, compute the average reward (empirical payout) of each arm $\widehat{r}_{a_1}, \ldots, \widehat{r}_{a_A}$ and return the action with the highest empirical payout $\arg\max_a \widehat{r}_a$. The purpose of this exercise is to study the number of samples required to output an arm that is at least ϵ-optimal with high probability. Intuitively, as n_{des} increases the empirical payout $\widehat{r}_a$ converges to its expected value $\overline{r}_a$ for every action a, and so choosing the arm with the highest empirical payout $\widehat{r}_a$ corresponds to approximately choosing the arm with the highest expected payout $\overline{r}_a$.

(a) (15 pts) Given: $r_a \in [0,1] \forall a$, and we sample every action $a \in \mathcal{A}$ exactly $n = n_{des}$ times. Just so that it's clear that r is a random variable, let $R_a = r_a$. We assume that R is i.i.d.

To show:

$$P\left(\exists a \in \mathcal{A} \quad s.t. \quad |\widehat{R}_a - \overline{R}_a| > \sqrt{\frac{\log(2/\delta)}{2n}}\right) < |\mathcal{A}|\delta.$$

We shall prove this directly through the union bound. The fact that this is a bandit problem doesn't matter for our proof, so for notational clarity, let $X_a = |\widehat{R}_a - \overline{R}_a|$, and let $\gamma = \sqrt{\frac{\log(2/\delta)}{2n}}$, which means we aim to show that $P(\exists a \in \mathcal{A} \quad s.t. \quad X_a > \gamma) < |\mathcal{A}|\delta$.

Saying that the probability that there exists a condition a where an event W is true, is the same as saying that the probability that at least one of the events $W_{a'}$ is true under some condition a', among all possible conditions $a \in A$. Note that this depends on the fact that $\mathcal{A}$ is a countable set (which is assumed by the question since a could be enumerated):

$$P(\exists a \in \mathcal{A} \quad s.t. \quad X_a > \gamma) = P\left((X_{a_1} > \gamma) \bigcup (X_{a_2} > \gamma) \bigcup \ldots \bigcup (X_{a_{|\mathcal{A}|}} > \gamma)\right) = P\left(\bigcup_{a \in \mathcal{A}} (X_a > \gamma)\right)$$

And thus,

$$P(X_a > \gamma) < \delta \qquad \text{Given, by Hoeffding's inequality}$$

$$\underbrace{P\left(\bigcup_{a \in \mathcal{A}} (X_a > \gamma)\right) \leq \sum_{a \in \mathcal{A}} P(X_a > \gamma)}_{\text{Union bound}} < \sum_{a \in \mathcal{A}} \delta = |\mathcal{A}|\delta \qquad QED$$

Alternatively, we could also prove this via the set complement, and by using Bernoulli's inequality, under the same assumptions:

$$P(\exists a \in \mathcal{A} \quad s.t. \quad X_a > \gamma) = 1 - P\Big((X_{a_1} \leq \gamma) \bigcap (X_{a_2} \leq \gamma) \bigcap ... \bigcap (X_{a_{|\mathcal{A}|}} \leq \gamma)\Big) = 1 - P\Big(\bigcap_{a \in \mathcal{A}} (X_a \leq \gamma)\Big)$$

And since:

$$P(X_a > \gamma) < \delta$$
$$P(X_a \leq \gamma) = 1 - P(X_a > \gamma) > 1 - \delta$$

Therefore:

$$P\Big(\bigcap_{a \in \mathcal{A}} (X_a \leq \gamma)\Big) = \prod_{a \in \mathcal{A}} P(X_a \leq \gamma) > \underbrace{(1-\delta)^{|\mathcal{A}|} \geq 1 - |\mathcal{A}|\delta}_{\text{Bernoulli's inequality: } (1+x)^r \geq 1 + rx}$$
$$1 - P\Big(\bigcap_{a \in \mathcal{A}} (X_a \leq \gamma)\Big) < |\mathcal{A}|\delta \quad QED$$

(b) (20 pts) In order to claim any sort of relationship between the true means $\bar{R_a}$ of any two arms, because the true means are not directly observable, we must work in the case where the empirical means $\hat{R_a}$ accurately estimate the true means. This means it can only happen when $(|\bar{R_a} - \hat{R_a}| \leq \gamma) \forall a$.

Working from part (a):

$$P\Big(\bigcup_{a \in A} (X_a > \gamma)\Big) < |A|\delta$$
$$P\Big(\bigcap_{a \in A} (X_a \leq \gamma)\Big) > 1 - |A|\delta \quad \text{Taking the set complement}$$

By definition of the argmax, $\bar{R_{a^*}} \geq \bar{R_{a^\dagger}}$, and $\hat{R_{a^\dagger}} \geq \hat{R_{a^*}}$.

One possible ordering on the real number line is:

$$\underbrace{\hat{R_{a^*}} - - - \hat{R_{a^\dagger}} - - - \overbrace{\bar{R_{a^\dagger}} - - - \bar{R_{a^*}}}^{\epsilon}}_{<\gamma}$$

The positions of $\hat{R_{a^\dagger}}$ and $\bar{R_{a^\dagger}}$ could be swapped with one another but it makes no difference in this case. What we want to show is that for some γ, the maximal ϵ is equal to γ itself, without violating the inequalities specified by the argmaxes. ϵ is maximal when $\bar{R_{a^\dagger}} = \hat{R_{a^*}}$.

Substituting $\epsilon = \gamma$ into the given equation in (b), we get $\delta' = |\mathcal{A}|\delta$

$$\epsilon = \sqrt{\frac{\log(2|\mathcal{A}|/\delta')}{2n_{des}}}$$
$$n_{des} = \frac{\log(2|\mathcal{A}|/\delta')}{2\epsilon^2}$$