CS 330: Problem Set 3

Ji Won Park

October 26, 2020 (Due)

Colab notebook is available at

`https://colab.research.google.com/drive/1JfCYRAB4eUTD-lNICjB2B-FRdY9oqDuU?usp=sharing`.

1

(a) Figure 1 shows the success rate without goal conditioning, which remains very low.

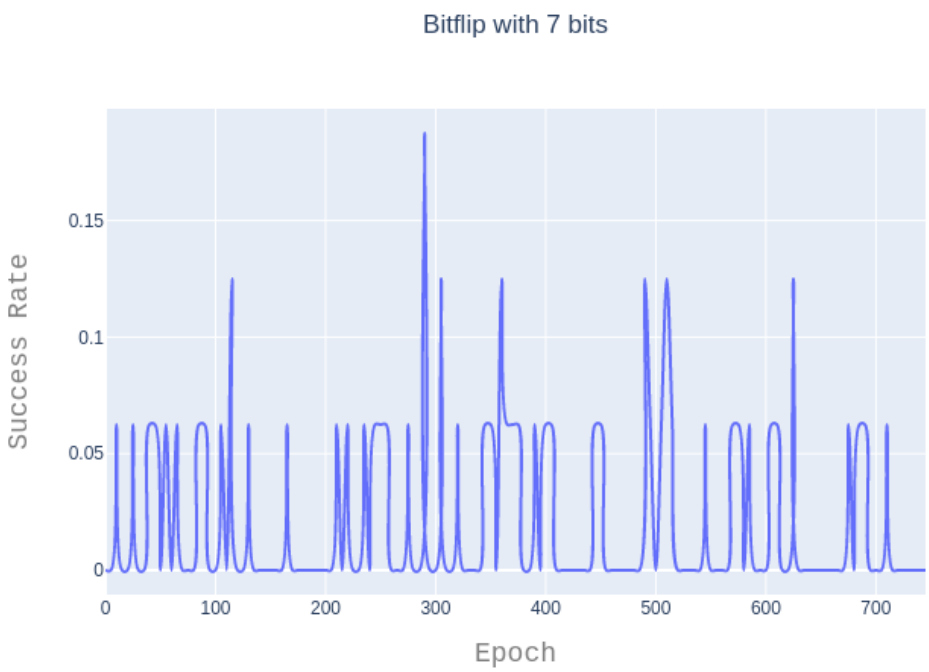

Figure 1: Success rate without goal conditioning

(b) See Colab. Note that I added a Boolean keyword argument to the `flip_bits` function, called `goal_conditioned`.

2

See Colab.

3

(a) Figure 2 shows the success rates with no HER and final HER, for 7 bits.

