cs234 hw3

Jon Sondag

2022-01-23

1 Policy Gradient Methods

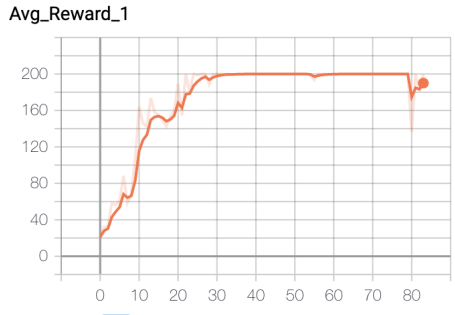
1.5

(a)(i) Cartpole with baseline, $r\_seed = 15$:

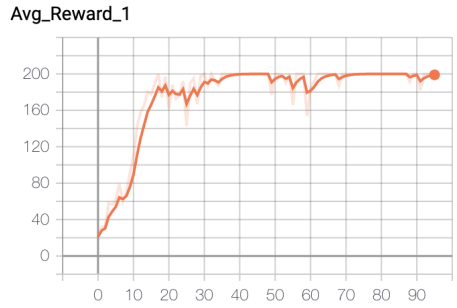
Cartpole no baseline, $r\_seed = 15$:

There's not a significant difference between the two reward graphs.

(a)(ii) Cartpole with baseline, $r\_seed = 12345456$:

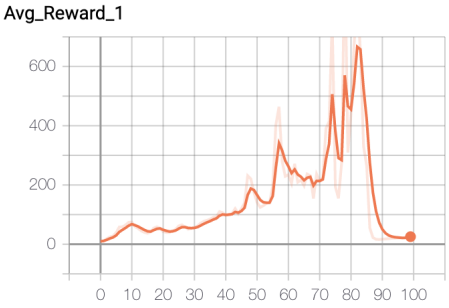


Cartpole no baseline, $r\_seed = 12345456$:

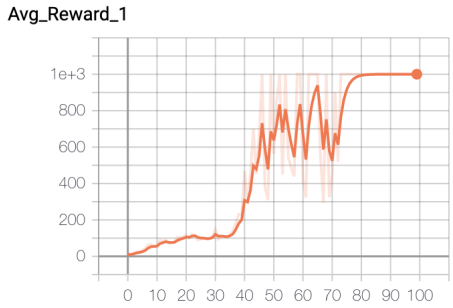


The no baseline reward chart shows slightly more wiggle but this may just be due to noise.

(b)(i) Pendulum with baseline, $r\_seed = 15$:



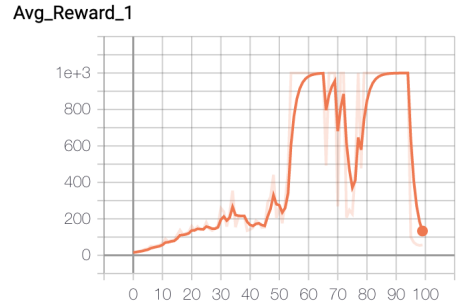
Pendulum no baseline, $r\_seed = 15$:



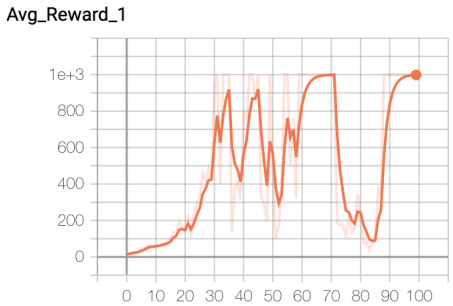
The no baseline reward ends at the maximum reward for the last 20+ batches.

The baseline reward ends up going down significantly over the last 10-20 batches. The baseline reward also takes longer to achieve the maximum reward, maybe in part due to its lower variance.

(b)(ii) Pendulum with baseline, $r\_seed = 8$:



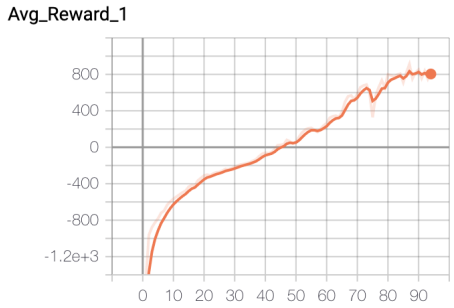
Pendulum no baseline, $r\_seed = 8$:



Again with this seed, the baseline reward takes longer to achieve the maximum reward, maybe in part due to its lower variance.

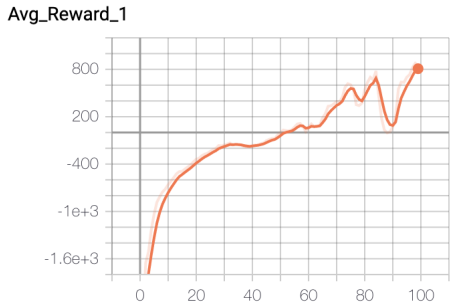
(c)(i) Cheetah with baseline, $r\_seed = 123$:

Cheetah no baseline, $r\_seed = 123$:

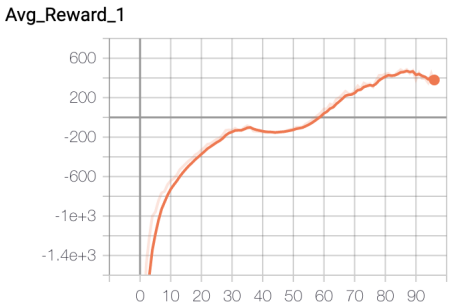


Surprisingly, the no baseline case has lower variance.

