Laboratory 6

Variant 3, Group 13

By Saumya Shah and Anas Tagui

Task

We created a Q-learning algorithm using the epsilon-greedy policy for `cliffwalking-v0`. The number of episodes was fixed. Three values of epsilon (initial exploration rate), gamma (discount factor), epsilon decay (Decay rate for epsilon), alpha (learning rate), and max steps were varied to examine their effects on the reward and convergence metrics.

For this task, the highest possible reward is -13 (number of steps required to go from the start location to the end location).

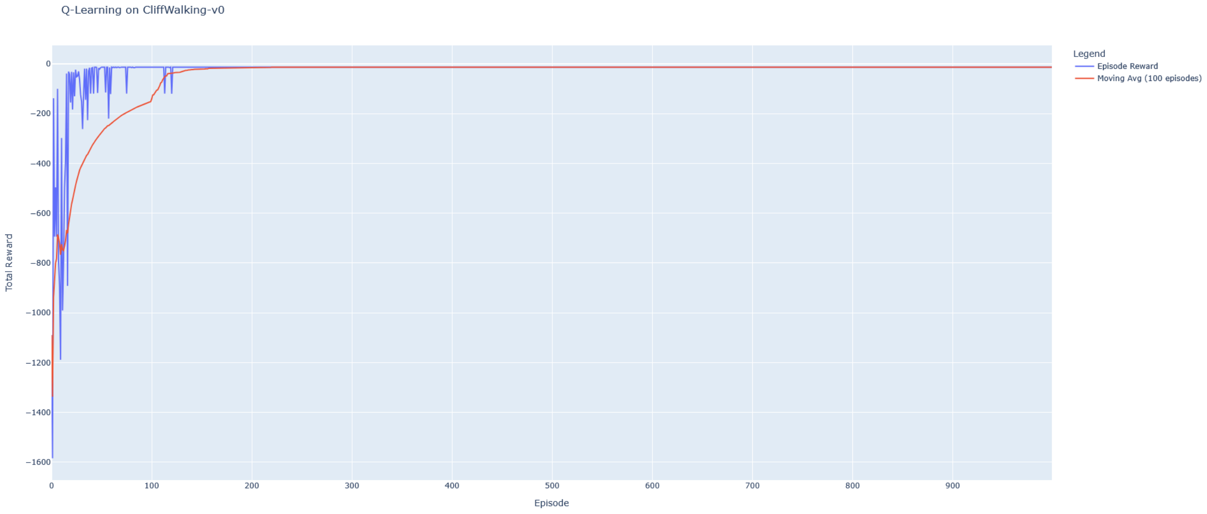
The baseline values used for the model are:
Alpha = 0.1
Gamma = 0.75
Epsilon = 1.0
Epsilon decay = 0.95
Max steps = 100

Observations

Alpha

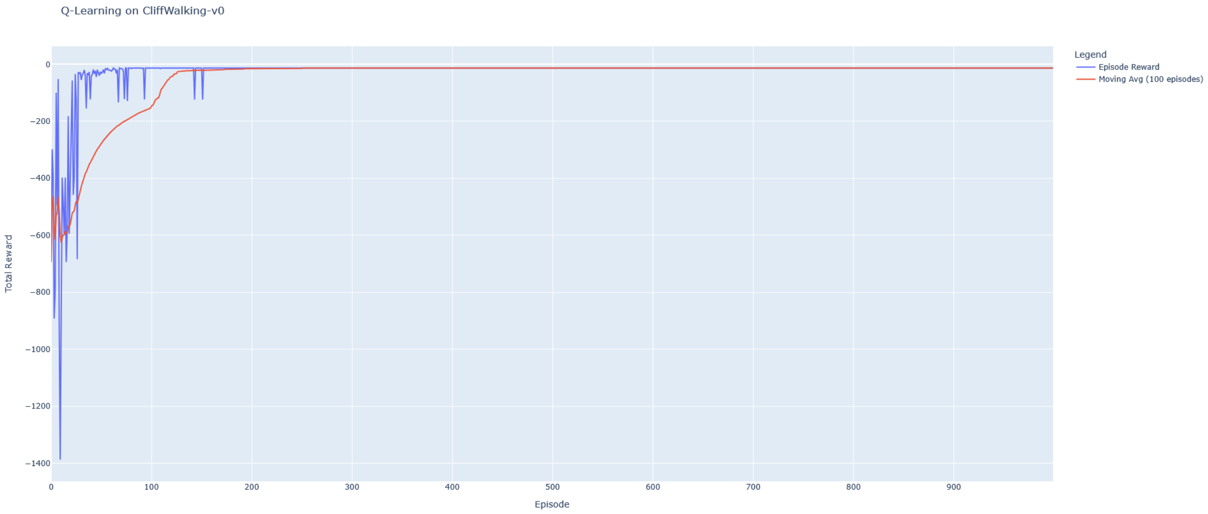
Three values of Alpha (1.0, 0.5, and 0.1) were tested. All other parameters were kept at their baseline values.

