CS330 HW3

Yiheng Li {yyhhli}

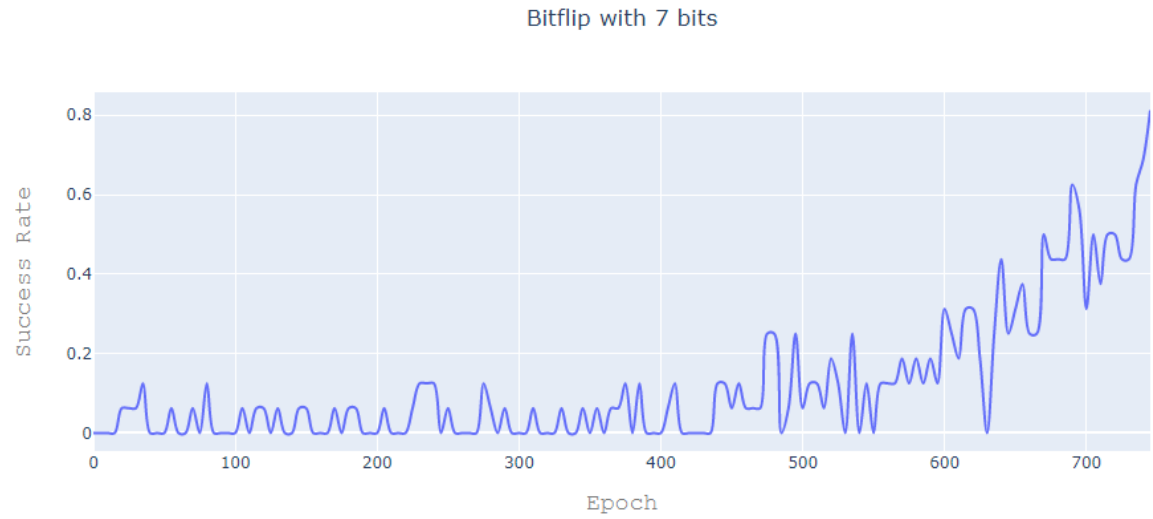
Link to my code

https://colab.research.google.com/drive/1QtQWnGgQKnq6yz-rUuwDaWlup3AGJTfs?usp=sharing

Problem 1

a) Plot the success rate of

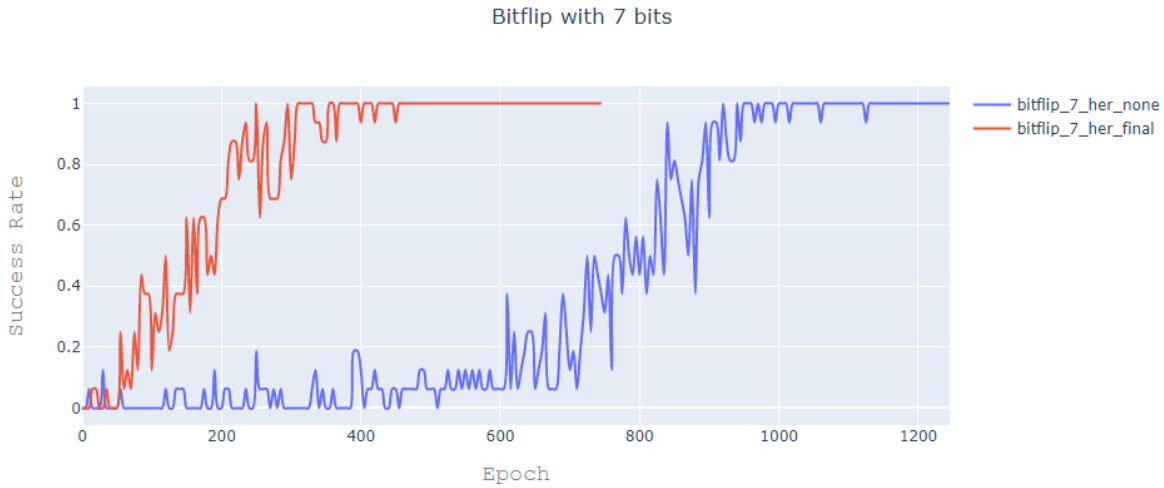
```
success_rate = flip_bits(num_bits=7, num_epochs=150, HER='None')
```



