Reinforcement Learning

Homework 1 Solution

Samuel Simão Canada Gomes, 76415

December 11, 2018

Solution for "Multi-armed bandits"

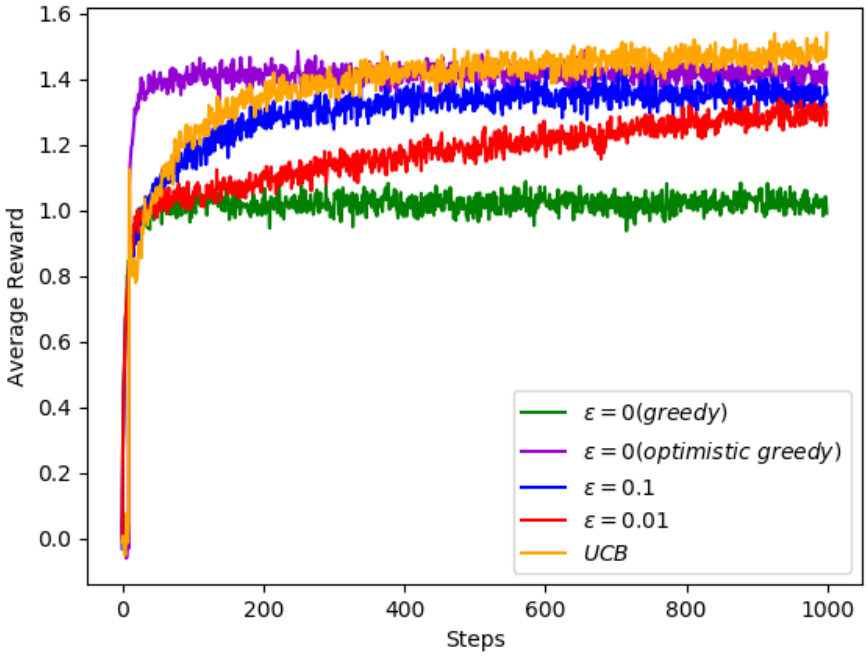

Figure 1: Chart comparing several RL action selection strategies.

The plotted charts are included in Figure 1. Considering the highest possible reward for this problem to be (≈ 1.55), several tendencies inherent to each policy can be observed:

- The greedy policy selects only the actions which seem best (have the higher action-value estimates) in each time step and therefore often gets stuck on exploiting sub-optimal actions. This makes the average rewards converge to a low value (≈ 1.0);
- In optimistic greedy, the high initial estimates promote early exploration as all the unvisited actions seem better than the ones already visited. This makes all the actions be visited several times and therefore the estimates converge to a higher value compared to the plain greedy strategy (≈ 1.41);

- As the ε greedy strategies consider exploration occasionally (according to the probability of exploration ε), they avoid being stuck exploiting a specific action. This allows them to outperform the plain greedy strategy, with average rewards of ≈ 1.2 to 1.3. However, such strategies cannot surpass the optimistic greedy, which only exploits certain high reward actions.

  With $\varepsilon = 0.1$, the exploration is faster, and therefore better actions are found early. This makes the early average rewards increase at a higher pace than with $\varepsilon = 0.01$. Nevertheless, because ε greedy with $\varepsilon = 0.01$ exploits each greedy action more frequently than ε greedy with $\varepsilon = 0.1$, after the 1000 steps, as better actions are visited, both ε greedy strategies get close rewards.

- UCB differs from the other strategies, because instead of only considering the action-value estimates, it has in account the uncertainty associated to those estimates. More specifically, it chooses the actions which have good estimates and are less visited, because despite not being the immediate optimal ones, these have the potential of leading to higher rewards in the future. A c tweaking parameter is included in the UCB algorithm to represent the importance given to the estimates uncertainties. In the simulations, a value of 2.0 was used for c. This is a commonly used value for this parameter.

  Like what happens in the optimistic greedy, UCB tests all actions early (but in UCB this is due to the next action calculations instead of high initial estimates). The initial choice of certain high rewarding actions explains a local maximum in the first steps. As the steps increase, the average reward converges to a nearly optimal, final value of ≈ 1.51. Such value is higher compared to the ones of the other strategies, namely the ε-greedy. Although UCB considers exploration like ε-greedy, such exploration is guided, not random. This makes the algorithm properly exploit the real best actions.

Solution for "The gambler problem"

In order to model the problem as an MDP:

- The states were considered to be the different possible amounts of money which can be kept by the gambler in each moment (101 possibilities as the gambler can keep from 0 $ to 100 $). The actions (transitions between the states) were modeled as the amounts of money the gambler can bet (99 possibilities as the gambler can bet from 1 $ up to 99 $).

