Reinforcement Learning: Homework 01

Guilherme Sant' Anna Varela

December 2019

1 Multi-armed bandits

The implementation of the experiment is such that: At the beginning of each trial the mean for each bandit is sampled from a Gaussian distribution with zero mean and unit variance those are fixed by the remainder of the trial and every subsequent pull will yield a random reward drawn from a Gaussian distribution with that same mean and unit variance. The trial consists of 1000 steps, of such pulls in which the agent observes the reward and updates it's estimates for the true mean. There are 10 bandits representing distinct choices. Five strategies are tested, which are policies for choosing the next action, for a total 1000 steps with the goal of maximizing the agent's total reward. These setup is repeated for 2000 iterations and the reward with respect to each policy is averaged for each time step.

A sketch of the solution is provided here and the full code in another file called *q1.py*

```
# The update estimate for each agent
def update(self, a, r):
    i = a - 1
    self.mu[i] = \
        (self.steps[i] * self.mu[i] + r) / (self.steps[i] + 1)
    self.steps[i] += 1

# The action choosen by the greedy agent
def greedy_action(self):
    return np.argmax(self.mu) + 1

# The action choosen by the eps greedy agent
def a(self):
    ga = self.greedy_action():
    if np.random.rand() < self.eps:
        ras = \
        [a for a in range(1, len(self.mu) + 1) if a != ga]
        return np.random.choice(ras)
    return ga

# The action choosen by the ucb agent
def a(self):
    t = np.sum(self.steps)
    ucb = self.mu + self.c * np.sqrt(np.log(t) / self.steps)
    return np.argmax(ucb) + 1
```

The strategies are described as follows: (1) **Greedy 0** consisting of initializing *a priori* estimates for the distribution's mean as zero and always choosing

the action that maximizes the reward thereafter. (2) **Greedy 5** consisting of initializing *a priori* estimates for the distribution's mean as 5 and always choosing the action that maximizes the reward thereafter. (3) **Greedy 10%** consisting of initializing *a priori* estimates for the distribution's mean as zero and 90% of the steps choosing the action that maximizes the reward and the other 10% making sub-optimal decisions. (4) **Greedy 1%** consisting of initializing *a priori* estimates for the distribution's mean as zero and 99% of the steps choosing the action that maximizes the reward and the other 1% making sub-optimal decisions. (5) **UCB** or the upper confidence consisting of an adaptive strategy that balances exploration and exploitation. The results can be seen on Figure 1 in regards to the reward collected in each step and in Figure 2 as the overall proportion of the time steps the optimal action was taken, both results are averaged over 2000 trials.

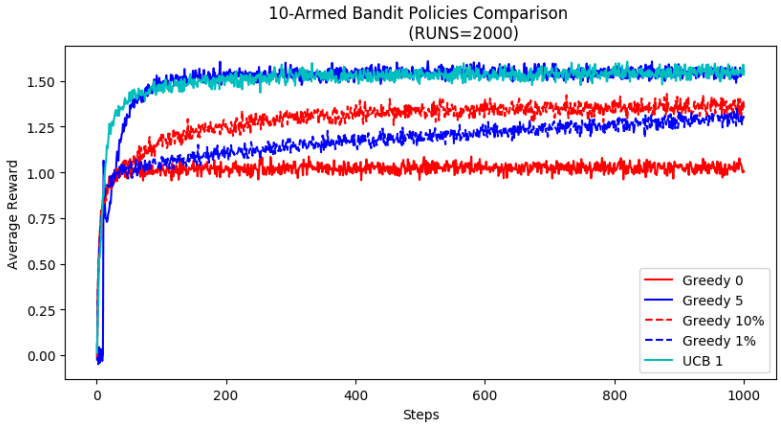

Figure 1: Steps vs Average reward along 2000 trial runs. The 5 strategies are tested and **UCB** converges faster.

Greedy 0 is the worst performing strategy as it doesn't sufficiently explore the search space towards the optimal action, it has the upside of being the simpler strategy but can get stuck on partially good yet sub optimal choices. For instance suppose that are only three bandits with means -1, 0 and 1, if the first bandit is chosen even if positive rewards are yield at first, eventually the trials will make the rewards converge to a negative mean and the agent might decide to try the second bandit. While the estimate for the second's bandit mean remains above the estimate for the first the agent will keep pulling the second lever indefinitely never trying the true optimal option which is the third.