Problem 3

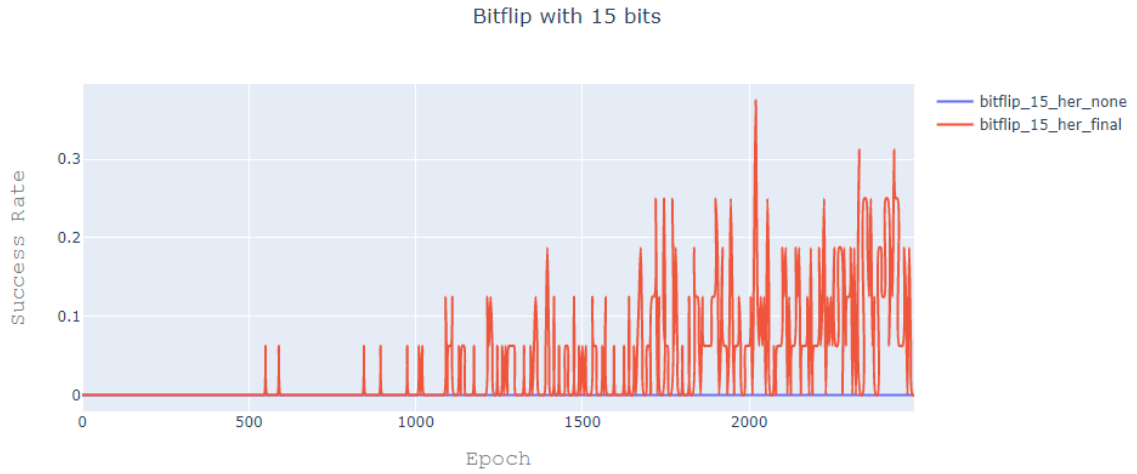
a) Plot the success rate of

```
success_rate = flip_bits(num_bits=7, num_epochs=250, HER='None')
success_rate = flip_bits(num_bits=7, num_epochs=150, HER='final')
```



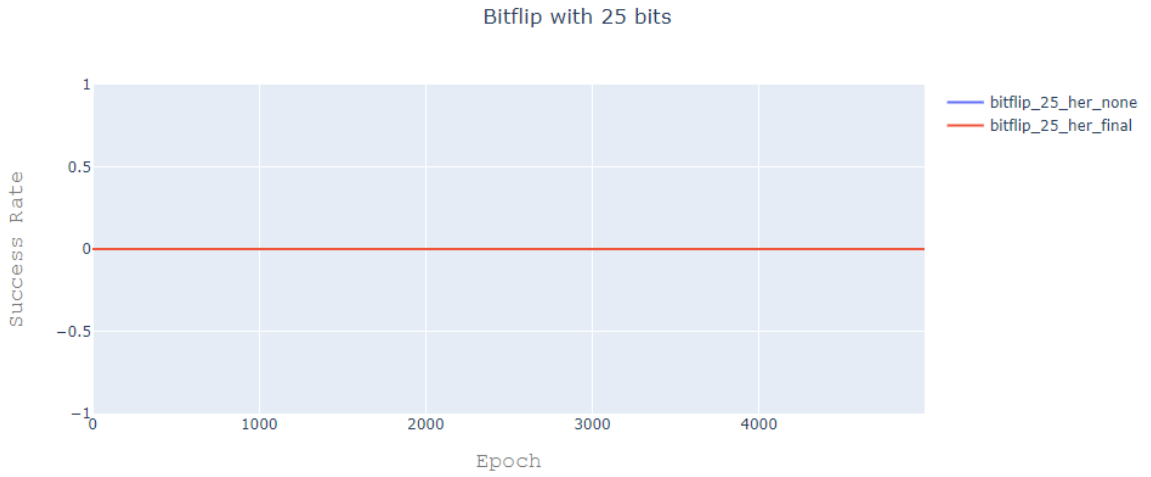
b) Plot the success rate of

```
success_rate = flip_bits(num_bits=15, num_epochs=500, HER='None')
success_rate = flip_bits(num_bits=15, num_epochs=500, HER='final')
```