Alpha = 1.0



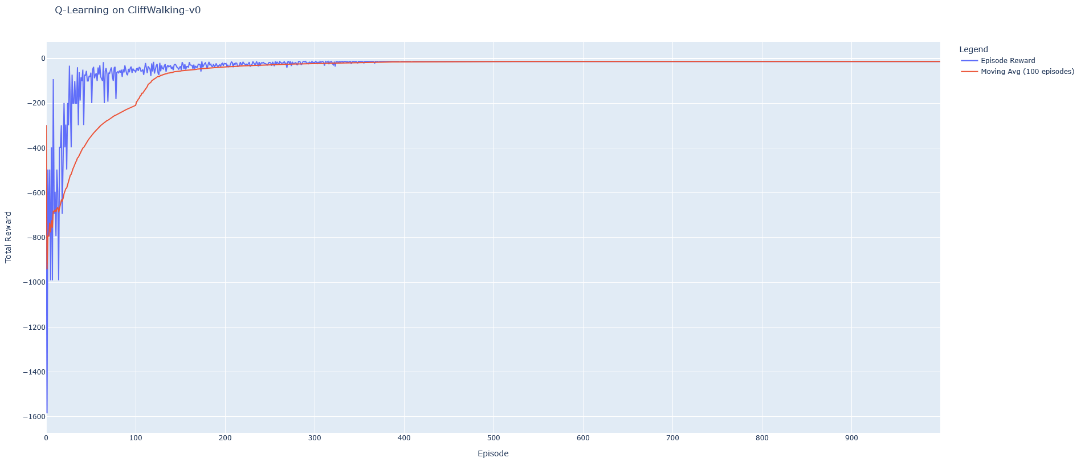
Fast updates caused high variance in learning, leading to unstable convergence.

Alpha = 0.5



Balanced learning rate that achieved the best convergence rate and stable learning.