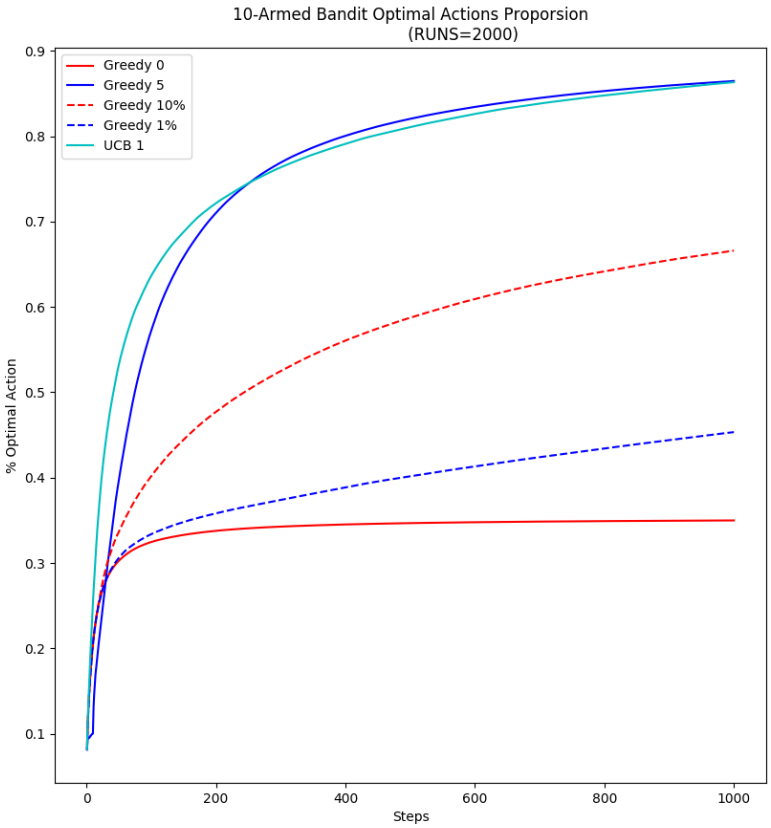

Figure 2: Steps vs % Optimal actions along 2000 trial runs. The 5 strategies are tested and **UCB** chooses more frequently the optimal action, earlier.

Greedy 5 also known as optimistic initial value is the top performing strategy, it's closed related to the **Greedy 0**, differing only by the initial estimate for the bandits' mean. It over-performs because it allows for more exploration during the first iterations, as long as the condition that the initial guess, 5 in this case, be sufficiently high in comparison to the true mean. And it promotes exploration by allowing the agent to be 'disappointed' during the first rounds as every reward seems too low with respect to the underlying optimistic assumption for the mean. The strategies' upside is it's simplicity but it requires one to at least have a guess over the upper bound for the quantity being estimated. Furthermore, for non stationary problems it fails to explore at the end of the trial.

Both **Greedy 10%** and **Greedy 1%** belong to the ϵ-Greedy set of strategies, in which ϵ denotes a probability that the agent will choose a sub-optimal action. While it promotes a constant probability for exploration, without any knowledge with regards to a bound for the quantity being estimated, it has been shown to converge slowly which is specially made worse on large action spaces where the probability of the maximum action being chosen is further divided by alternative sub-optimal actions. That's the most plausible interpretation as why **Greedy 10%** is better **Greedy 1%** and both still show a positive slope towards choosing the best action indicating improvements, neither is the best performing.

The last strategy is called **UCB** which stands for Upper Confidence Bound, it's also the top performing method and as with **Greedy 5** it promotes a higher level of exploration during the beginning while showing a high level of performance during the later stages of the trial. It does so by transforming the action value function by inserting a bias to the least tried actions and reducing such a bias every time such actions are explored – the bias is also regulated by the overall number or trials t and is proportional to constant factor c. Unlike **Greedy 5** optimistic initial values it doesn't require bound assumptions over quantities and it also allows exploration on later stages of the trial.