- The transition probabilities for each action were defined using the following values: (0.4 for every transition to *future kept money = kept money + bet amount* and 0.6 for every transition to *future kept money = kept money - bet amount*). A probability of 1.0 was given to transitions to *future money = kept money* (transitions to the same state) for every impossible action. Impossible actions for this problem include when the gambler bets more than what he has (*bet amount > kept money*) or when the money he bets possibly makes the kept money be higher than 100 $ (*kept money + bet amount > 100 $*).

- The rewards were distributed as: 1.0 if *kept money = 100 $* (the gambler wins the game) and 0.0 otherwise[1].

The results from running the asked simulations are included in the charts of Figures 2 and 3. Figure 2 shows big jumps in the value function estimates of lower iterations, as the value function which is being computed is still a rough approximation of the real value function. In the first iteration (shown in blue), only in the state *kept money = 50 $* can the function increment its value, because betting 50 $ is the only direct way of receiving a reward (winning the game)[2]. In the second iteration (shown in red), the function also increments its value in other states like *kept money = 25 $*, as the algorithm accounts for the possibility of the gambler getting a reward later[3]. In subsequent iterations, the values are updated with an evergrowing future possible rewards in mind, which makes the estimated value function smooth out and approximate to the real value function. The value function computed at the last iteration is included in Figure 2, in black. Based on such function, the optimal policy (optimal bet amounts for each money currently kept by the gambler) was also computed. It is presented by the chart of Figure 3.

[1] As the problem defined, we only want the gambler to decide how much to bet, not in which outcome to bet. This can be ensured by rewarding the gambler only when he wins, not when specific interim states are reached.

[2] The function value in this state corresponds to 1.0*0.4 = 0.4: the value of the final state which equals to the reward obtained when winning the game (1.0) multiplied by the probability of betting *50 $* and receiving *50 $* (0.4).

[3] The function value in this state corresponds to 0.4*0.4 = 0.16: the value of the state *kept money = 50 $* (0.4) multiplied by the probability of reaching it, by betting *25 $* and receiving *25 $* (0.4).

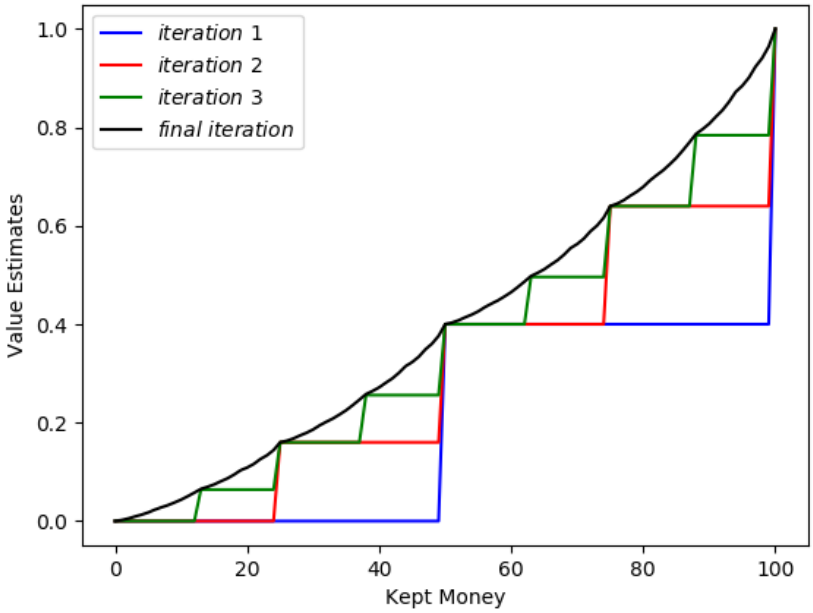

Figure 2: Value function approximation computed by the value iteration algorithm applied to the gambler problem.

