Different Reinforcement Learning Algorithms to Solve Grid World

Matteo Espositoa, William Ngoa, Spyros Orfanosa

*aConcordia University, Montreal, Quebec*

# Abstract

Reinforcement Learning (RL) is widely used in different research areas to tackle problems such as resource management in computers, traffic light control, and playing games like chess at a super-human level. In this project, we tackle the problem of solving a two-dimensional maze given a specific starting point. To accomplish this, three different policies were devel- oped using RL algorithms known as SARSA, Q-Learning, and n-step SARSA. Each learning method has its theoretical tradeoffs from one another and thus we are comparing their perfor- mances by evaluating their Mean-Squared Error, average training time and average number of episodes until they reach optimality.Out of the 3 policies, conclude based on runtime, MSE, speed of convergence and (optionally) scalability (i.e. different maze sizes if possible,

. . . )

# Introduction

Our project, commonly referred to as ”Grid World”, tackles the problem of exiting a two-dimensional maze. This problem consists of a starting point, an exit point, and a board in which the agent (player) can take actions to move up, down, left or right. This is quite a simple task for a human as one can simply inspect the entirety of the maze and, at a glance, deduce an exit strategy. However, a machine cannot readily ”see” the maze or ”think” of actions to take, so how can it learn to solve a maze?

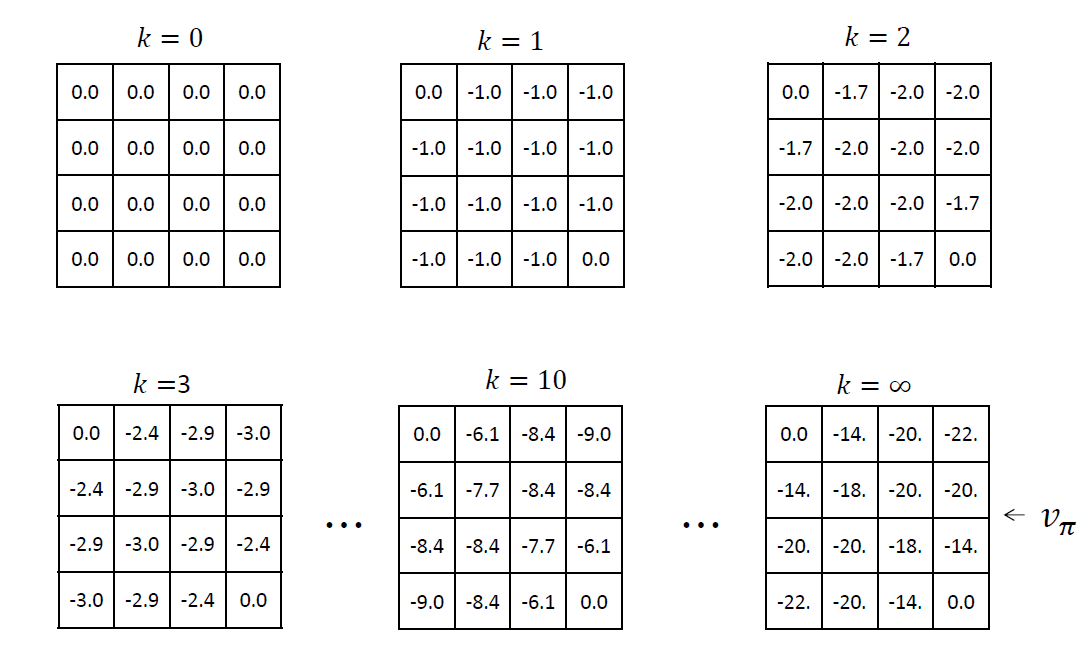
To allow a machine to “understand” the problem at hand, we must first reformulate the problem into something it can understand: a sequential decision problem. A sequential decision problem consists of a set of states, a set of actions, and a reward for taking some action given the current state.

By formulating the problem in such a way, we may use Reinforcement Learning (RL) techniques. RL is learning what to do: how to map situations to actions so as to maximize some notion of cumulative reward. In other words, RL is all about discovering the best action given your current state. To do this, an objective that is designed as a numerical reward by learning from an environment. Most reinforcement learning problems can be expressed as sequential decision problems, as they must include a set of the possible states of the environment, a set of possible actions and an objective.

