

May 25, 2024

HOMEWORK 3 — Report

1 Questions

1.1 Agent:

Agent corresponds to the entity that interacts with the environment. It is responsible for taking actions, observing the environment, and receiving rewards. The agent is the entity that learns the optimal policy to maximize the cumulative reward. When compared to supervised learning, the agent is not provided with the correct output, but it learns the optimal policy through trial and error. That is concept of error is not exactly available in reinforcement setting.

1.2 Environment:

Environment means the world that the agent interacts with. It is the entity that the agent observes and takes actions. The environment is responsible for providing the agent with the current state, reward, and possible actions. The environment is also responsible for updating the state of the agent based on the action taken by the agent. When compared to supervised learning, the environment corresponds to the both dataset and loss function that the system trained and run on.

1.3 Reward:

Reward is the feedback that the agent receives from the environment. It is the scalar value that the agent receives after taking an action. The reward is used to evaluate the action taken by the agent. The agent aims to maximize the cumulative reward by learning the optimal policy. When compared to supervised learning, the reward corresponds to the loss function that the system tries to minimize.

1.4 Policy:

Policy is the strategy that the agent uses to take actions. It is the mapping from the state to the action. If we compare with supervised learning, the policy corresponds to the model that the system uses to predict the output.

1.5 Exploration:

The exploration corresponds to the process of trying different actions to learn the optimal policy. The agent explores the environment by taking different actions and observing the reward. The exploration is necessary to learn the optimal policy. When we try to map exploration to supervised learning step we may encounter more than one phenomena. For example, we can say that exploration corresponds to the training process of the model. On the other hand we can say the exploration corresponds to the data augmentation process or other randomization steps taken during training.

1.6 Exploitation:

Exploitation corresponds to the process of taking the best action based on the learned policy. The agent exploits the environment by taking the action that maximizes the reward. The exploitation is necessary to maximize the cumulative reward. When we try to map exploitation to supervised learning step, we can say that exploitation corresponds to the inference process of the model.

2 Experimental Work

The proper implementation for maze, temporal difference learning and Q-learning are implemented and provided in appendix.

The experimental work section is divided into four main part where each part has TD learning and Q learning related experiments seperately. First the default parameter outputs are presented. Then the effect of alpha parameter is investigated. After that the effect of gamma parameter is investigated. Lastly the effect of epsilon parameter is investigated.

2.1 Temporal Difference Learning Default Parameters

The default parameters for the temporal difference learning are set as follows:

- Alpha: 0.1
- Gamma: 0.95
- Epsilon: 0.2
- Episodes: 10000

According to these settings the training is done. The policy maps are provided in Figure 1. The value function plots are provided in Figure 2. The convergence plots are provided in Figure 3.

Basically we can say that the agent learns the optimal policy and value function.

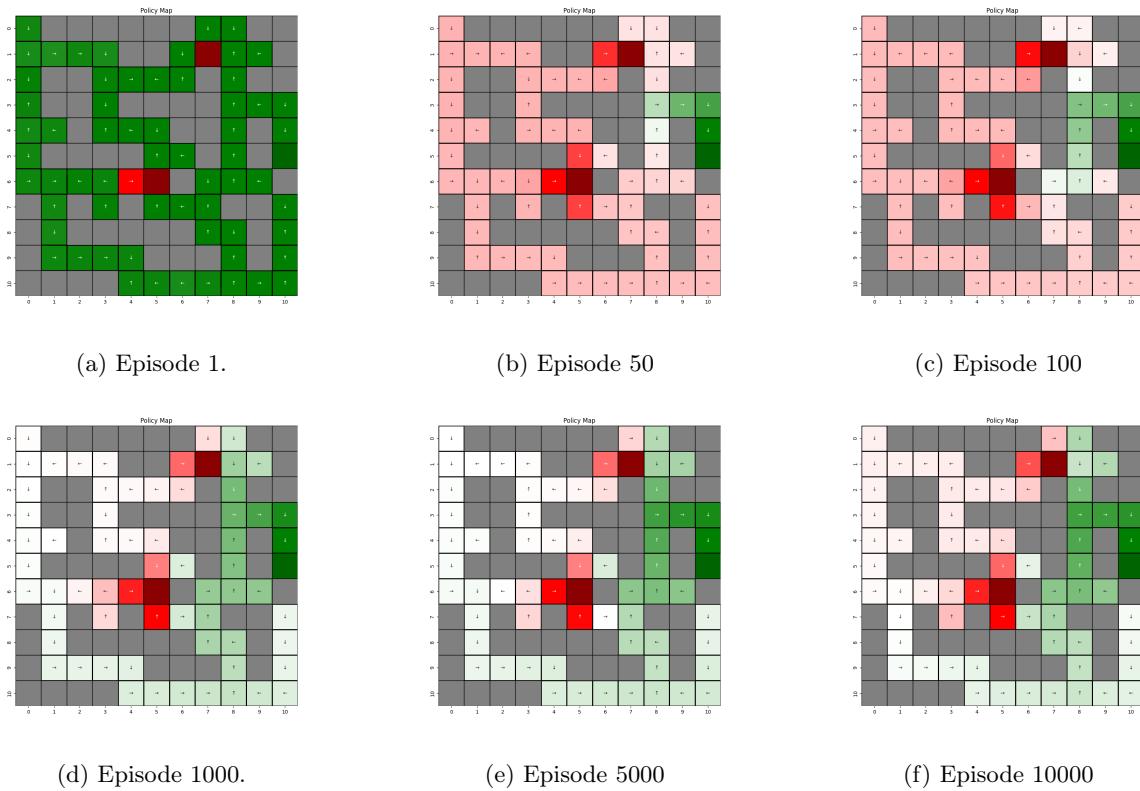


Figure 1: Evolution of policy maps throughout episodes.

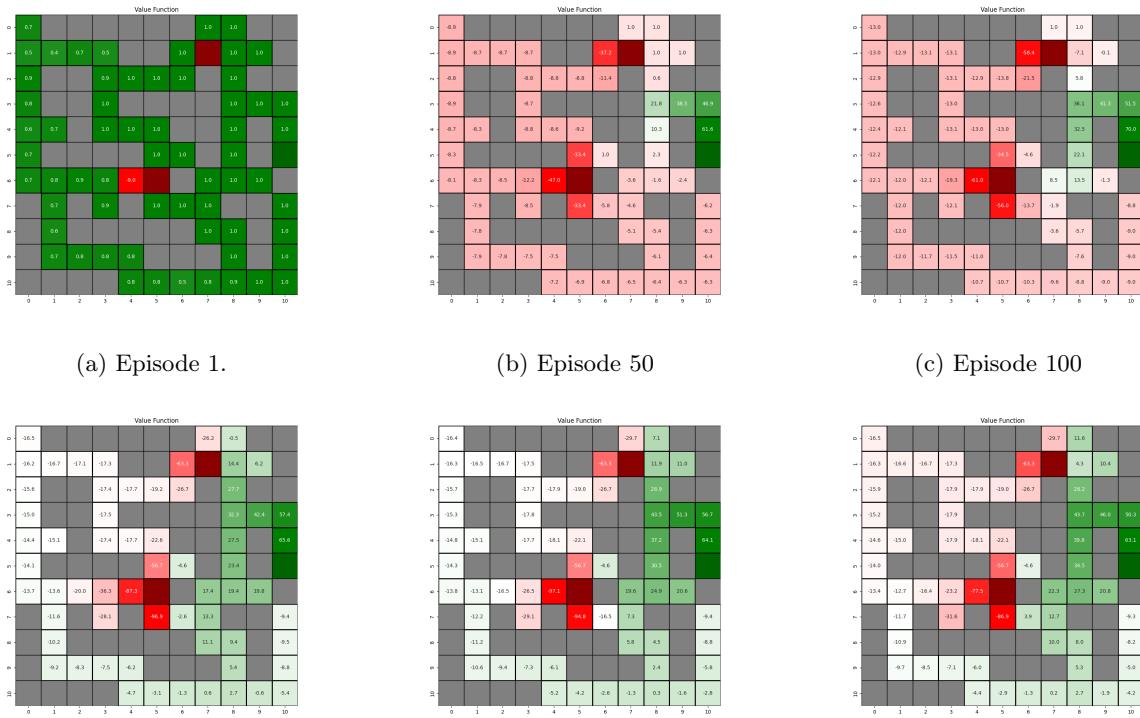


Figure 2: Evolution of value function throughout episodes.

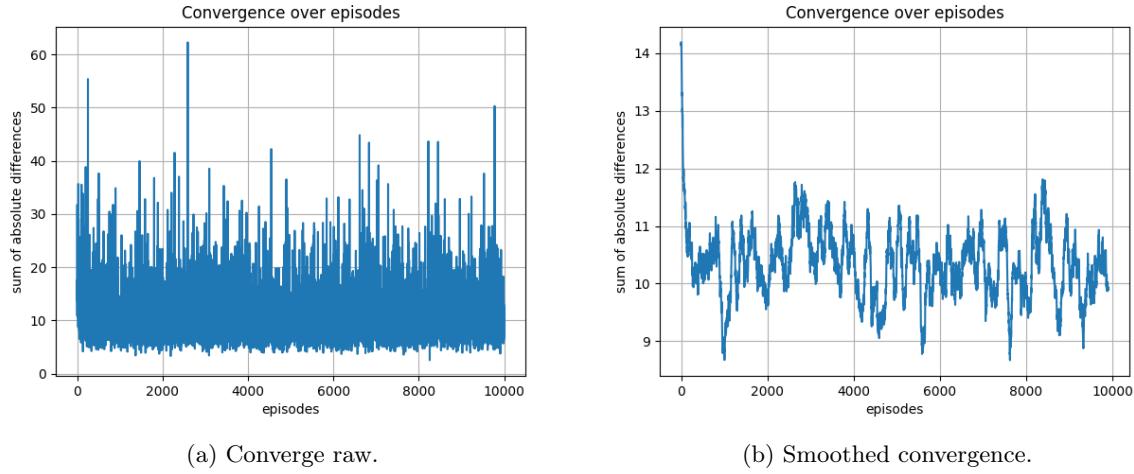


Figure 3: Converge of value function.

2.2 Q-Learning Default Parameters

As same as the temporal difference learning, the default parameters for the Q-learning are set. Then the training is done. The policy maps are provided in Figure 4. The value function plots are provided in Figure 5. The convergence plots are provided in Figure 6.

So, again we can say that the agent learns the optimal policy and value function at the end.

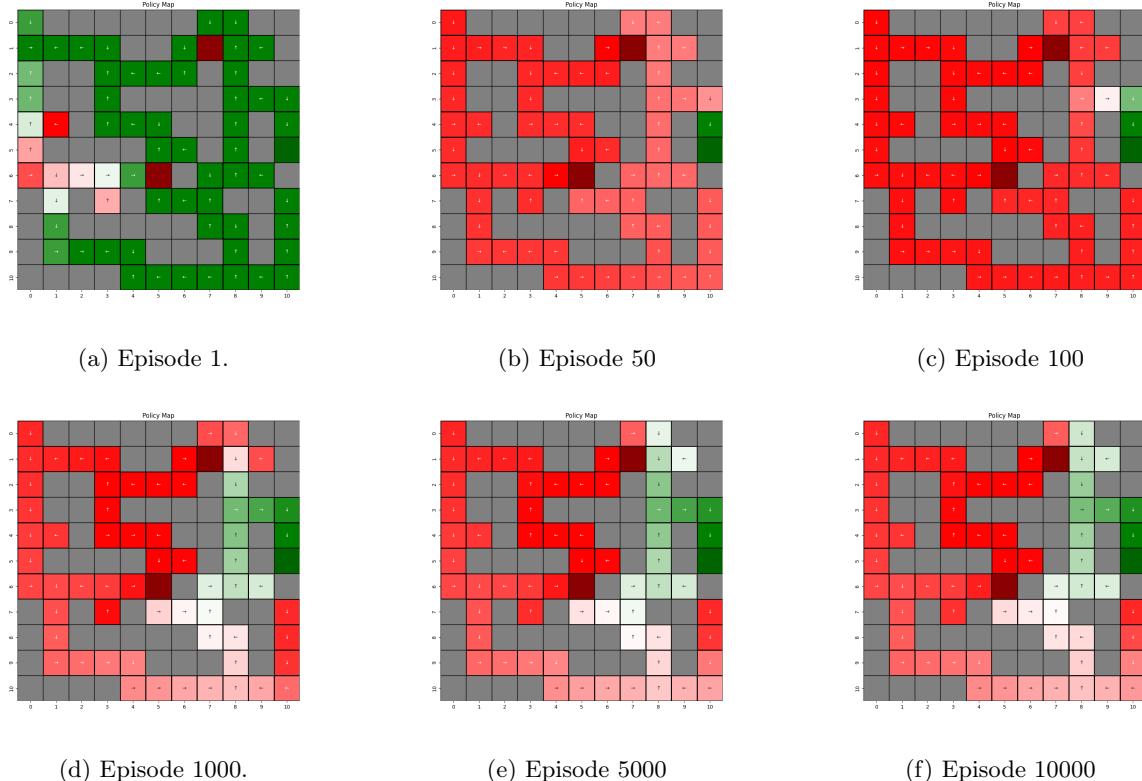


Figure 4: Evolution of policy maps throughout episodes.

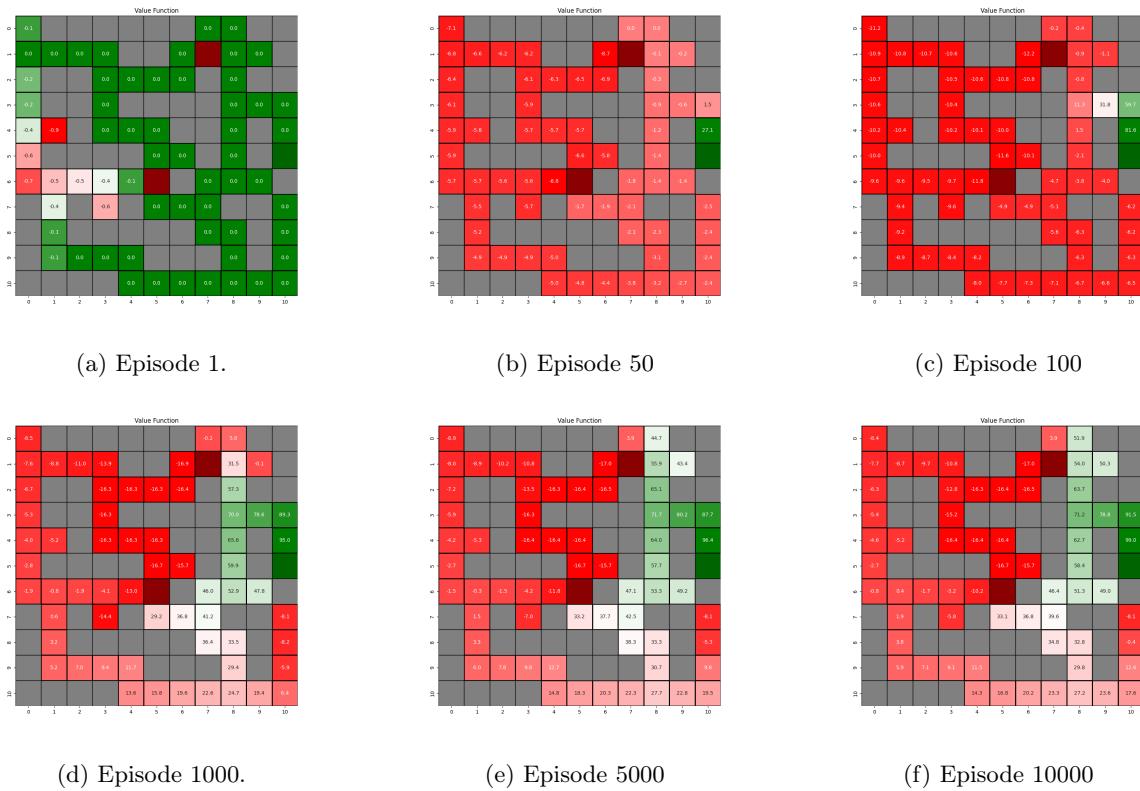


Figure 5: Evolution of value function throughout episodes.

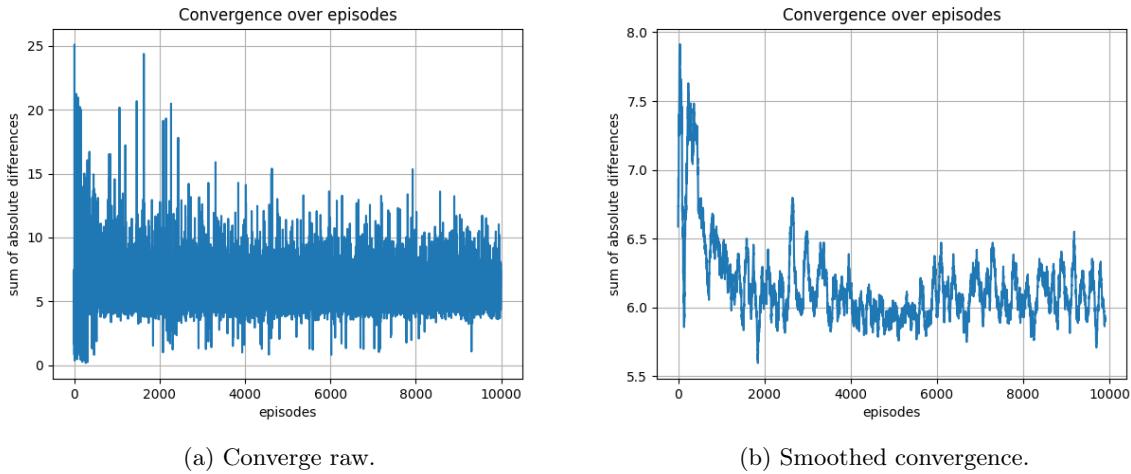


Figure 6: Converge of value function.

2.3 Effect of Alpha in Temporal Difference Learning

Let us first provide the necessary output for each alpha parameter and then discuss the results at the end of this section.

Figure 7 provides the policy maps for the alpha parameter set to 0.001. Figure 8 provides the value function plots for the alpha parameter set to 0.001. Figure 9 provides the convergence plots for the alpha parameter set to 0.001.

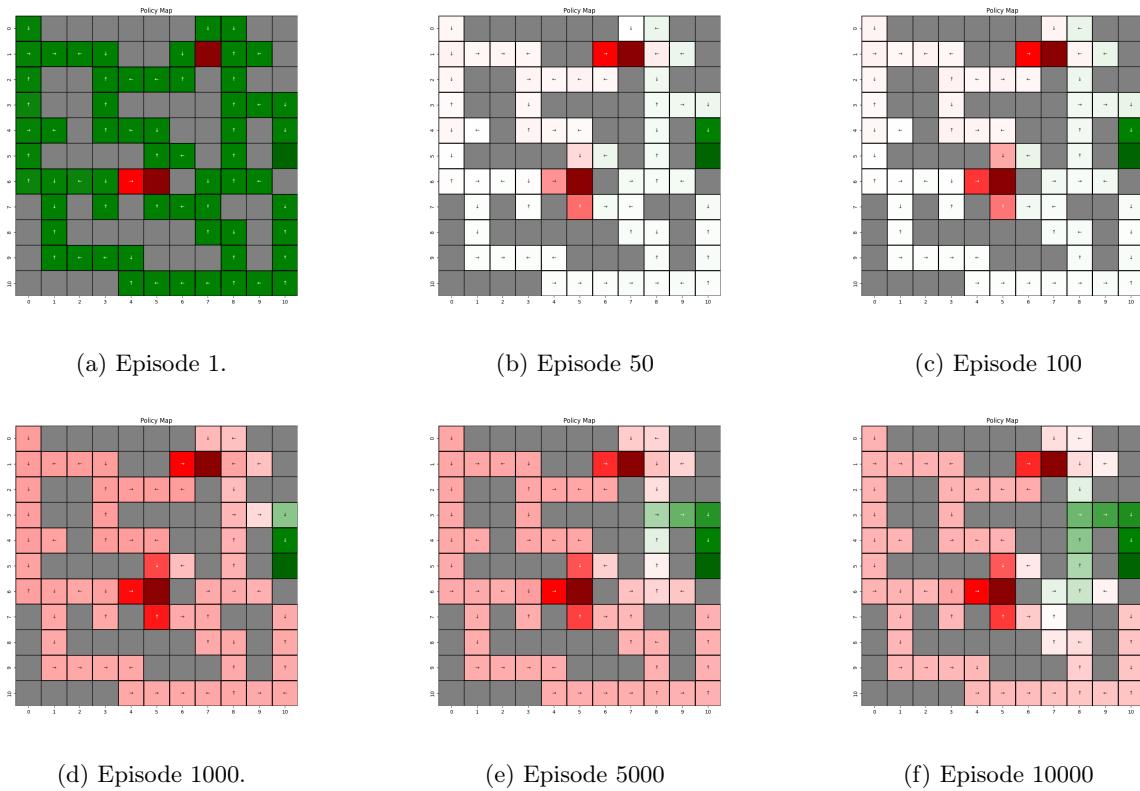


Figure 7: Evolution of policy maps throughout episodes.

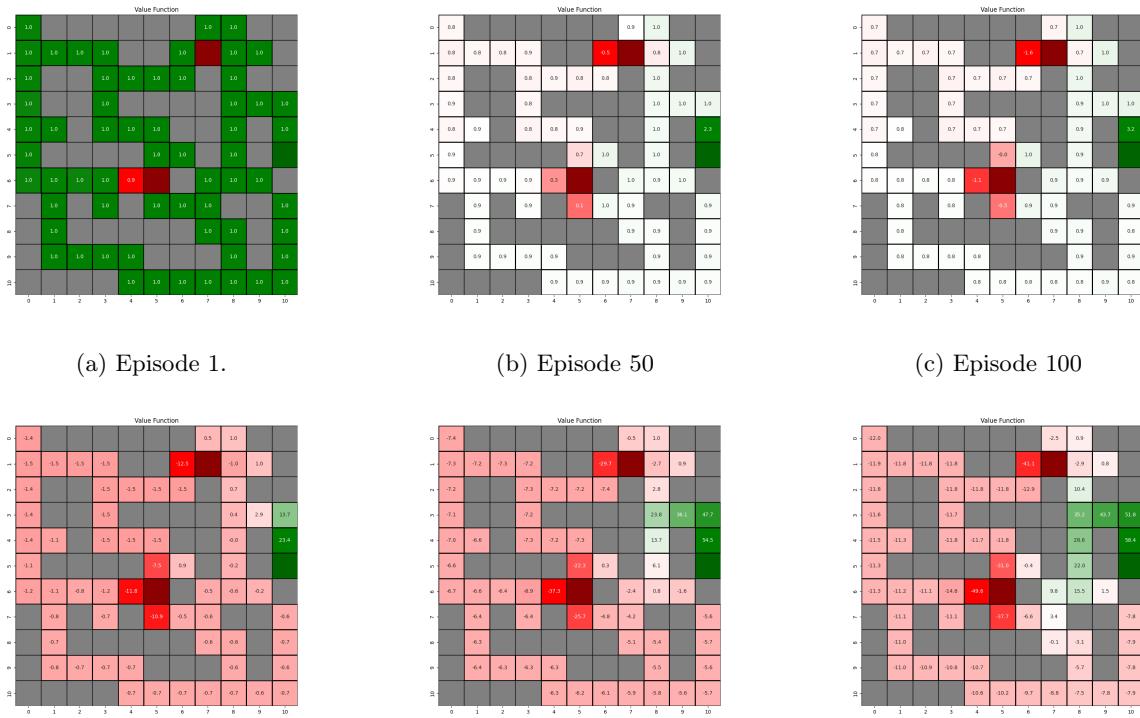


Figure 8: Evolution of value function throughout episodes.