2 Gambler's Problem

A sketch of the solution is provided here and the full code in another file called *q2.py*

```
THRESHOLD = 1e-8
GOAL = 100
GAMMA = 1      # this is the discount factor
P_H = 0.4      # probability that a coin toss will be heads
P_T = 1 - P_H  # probability that a coin toss will be tails

# States: that are effectively visited by the agent
# S = (1, 2, 3, ..., 99)
S = np.arange(1, GOAL)
# Value: # states + "2 special states"
# V = (0, 1, 2, ..., 99, 100)
V = np.zeros((GOAL + 1,), dtype=np.float)
# Rewards: always zero except on GOAL
R = np.zeros((GOAL + 1,), dtype=np.float)
R[GOAL] = 1

while delta > THRESHOLD or sweeps < 32:
delta = 0
# Iterate forwards
for state in S:
    v = V[state]
    # maximum bet to reach target
    max_stake = min(state, GOAL - state)
    stakes = np.arange(1, max_stake + 1)

    H = np.array([
       R[state + stake] + GAMMA * V[state + stake]
       for stake in stakes
    ], dtype=np.float)

    T = np.array([
       R[state - stake] + GAMMA * V[state - stake]
       for stake in stakes
    ], dtype=np.float)

    E = roundup(P_H * H + P_T * T)

    V[state] = np.max(E)
    PI[state - 1] = stakes[np.argmax(E)]
    delta = max(delta, np.abs(V[state] - v))
```

The results from the algorithm can be seen on Figures 3 as the value function or probability of winning and 3 as the stakes the gambler has to make at each turn. Together they illuminate the relationship between the wealth, stakes and probability of winning. In regards to gambles made on such scenarios, it has been shown that if the odds are stacked against the player, our case $ph = 0.4$ it's worth to play as few turns as possible, by making larger bets i.e the gambler might get luck in the short-term but eventually in the long run the probable outcome will be ruin. Conversely if the odds are in favor it pays out to play as many turns as possible, by preserving the capital and making small bets, it's because in the short-term we might get unlucky, but in the long term the gambler will probably benefit from the favorable odds.

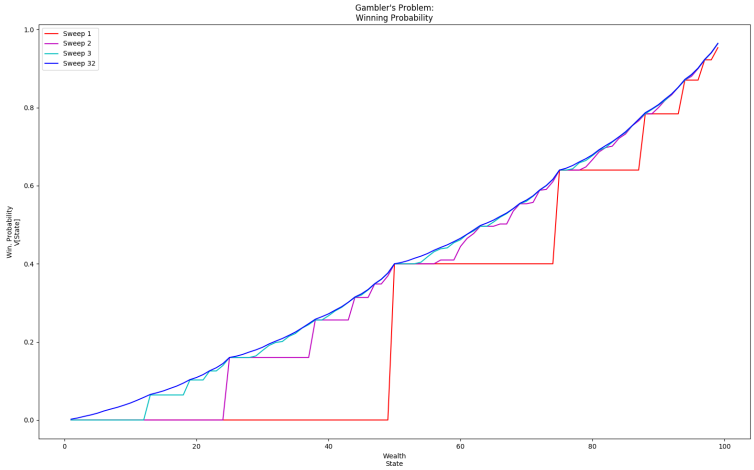

Figure 3: Wealth vs Winning Probability. The states represent the current level of wealth and the value function represents the probability of reaching the goal.

In this experiment the gambler plays conservatively most of the times and aggressively at each power of two from the goal $1/8, 1/4, 1/2, 3/4$. At wealth 50 he bets it all (Figure 4) and has a probability of winning equal to the probability of the favorable flip, $ph = 0.4$, (Figure 3). The strategy seems to get to some 'milestone' states 12, 25, 50, 75 where the agents either bets it all or a large portion of it's wealth to either get to the next aggressive state or reach the goal outright.