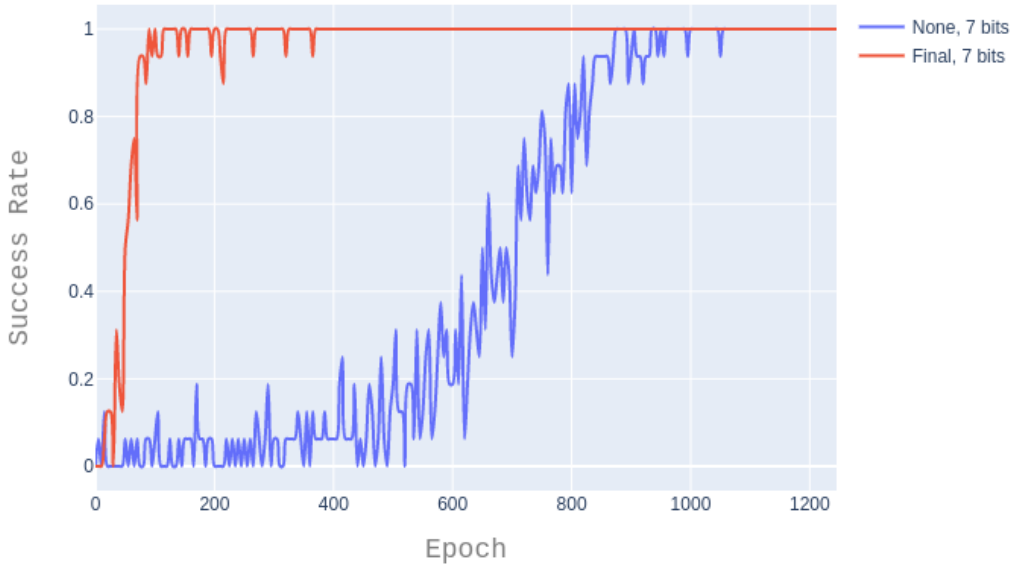

Figure 2: Success rates with no HER and final HER, for 7 bits

(b) Figure 3 shows the success rates with no HER and final HER, for 15 bits.

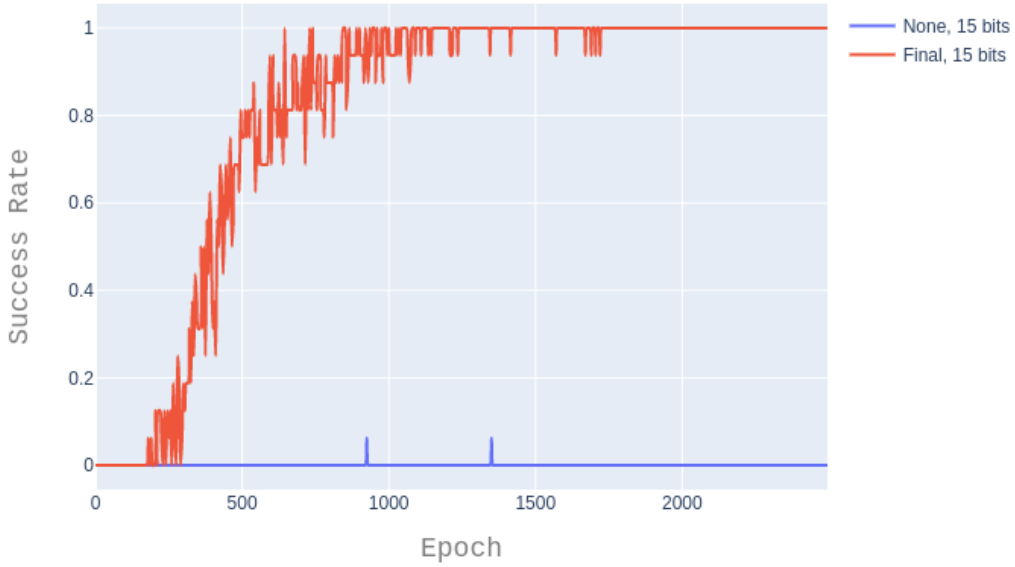

Figure 3: Success rates with no HER and final HER, for 15 bits

(c) Figure 4 shows the success rates with no HER and final HER, for 25 bits.