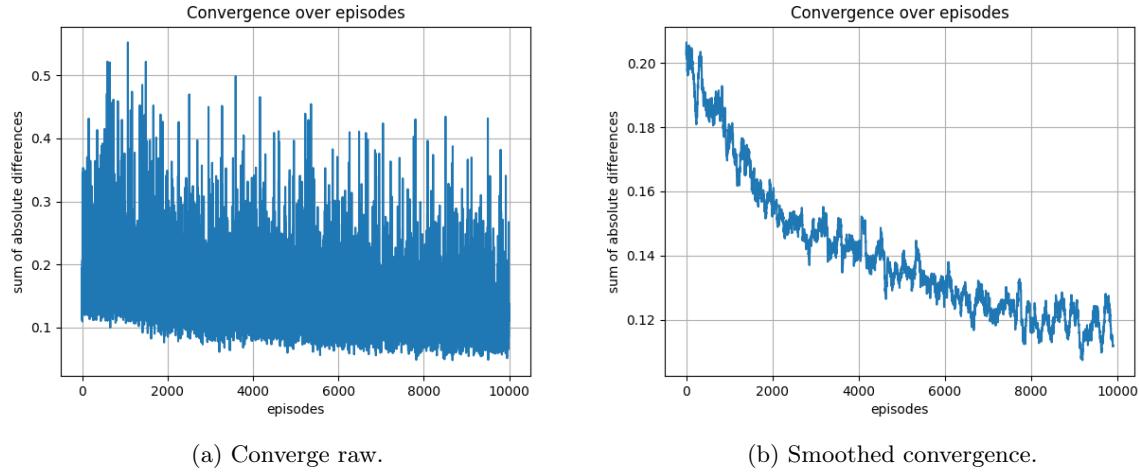


Figure 9: Converge of value function.

Figure 10 shows the policy maps for the alpha parameter set to 0.5. Figure 11 illustrates the value function plots for the alpha parameter set to 0.5. Figure 12 provides the convergence plots for the alpha parameter set to 0.5.

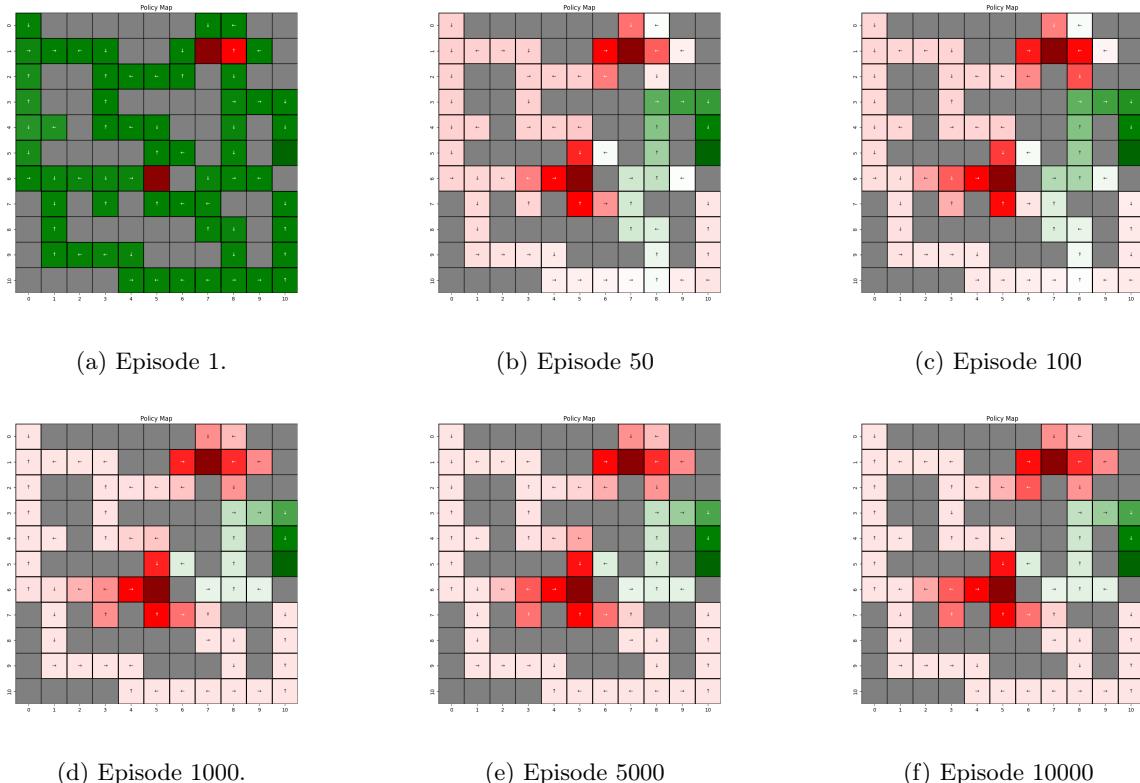


Figure 10: Evolution of policy maps throughout episodes.

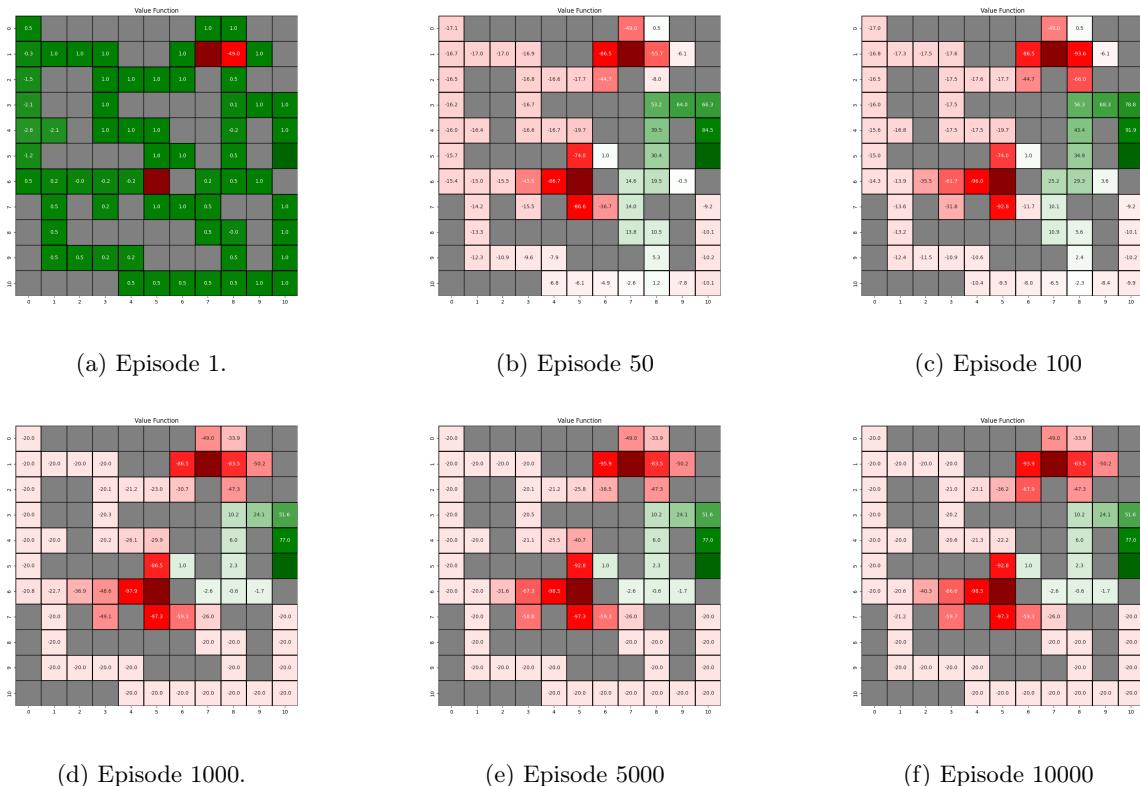


Figure 11: Evolution of value function throughout episodes.

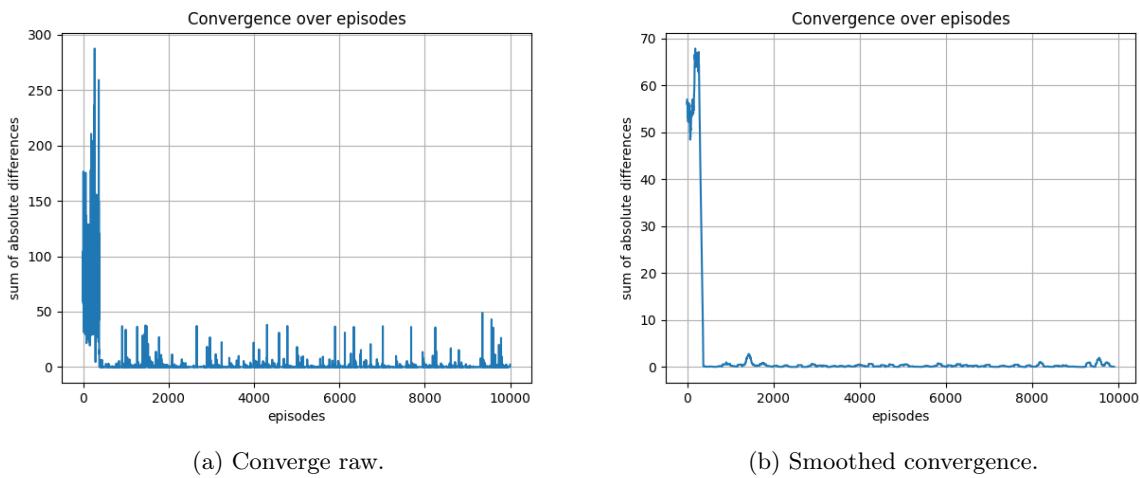


Figure 12: Convergence of value function.

Lastly, Figure 13 shows the policy maps for the alpha parameter set to 1. Figure 14 illustrates the value function plots for the alpha parameter set to 1. Figure 15 provides the convergence plots for the alpha parameter set to 1.

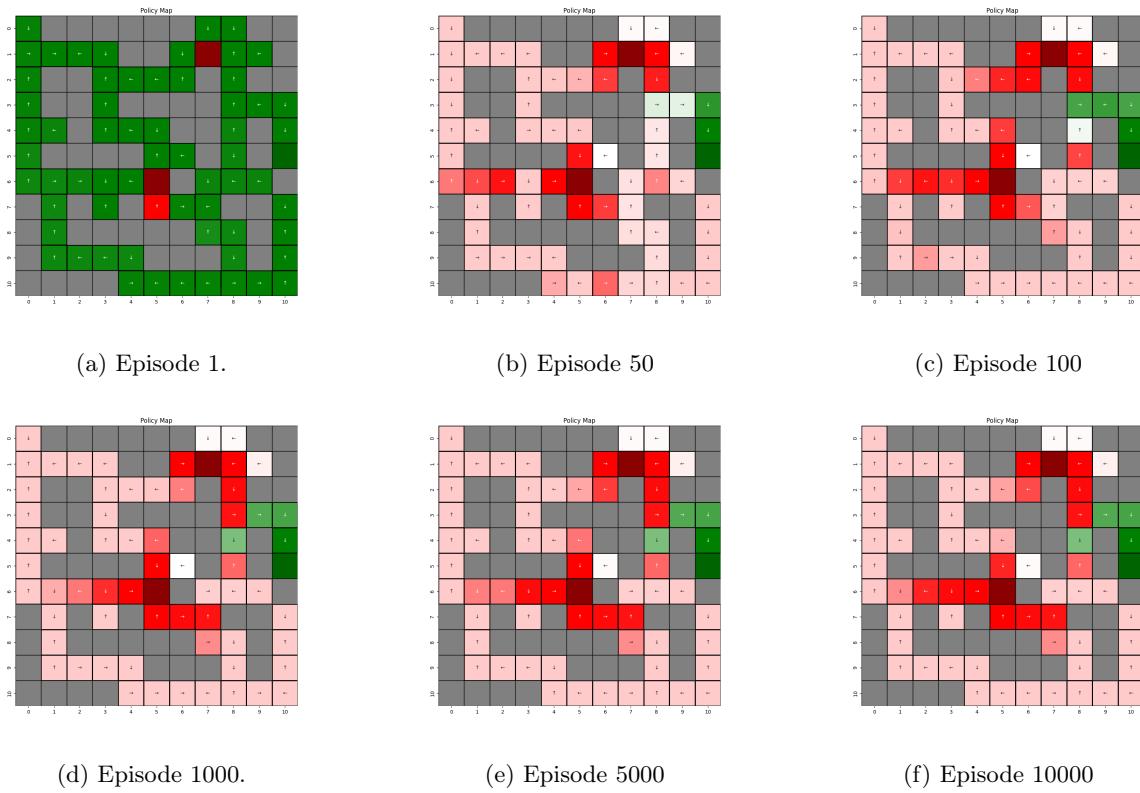


Figure 13: Evolution of policy maps throughout episodes.

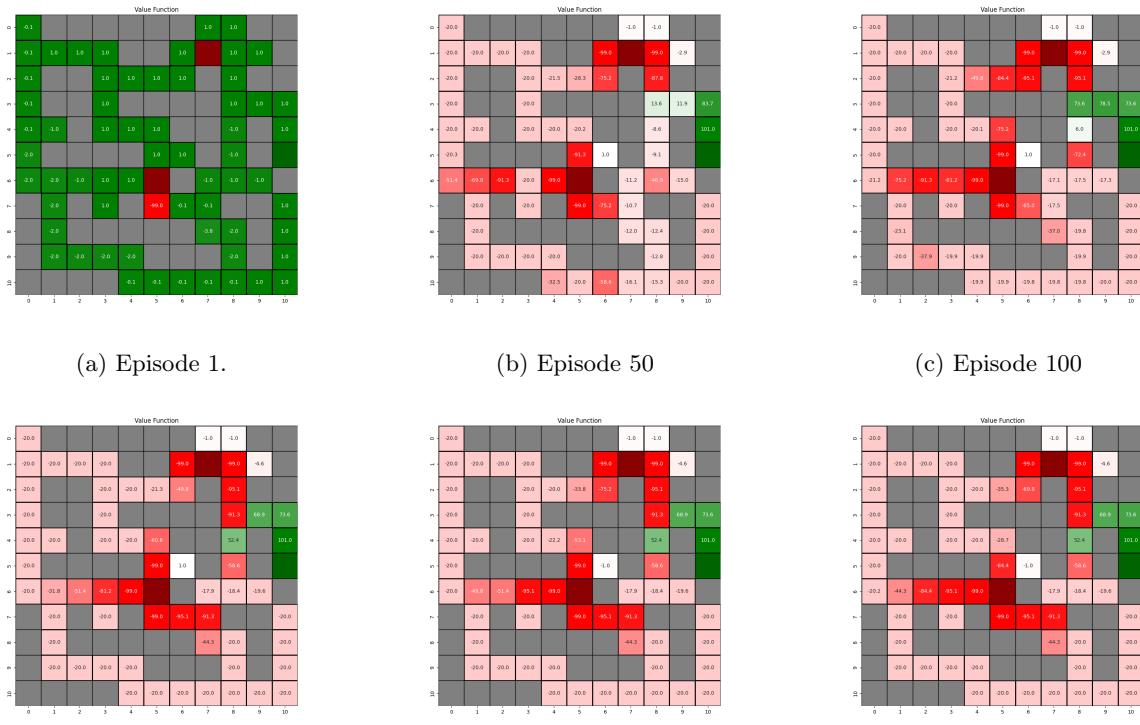


Figure 14: Evolution of value function throughout episodes.

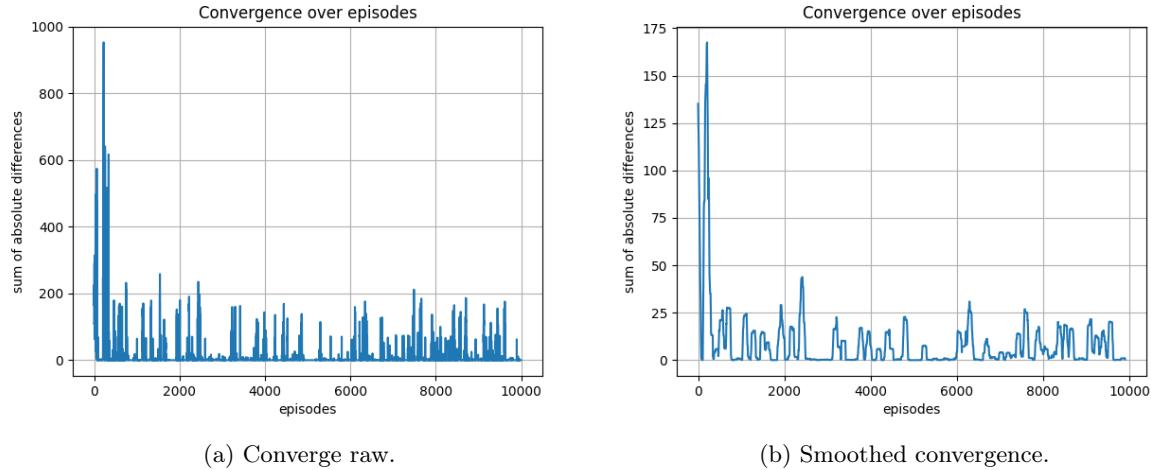


Figure 15: Converge of value function.

What we can interpret from the sweep of alpha value in temporal difference learning as follows. When the alpha value is set to 0.001, the agent try learns the optimal policy and value function as expected. However, the convergence is slower compared to the other alpha values and proper convergence is not observed. When the alpha value is set to 0.5, the agent learns the optimal policy and value function as expected. The convergence is faster compared to the alpha value set to 0.001. Lastly, when the alpha value is set to 1, the agent can not learn the optimal policy and value function since the exploration is too much. The convergence is faster compared to the alpha value set to 0.5 but it is hard to say this is a stable convergence. The reason is, the convergence is not as smooth as the alpha value set to 0.5. So the defaul value and 0.5 are better choices for the alpha parameter in temporal difference learning.

2.4 Effect of Alpha in Q-Learning

Similarly, we can analyze the effect of alpha parameter in Q-learning. The results are provided below.

Figure 16 shows the policy maps for the alpha parameter set to 0.001. Figure 17 illustrates the value function plots for the alpha parameter set to 0.001. Figure 18 presents the convergence plots for the alpha parameter set to 0.001.

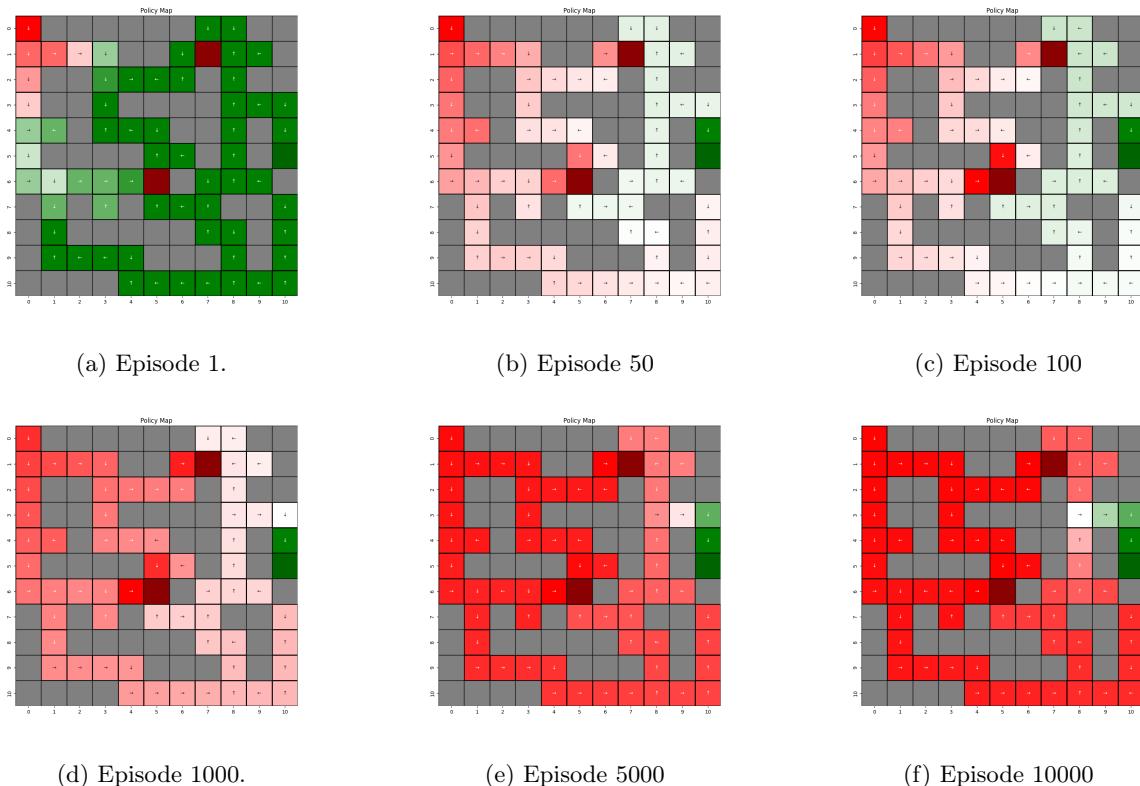


Figure 16: Evolution of policy maps throughout episodes.