Alpha = 0.1

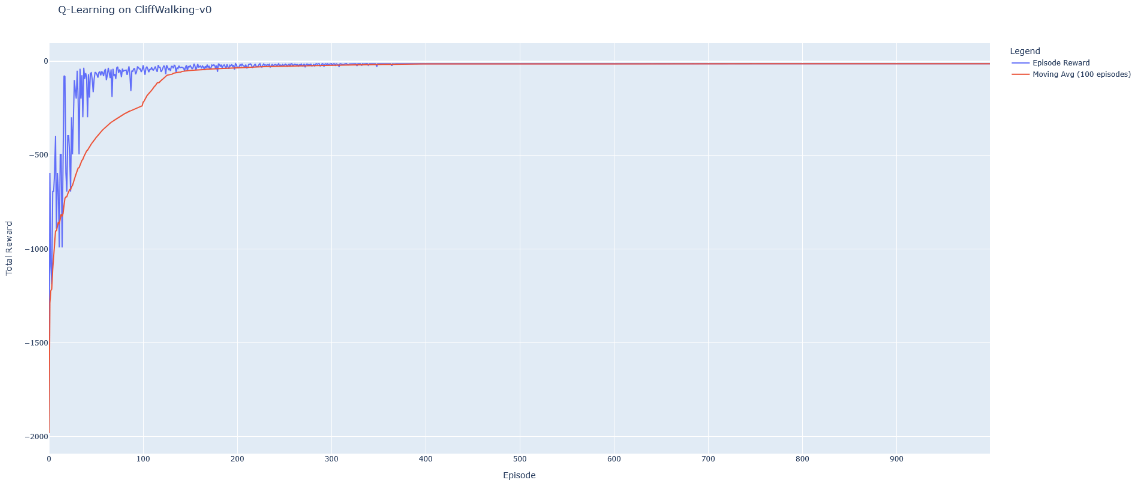


Updates were too small, causing very slow learning and convergence.

Gamma

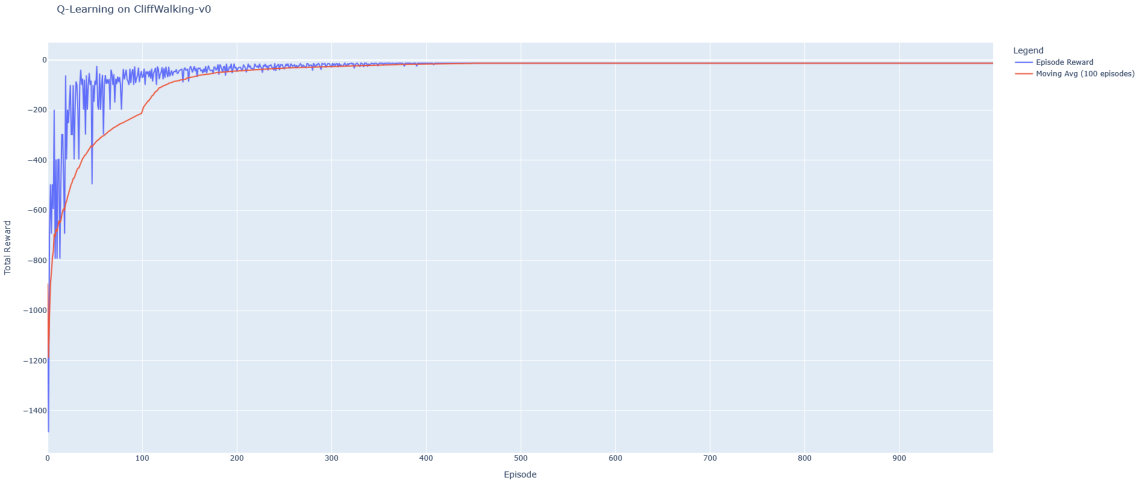
Three values of Gamma (1.0, 0.5, 0) were tested. All other parameters were kept at their baseline values.

Gamma = 1.0



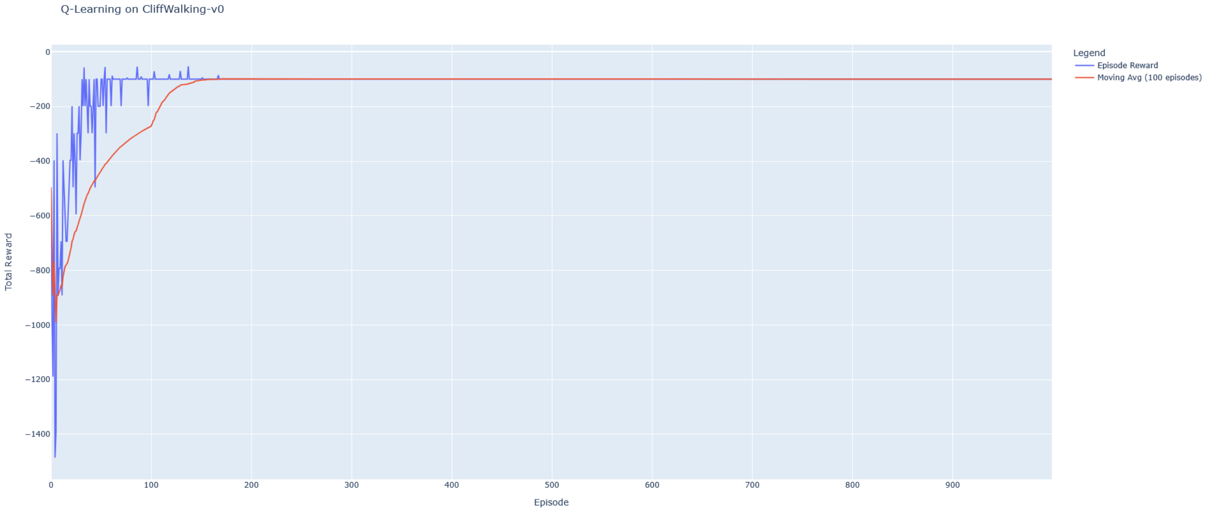
Agent prioritised long-term rewards, resulting in efficient paths but slight instability during training.

