Experience replay

running new samples each time is:

- wasteful
- biassed by the current strategy
- not giving the maximum amount of information

EXPERIENCE REPLAY

the buffer has fixed size

A buffer stores step: (s, a, r, s')

Two actions:

(1) sample a trajectory from the environment and add each step independently to the buffer

(2) sample from the buffer to update

PRIORITISED EXPERIENCE REPLAY in Q-LEARNING:

when we get a step

(s, a, r, s') we perform the update:

$$q(s,a) = q(s,a) + \alpha\left(r + \gamma \cdot \max_{a' \in A} q(s',a') - q(s,a)\right)$$

$$= B$$

indicator of how surprising is (s, a, r, s')

Algorithm:

(1) Sample trajectories and add each step (s, a, r, s') to the buffer with bias B

(2) Sample from the buffer with probability

$$\frac{\exp(B)}{\sum_{B'} \exp(B')}$$

and update with the corresponding step

To do: breaking trajectories into

Important! [illegible]

step removes the strategy bias!