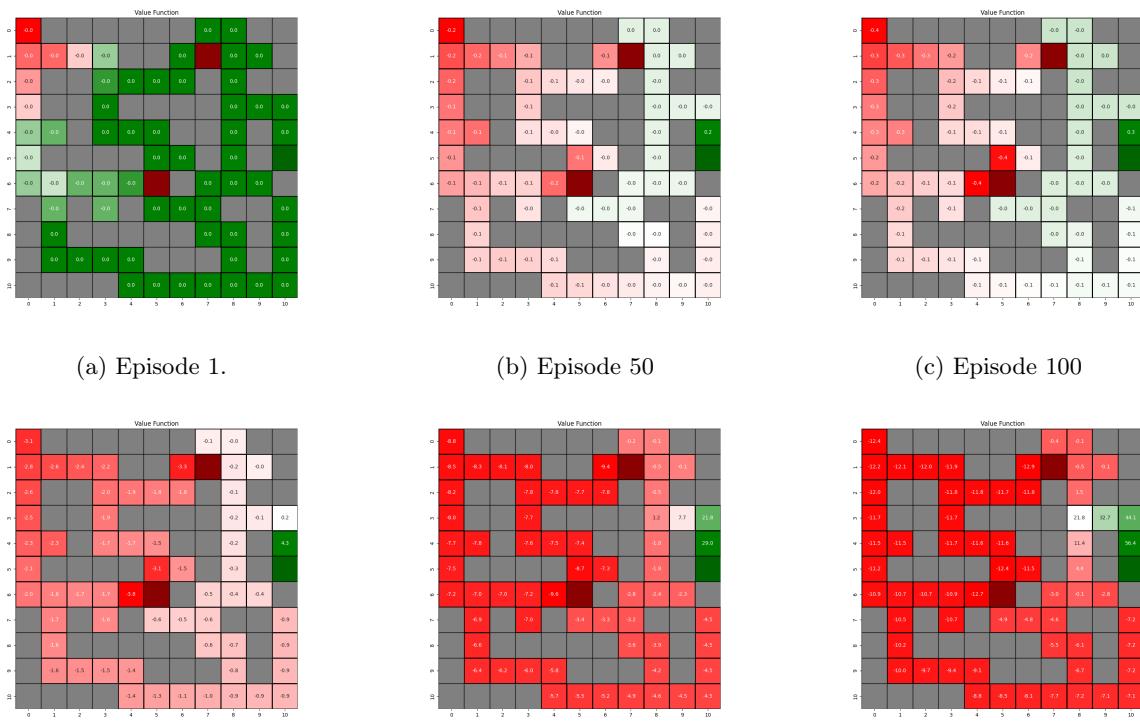


Figure 17: Evolution of value function throughout episodes.

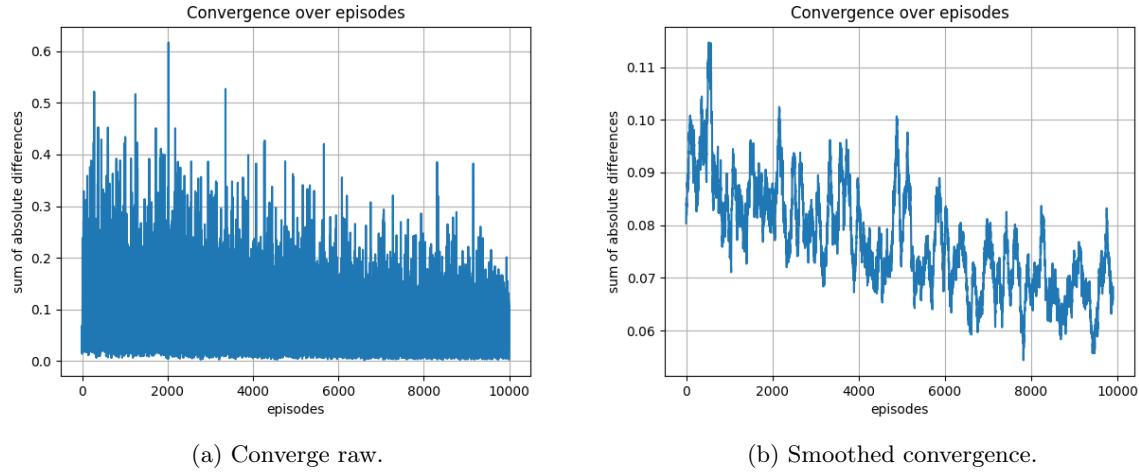


Figure 18: Converge of value function.

Figure 19 shows the policy maps for the alpha parameter set to 0.5. Figure 20 illustrates the value function plots for the alpha parameter set to 0.5. Figure 21 provides the convergence plots for the alpha parameter set to 0.5.

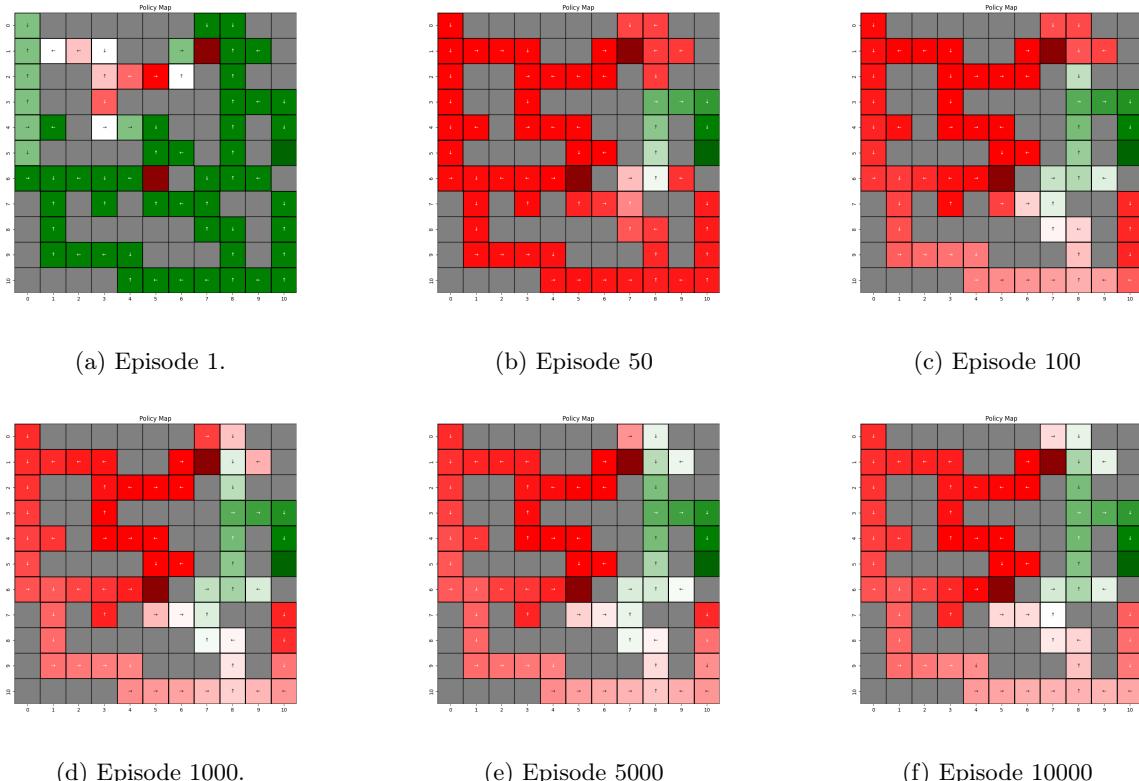


Figure 19: Evolution of policy maps throughout episodes.

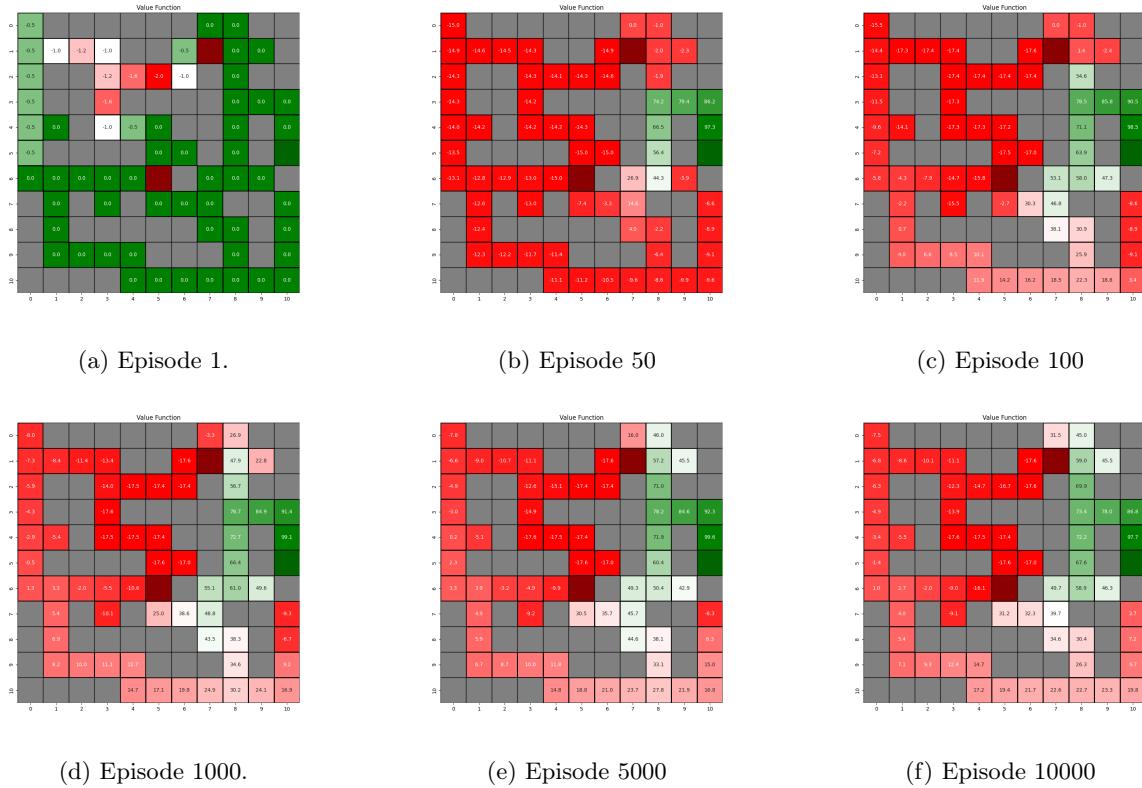


Figure 20: Evolution of value function throughout episodes.

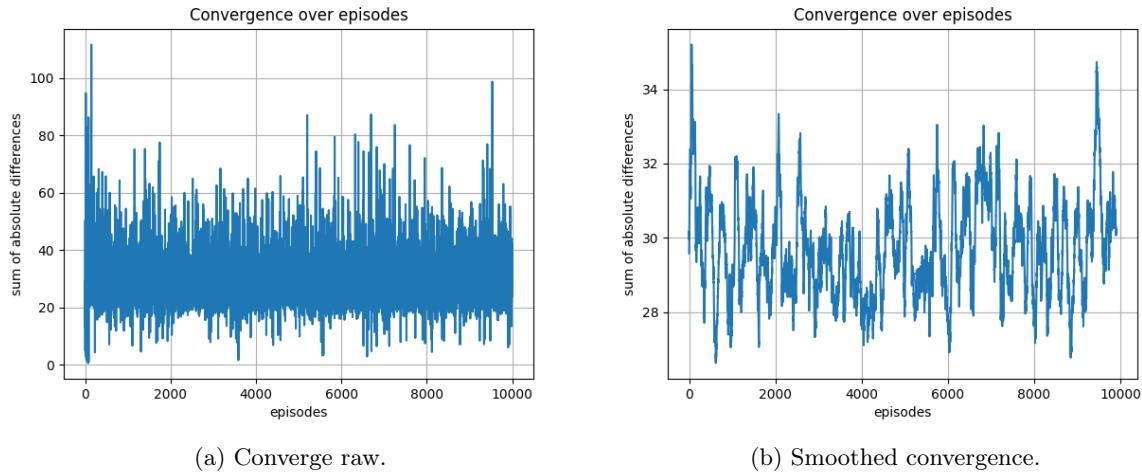


Figure 21: Converge of value function.

Figure 22 shows the policy maps for the alpha parameter set to 1. Figure 23 illustrates the value function plots for the alpha parameter set to 1. Figure 24 provides the convergence plots for the alpha parameter set to 1.

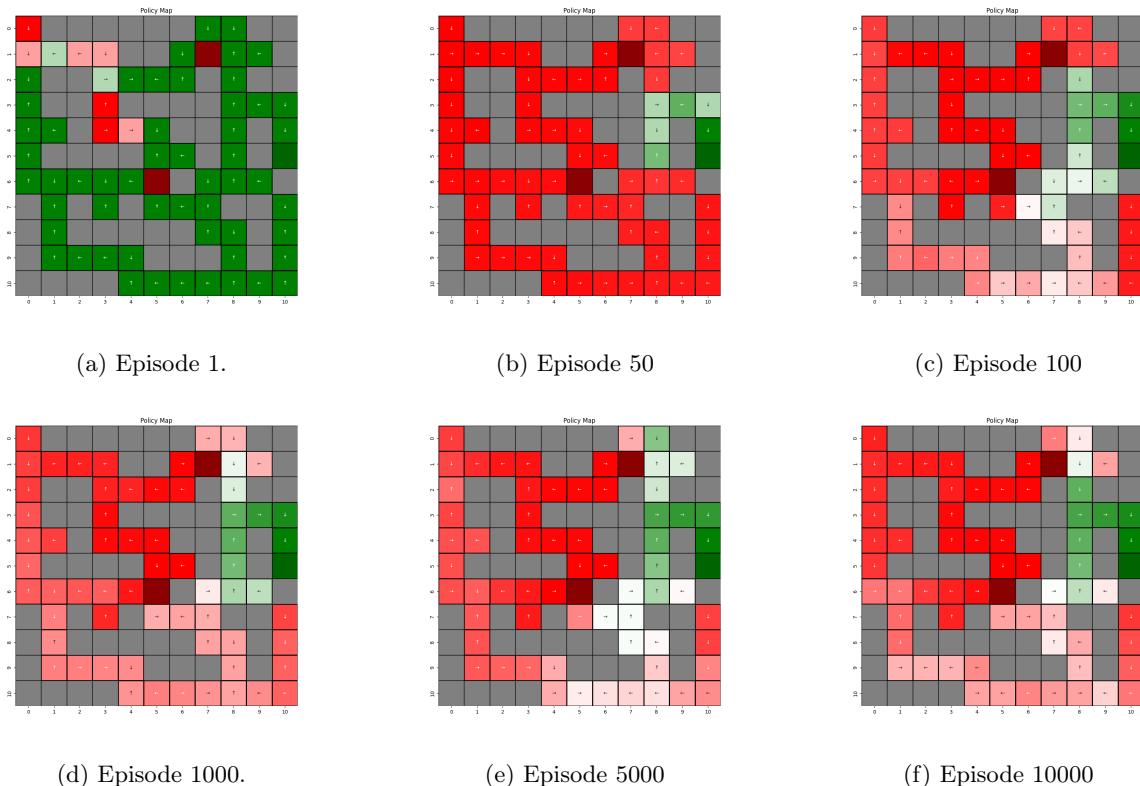


Figure 22: Evolution of policy maps throughout episodes.

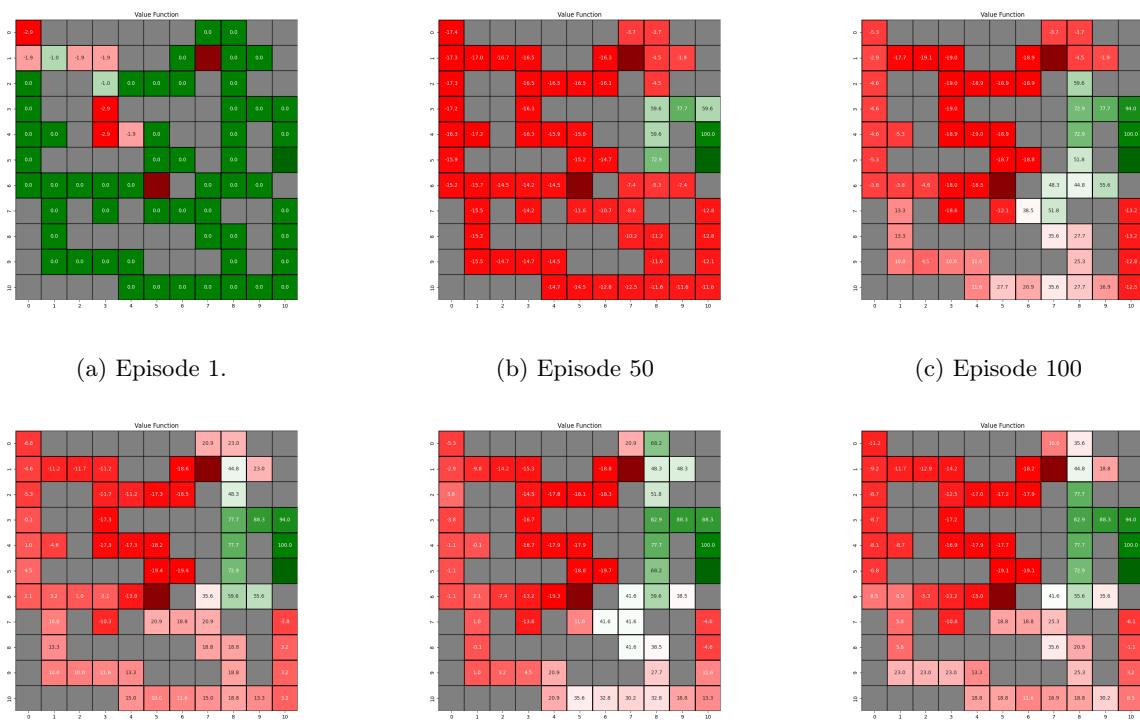


Figure 23: Evolution of value function throughout episodes.

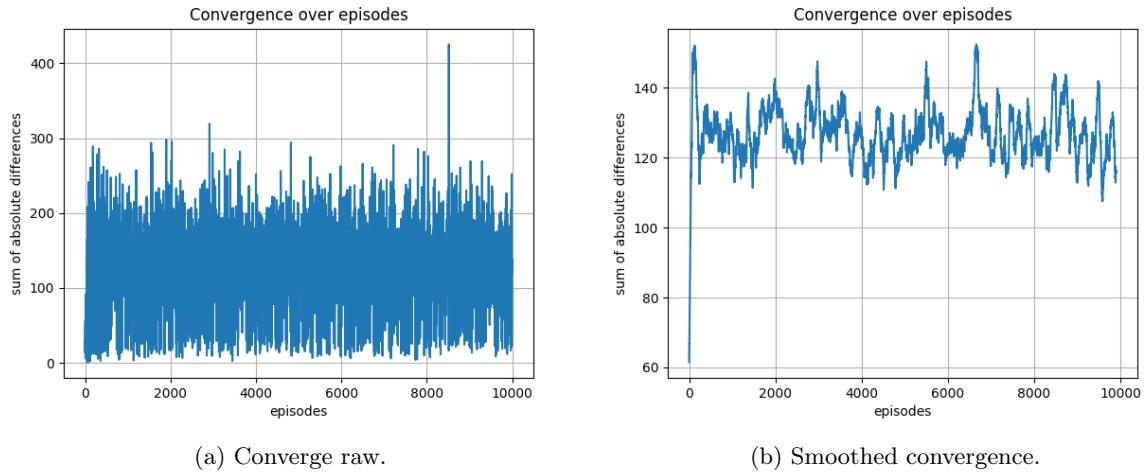


Figure 24: Converge of value function.

2.5 Effect of Gamma in Temporal Difference Learning

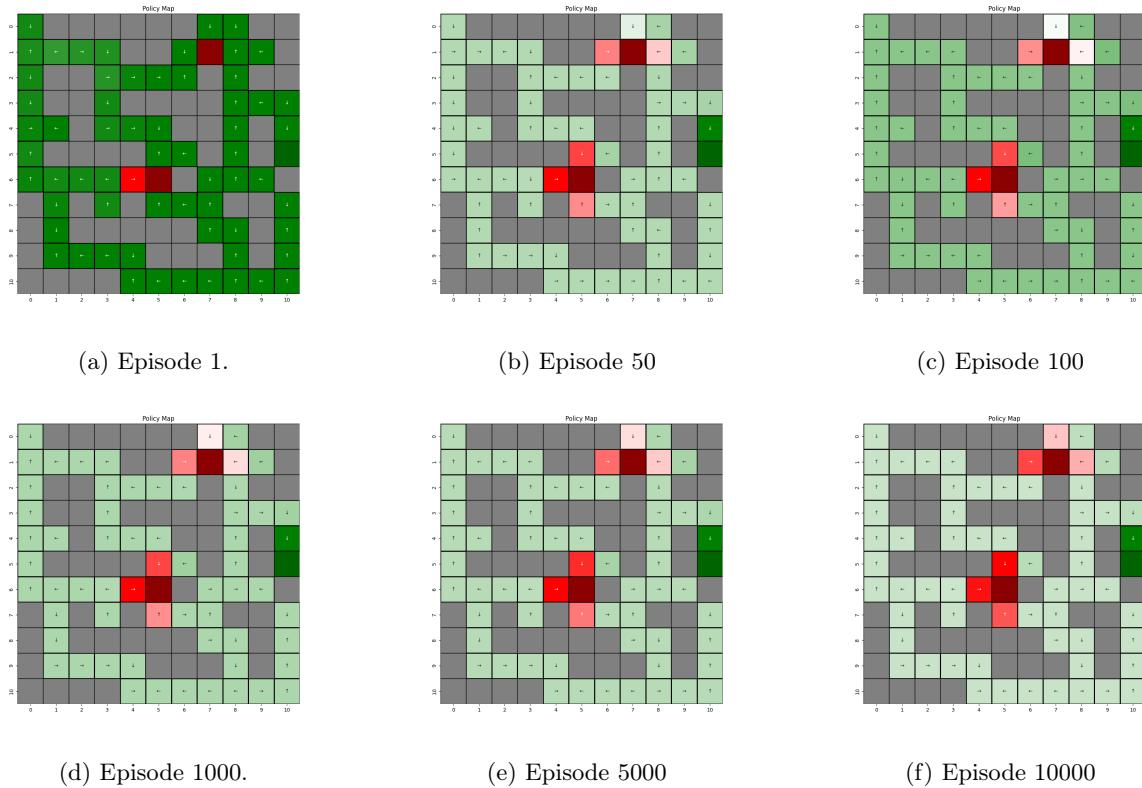


Figure 25: Evolution of policy maps throughout episodes.

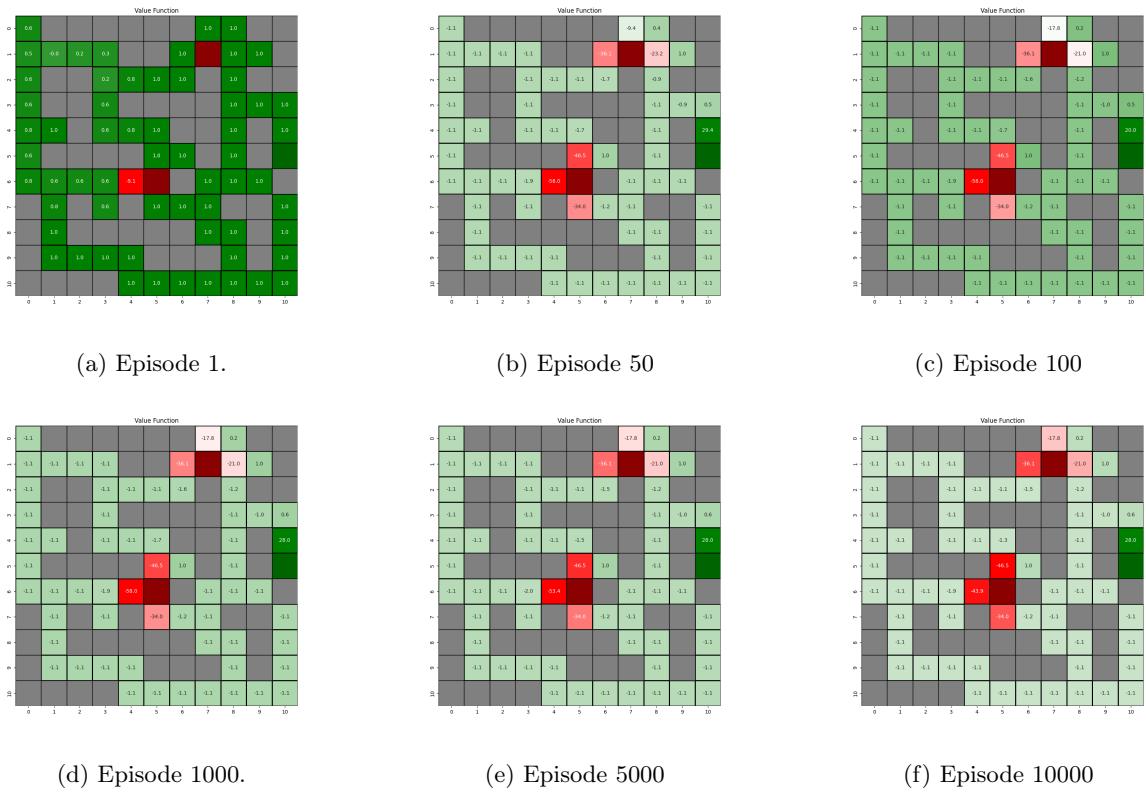


Figure 26: Evolution of value function throughout episodes.