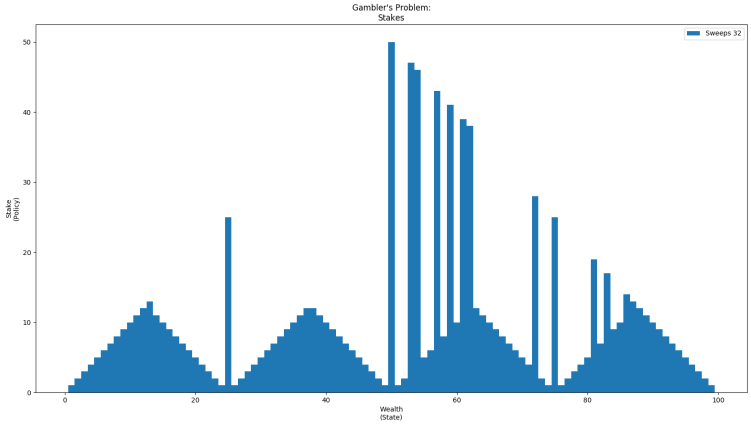

Figure 4: Wealth vs stakes. The states represent the current level of wealth and the actions represent the stake at each turn.

The one feature is the numeric instability for the states larger than 50, Figure 3 should be symmetric around that state. The it's effect is such that the agent takes more aggressive bets at higher levels of wealth.

3 Convergence of value iteration

Considering the framework is given for question 3 the compact version of v_π is defined by the following recursive equation:

$$\mathbf{v}_\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}_\pi \tag{1}$$

Our proof is by construction, first we show that there's a sequence $\{w\}_g$ which is Cauchy and following the result of Banach fixed point theorem it converges to a unique fixed point in space which is v_π itself.

Take the following two arbitrary points in space $\mathbf{v}^k, \mathbf{u}^q \in \mathbb{R}^{|\mathcal{S}|}$ and define the following recursive sequences:

$$\mathbf{v}^{k+1} = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}^k \tag{2}$$

$$\mathbf{u}^{q+1} = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{u}^q \tag{3}$$

We define $\{w\}_g$ as the following:

$$\{w\}_g = \{v\}_g - \{u\}_g \tag{4}$$

Replacing definitions (2) and (3) on (4):

$$\{w\}_g = \gamma \mathbf{P}_\pi \left(\{v\}_g - \{u\}_g\right)$$

Where $\{w\}_g$ follows a stricter version of definition (9), since $\mathbf{r}_\pi = 0$. We take the mapping $\mathbf{T} = \gamma \mathbf{P}_\pi$ then:

$$\begin{aligned} ||\mathbf{T}\{w\}_g||_2 &= ||\gamma \mathbf{P}_\pi \{w\}_g||_2 \\ &= \gamma ||\mathbf{P}_\pi \{w\}_g||_2 && (5) \\ &\leq \gamma ||\mathbf{P}_\pi||_2 ||\{w\}_g||_2 && (6) \\ &\leq \gamma ||\{w\}_g||_2 && (7) \end{aligned}$$

Where (5) follows from the constraint $0 < \gamma < 1$, (6) follows from the dot product property and finally, (7) from the fact that $||\mathbf{P}_\pi||_2$ is a probability matrix which the determinant is in the unit circle. Then:

$$||\mathbf{T}\{w\}_g||_2 = ||\mathbf{T}\{v\}_g - \mathbf{T}\{u\}_g)||_2 \leq \gamma ||\{v\}_g - \{u\}_g)||_2$$

$\mathbf{T}$ is a contraction, it has one and only one fixed point, this point must be $\mathbf{v}_\pi$.

4 Temporal Difference Methods

The proof follows by construction more or less in the manner seem on question 3. Our starting point is taking the definition for the temporal operator $\mathbf{T}^{\lambda}$

$$\mathbf{T}^{(\lambda)}\mathbf{v} = \sum_{n=0}^{\infty}(\lambda\gamma\mathbf{P})^{n}\left[\mathbf{r} + \gamma\mathbf{P}\mathbf{v} - \mathbf{v}\right] + \mathbf{v} \tag{8}$$

We proceed to expand the terms using the facts presented i.e:

$$\begin{aligned}
&||\mathbf{P}||_2 < 1 \\
&0 < \gamma < 1 \\
&0 < \lambda < 1 \\
&\sum_{n=0}^{\infty}(\lambda\gamma\mathbf{P})^{n} = \left[\mathbf{I} - (\lambda\gamma\mathbf{P})^{n}\right]^{-1}
\end{aligned}$$