Gamma = 0.5



Agent favoured short-term rewards. Learning was stable, but the agent often chose suboptimal paths.

Gamma = 0

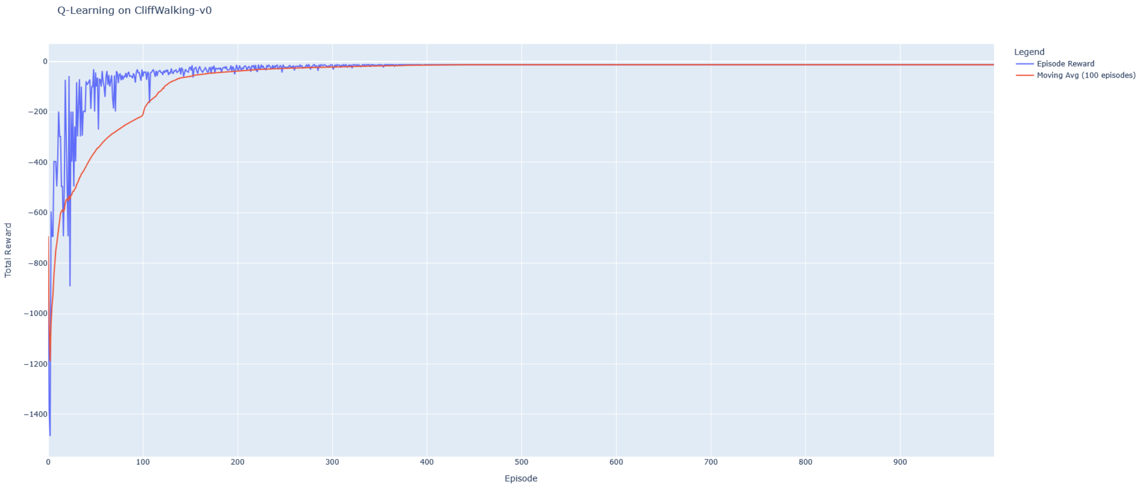


Agent completely ignored future rewards, leading to random and inefficient behaviour.