*Email addresses:* [ds.mesposito@gmail.com](mailto:ds.mesposito@gmail.com) (Matteo Esposito), [ngowilliam96@gmail.com](mailto:ngowilliam96@gmail.com) (William Ngo), [spyros.orfanos1@gmail.com](mailto:spyros.orfanos1@gmail.com) (Spyros Orfanos)

*Preprint submitted to Dr.Frederic Godin, Concordia University April 10, 2019*

The goal of this project is to compare the rate of convergence of the estimated value function to the optimal policy’s value function of three RL methods: SARSA, Q-Lesarning, and n-step SARSA.

Figure 1: Gridworld problem iterative policy evaluation. Here, *vπ* is the state-value function, *k* is the number of iterations. The figure depicts the convergence of the policy evaluation as *k* → ∞.

* 1. *Review of LiteratureCould condense/remove.*

Given that the issue of optimal path finding/maze solving is a widely applicable problem, various machine learning, notably, reinforcement learning approaches have been taken in an effort to improve and/or complement current methods.

In Mnih et al. (2016), researchers from Google and MILA compared asnychronous vari- ants of four standard reinforcement learning approaches. It was concluded that the asny- chronous advantage actorcritic (A3C) method yields the most promising results.

In an advantage actor-critic implementation, the actor is the policy, and the critic the value function. In practice, the agent will use the critic, or value estimate, to update the policy, or actor more efficiently than policy gradient methods. The critic is typically a single linear output connected to a convolutional neural network that has one softmax output for the actor/policy. This method also introduces the concept of an advantage function, which is defined as *A*(*s, a*) = *Q*(*s, a*) *V* (*s*) where *Q*(*s, a*) is the state value of *s* given action *a* and *V* (*s*), the average value of state *s*. In practice, the discounted returns are used as an estimate of *Q*(*s, a*). Using this function versus discounted returns allows the agent to determine how much better its selected actions turned out to be than expected. This grants the algorithm the ability to focus on where the network’s predictions were lacking in precision.

−

By bypassing the use of a GPU, or any other specialized hardware, this method was able to outperform the state-of-the-art algorithms in Atari 2600 applications making use of a single multi-core CPU, and mastered a variety of 2D and 3D games, including the exploration of 3D mazes purely from visual inputs. According to the article, it was its ability to train feedforward and recurrent agents as well as its success in discrete and continuous action spaces that makes it the most general and successful reinforcement learning agent to date. [1]

In other articles such as Halici (2001), the use of random neural networks is proposed which differs from Q-learning and temporal difference learning since it does not explicitly maintain estimates of reinforcements for each (state, action) pair. Such an estimate is somehow implicit in the connection weights. [2]. Whereas in Lei et al. (2018), path planning for mobile robots is the motivation behind the study. Here, the application of a double Q- network (DDQN) created by DeepMind is used for a robot’s path planning in an unknown environment. To make due for the fact that conventional Q-learning cannot effectively plan a path in random dynamic environment because of the lack of generalization ability and a large Q table and to tackle the issue of dimensionality in a high dimensional state space, the optimal action value function Q in Q-learning can be parameterized by an approximate value function *Q*(*s, a*; *θ*) *Q*∗(*s, a*) where *θ* is the Q-network parameter [3]. The introduction of this weight allows for the fair evalution of the value of the policy at hand.

≈

# Problem Description

As described above, our problem, ”Grid World”, is a sequential decision problem as the set of states is the possible positions in the maze, the set of actions is deciding which direction to go.

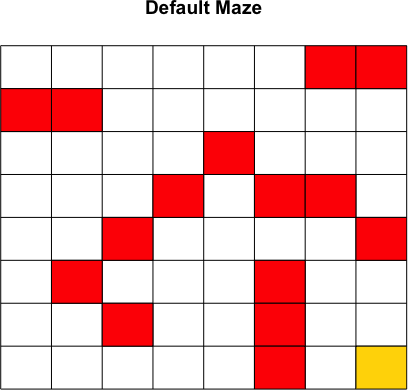
Here, the set of states consists of all valid board positions (*x, y*) where *x, y* 1*,* 2*, . . . ,* 8 . The set of actions consists of moving Right (*action* = 1)*,* Down (*a* = 2)*,* Left (*a* =

{

∈ { }

1. or Up (*a* = 4) where the objective would be to find the exit of the maze which is at point (8*,* 8), starting at point (1*,* 1) in our test cases. To give our algorithms an easier time identifying the different elements of the maze environment, we’ve attributed valid spaces, walls and exit tile to integer values of 0,1 and 10 respectively in a 2D-array forming our Maze class. Through the use of object-oriented programming and the R6 package, we can access and modify many properties of our maze object including but not limited to: the possible actions at a given point in time, or a vector of the visited states.