We can re-write definition 8 as follows:

$$\begin{aligned}
\mathbf{T}^{(\lambda)}\mathbf{v} &= \sum_{n=0}^{\infty}(\lambda\gamma\mathbf{P})^{n}\left[\mathbf{r} + \gamma\mathbf{P}\mathbf{v} - \mathbf{v}\right] + \mathbf{v} \\
&= \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\left[\mathbf{r} + \gamma\mathbf{P}\mathbf{v} - \mathbf{v}\right] + \mathbf{v} \\
&= \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\mathbf{r} + \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\left[\gamma\mathbf{P}\mathbf{v} - \mathbf{v}\right] + \mathbf{v} \\
&= \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\mathbf{r} + \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\left[\gamma\mathbf{P}\mathbf{v} - \mathbf{I}\right]\mathbf{v} + \mathbf{v} \\
&= \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\mathbf{r} + \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\left[\gamma\mathbf{P} - \mathbf{I} + \mathbf{I} - (\lambda\gamma\mathbf{P})\right]\mathbf{v} \\
&= \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\mathbf{r} + \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\left[\mathbf{I} - \lambda\mathbf{I}\right]\gamma\mathbf{P}\mathbf{v}
\end{aligned} \tag{9}$$

We define in a analagous manner the following transformation:

$$\mathbf{T}^{(\lambda)}\mathbf{u} = \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\mathbf{r} + \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\left[\mathbf{I} - \lambda\mathbf{I}\right]\gamma\mathbf{P}\mathbf{u} \tag{10}$$

Subtracting the transformation (10) from (9):

$$\mathbf{T}^{(\lambda)}\mathbf{v} - \mathbf{T}^{(\lambda)}\mathbf{u} = \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\left[\mathbf{I} - \lambda\mathbf{I}\right]\gamma\mathbf{P}(\mathbf{v} - \mathbf{u}) \tag{11}$$

$\mathbf{T}^{(\lambda)}$ is a contraction with respect to the sup-norm if and only if:

$$||\mathbf{T}^{(\lambda)}\mathbf{v} - \mathbf{T}^{(\lambda)}\mathbf{u}||_{sub} \leq \gamma||\mathbf{v} - \mathbf{u}||_{sub} \tag{12}$$

Then

$$\begin{aligned}
||\mathbf{T}^{(\lambda)}\mathbf{v} - \mathbf{T}^{(\lambda)}\mathbf{u}||_{sub} &= ||\left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\left[\mathbf{I} - \lambda\mathbf{I}\right]\gamma\mathbf{P}(\mathbf{v} - \mathbf{u})||_{sub} && (13)\\
&= \gamma||\left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\left[\mathbf{I} - \lambda\mathbf{I}\right]\mathbf{P}(\mathbf{v} - \mathbf{u})||_{sub} && (14)\\
&\leq \gamma||\left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\left[\mathbf{I} - \lambda\mathbf{I}\right]||_{sub}||\mathbf{P}(\mathbf{v} - \mathbf{u})||_{sub} && (15)\\
&\leq \gamma||\mathbf{P}(\mathbf{v} - \mathbf{u})||_{sub} && (16)\\
&\leq \gamma||\mathbf{P}||_{sub}||(\mathbf{v} - \mathbf{u})||_{sub} && (17)\\
&\leq \gamma||\mathbf{v} - \mathbf{u}||_{sub} && (18)
\end{aligned}$$

We have to show that sup-norm of (15) is less than unity to that lets define the following function $f : (0,1) \rightarrow \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, we then proceed to show that $||f(\lambda)||_{sup} < 1$

$$f(\lambda) = \left[\mathbf{I} - (\lambda\gamma\mathbf{P})\right]^{-1}\left[\mathbf{I} - \lambda\mathbf{I}\right] \tag{19}$$

The following is true:

$$f(0^{+}) = \lim_{\lambda_n \rightarrow 0} = \mathbf{I} \tag{20}$$

$$f(1^{-}) = \lim_{\lambda_n \rightarrow 1} = 0 \tag{21}$$

$$\forall\lambda \rightarrow \frac{df}{d\lambda}(\lambda) < 0 \tag{22}$$

From (20) the limit from right down to 0 is $\mathbf{I}$, the limit from the left up to 1 is 0 (21) and finally the function is always decreasing (22) and hence the maximum is given at limit (20). Completing the proof.