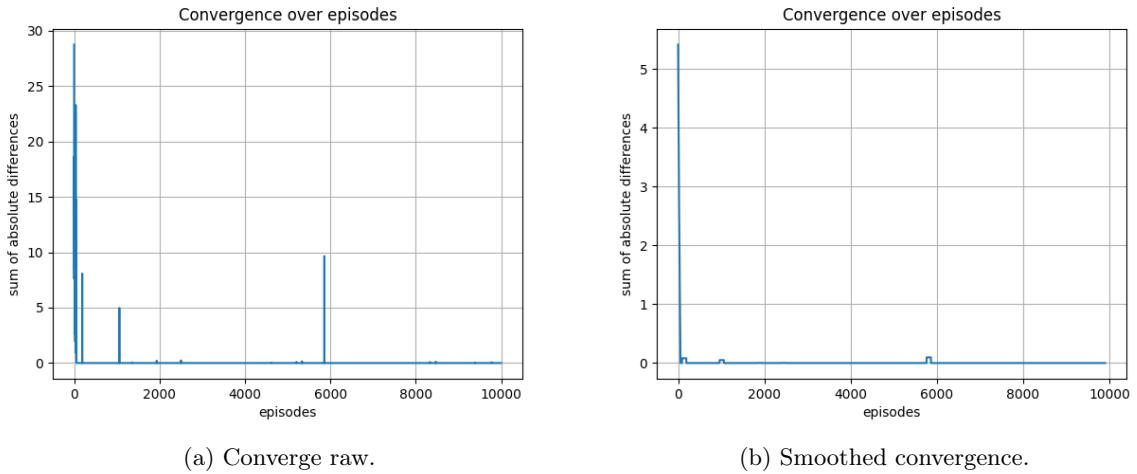


Figure 27: Convergence of value function.

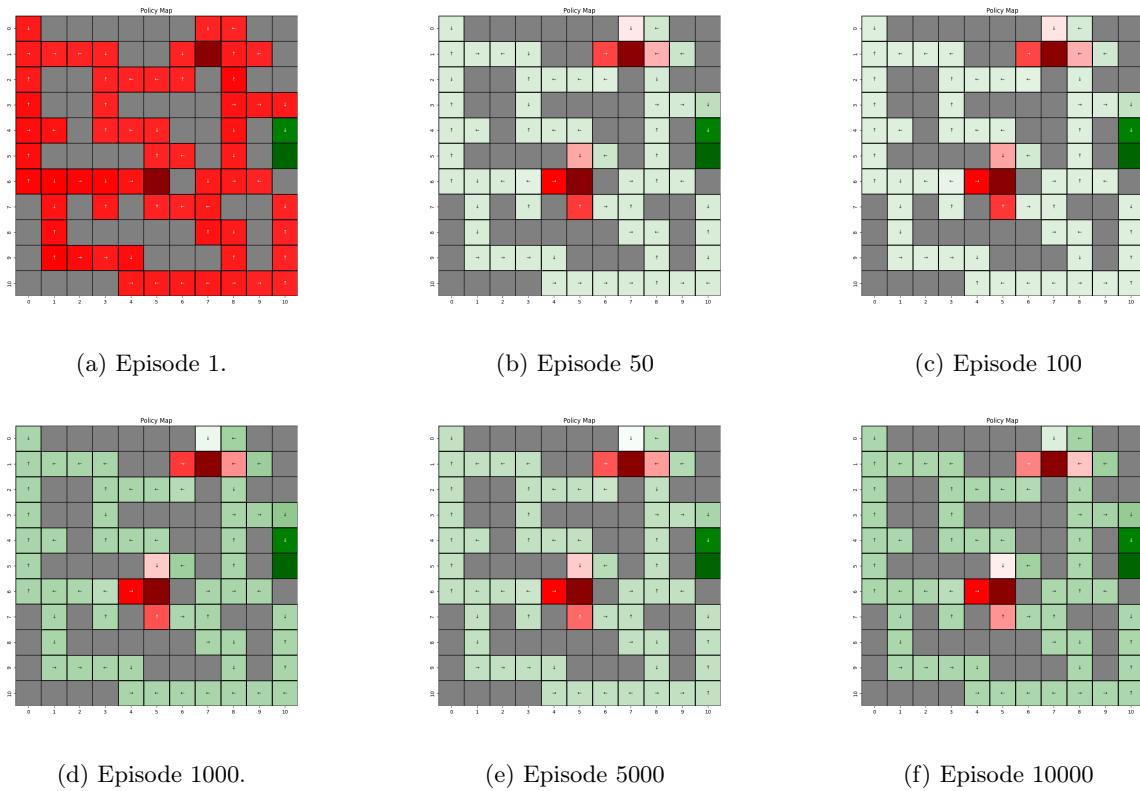


Figure 28: Evolution of policy maps throughout episodes.

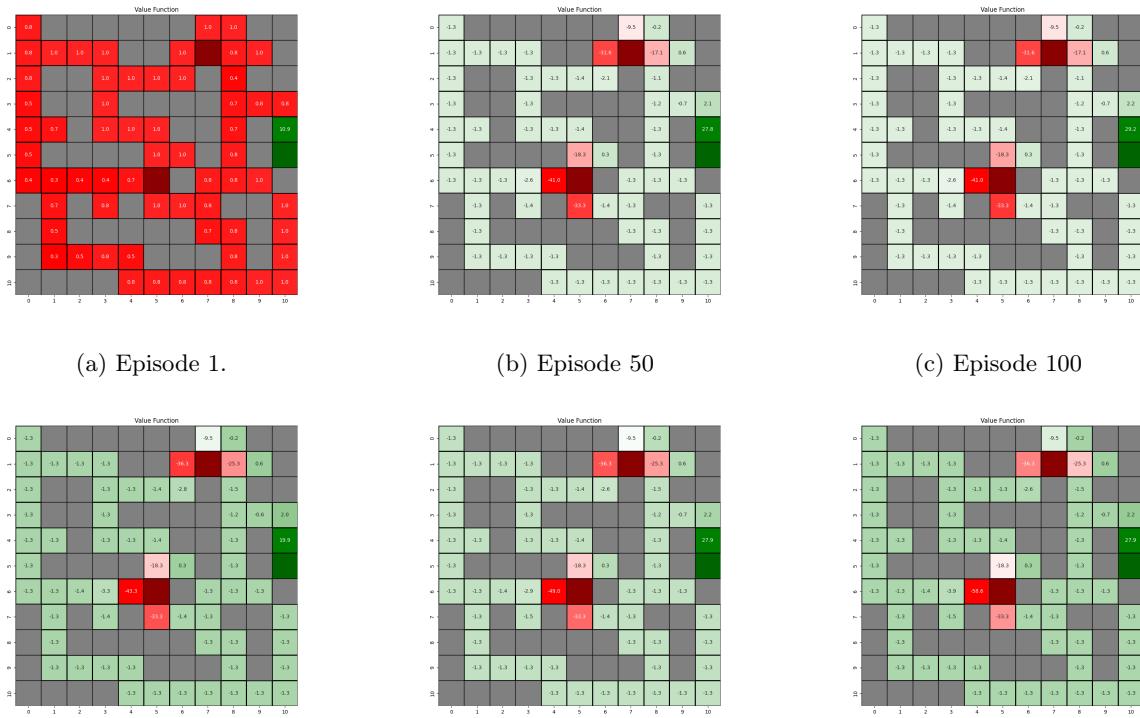


Figure 29: Evolution of value function throughout episodes.

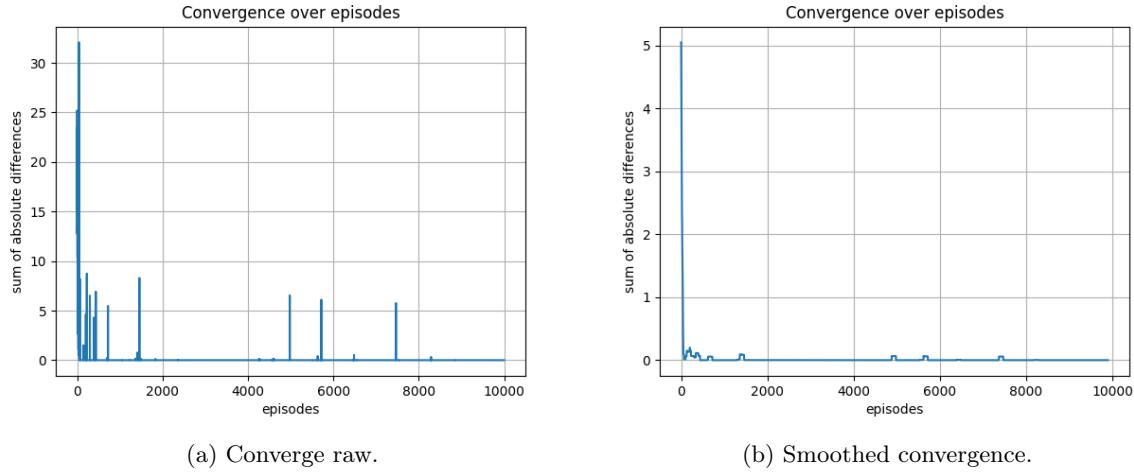


Figure 30: Converge of value function.

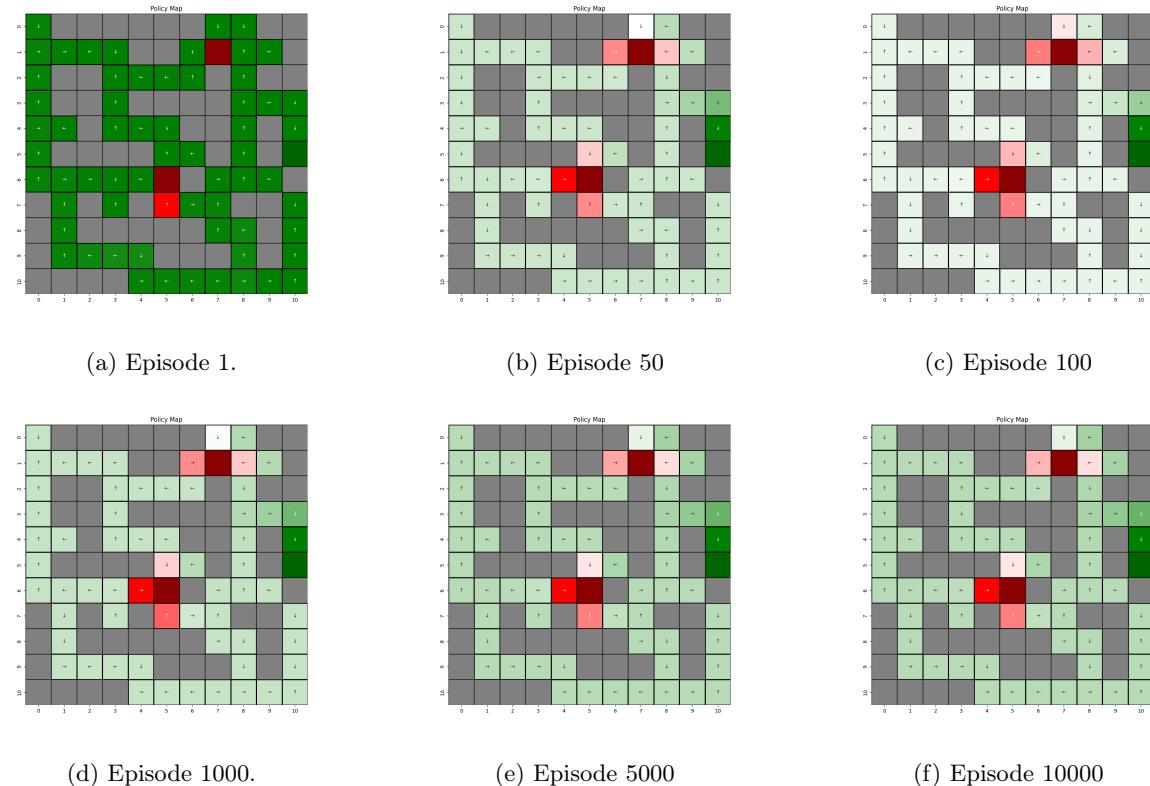


Figure 31: Evolution of policy maps throughout episodes.

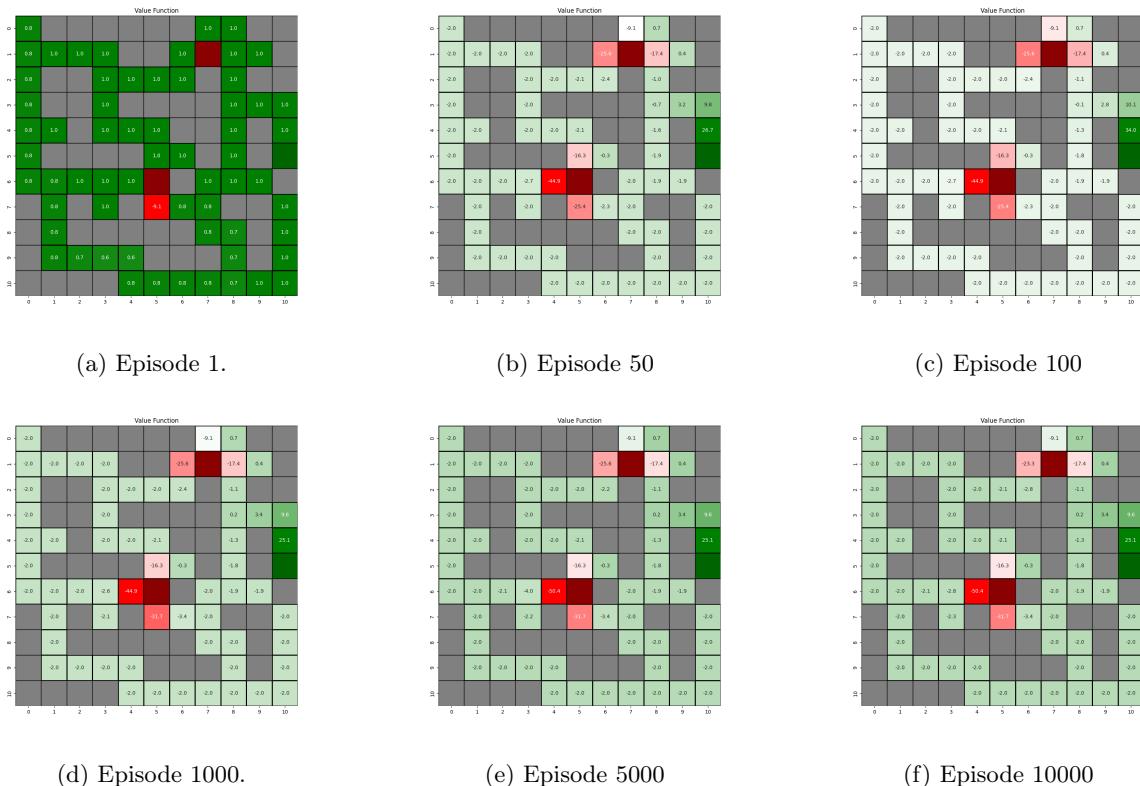


Figure 32: Evolution of value function throughout episodes.

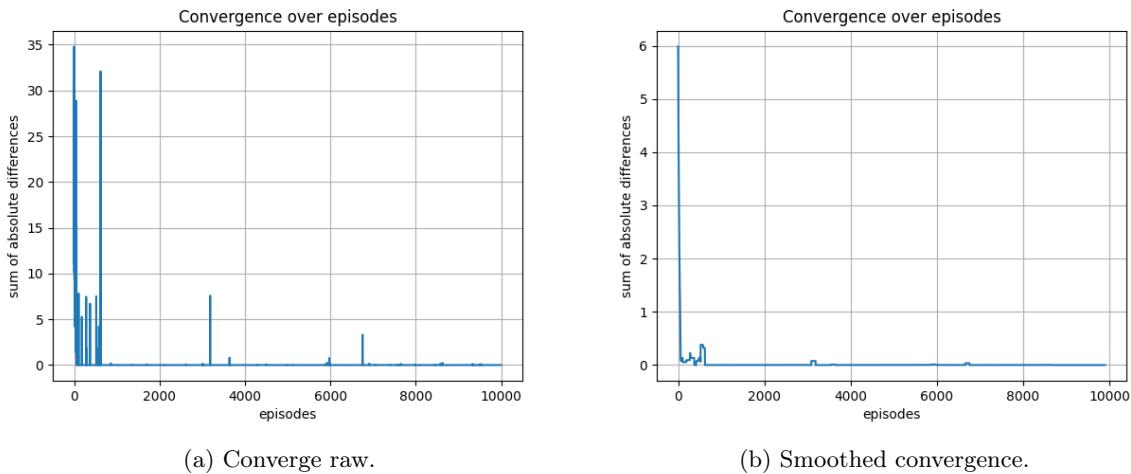


Figure 33: Convergence of value function.

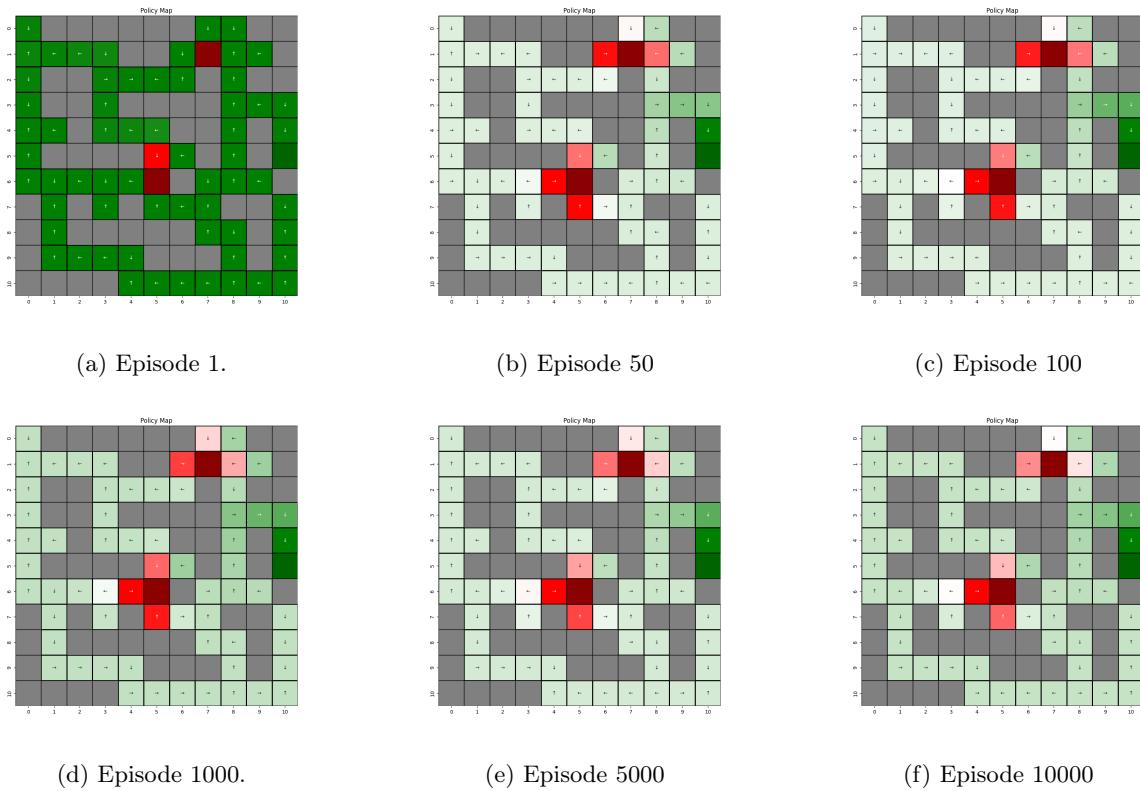


Figure 34: Evolution of policy maps throughout episodes.

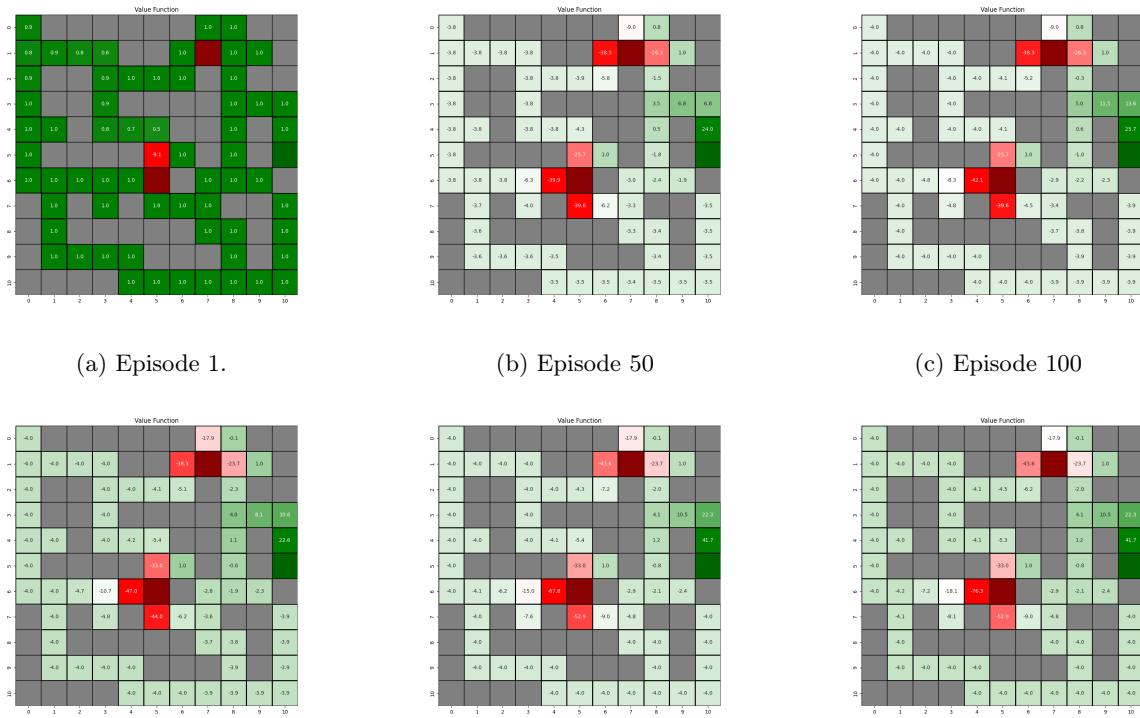


Figure 35: Evolution of value function throughout episodes.