(c)(ii) Cheetah with baseline, $r\_seed = 15$:

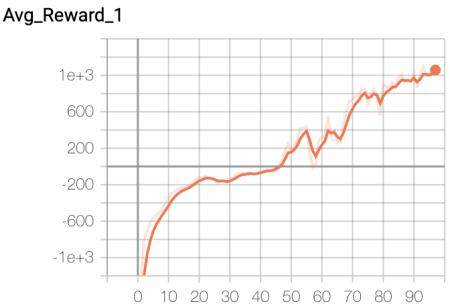


Cheetah no baseline, $r\_seed = 15$:

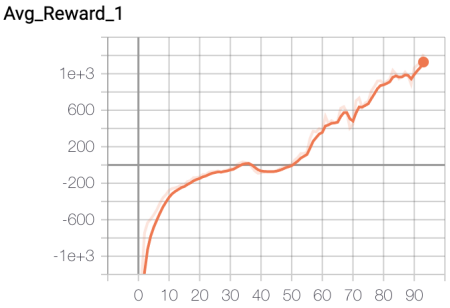


Surprisingly, the no baseline case has lower variance.

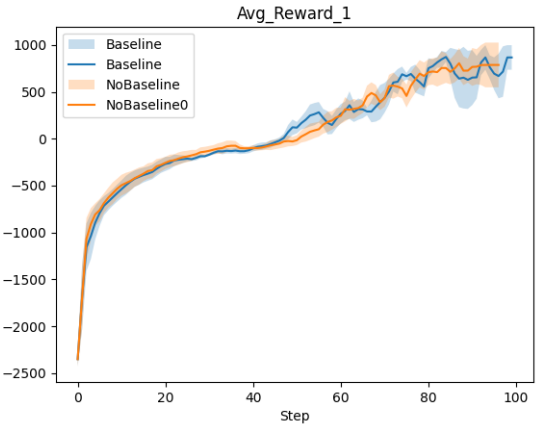
(c)(iii) Cheetah with baseline, $r\_seed = 12345456$:



Cheetah no baseline, $r\_seed = 12345456$:



Cheetah combined performance:



The no baseline cases have lower variance across the three random seeds. There is a bunch of randomness in the environment but it is surprising to see this result happen so consistently.

2 Best Arm Identification in Multi-Armed Bandit

(a) As stated in the problem, by Hoeffding's inequality:

$$\Pr\left(|\hat{X} - \bar{X}| > \sqrt{\frac{log(2/\delta)}{2n_e}}\right) < \delta$$

$$\text{So } \Pr\left(|\hat{r}_a - \bar{r}_a| \leq \sqrt{\frac{log(2/\delta)}{2n_e}}\right) \geq 1 - \delta$$

For each arm this is an independent probability so we can multiply events to get the joint probability:

$$\Pr\left(\forall a \in \mathcal{A}, |\hat{r}_a - \bar{r}_a| \leq \sqrt{\frac{log(2/\delta)}{2n_e}}\right) \geq (1-\delta)^{|\mathcal{A}|}$$

Therefore:

$$\Pr\left(\exists a \in \mathcal{A} \text{ s.t. } |\hat{r}_a - \bar{r}_a| > \sqrt{\frac{log(2/\delta)}{2n_e}}\right) < 1 - (1-\delta)^{|\mathcal{A}|} \leq |\mathcal{A}|\delta$$

Where the last inequality holds in the problem since $\delta > 0, |\mathcal{A}| >= 1$

(b) The most difficult case is where the suboptimal $|\mathcal{A}| - 1$ arms all have mean $\bar{r}_a = \bar{r}_{a^*} - \epsilon$.
If all arms differ from their actual mean by at most $\frac{\epsilon}{2}$ we will be all set. Based on the result from part (a),
$\Pr\left(\exists a \in \mathcal{A} \text{ s.t. } |\hat{r}_a - \bar{r}_a| > \sqrt{\frac{log(2/\delta)}{2n_e}}\right) < |\mathcal{A}|\delta$
We only care about the difference in one direction since the optimal arm's observed mean can't be too low and all other arms' means can't be too high. Therefore in our case the relevant formula is:
$\Pr\left(\exists a \in \mathcal{A} \text{ s.t. } \hat{r}_a - \bar{r}_a > \sqrt{\frac{log(2/\delta)}{2n_e}}\right) < \frac{|\mathcal{A}|\delta}{2}$
So in our case let $\delta' = \frac{|\mathcal{A}|\delta}{2}$, so $\delta = \frac{2\delta'}{|\mathcal{A}|}$. Then,

$$\frac{\epsilon}{2} > \sqrt{\frac{log(\frac{|\mathcal{A}|}{\delta'})}{2n_e}}$$

$$\frac{\epsilon^2}{4} > \frac{log(\frac{|\mathcal{A}|}{\delta'})}{2n_e}$$

Then,

$$n_e > \frac{2log(\frac{|\mathcal{A}|}{\delta'})}{\epsilon^2}$$