Reinforcement Learning
Exercise 8 - Solution

Jonathan Schnitzler - st166934
Eric Choquet - st160996

June 28, 2024

1 n-step TD compared to planing

a) Improve TD(0) with n-step TD The difference of the n-step temporal difference to the Dyna-Q planning, is that only the reward of the path which was taken can be accounted for. Therefore, unlike the image it is not possible after one episode to have a policy for all tiles, but instead only for the taken path. In contrast, Dyna-Q planning revisits arbitrary a virtual tile, which allows a richer interpretation. A visual interpretation is depicted in Figure **??**.

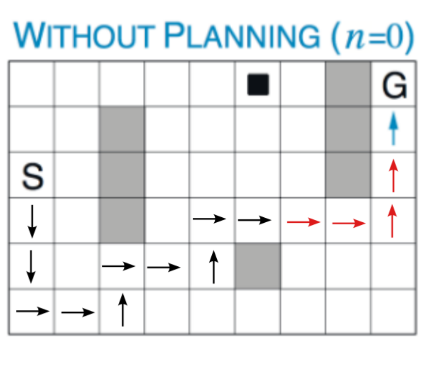

Figure 1: The red arrows indicate 4-step TD, while the black arrows just continue for an arbitrary n, where the sampled episode is also arbitrary

b) Recursive lambda-return The λ-return is defined as

$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} G_{t:t+n} \tag{1}$$

with

$$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n}) \tag{2}$$

To define it recursively, we want to reframe the problem involving G_t^λ

$$G_{t+1}^\lambda = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_{t+1:t+n+1} \tag{3}$$

Here we can note that

$$G_{t+1:t+n+1} = R_{t+2} + \gamma R_{t+3} + \ldots + \gamma^{n-1}R_{t+n+1} + \gamma^n V(S_{t+n+1}) \tag{4}$$

We can rewrite $G_{t+1:t+n+1}$ as

$$G_{t+1:t+n+1} = \frac{1}{\gamma}\left(G_{t:t+n} - R_{t+1} + \gamma^n R_{t+n+1} - \gamma^n V(S_{t+n})\right) + \gamma^n V(S_{t+n+1}) \tag{5}$$

Then we can rewrite the λ-return as

$$
\begin{aligned}
G_{t+1}^\lambda &= (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\left[\frac{1}{\gamma}\left(G_{t:t+n} - R_{t+1} + \gamma^n R_{t+n+1} - \gamma^n V(S_{t+n})\right) + \gamma^n V(S_{t+n+1})\right] \\
G_{t+1}^\lambda &= \frac{1}{\gamma}(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_{t:t+n} \\
&\quad + (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\left[\frac{1}{\gamma}\left(-R_{t+1} + \gamma^n R_{t+n+1} - \gamma^n V(S_{t+n})\right) + \gamma^n V(S_{t+n+1})\right] \\
G_{t+1}^\lambda &= \frac{1}{\gamma}G_t^\lambda + (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\left[-\frac{R_{t+1}}{\gamma} + \gamma^{n-1}R_{t+n+1} - \gamma^{n-1}V(S_{t+n}) + \gamma^n V(S_{t+n+1})\right]
\end{aligned}
$$

Could be extended further, but this should be sufficient for today.

2 n-step Sarsa on the Frozen Lake

The implemented algorithm in python with the Gym FrozenLake Environment can be found attached to this file. The 8x8 map is the following

S	F	F	F	F	F	F	F
F	F	F	F	F	F	F	F
F	F	F	H	F	F	F	F
F	F	F	F	F	H	F	F
F	F	F	H	F	F	F	F
F	H	H	F	F	F	H	F
F	H	F	F	H	F	H	F
F	F	F	H	F	F	F	G

Table 1: Where F is frozen, H is hole, S is start and G is goal

The value function after 250000 episodes with $\alpha = 0.025$ and $\epsilon = 0.1$ is depicted in Figure 2 and the state-action value function in Figure 3. The policy is given in Table 2.

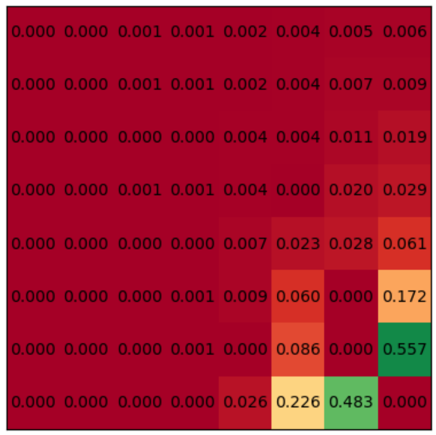

Figure 2: Value function after 250000 episodes

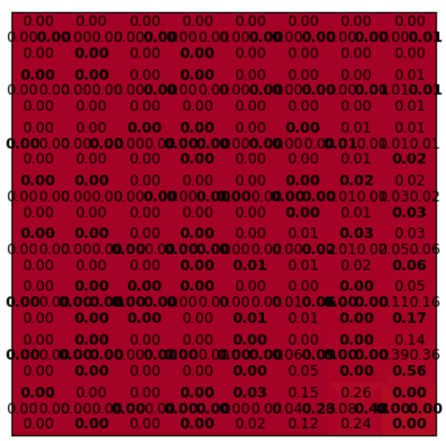

Figure 3: State-Action value function after 250000 episodes

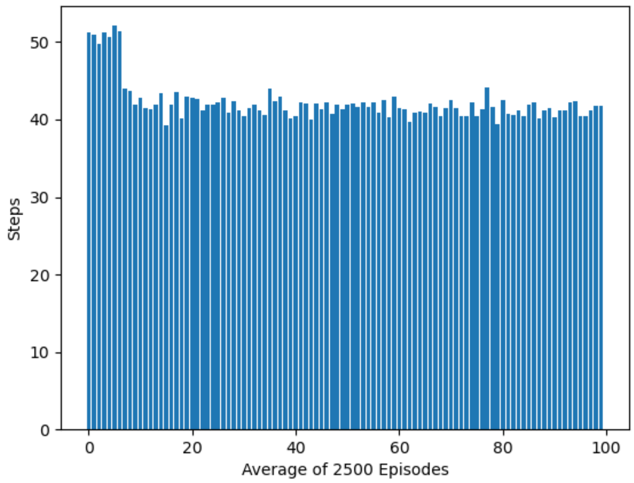

Figure 4: Average episode length through the training process

$\rightarrow$	$\downarrow$	$\rightarrow$	$\downarrow$	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\rightarrow$
$\uparrow$	$\uparrow$	$\rightarrow$	$\uparrow$	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\rightarrow$
$\leftarrow$	$\rightarrow$	$\uparrow$	H	$\rightarrow$	$\uparrow$	$\leftarrow$	$\downarrow$
$\uparrow$	$\uparrow$	$\rightarrow$	$\rightarrow$	$\leftarrow$	H	$\uparrow$	$\downarrow$
$\uparrow$	$\uparrow$	$\leftarrow$	H	$\downarrow$	$\rightarrow$	$\uparrow$	$\downarrow$
$\leftarrow$	H	H	$\uparrow$	$\downarrow$	$\rightarrow$	H	$\downarrow$
$\leftarrow$	H	$\rightarrow$	$\leftarrow$	H	$\rightarrow$	H	$\downarrow$
$\uparrow$	$\downarrow$	$\leftarrow$	H	$\uparrow$	$\rightarrow$	$\rightarrow$	G

Table 2: Where the arrow indicates the walking direction