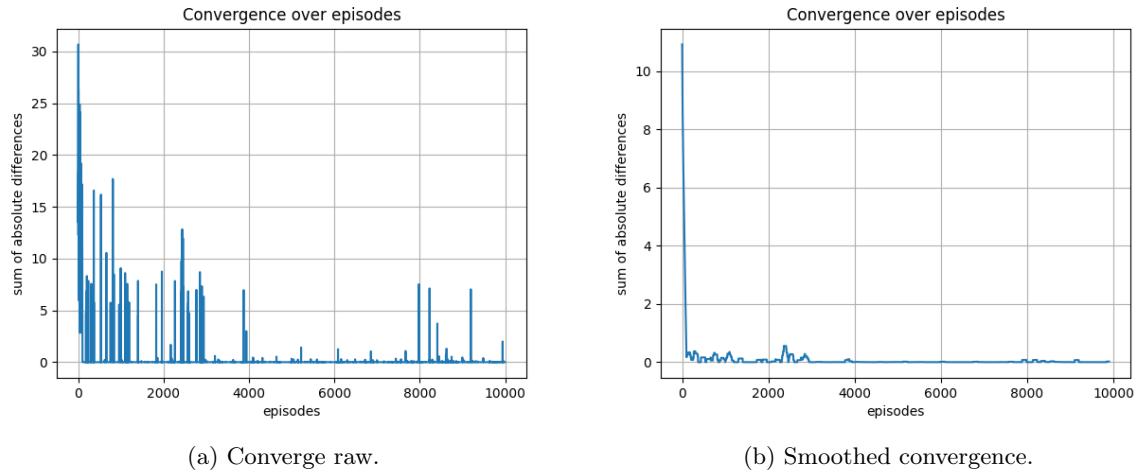


Figure 36: Converge of value function.

2.6 Effect of Gamma in Q-Learning

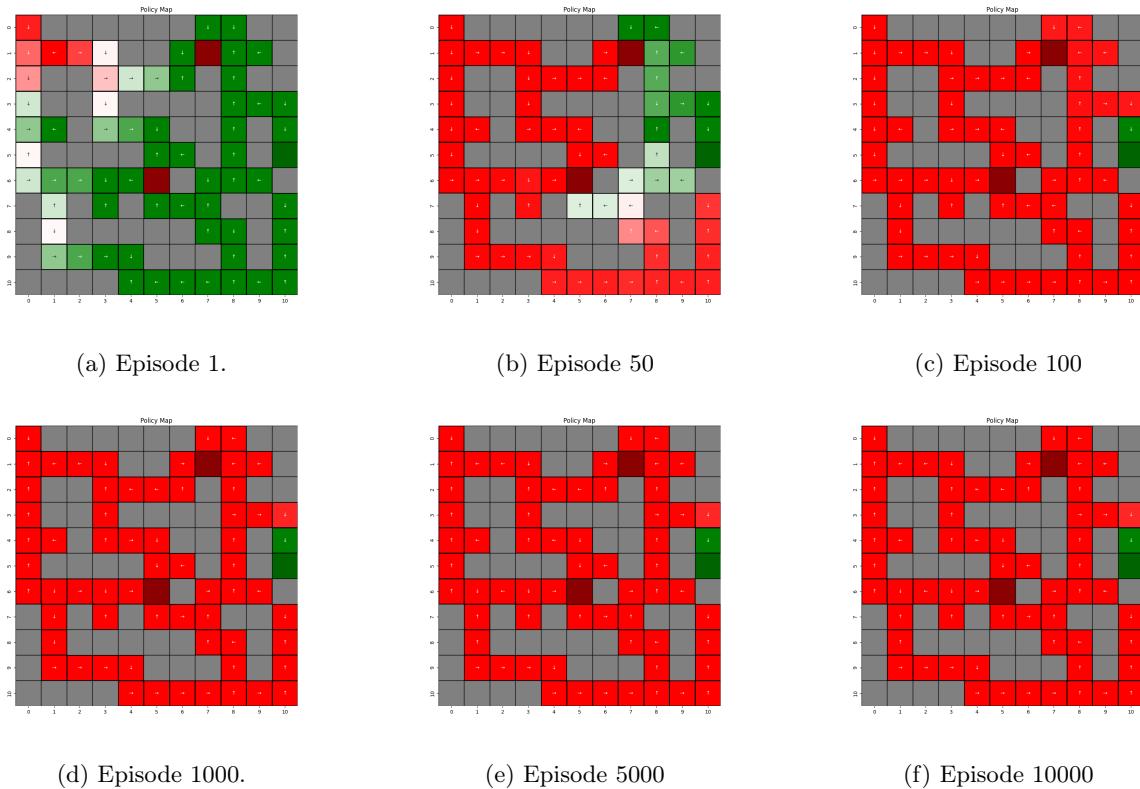


Figure 37: Evolution of policy maps throughout episodes.

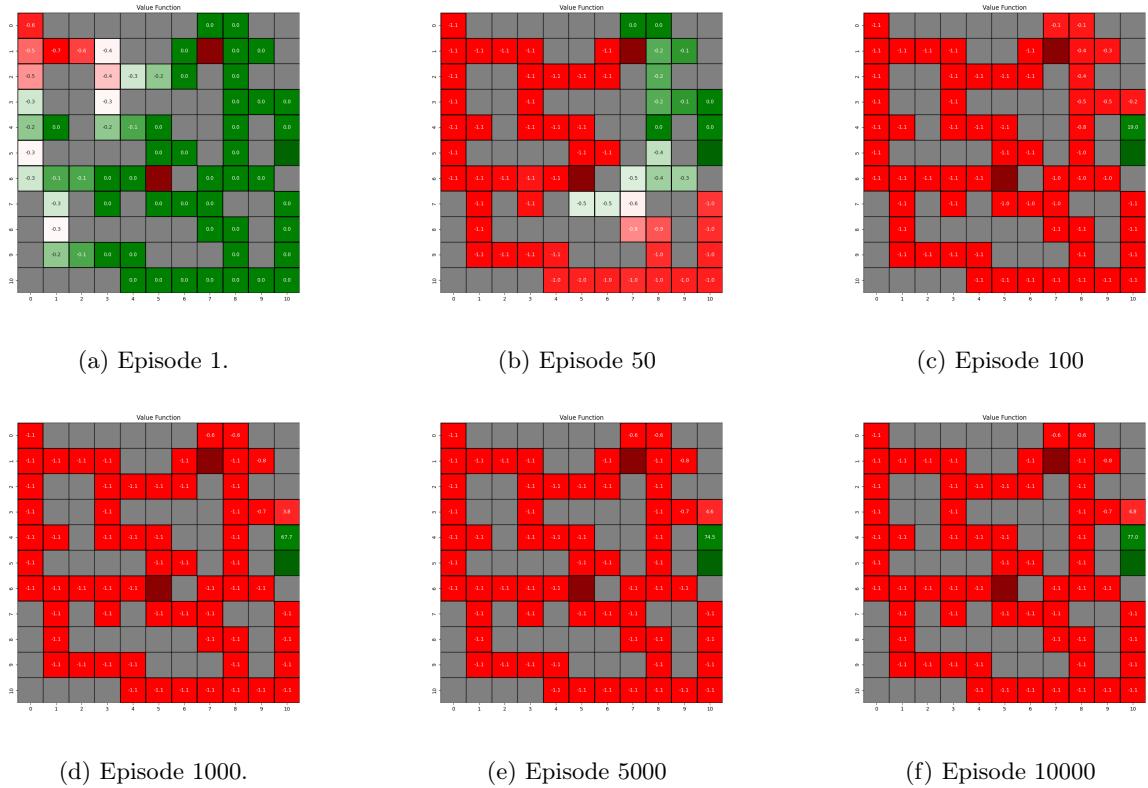


Figure 38: Evolution of value function throughout episodes.

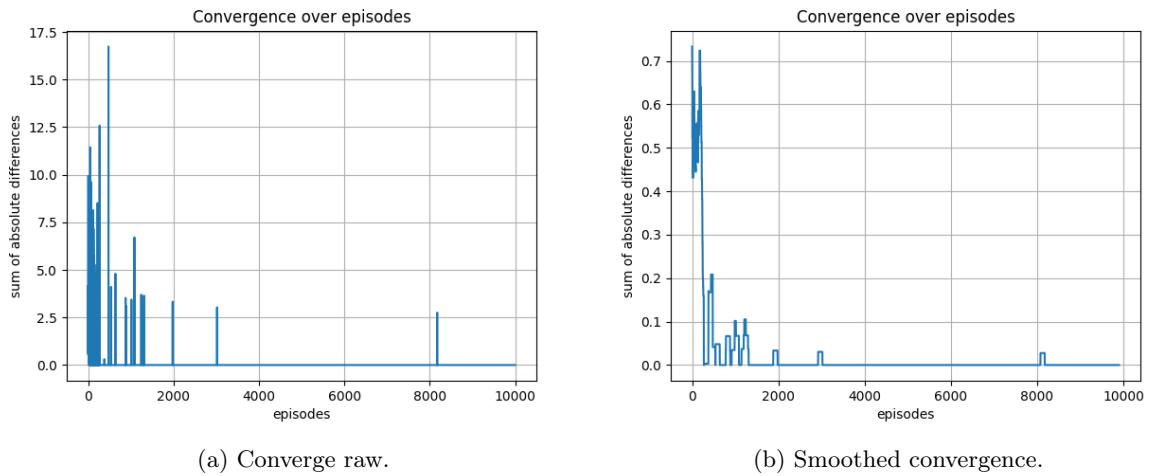


Figure 39: Convergence of value function.

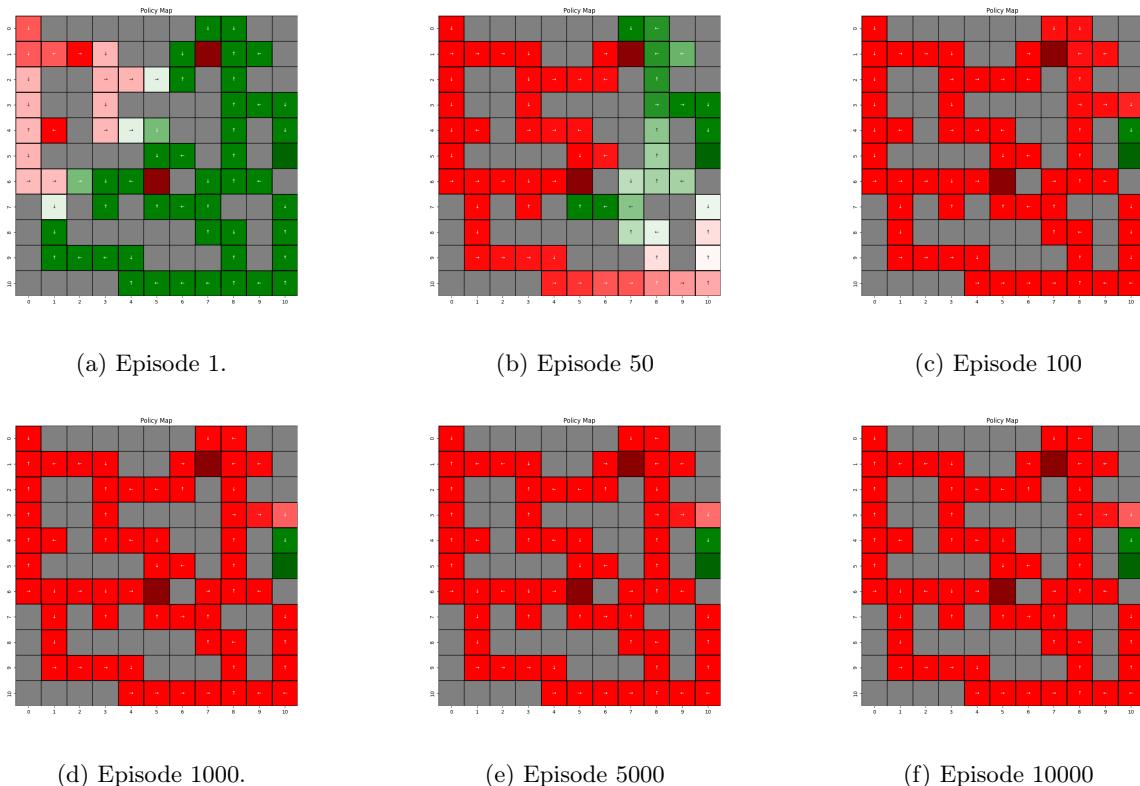


Figure 40: Evolution of policy maps throughout episodes.

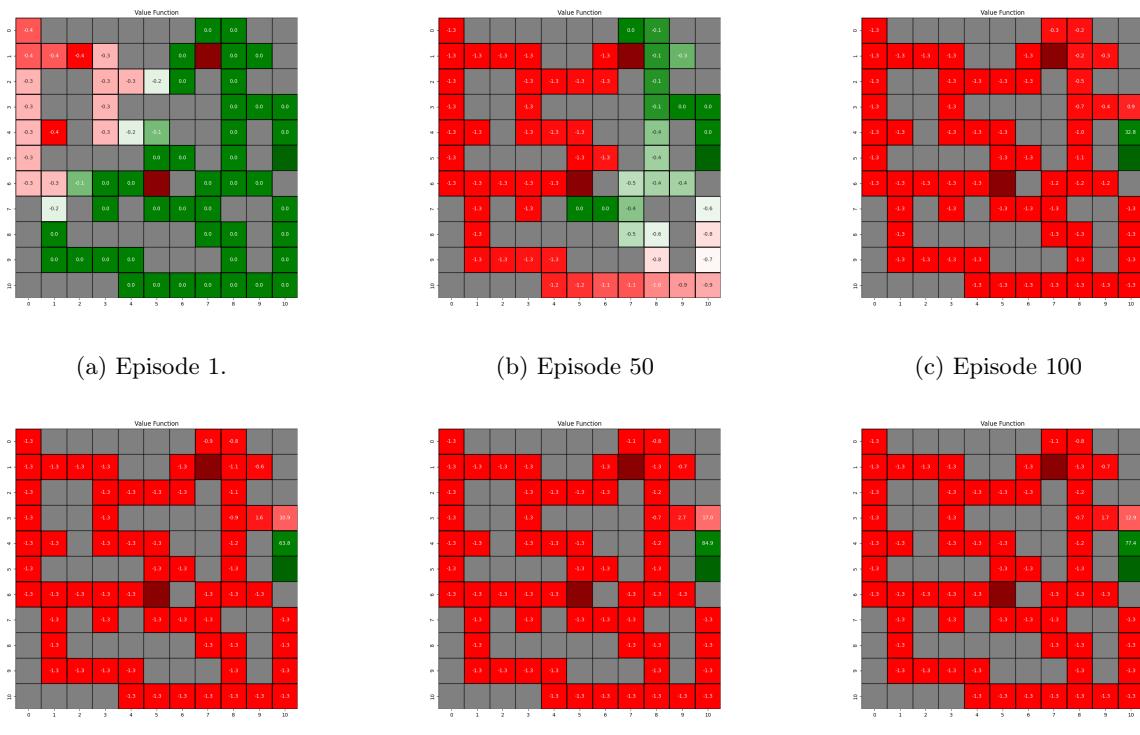


Figure 41: Evolution of value function throughout episodes.

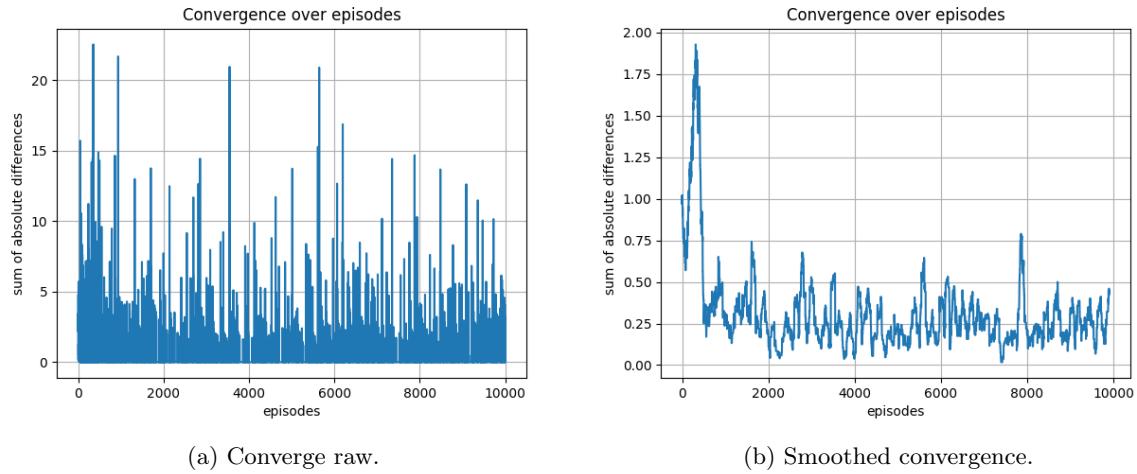


Figure 42: Converge of value function.

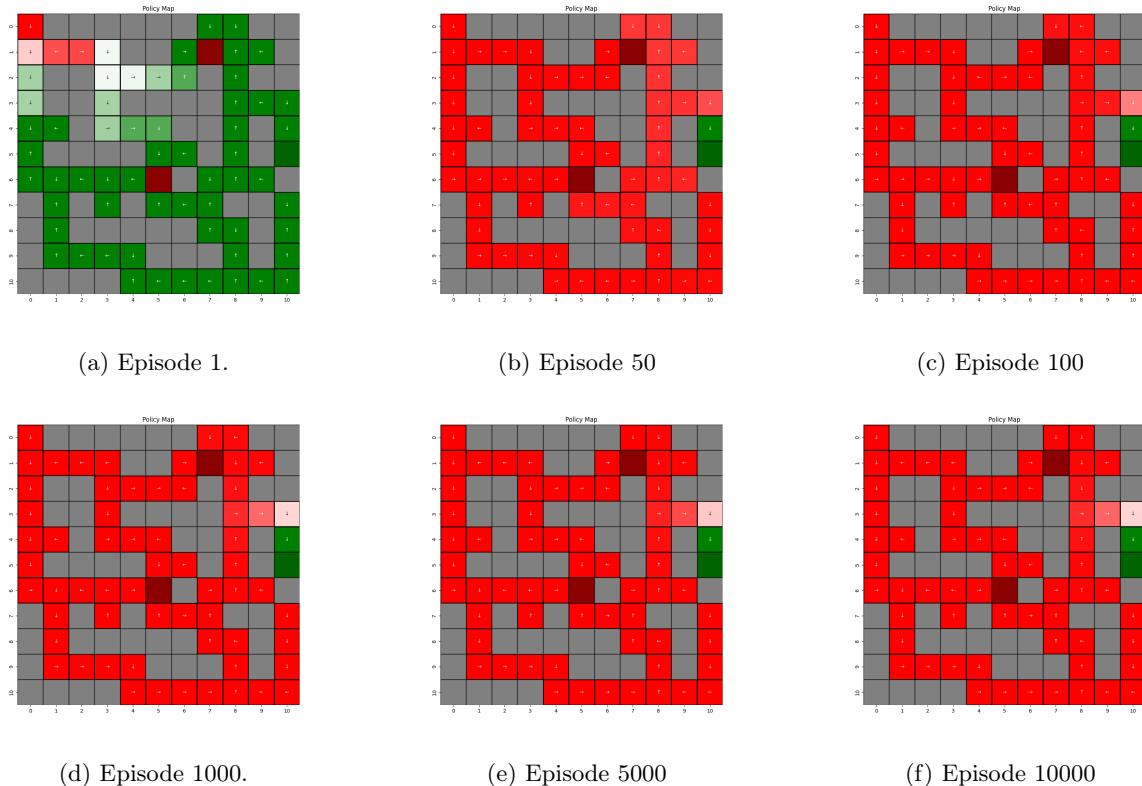


Figure 43: Evolution of policy maps throughout episodes.

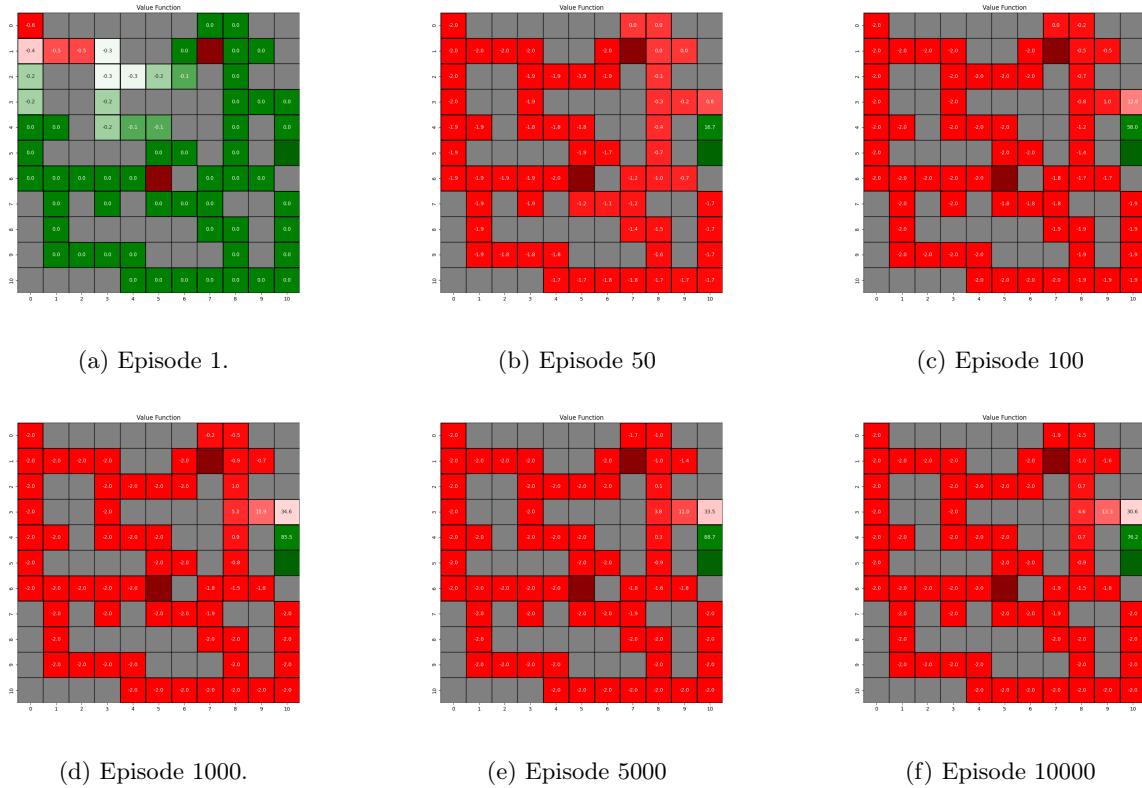


Figure 44: Evolution of value function throughout episodes.

