Carolyn Kao (chkao831@stanford.edu)

CME 241, Winter 2021

HW13

1. MC Prediction with GLIE pseudo

```
Sample kth episode using $\pi : \{S_1, A_1, R_2, \ldots, S_T\} \sim \pi$
For each state $S_t$ and $A_t$ in the episode,
    • $N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$
    • $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)} (G_t - Q(S_t, A_t))$
Improve policy based on new action-value function
$\epsilon \leftarrow 1/k$
$\pi \leftarrow \epsilon-$greedy $(Q)$
```

2. Sarsa pseudo

```
Initialize $Q(s, a), \forall s \in S, a \in A(s)$
$Q(\text{termianl}, .) = 0$
For each episode:
    • Initialize $S$
    • Choose $A$ from $S$ using $\epsilon-$greedy
    • For each step of episode:
        – Take action $A$, observe $R, S'$
        – Choose $A'$ from $S'$ using $\epsilon-$greedy
        – $Q(S, A) = Q(S, A) + \alpha[R + \gamma Q(s', A') - Q(S, A)]$
    • until $S$ is terminal
```

4. Thoughts outline:
By each end of day, we sketch out the following sequence in order as the following, in which sub points are the states,

- sell stock from previous day
    - observe t
    - observe net cash by selling
    - observe backlogged withdrawals
    - observe value of a single share of the stock
    - observe money currently owed to the bank

- make a decision to increase or decrease the quantity of cash-borrowing
    - observe t
    - observe net cash by selling
    - observe backlogged withdrawals
    - observe value of a single share of the stock
    - observe money currently owed to the bank
- make a decision to purchase a certain quantity of stock for the next day
    - observe t
    - observe net cash by selling
    - observe backlogged withdrawals
    - observe value of a single share of the stock
    - observe money currently owed to the bank

Then, the action is consist of the further increase or decrease in our liability and the number of shares of stock to purchase on the next day.

Furthermore, the Reward is 0 for $t \in [1, T-1]$, otherwise, it's defined by the bank owner's risk-aversion throughout time of liability til T.

Formulating this as MDP might requires to consider the fact of large action space and large state space, resulting in the potential deployment of the policy gradient algorithm and actor-critic algo further.