}



For every action we assign a reward strictly dependent on the environment (set of states), with the following structure:

invalidMoveReward = -0.75 (Hitting a wall - denoted as a red tile, or going beyond the border)

•

* + unseenStateReward = -0.05 (Visiting a new, unseen tile that isn’t the exit tile)
  + seenStateReward = -0.30 (Revisiting a tile)
  + winningGameReward = 5 (Objective: Visiting the exit tile - denoted as a yellow tile)

One thing to note is that if the agent attempts to hit a wall or the outer bound of the maze then it stays at its current position and gets attributed a reward of -0.30. Given the structure of our test maze, our optimal return would be *G*0 = 3*.*85, which will be denoted as the green convergence asymptote curve in our return average plots presented in the following section of this report.

# Analyses

In approaching this problem, as in almost all reinforcement problems, choosing a suitable value of *ε* was a challenge, as we wanted to encourage the agent to use the highest return policy at any point in time as well as take time to explore alternate solutions to uncover a potentially better policy in the long run.

To address this challenge, we implore the following logic for defining our epsilon choice at any given episode . . .Write epsilon selection logic...

Another consideration of any machine leaning project is the bias(underfitting) - vari- ance(overfitting) tradeoff. Since we are only dealing with temporal difference methods, we will only speak about the bias anc variance present in these types of reinforcement leaning algorithms. Generally speaking, TD methods have low variance and are somewhat biased. This low variance is due to the fact that in the update equation for the value function of TD methods, the term/target value *Rt*+1 + *γVπ*(*St*+1) shows that the state-value update only depends on the immediate reward *Rt*+1. We can contrast this to high variance Monte Carlo methods, which have target value *Gt* = *Rt*+1 +*γRt*+2 +*. . .*, which clearly depend on a greater number of rewards.

The presence of bias in the TD methods is related to the fact that the value function estimate is what is being used in the update function and it relies on an initial value we set for it, and therefore is not representative of the actual value of the value function. This can explain why the the beginning of our return curves are not monotonically increasing. With time, the estimate of the value function becomes a decreasingly biased estimator of the actual value function.

* 1. *Random Policy*

To better visualize the effect of the aforementioned RL methods, we chose to include the use of a random policy. Therefore, we sample randomly from the list of possible moves our agent can perform and observe the state values that are generated. One thing to note is that this approach does not guarantee convergence/successful exiting of the maze provided and out of 10000 episodes only 10.75% of all episodes exit the maze successfully/terminate, yielding an average reward of -115.596.

Table 1: Results of random policy after 10000 episodes

|  |  |
| --- | --- |
| Number of episodes that ended | Average Return of the policies that ended |
| 1075 | -115.569 |