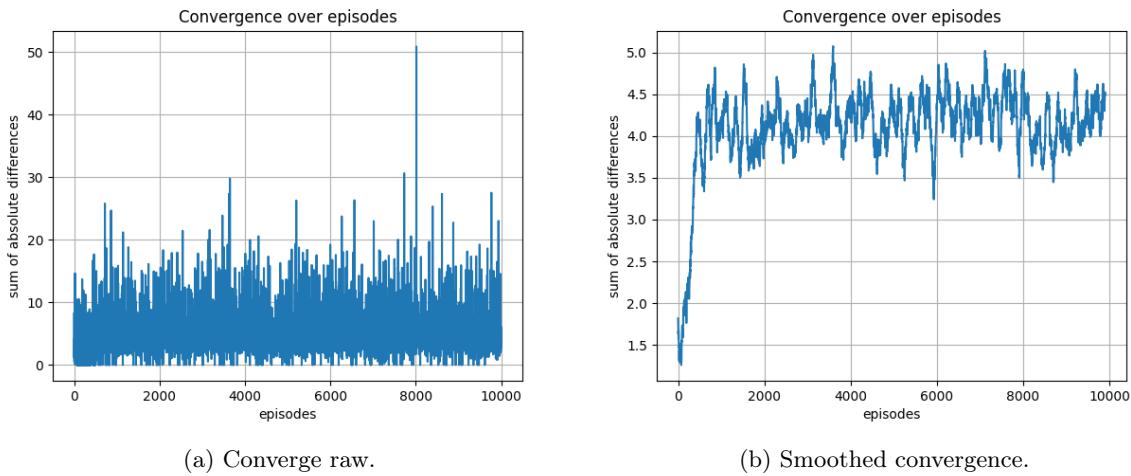


Figure 45: Convergence of value function.

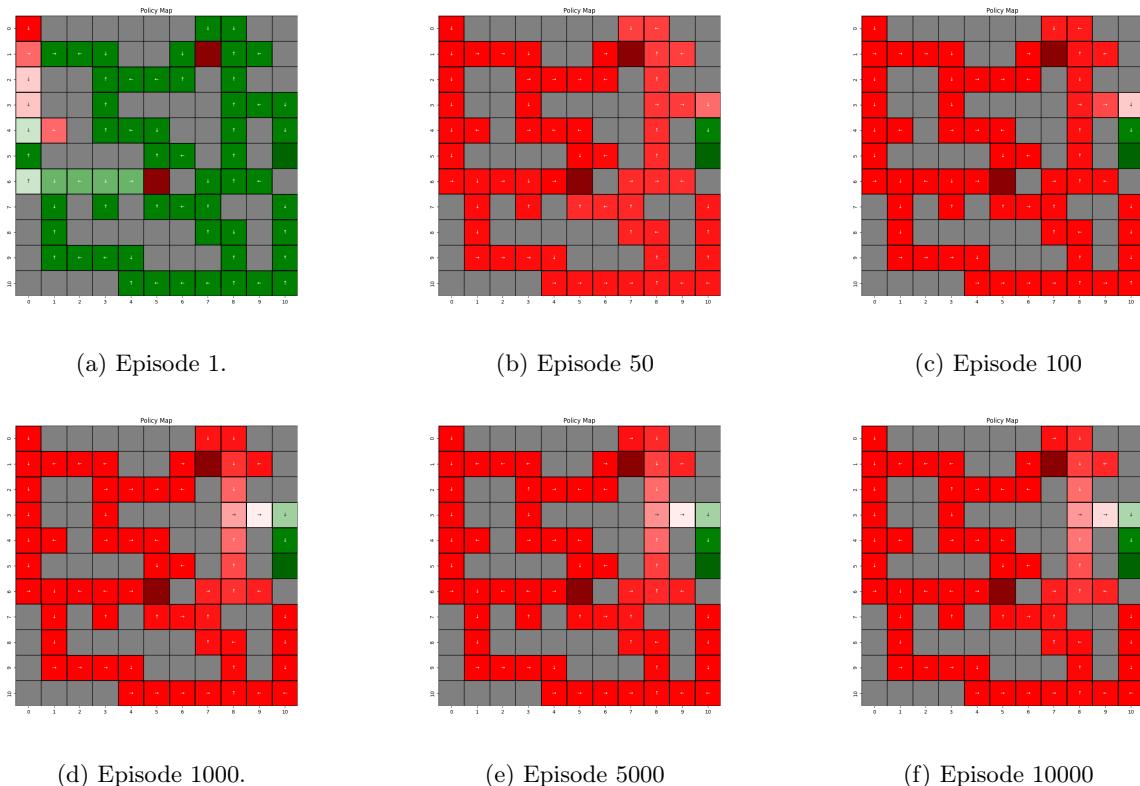


Figure 46: Evolution of policy maps throughout episodes.

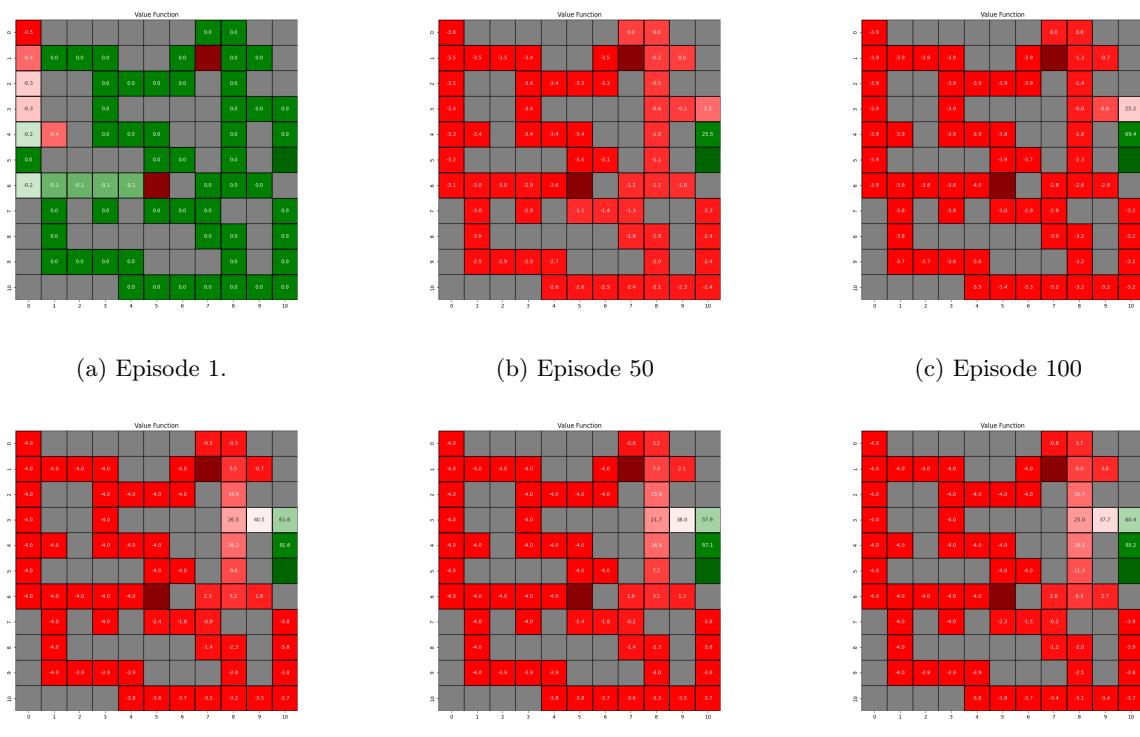


Figure 47: Evolution of value function throughout episodes.

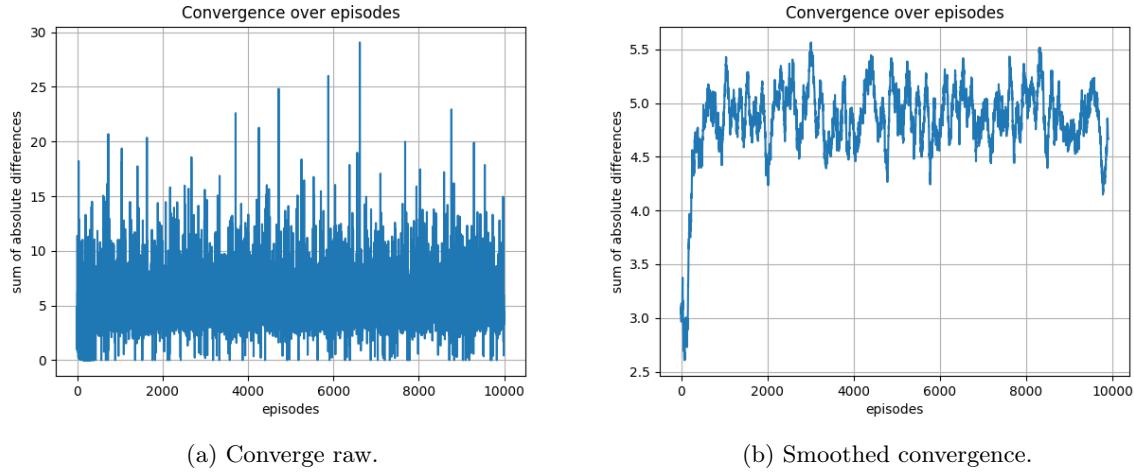


Figure 48: Converge of value function.

2.7 Effect of Epsilon in Temporal Difference Learning

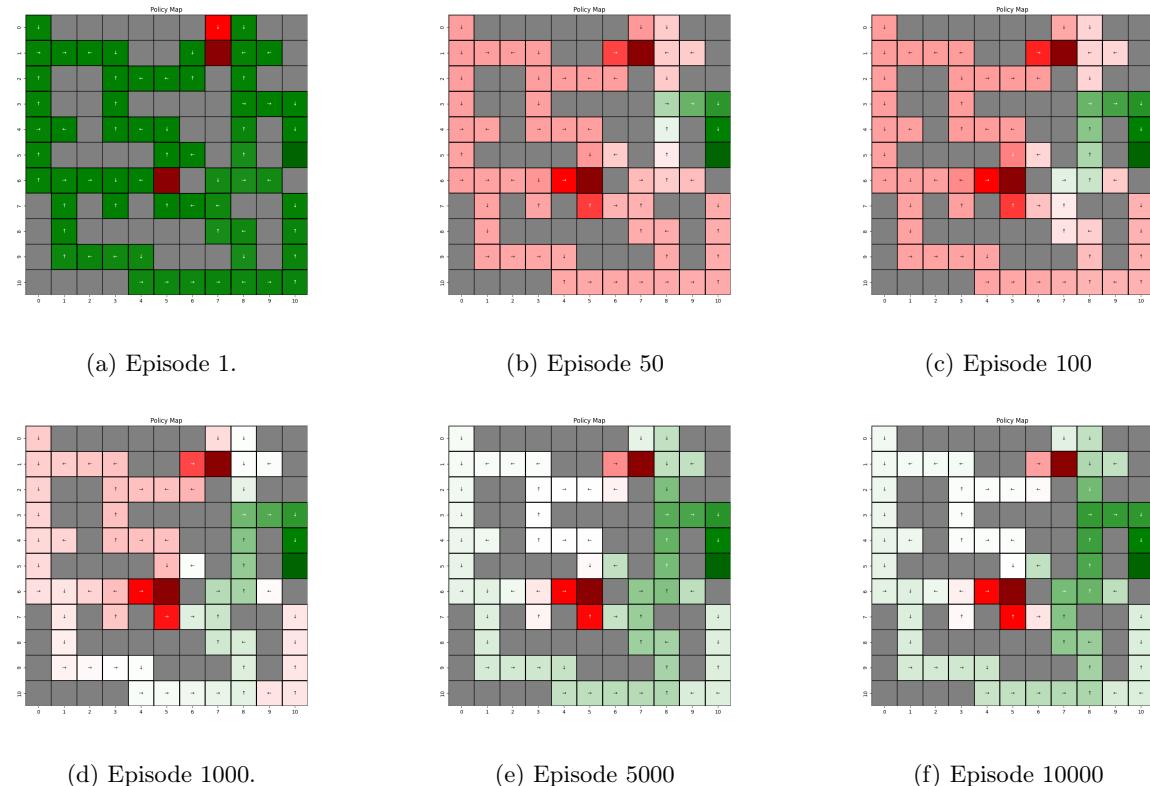


Figure 49: Evolution of policy maps throughout episodes.

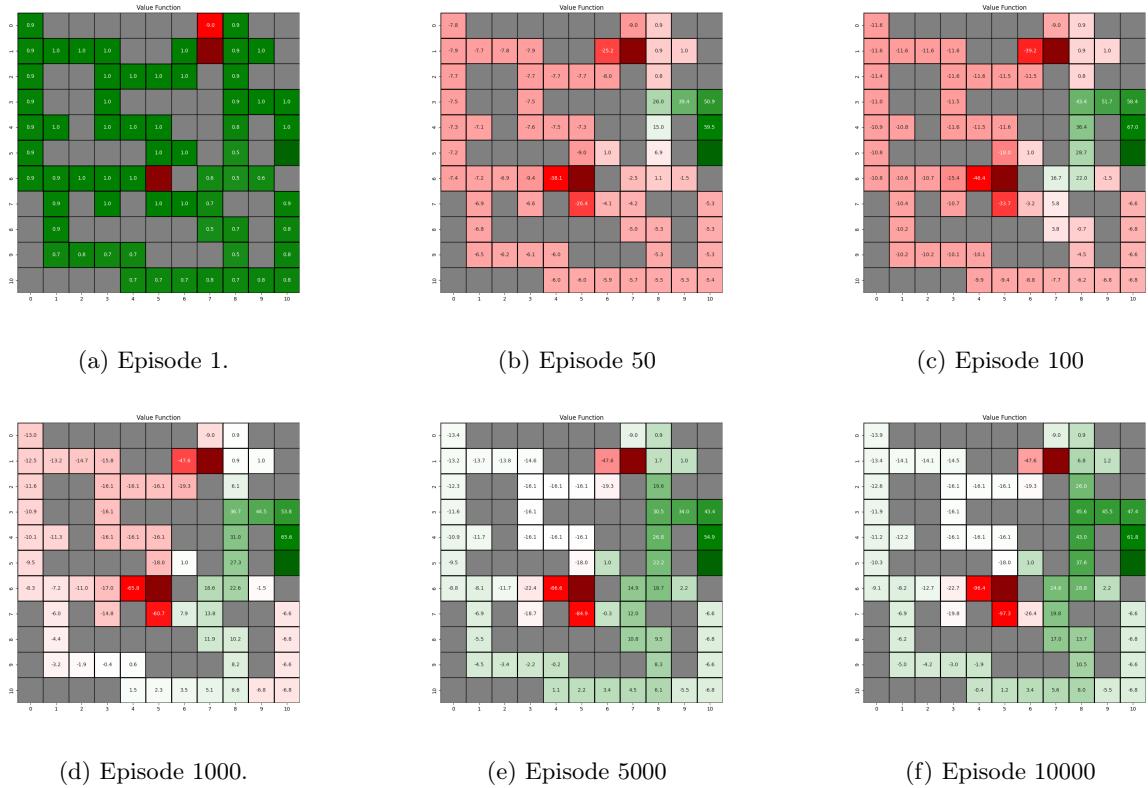


Figure 50: Evolution of value function throughout episodes.

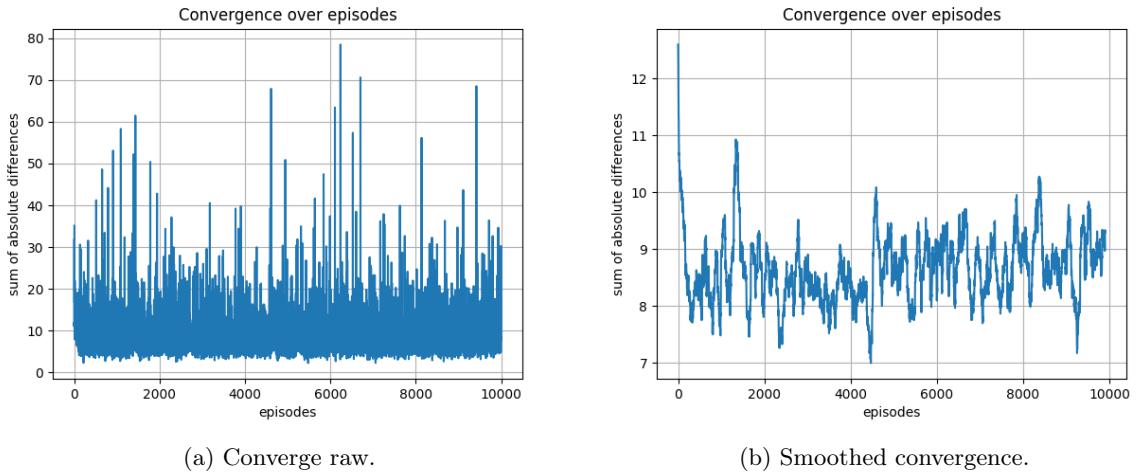


Figure 51: Convergence of value function.

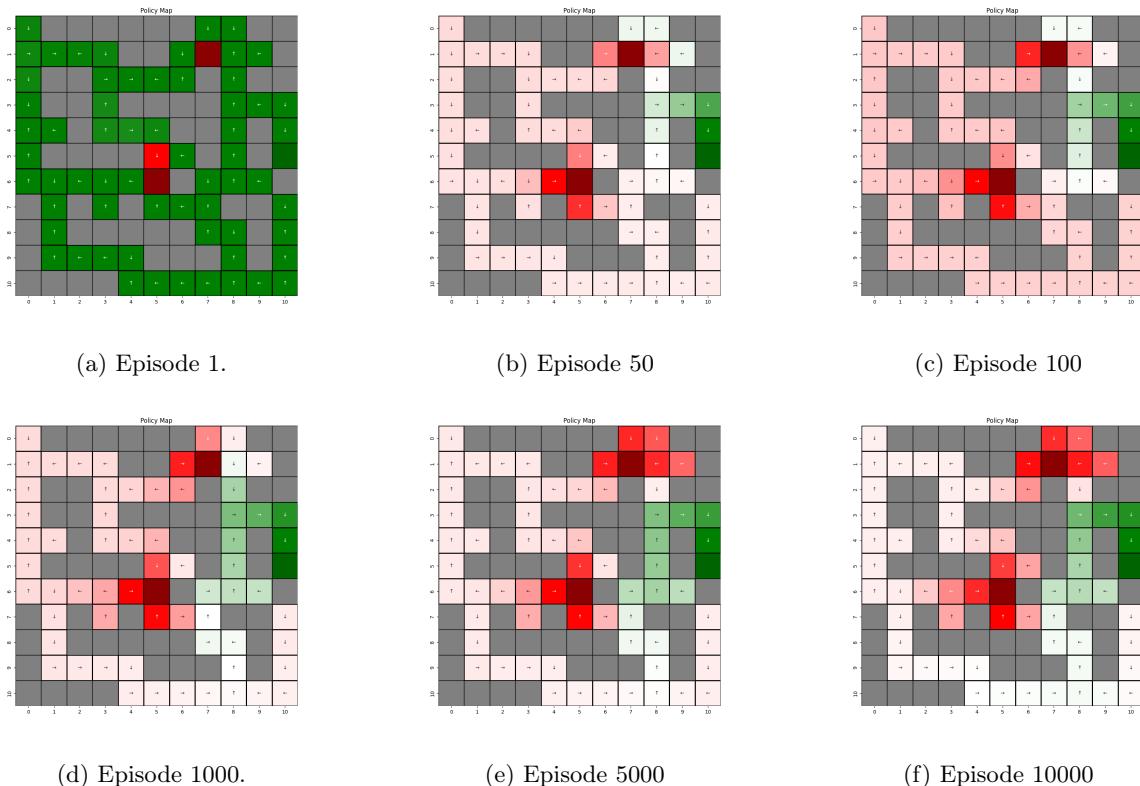


Figure 52: Evolution of policy maps throughout episodes.

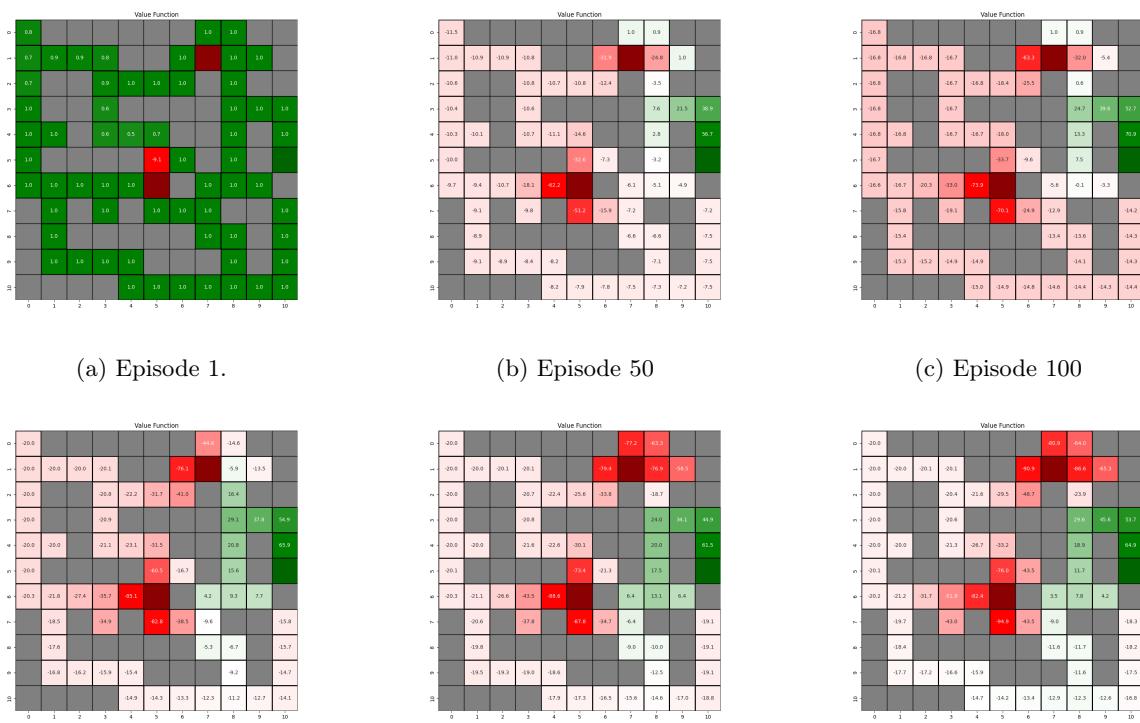


Figure 53: Evolution of value function throughout episodes.