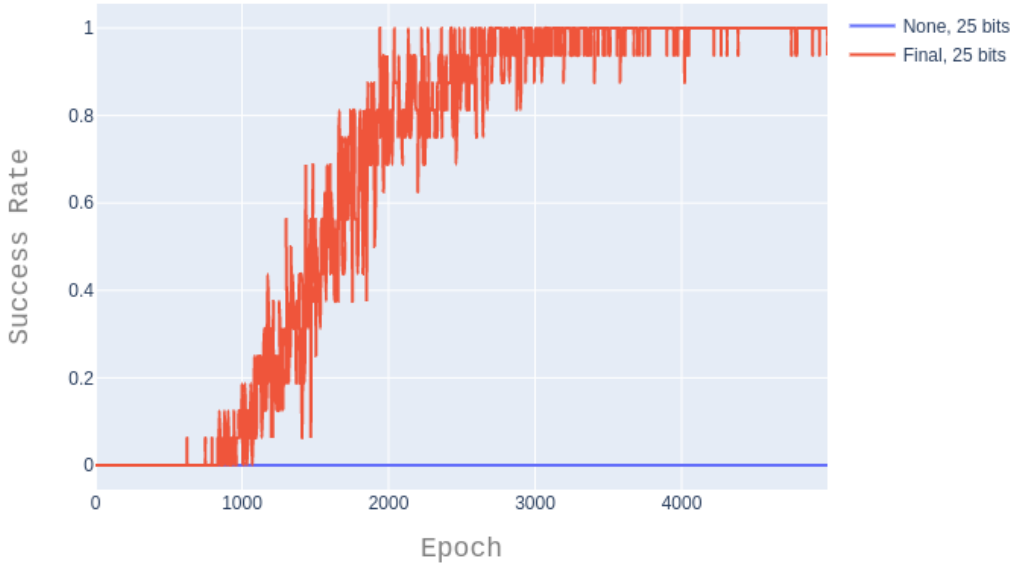

Figure 4: Success rates with no HER and final HER, for 25 bits

(d) Figure 5 shows the success rates with no HER, final HER, future HER, and random HER, for 15 bits. Figure 6 shows the same plot with smoothing applied.

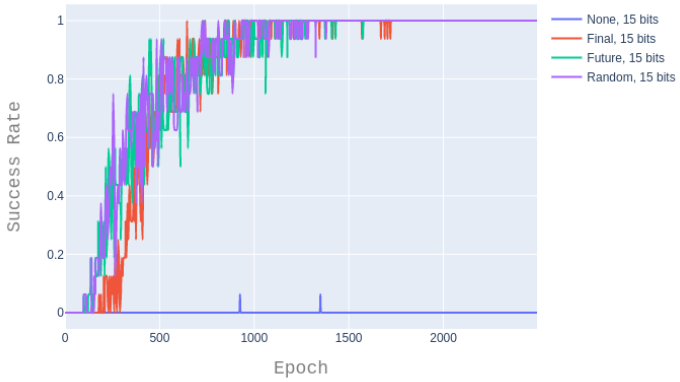

Figure 5: Success rates with no HER, final HER, future HER, and random HER, for 15 bits

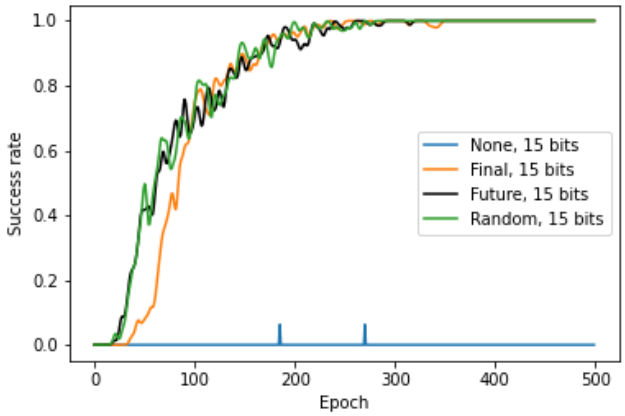

Figure 6: (Smoothed using Gaussian filters) Success rates with no HER, final HER, future HER, and random HER, for 15 bits

(e) In Part (a), there are only 7 bits so the run without HER is eventually able to reach a success rate of 1, just as the run with final HER does. But it does so more slowly, because it's rarer for the model to encounter tuples with non-negative rewards. The rate of tuples with non-negative rewards for the no-HER run is $2^{-7} \approx 0.008$ compared to $\frac{1}{7} \approx 0.14$ or better with HER.

In Part (b), with an increased number of bits (15), the success rate for the run without HER falls to 0, because the rate of tuples with non-negative rewards has decreased to $2^{-15} \approx 0.00003$ from $2^{-7} \approx 0.008$. The run with final HER is still able to reach a success rate of 1, though a bit slowly and with more variance than in the case of 7 bits given its non-negative-reward rate of order $\frac{1}{15} \approx 0.07$. The relabeling scheme exposes the model to more tuples with non-negative rewards from which to learn, so leads to faster/better learning.