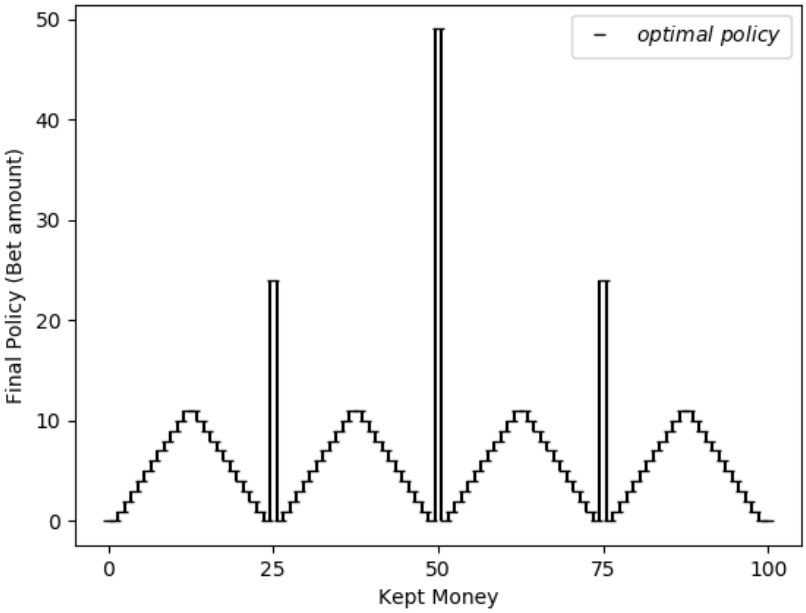

Figure 3: The estimated optimal policy, which corresponds the optimal bet amounts for each money currently kept by the gambler.

Solution for "Convergence of value iteration"

First, we will assume that the value iteration update function can be defined as an operator $\mathbf{T}$. If we consider two $\mathbf{v}_\pi$ value function estimates, $\mathbf{x}$ and $\mathbf{y}$, we can prove that $\mathbf{T}$ is a contraction by proving that $\|\mathbf{T_x} - \mathbf{T_y}\| \leq \alpha \|\mathbf{x} - \mathbf{y}\|$, with an $\alpha < 1$. We can write $\mathbf{T_x}$ and $\mathbf{T_y}$ as:

$\mathbf{T_x} = \mathbf{r}_\pi + \gamma * \mathbf{P}_\pi * \mathbf{x}$
$\mathbf{T_y} = \mathbf{r}_\pi + \gamma * \mathbf{P}_\pi * \mathbf{y}$

Now, we can reduce $\|\mathbf{T_x} - \mathbf{T_y}\|$:

$\|\mathbf{T_x} - \mathbf{T_y}\| \leftrightarrow$
$\|(\mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{x}) - (\mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{y})\| \leftrightarrow$
$\|\mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{x} - \mathbf{r}_\pi - \gamma \mathbf{P}_\pi \mathbf{y}\| \leftrightarrow$
$\|\cancel{\mathbf{r}_\pi} + \gamma \mathbf{P}_\pi \mathbf{x} - \cancel{\mathbf{r}_\pi} - \gamma \mathbf{P}_\pi \mathbf{y}\| \leftrightarrow$
$\|(\gamma \mathbf{P}_\pi \mathbf{x} - \gamma \mathbf{P}_\pi \mathbf{y})\| \leftrightarrow$
$\|\gamma \mathbf{P}_\pi (\mathbf{x} - \mathbf{y})\|$

Using the following properties of the norms:

> For any scalar α and all matrices A in the space K^{mxn} of a field K: $\|\alpha A\| \leq \mid \alpha \mid \|B\|$
> For all matrices A and B in the space K^{nxn} of a field K: $\|AB\| \leq \|A\| \|B\|$

, we can write: $\|\gamma \mathbf{P}_\pi (\mathbf{x} - \mathbf{y})\| \leq \mid \gamma \mid \|\mathbf{P}_\pi\| \|\mathbf{x} - \mathbf{y}\|$, which is in the contraction form $\|\mathbf{T_x} - \mathbf{T_y}\| \leq \alpha \|\mathbf{x} - \mathbf{y}\|$, with $\alpha = \mid \gamma \mid \|\mathbf{P}_\pi\|$. By the contraction definition, we have that $\mid \gamma \mid \|\mathbf{P}_\pi\|$ must be < 1, which is a true statement *for discounted finite MDPs* as: $\mathbf{P}_\pi$ is the probability matrix with respect to policy π having all its values bound between 0 and 1 ($\|\mathbf{P}_\pi\| \leq 1$), and $\mid \gamma \mid < 1$.

Using Banach Fixed Point theorem, we know that $\mathbf{T}$ converges to a value $\mathbf{x}^*$ such that $\mathbf{T}\mathbf{x}^* = \mathbf{x}^*$. As, by definition we get:

$\mathbf{v}_\pi = \mathbf{r}_\pi + \gamma * \mathbf{P}_\pi * \mathbf{v}_\pi \leftrightarrow$
$\mathbf{v}_\pi = \mathbf{T}\mathbf{v}_\pi$

, which leads us to conclude that $\mathbf{x}^*$ is what we are looking for, the true value function $\mathbf{v}_\pi$.