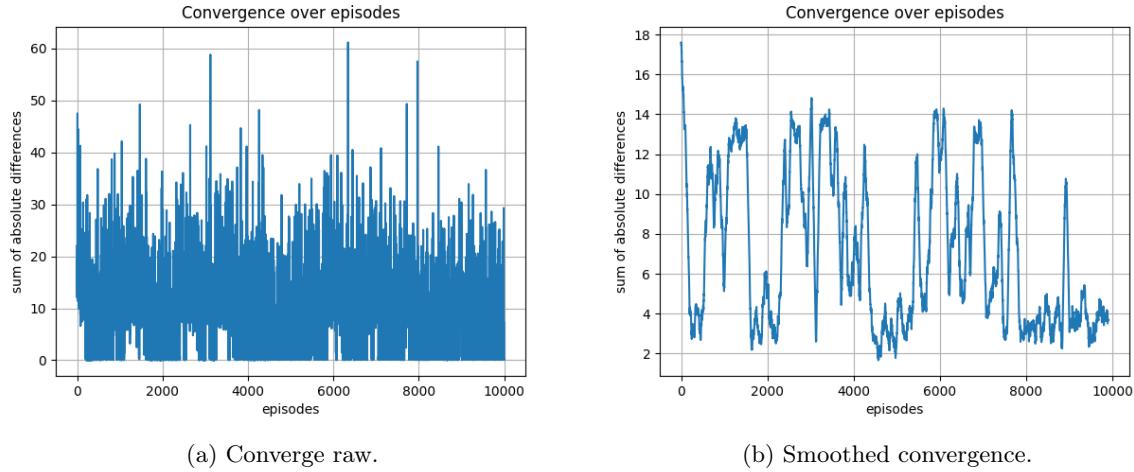


Figure 54: Converge of value function.

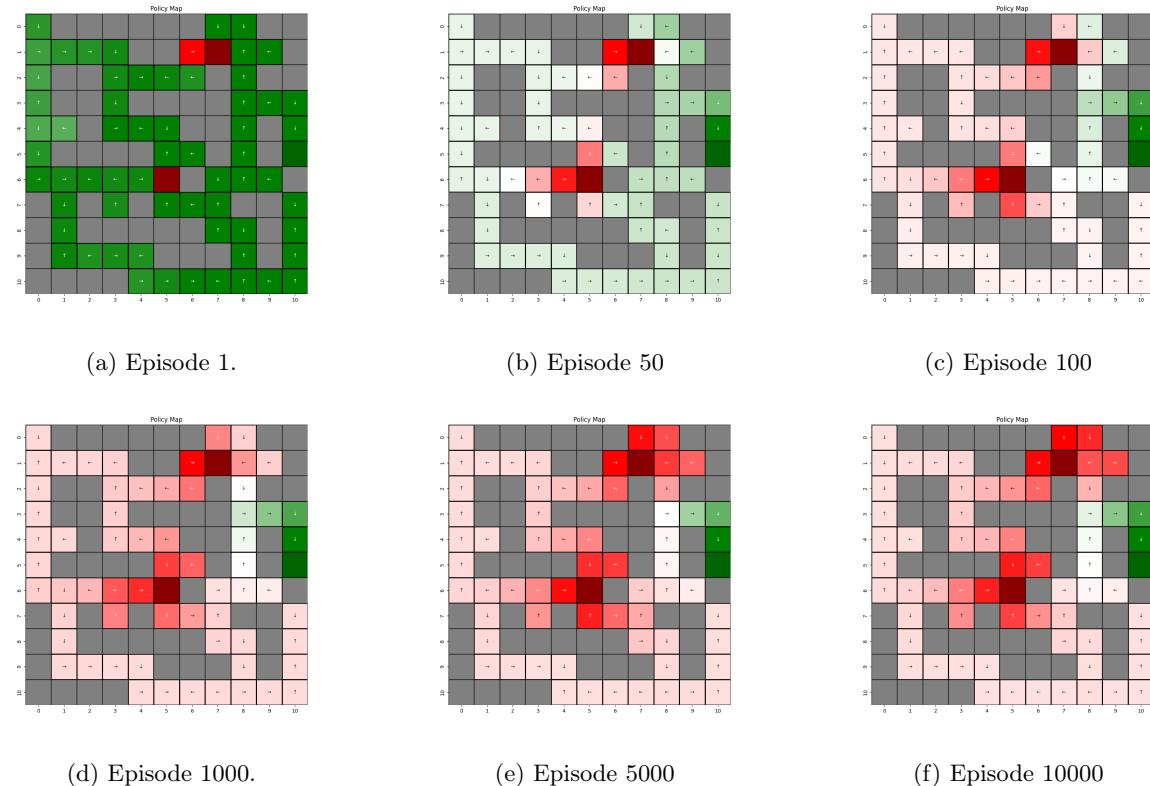
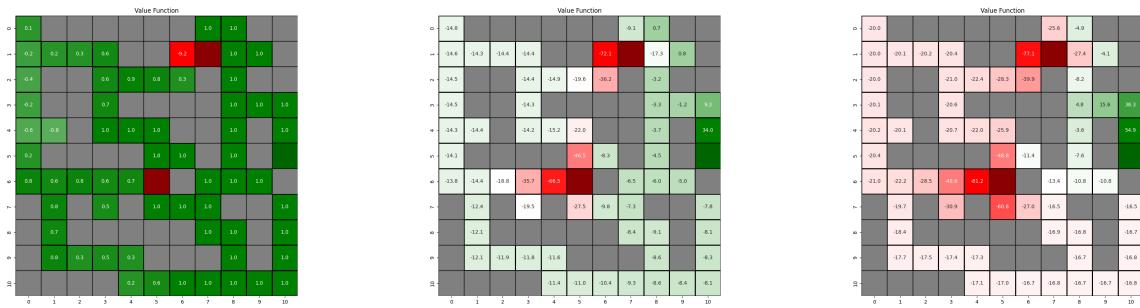


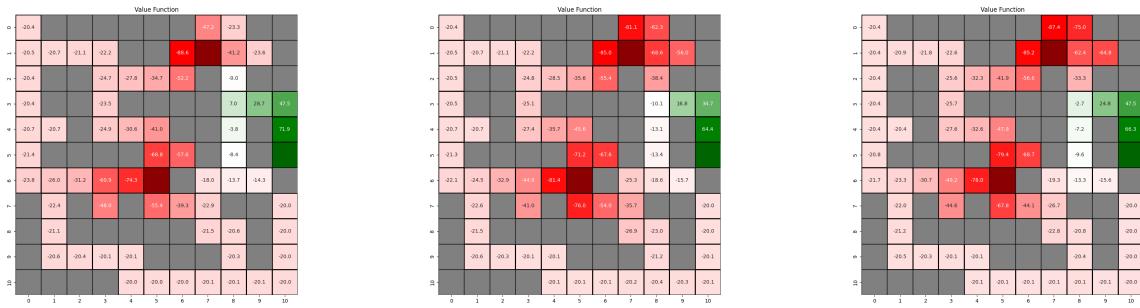
Figure 55: Evolution of policy maps throughout episodes.



(a) Episode 1.

(b) Episode 50

(c) Episode 100



(d) Episode 1000.

(e) Episode 5000

(f) Episode 10000

Figure 56: Evolution of value function throughout episodes.

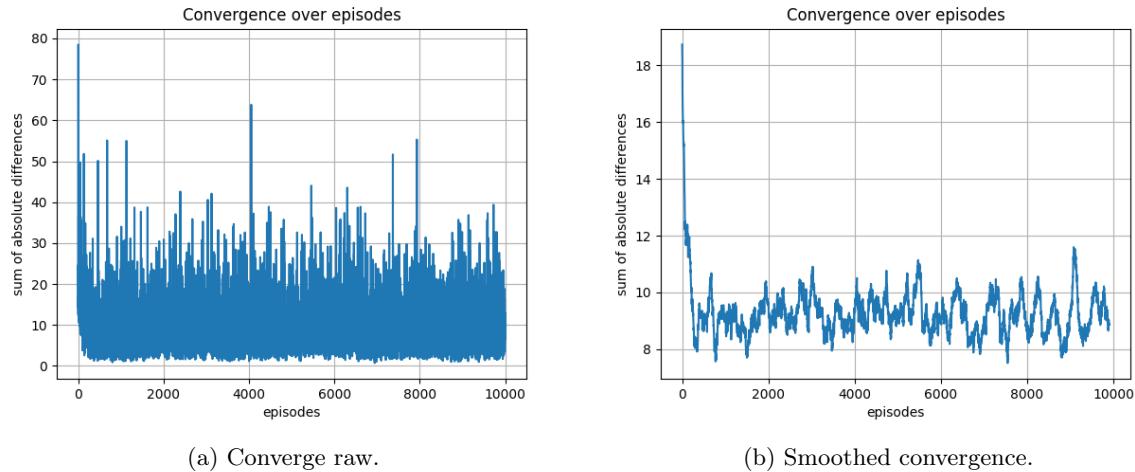


Figure 57: Convergence of value function.

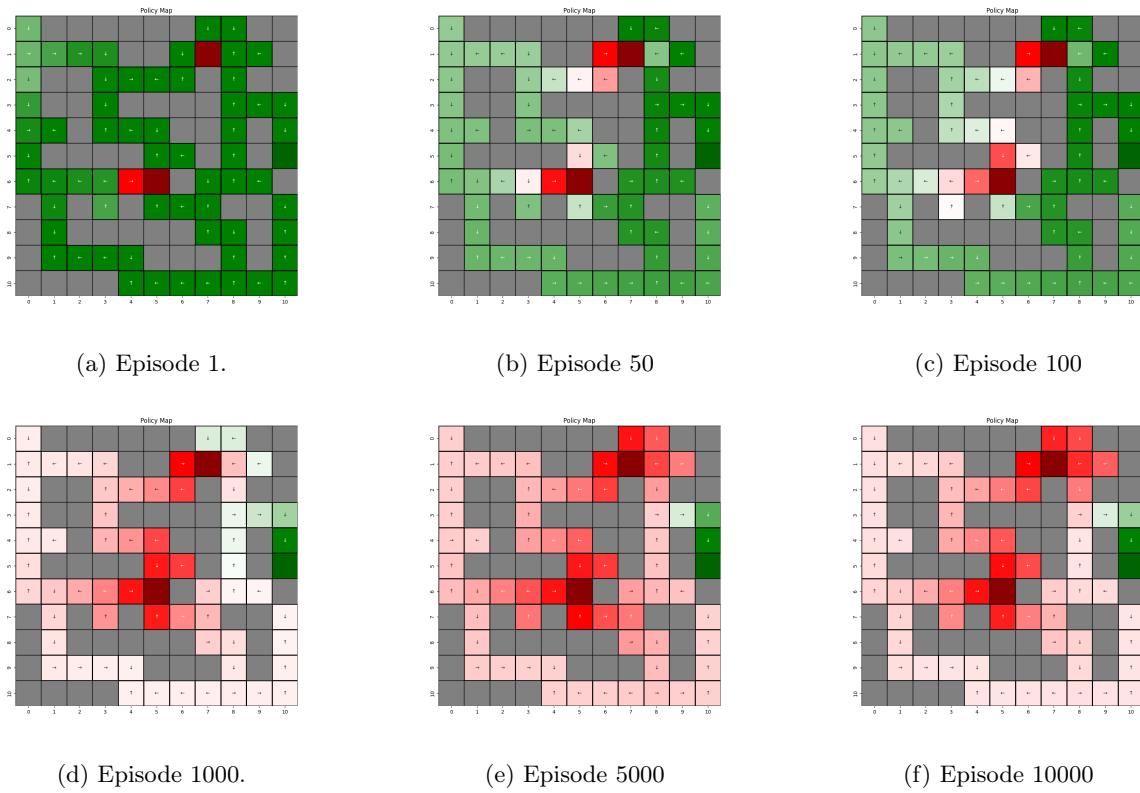


Figure 58: Evolution of policy maps throughout episodes.

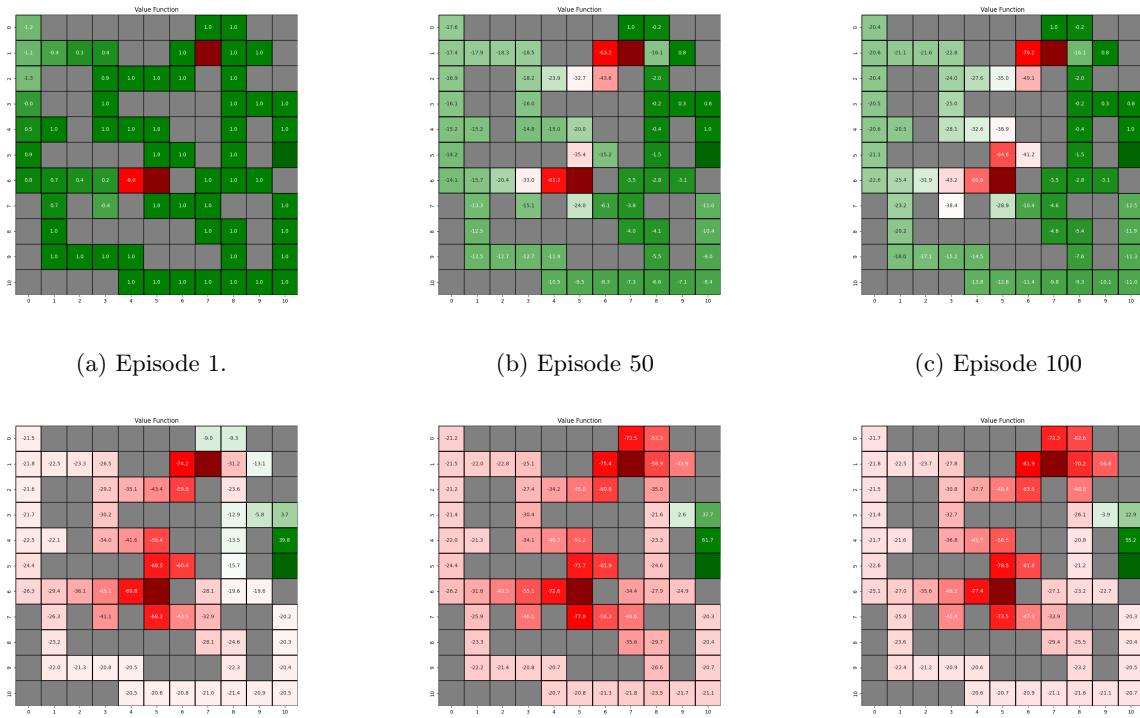


Figure 59: Evolution of value function throughout episodes.

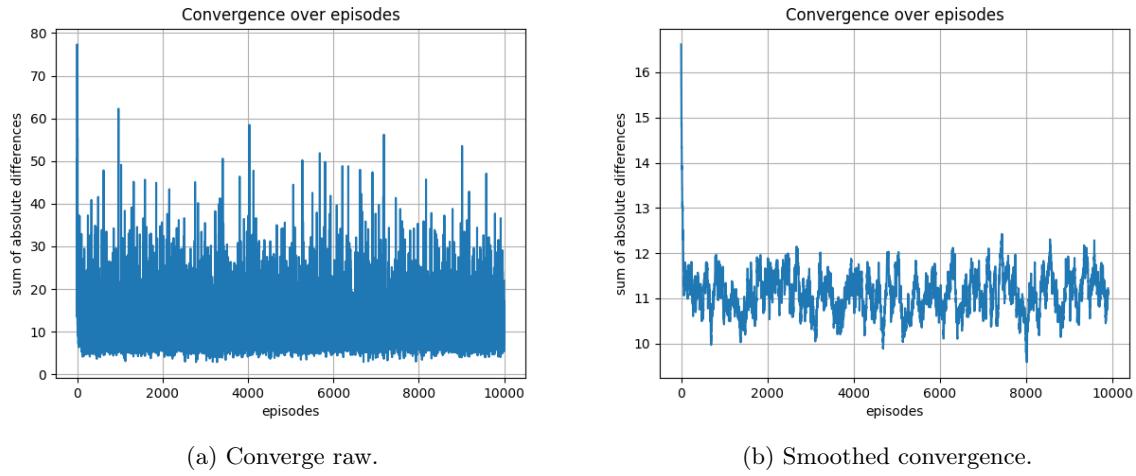


Figure 60: Converge of value function.

2.8 Effect of Epsilon in Q-Learning

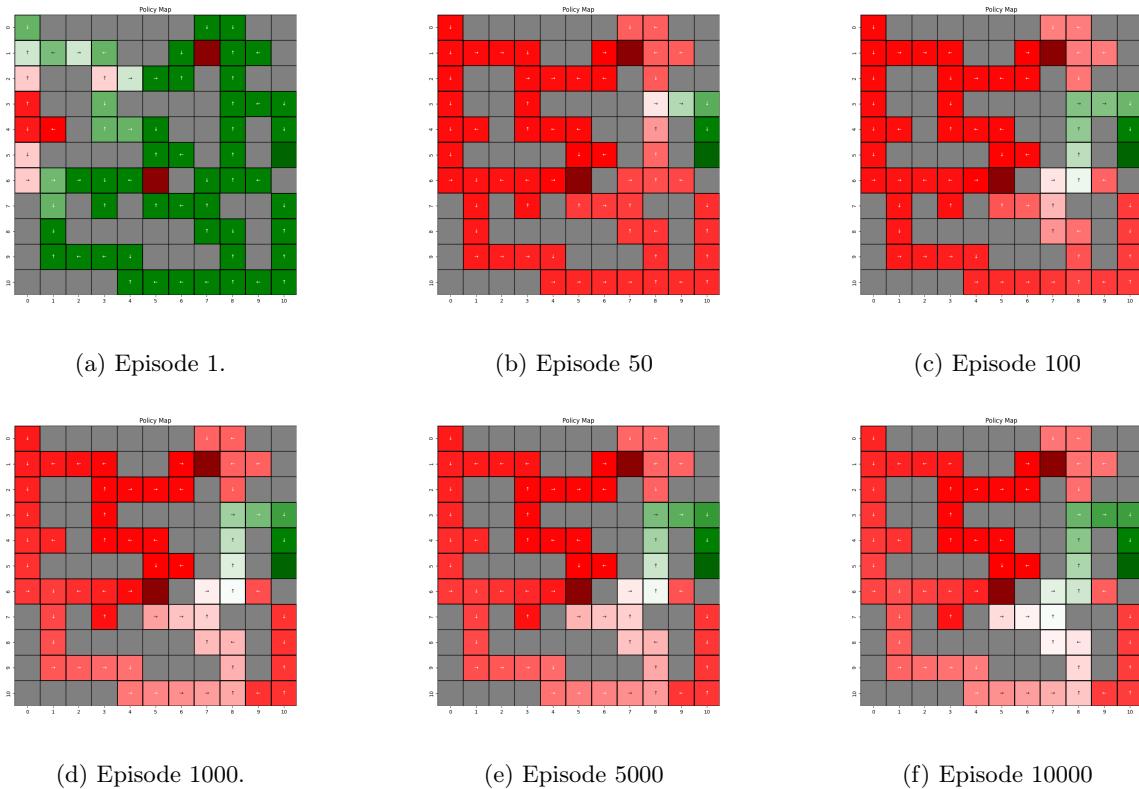


Figure 61: Evolution of policy maps throughout episodes.

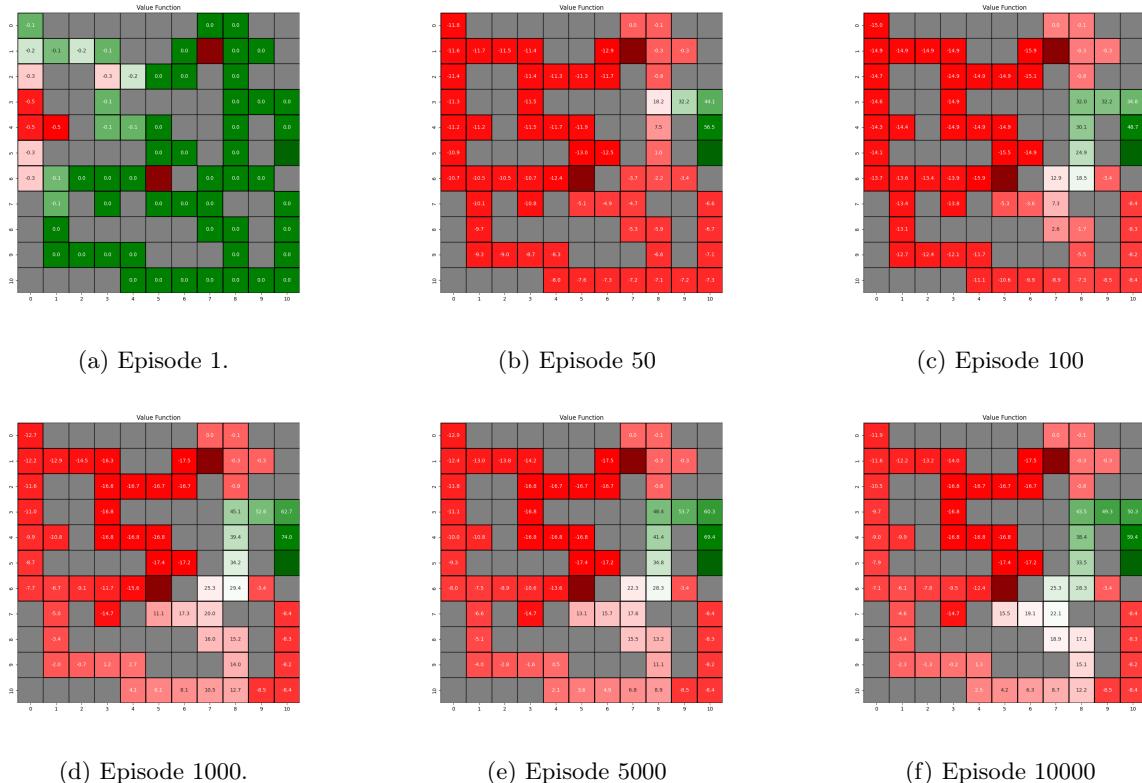


Figure 62: Evolution of value function throughout episodes.

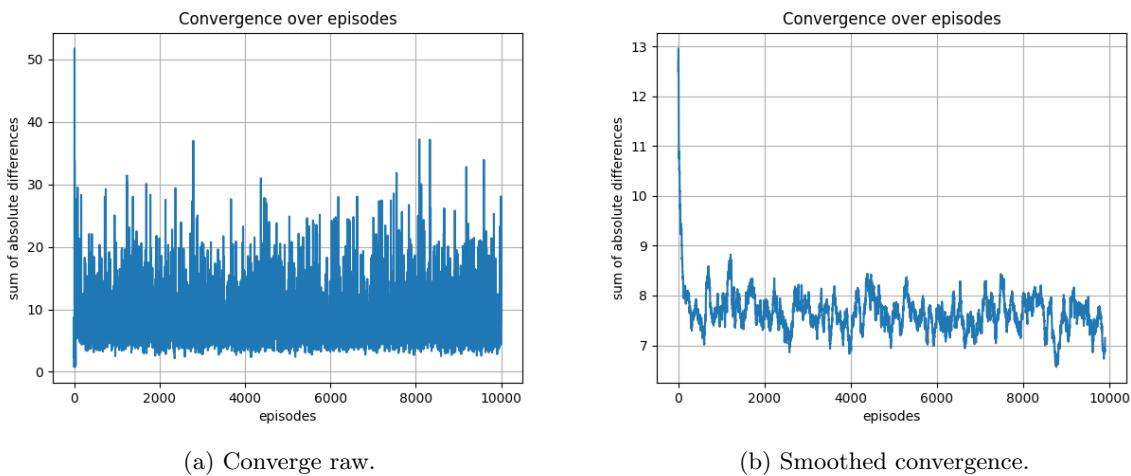


Figure 63: Convergence of value function.