Epsilon

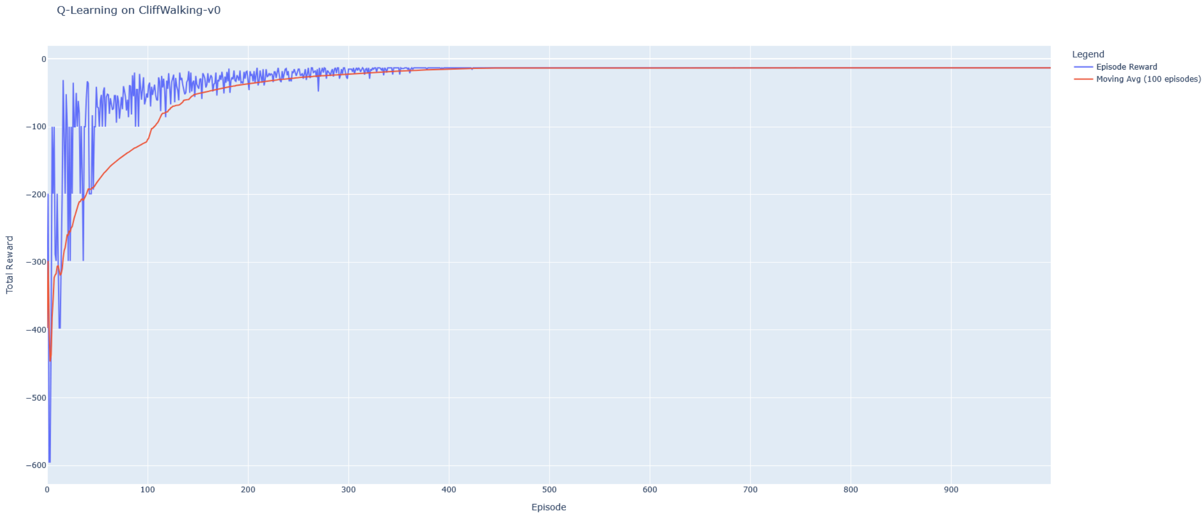
Three values of Epsilon (1.0, 0.5, 0) were tested. All other parameters were kept at their baseline values.

Epsilon = 1.0



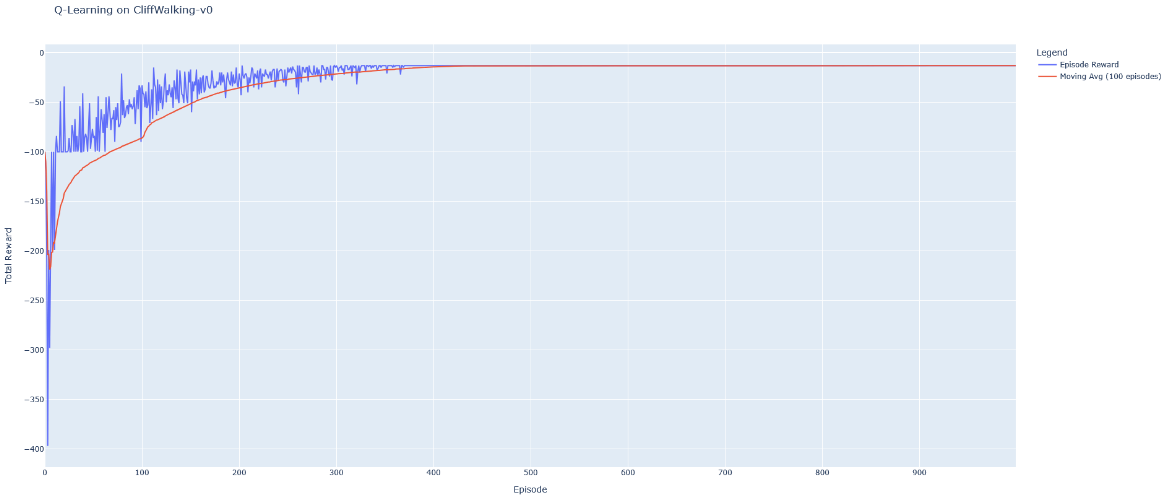
High exploration led to poor early performance but helped avoid local optima in the long run.

Epsilon = 0.5



Balanced exploration and exploitation, leading to faster convergence and better cumulative reward.