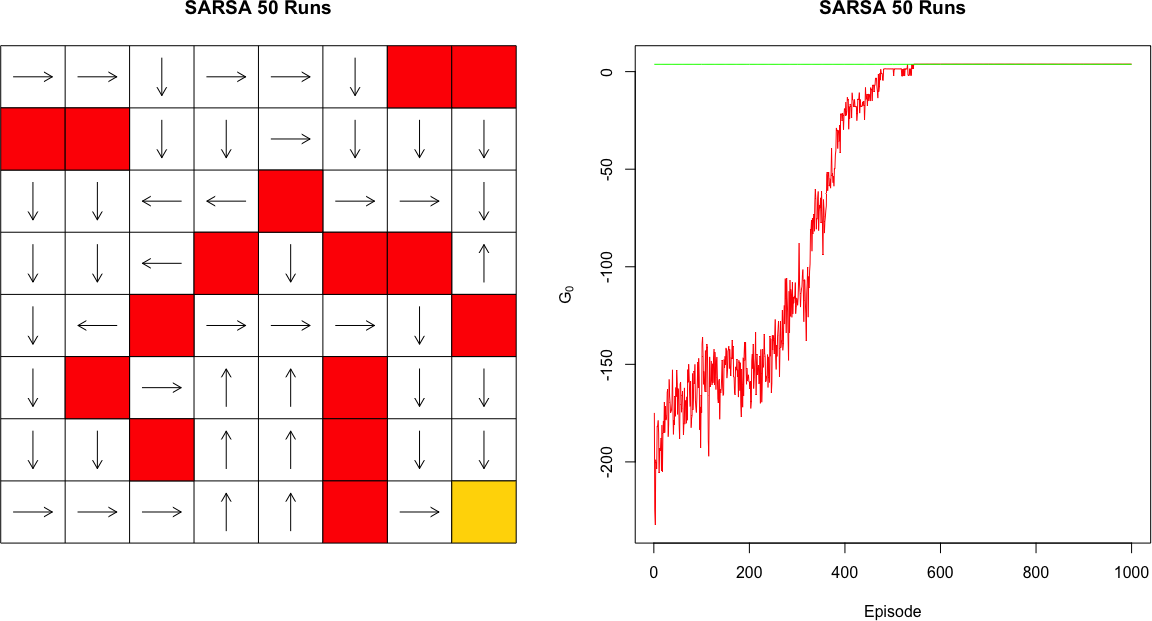
* 1. *SARSA*

Our second algorithm is the SARSA (State-Action-Reward-State-Action) method. The SARSA (State-Action-Reward-State-Action, one-step SARSA or SARSA(0)) method is an on-policy temporal difference control method. It enables the agent to update the (state,value) or *Qπ*(*s, a*) value after every single action under a given policy *π*. Given an epsilon value, the agent will either tend to explore alternate actions in an effort to find different trajectories to the end goal or exploit the current optimal action at a given state (*s greedy*).

−

Its update rule is defined as follows:

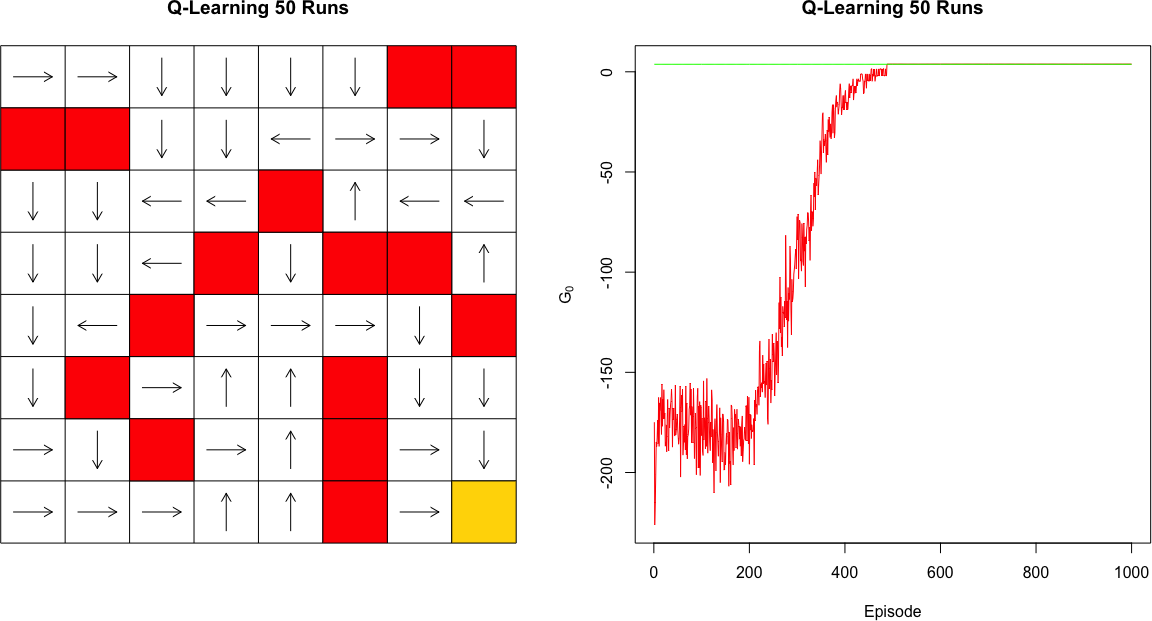
*Qt*+1(*St, At*) ← *Qt*(*St, At*) + *α*[*Rt*+1 + *γQt*(*St*+1*, At*+1) − *Qt*(*St, At*)]



* 1. *Q-Learning*

Q-Learning is an off-policy temporal difference control policy. It is exactly like SARSA, the only difference being that it doesn’t follow a policy to find the next action A’ but rather chooses the action in a greedy fashion. Similar to SARSA its aim is to evaluate the Q val- ues. In this case, the learned action-value function, Q, directly approximates q\*, the optimal action-value function, independent of the policy being followed. This dramatically simplifies the analysis of the algorithm and enabled early convergence proofs. The policy still has an effect in that it determines which state–action pairs are visited and updated. However, all that is required for correct convergence is that all pairs continue to be updated. As we ob- served in Chapter 5, this is a minimal requirement in the sense that any method guaranteed to find optimal behavior in the general case must require it. Under this assumption and a variant of the usual stochastic approximation conditions on the sequence of step-size param- eters, Q has been shown to converge with probability 1 to q\*. The Q-learning algorithm is shown below in procedural form.