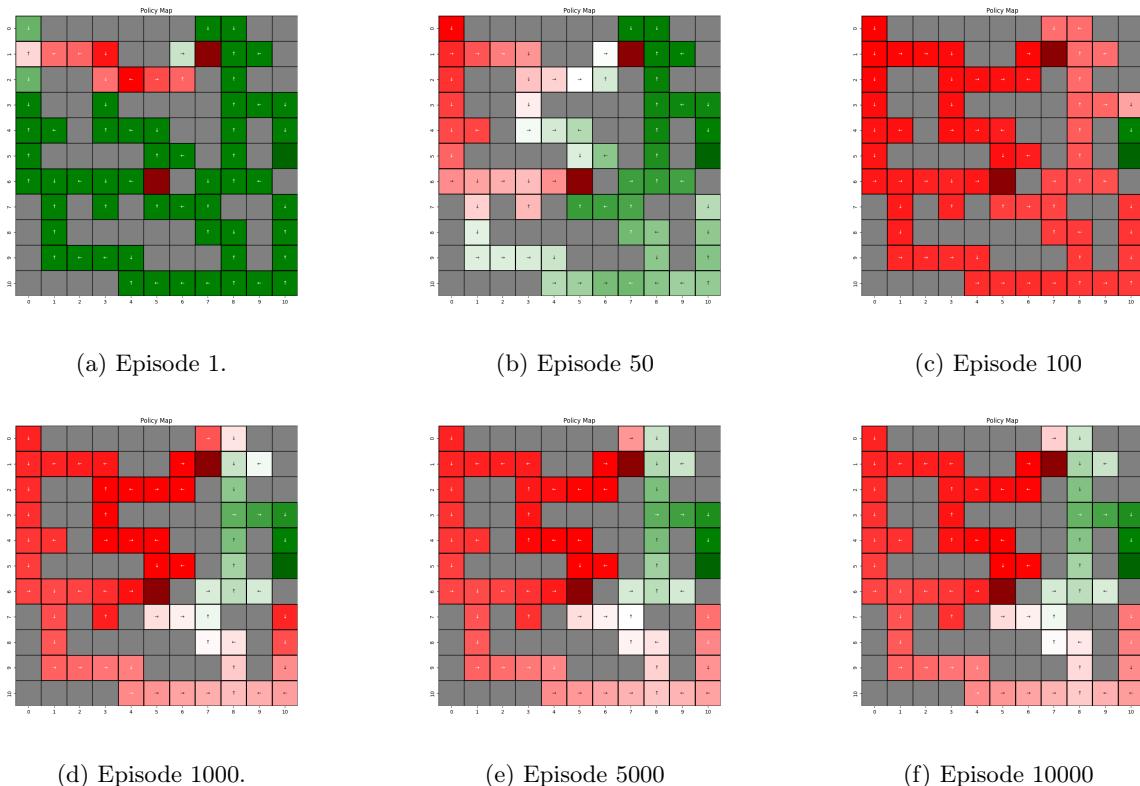


Figure 64: Evolution of policy maps throughout episodes.

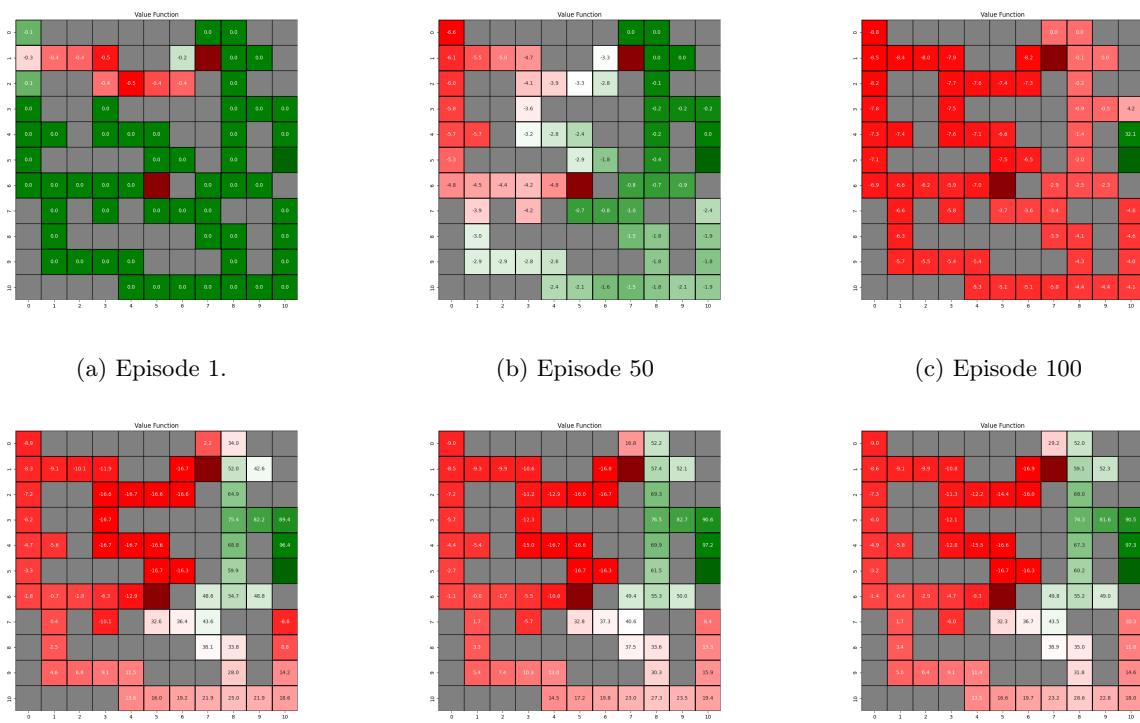


Figure 65: Evolution of value function throughout episodes.

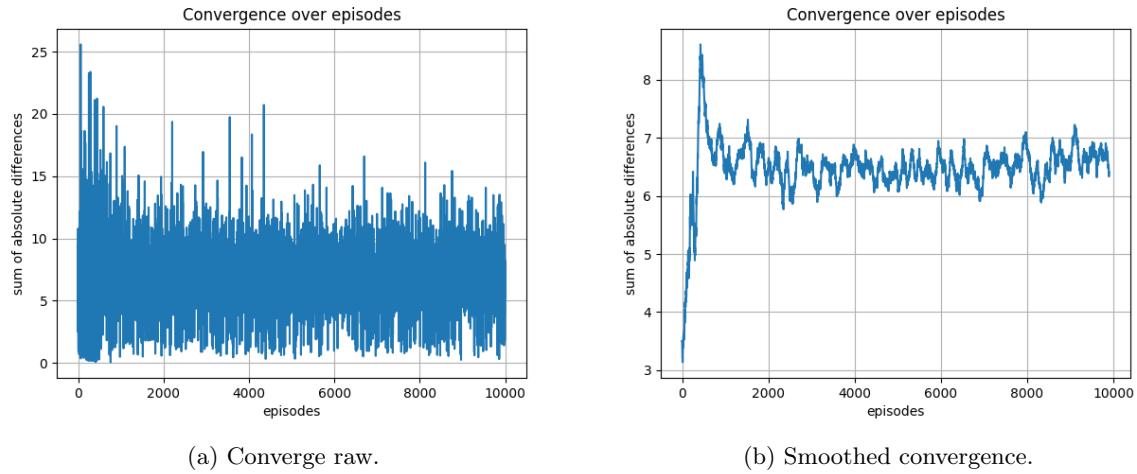


Figure 66: Converge of value function.

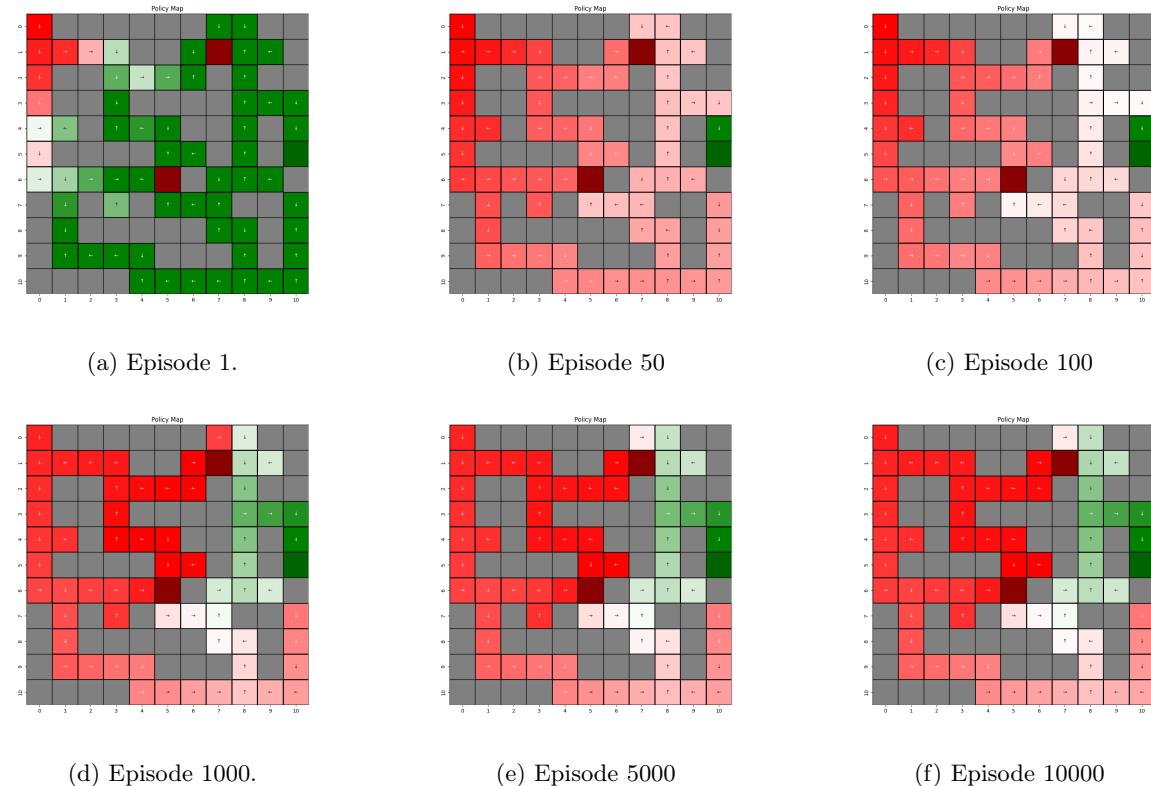


Figure 67: Evolution of policy maps throughout episodes.

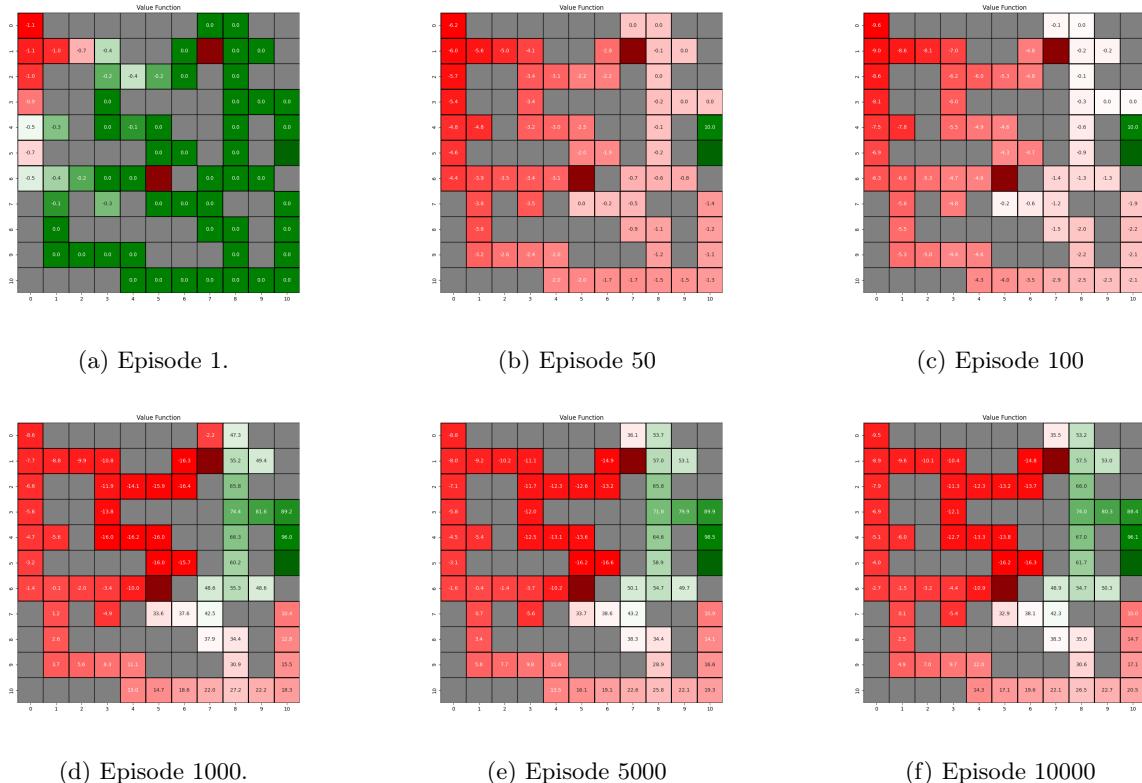


Figure 68: Evolution of value function throughout episodes.

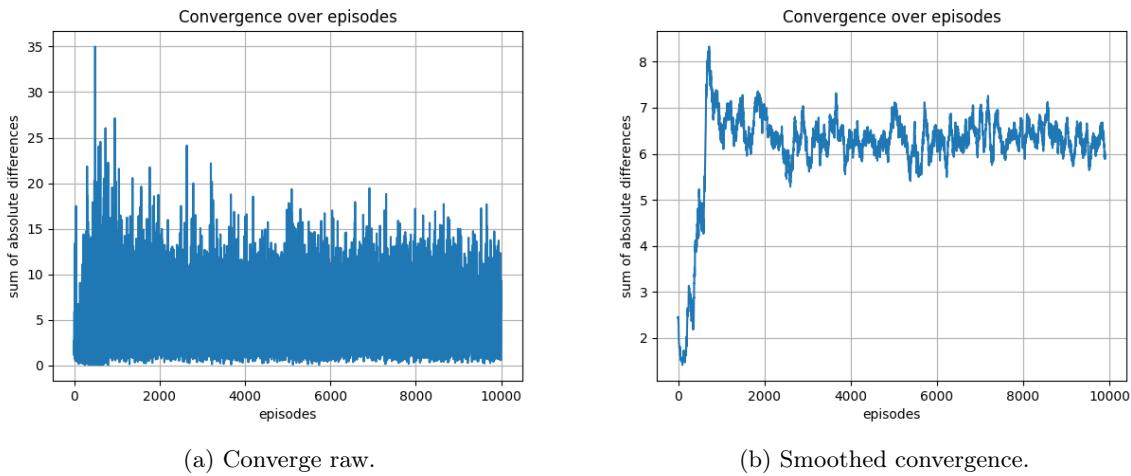


Figure 69: Convergence of value function.

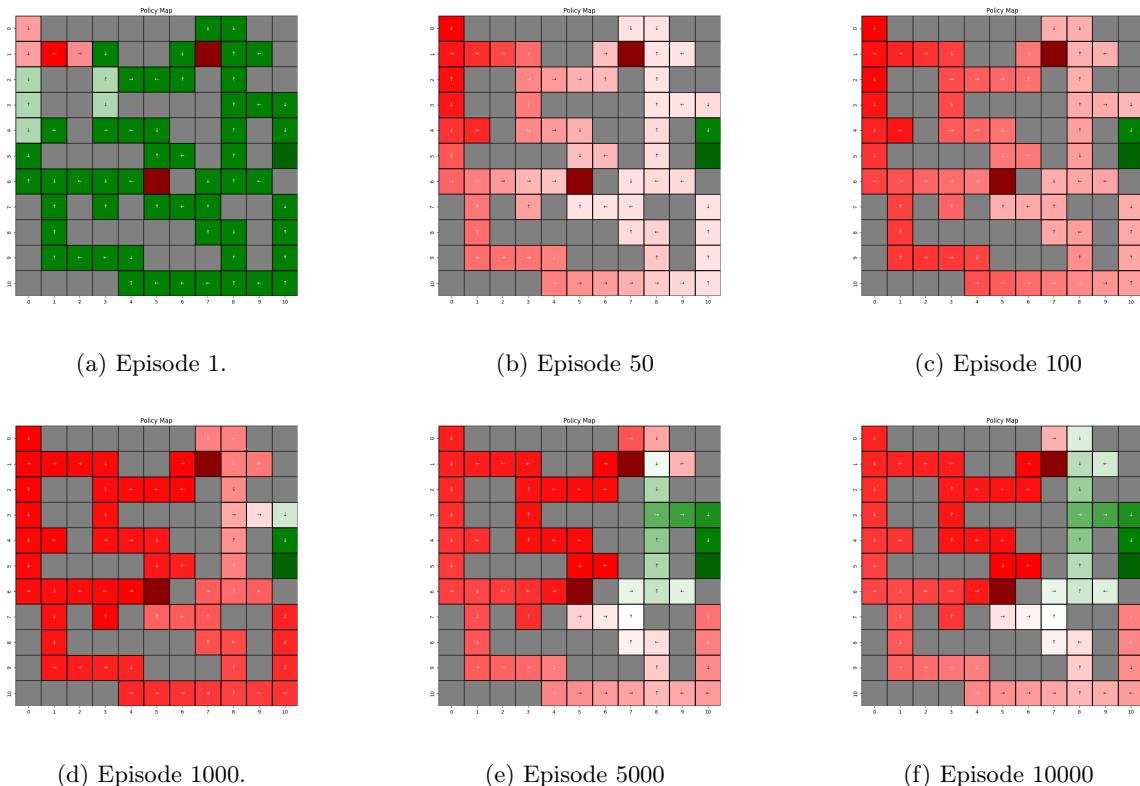


Figure 70: Evolution of policy maps throughout episodes.

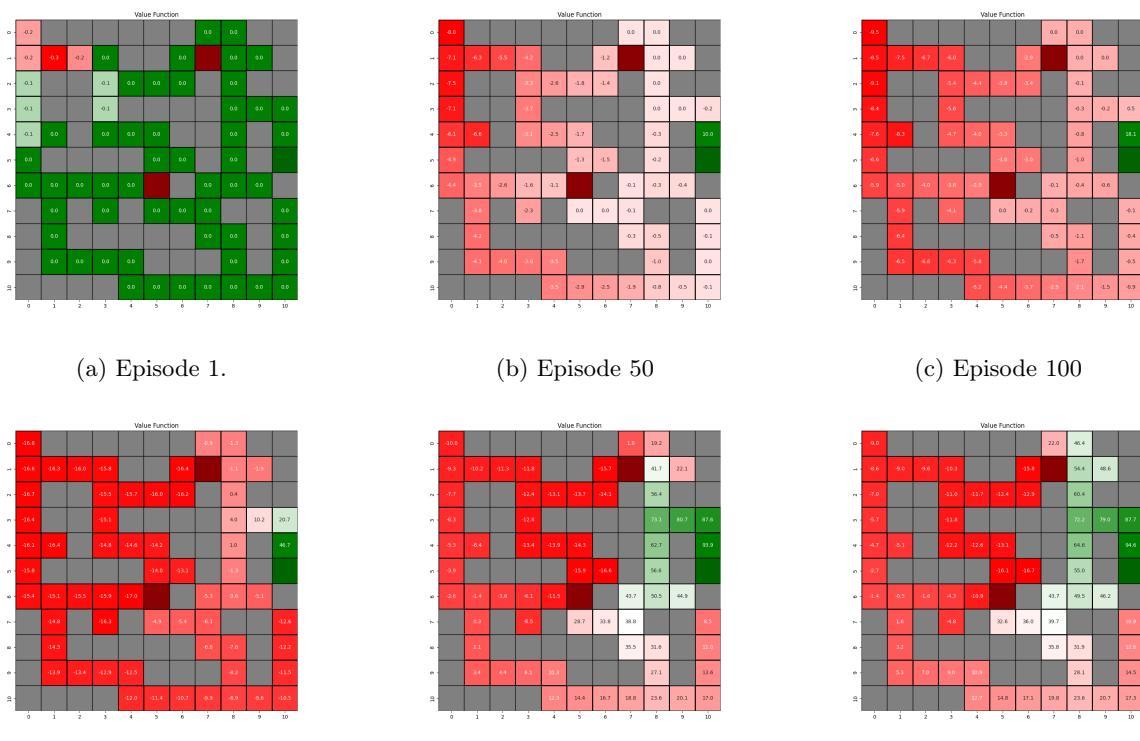


Figure 71: Evolution of value function throughout episodes.

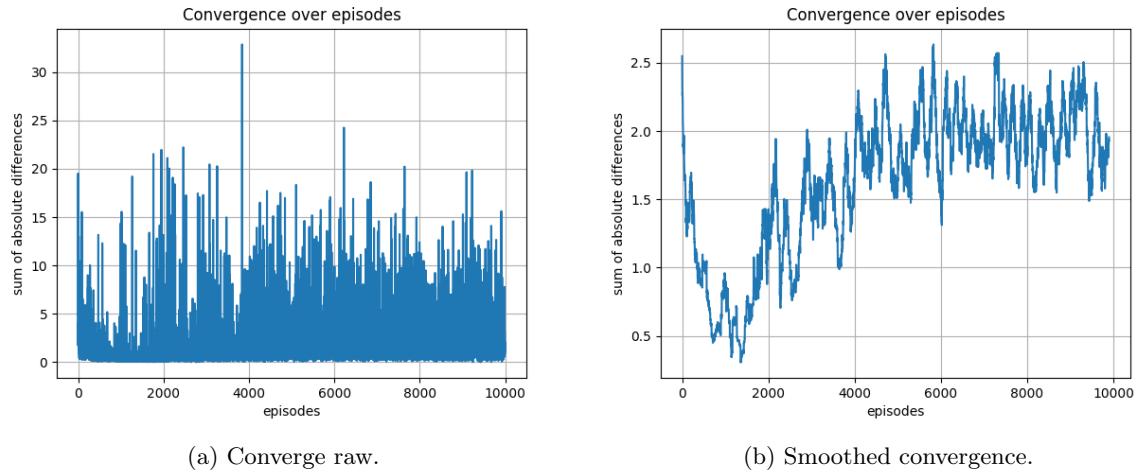


Figure 72: Converge of value function.

3 Discussions

3.1 Q1

3.2 Q2

3.3 Q3

3.4 Q4

3.5 Q5

3.6 Q6

3.7 Q7

3.8 Q8

3.9 Q9

3.10 Q10

Appendix

The code set used throughout this homework is provided as follows.

Submitted by Ahmet Akman 2442366 on May 25, 2024.