Solution for "The temporal difference operator"

This time, we will consider the operator $\mathbf{T}_v^{(\lambda)}$. If we consider two $\mathbf{v}_\pi$ value function estimates, $\mathbf{x}$ and $\mathbf{y}$, we can prove that $\mathbf{T}_v^{(\lambda)}$ is a contraction by proving that $\left\|\mathbf{T}_\mathbf{x}^{(\lambda)} - \mathbf{T}_\mathbf{y}^{(\lambda)}\right\| \leq \alpha \left\|\mathbf{x} - \mathbf{y}\right\|$, with an $\alpha < 1$. Therefore we can start by defining:

$\mathbf{T}_\mathbf{x}^{(\lambda)} = \sum_{n=0}^{\infty}[(\lambda\gamma\mathbf{P}_\pi)^n(\mathbf{r}_\pi + \gamma\mathbf{P}_\pi\mathbf{x} - \mathbf{x})] + \mathbf{x}$
$\mathbf{T}_\mathbf{y}^{(\lambda)} = \sum_{n=0}^{\infty}[(\lambda\gamma\mathbf{P}_\pi)^n(\mathbf{r}_\pi + \gamma\mathbf{P}_\pi\mathbf{y} - \mathbf{y})] + \mathbf{y}$

Now we reduce the expression $\left\|\mathbf{T}_\mathbf{x}^{(\lambda)} - \mathbf{T}_\mathbf{y}^{(\lambda)}\right\|$:

$\left\|\mathbf{T}_\mathbf{x}^{(\lambda)} - \mathbf{T}_\mathbf{y}^{(\lambda)}\right\| \leftrightarrow$
$\|(\sum_{n=0}^{\infty}[(\lambda\gamma\mathbf{P}_\pi)^n(\mathbf{r}_\pi + \gamma\mathbf{P}_\pi\mathbf{x} - \mathbf{x})] + \mathbf{x}) - (\sum_{n=0}^{\infty}[(\lambda\gamma\mathbf{P}_\pi)^n(\mathbf{r}_\pi + \gamma\mathbf{P}_\pi\mathbf{y} - \mathbf{y})] + \mathbf{y})\| \leftrightarrow$
$\|\sum_{n=0}^{\infty}[(\lambda\gamma\mathbf{P}_\pi)^n[(\mathbf{r}_\pi + \gamma\mathbf{P}_\pi\mathbf{x} - \mathbf{x}) - (\mathbf{r}_\pi + \gamma\mathbf{P}_\pi\mathbf{y} - \mathbf{y})]] + (\mathbf{x} - \mathbf{y})\| \leftrightarrow$
$\|\sum_{n=0}^{\infty}[(\lambda\gamma\mathbf{P}_\pi)^n[(\cancel{\mathbf{r}_\pi} + \gamma\mathbf{P}_\pi\mathbf{x} - \mathbf{x}) - (\cancel{\mathbf{r}_\pi} + \gamma\mathbf{P}_\pi\mathbf{y} - \mathbf{y})]] + (\mathbf{x} - \mathbf{y})\| \leftrightarrow$
$\|\sum_{n=0}^{\infty}[(\lambda\gamma\mathbf{P}_\pi)^n[(\gamma\mathbf{P}_\pi\mathbf{x} - \mathbf{x}) - (\gamma\mathbf{P}_\pi\mathbf{y} - \mathbf{y})]] + (\mathbf{x} - \mathbf{y})\| \leftrightarrow$
$\|\sum_{n=0}^{\infty}[(\lambda\gamma\mathbf{P}_\pi)^n(\gamma\mathbf{P}_\pi - \mathbf{I})(\mathbf{x} - \mathbf{y})] + (\mathbf{x} - \mathbf{y})\|$

Using the following properties of the norms:

> For any scalar α and all matrices A in the space K^{mxn} of a field K: $\|\alpha A\| \leq \mid \alpha \mid \|B\|$
>
> For all matrices A and B in the space K^{nxn} of a field K: $\|AB\| \leq \|A\| \, \|B\|$
>
> (Triangle inequality)
> For all matrices A and B: $\|A + B\| \leq \|A\| + \|B\|$

and the fact that $0 < \lambda, \gamma < 1$, we can write:

$\|\sum_{n=0}^{\infty}[(\lambda\gamma\mathbf{P}_\pi)^n(\gamma\mathbf{P}_\pi - \mathbf{I})(\mathbf{x} - \mathbf{y})] + (\mathbf{x} - \mathbf{y})\| \leq$
$\|\sum_{n=0}^{\infty}[(\lambda\gamma\mathbf{P}_\pi)^n(\gamma\mathbf{P}_\pi - \mathbf{I})(\mathbf{x} - \mathbf{y})]\| + \|(\mathbf{x} - \mathbf{y})\| \leq$
$\sum_{n=0}^{\infty}[\|(\lambda\gamma\mathbf{P}_\pi)^n(\gamma\mathbf{P}_\pi - \mathbf{I})(\mathbf{x} - \mathbf{y})\|] + \|(\mathbf{x} - \mathbf{y})\| \leq$
$\sum_{n=0}^{\infty}[(\lambda\gamma \left\|\mathbf{P}_\pi\right\|)^n(\gamma \left\|\mathbf{P}_\pi\right\| - \|\mathbf{I}\|) \left\|(\mathbf{x} - \mathbf{y})\right\|] + \|(\mathbf{x} - \mathbf{y})\|$

As $\|\mathbf{I}\| = 1$ and $\|\mathbf{P}_\pi\| \leq 1$, we get:

$\sum_{n=0}^{\infty}[(\lambda\gamma \left\|\mathbf{P}_\pi\right\|)^n(\gamma \left\|\mathbf{P}_\pi\right\| - \|\mathbf{I}\|) \left\|(\mathbf{x} - \mathbf{y})\right\|] + \|(\mathbf{x} - \mathbf{y})\| \leq$
$\sum_{n=0}^{\infty}[(\lambda\gamma)^n(\gamma - 1) \left\|(\mathbf{x} - \mathbf{y})\right\|] + \|(\mathbf{x} - \mathbf{y})\| \leftrightarrow$
$(\gamma - 1) \left\|(\mathbf{x} - \mathbf{y})\right\| \sum_{n=0}^{\infty}(\lambda\gamma)^n + \|(\mathbf{x} - \mathbf{y})\|$

As (by the problem definition) $\lambda\gamma < 1$, we can use the closed form of the geometric series:

$$\sum_{n=0}^{\infty} c^k \rightarrow \frac{1}{1-c}, \text{ for } 0 \leq c < 1$$

, to reduce our calculations:

$(\gamma - 1) \|(\mathbf{x} - \mathbf{y})\| \sum_{n=0}^{\infty} (\lambda\gamma)^n + \|(\mathbf{x} - \mathbf{y})\| \leftrightarrow$
$(\gamma - 1) \|(\mathbf{x} - \mathbf{y})\| \frac{1}{1-\lambda\gamma} + \|(\mathbf{x} - \mathbf{y})\|$

Multiplying and dividing the right $\|(\mathbf{x} - \mathbf{y})\|$ by $1 - \lambda\gamma$, we get:

$(\gamma - 1) \|(\mathbf{x} - \mathbf{y})\| \frac{1}{1-\lambda\gamma} + \|(\mathbf{x} - \mathbf{y})\| \leftrightarrow$
$\frac{(\gamma-1)\|(\mathbf{x}-\mathbf{y})\| + \|(\mathbf{x}-\mathbf{y})\| - \lambda\gamma * \|(\mathbf{x}-\mathbf{y})\|}{1-\lambda\gamma} \leftrightarrow$
$\frac{\gamma-1+1-\lambda\gamma}{1-\lambda\gamma} \|(\mathbf{x} - \mathbf{y})\| \leftrightarrow$
$\frac{\gamma-\cancel{1}+\cancel{1}-\lambda\gamma}{1-\lambda\gamma} \|(\mathbf{x} - \mathbf{y})\| \leftrightarrow$
$\frac{\gamma-\lambda\gamma}{1-\lambda\gamma} \|(\mathbf{x} - \mathbf{y})\|$

From this result, we have that $\|\mathbf{T}_\mathbf{x} - \mathbf{T}_\mathbf{y}\| \leq \frac{\gamma-\lambda\gamma}{1-\lambda\gamma} \|\mathbf{x} - \mathbf{y}\|$, which is the contraction form. Now, remembering the definition of a contraction, we only have to prove that $\frac{\gamma-\lambda\gamma}{1-\lambda\gamma}$ is < 1. We do this by assuming the numerator is smaller than the denominator:

$\gamma - \lambda\gamma < 1 - \lambda\gamma \leftrightarrow$
$\gamma - \cancel{\lambda\gamma} < 1 - \cancel{\lambda\gamma} \leftrightarrow$
$\gamma < 1$

, which is a true statement, considering *discounted finite MDPs*. Therefore, it is proven that the operator $\mathbf{T}^{(\lambda)}$ is a contraction.