In Part (c), when the number of bits increases even more to 25, the success rate for the run without HER flatlines at 0, the rate of tuples with non-negative rewards being even smaller ($2^{-25} \approx 3e-8$). The run with final HER reaches a success rate of 1 more slowly and with even higher variance than in the case of 15 bits, given its non-negative-reward rate of order $\frac{1}{25} \approx 0.04$.

In Part (d), we compare the run without HER with various HER relabeling schemes. The no-HER run is not able to learn, because of low rate of tuples with non-negative rewards to learn from ($2^{-15} \approx 0.00003$). The three HER schemes are comparable and all reach a success rate of 1, given the non-negative-reward rate of around $\frac{1}{15} \approx 0.07$. The future HER and random HER actually optimize slightly better than the final HER because the final HER only has one relabeled tuple whereas the random and future HER had 4 relabeled tuples in our setting, so the rates are higher for the random and future HER runs by a factor of 4. Another interpretation is that the most informative goals to replay are the states that are achieved in the future and ones that are encountered in the overall training procedure.

4

See Colab. Note that I added a Boolean keyword argument to the `run_sawyer` function, called `goal_conditioned`.

5

See Colab. Figure 5 shows the success rates with no HER and final HER, for Sawyer. As expected, the final-HER run learns better and faster (final success rate of around 0.9%) than the no-HER run (final success rate of around 0.3%). Although the reward function is now Euclidean, the explanation is similar to one given in Problem 3. The re-labeling scheme of HER encodes a concept of goal similarity, whereby the model can receive positive feedback on how much it has been able to achieve by the end of its episode even if it hasn't been able to completely reach the ultimate goal of moving to a specified set of coordinates.

The difference between no-HER and final-HER runs is not as stark as in the Bit Flip environment, however, because the reward function is continuous, being a Euclidean distance metric, rather than sparse. So there is still some learning signal in the tuples that are very far from being successful.

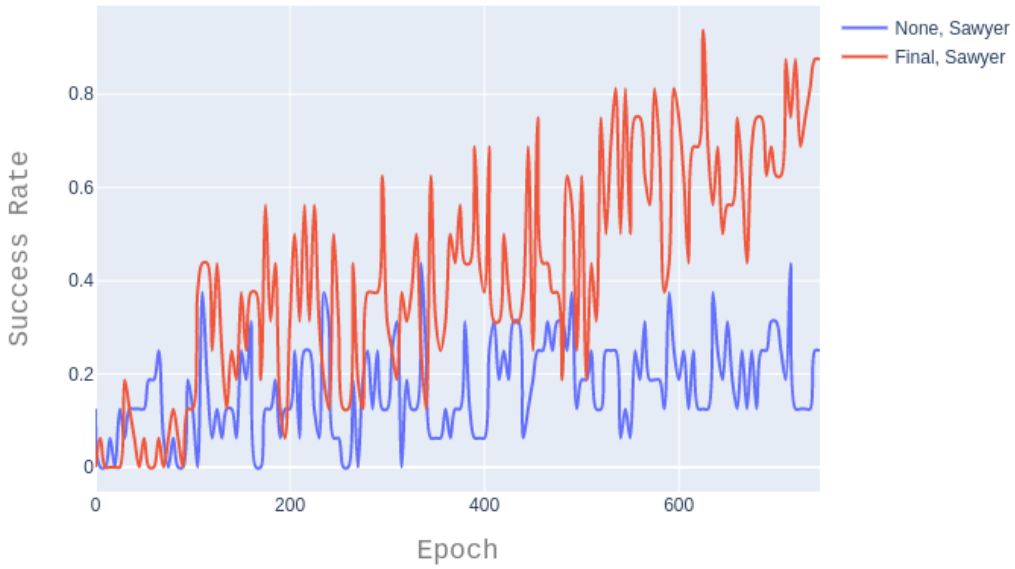

Figure 7: Success rates with no HER and with final HER, Sawyer