Epsilon = 0

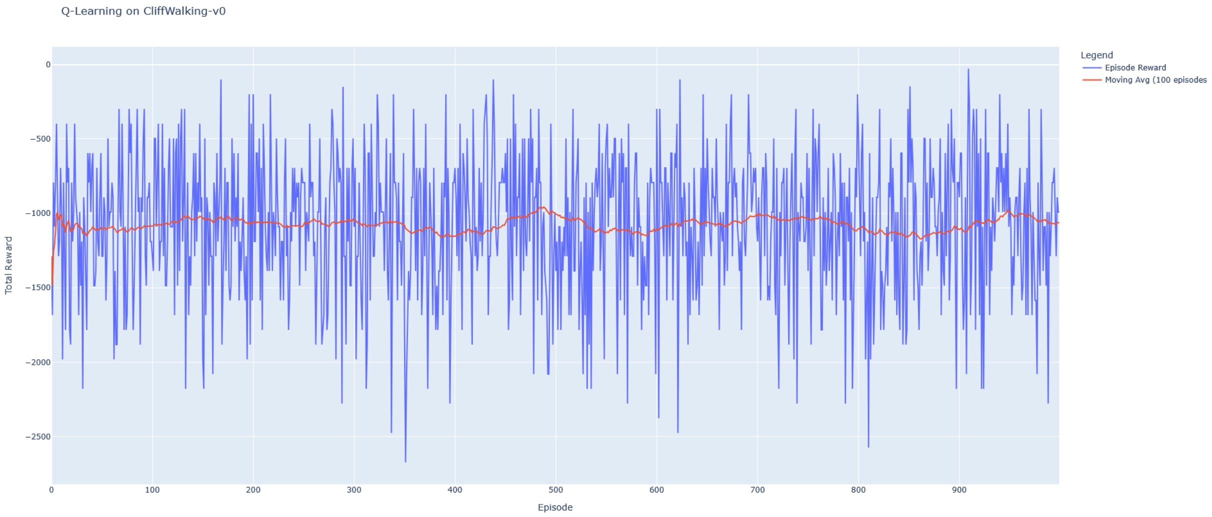


No exploration. Led to slightly slower convergence.

Epsilon Decay

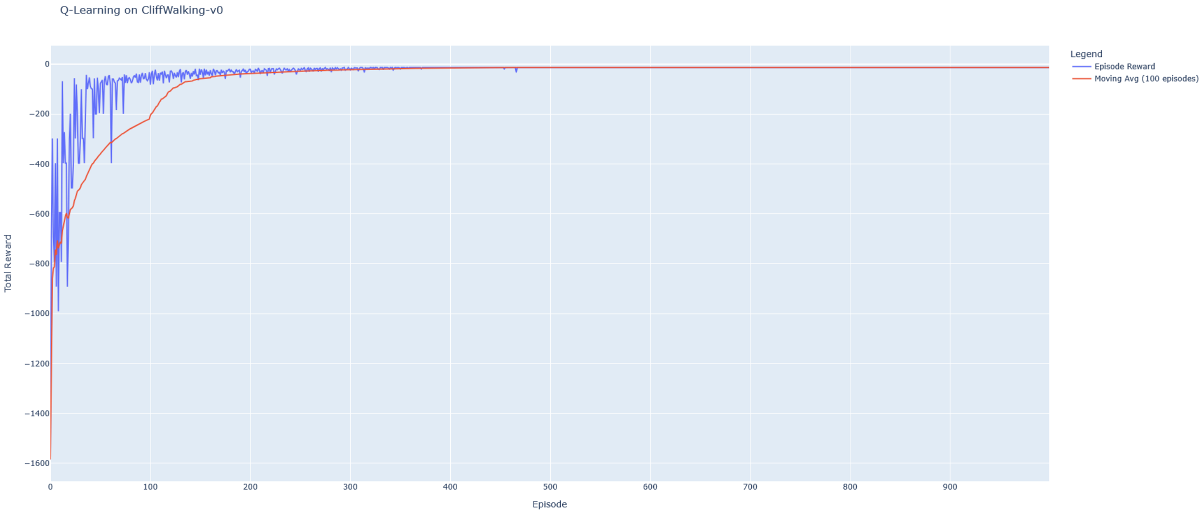
Three values of Epsilon Decay (1.0, 0.95, 0.5) were tested. All other parameters were kept at their baseline values.