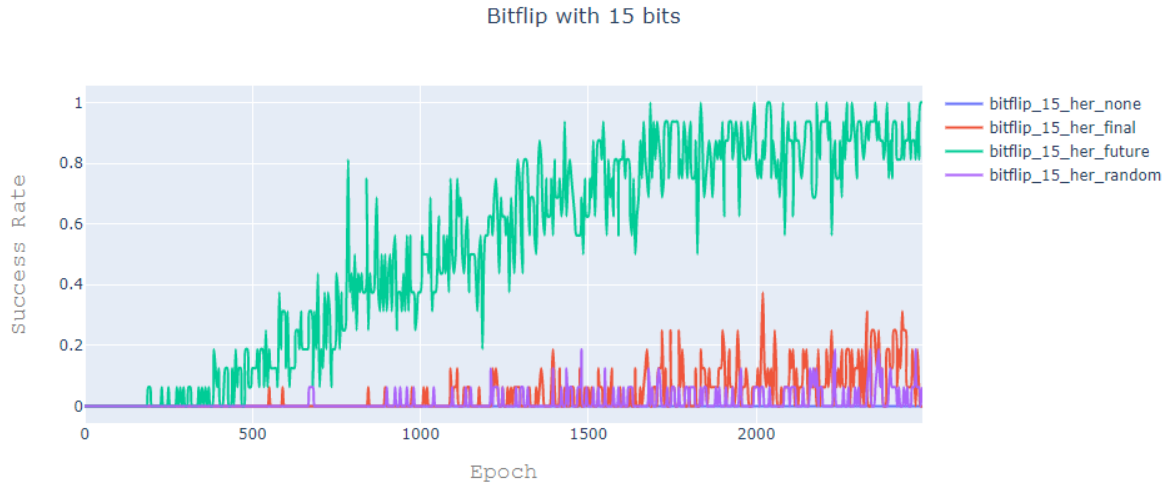


c) Plot the success rate of

```
success_rate = flip_bits(num_bits=25, num_epochs=1000, HER='None')
success_rate = flip_bits(num_bits=25, num_epochs=1000, HER='final')
```



d) Compare three versions of HER.



e) Explanation

For part (a). When bit is 7, both `HER=None` ('none' strategy) and `HER=final` ('final' strategy) can learn to achieve good performance at the end. However, `HER=None` takes more epochs to achieve similar performance of `HER=final`

For part (b). When bit is 15, 'none' cannot get any improvements during training. cause one experience of reward would take 2^{15} samples to get, which is too low a probability. And 'final' can still learn something, but the performance is worse compared to bit=7. The success rate for 'final' is never better than 0.4.

For part (c). When bit is 25, neither 'none' nor 'final' can learn. Because even for 'final', to reach the goal state of 'final' would take average approximately 12 step for any middle state. The success rates for both methods are always 0.

For part (d). Performance: 'future' > 'final' > 'random' > 'none'. Random method is somewhat detrimental to the learning process compared to 'final' or 'future', since it may label state that are not desired as goal state, but it still let the model to learn something compared to 'none'. 'Future' is better than 'final' because when bit is 15, 'future' can accelerate the learning more dramatically since it provides more "positive samples".

Problem 5

Due to HER, in each episode, the final state was relabeled and the reward is recalculated. Thus the model can learn not only how to succeed, but also how to get to some intermediate states, which is helpful for quicker converge of the model.