Contrary to SARSA which transitioned to an action-state *At*+1*, St*+1 to make the update,

this methods considers transitions from an action-state pair *St, At* to a state *St*+1;. *Qt*+1(*St, At*) ← *Qt*(*St, At*) + *α*[*Rt*+1 + *γ* max *Qt*(*St*+1*, At*+1) − *Qt*(*St, At*)]

*a*

* + 1. *n-Step SARSA*

Finally, The main idea of the n-step SARSA is to simply switch states for actions (state–action pairs) and then use an *s*-greedy policy.

Its algorithm is as follows:

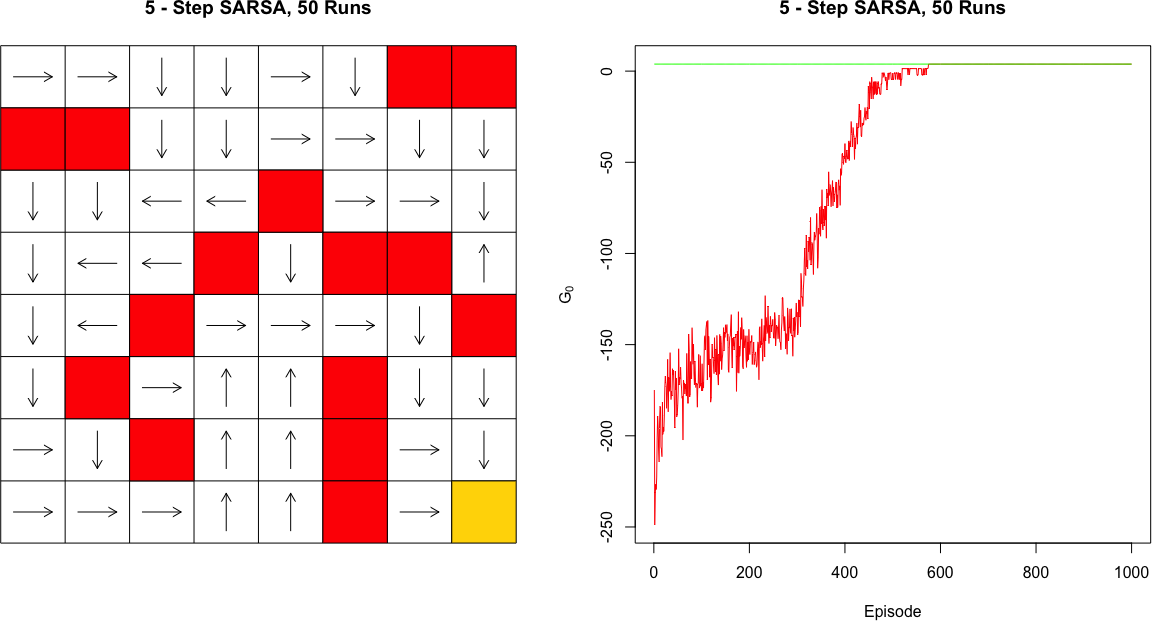
*Gt*:*t*+*n* = *Qt*−1(*St, At*) +

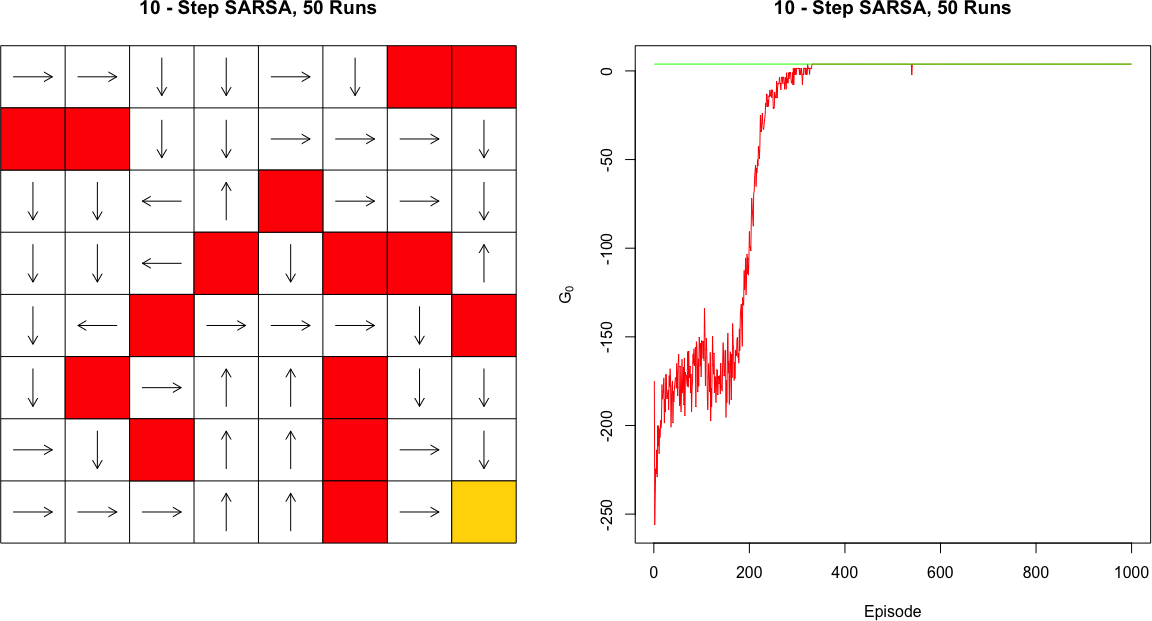
min(*t*+*n,T* )−1

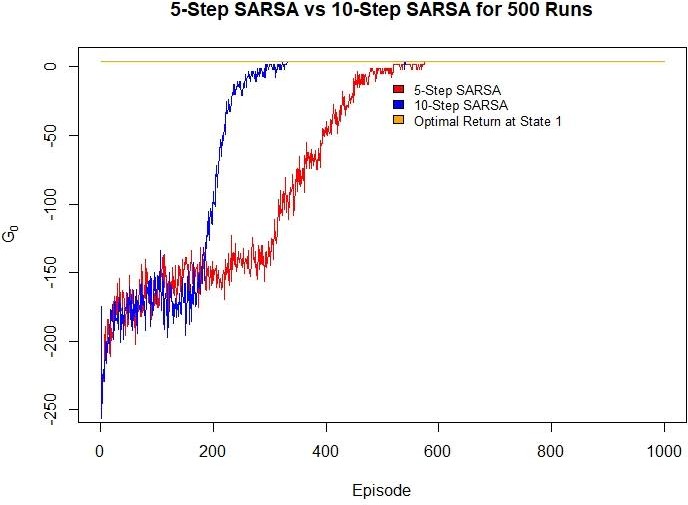
*k*=*t*

Σ

*γk*−*t*[*Rk*+1 + *γQk*(*Sk*+1*, Ak*+1) − *Qk*−1(*Sk, Ak*)]







From the above figures we can conclude that we will never user 5-step sarsa as we have shown 10-step is better for this problem,add to this...

# Conclusion

Summarize the problem you tackled, how you did, and provide the most salient results from your analysis.

•

Potential future avenues of research (additional analysis which could be performed) could be stated here.

•

(From multiple sources online) Q learning converges before both SARSA and n-Step SARSA.

⇒

|  |  |  |
| --- | --- | --- |
| **Method** | **Average# Ep. Before Convergence/Optimality** | **Runtime (s)** |
| SARSA | - | - |
| Q-Learning | - | - |
| n-Step SARSA | - | - |

# References

[1]V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, K. Kavukcuoglu, Asynchronous methods for deep reinforcement learning, CoRR abs/1602.01783 (2016). arXiv:1602.01783.

URL <http://arxiv.org/abs/1602.01783>

[2]U. Halici, Reinforcement learning with internal expectation in the random neural networks for cascaded decisions, Biosystems 63 (1-3) (2001) 21–34 (2001). doi:10.1016/s0303-2647(01)00144-7.

[3]X. Lei, Z. Zhang, P. Dong, Dynamic path planning of unknown environment based on deep reinforcement learning, Journal of Robotics 2018 (2018) 1–10 (2018). doi:10.1155/2018/5781591.