Epsilon decay = 1.0



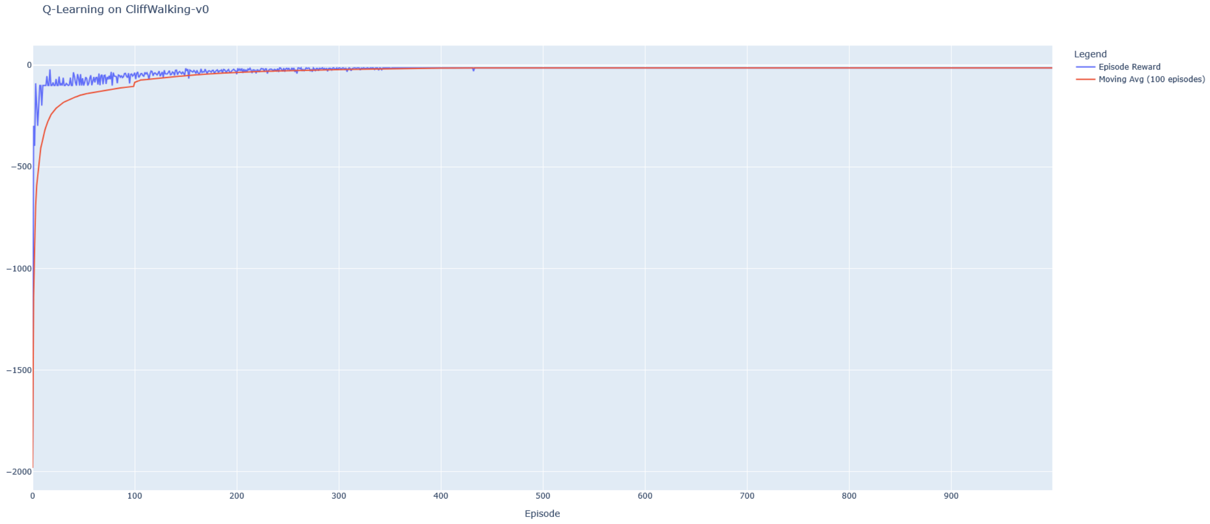
Constant high exploration throughout training. No convergence achieved.

Epsilon decay = 0.95



Smooth transition from exploration to exploitation. Achieved the best performance.

Epsilon decay = 0.5

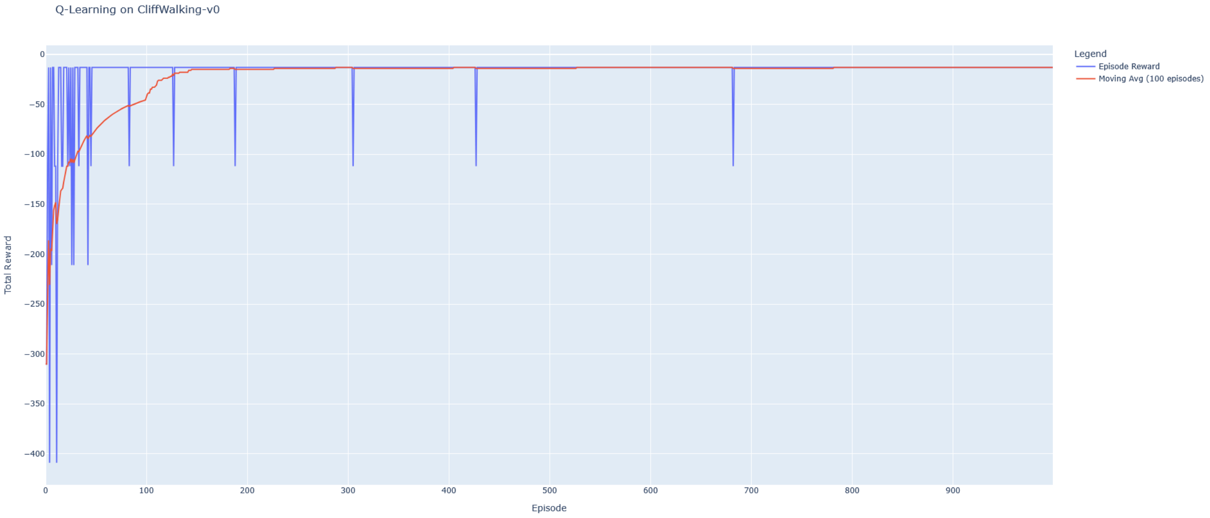


Exploration dropped too quickly, causing premature convergence to poor policies.

Max Steps

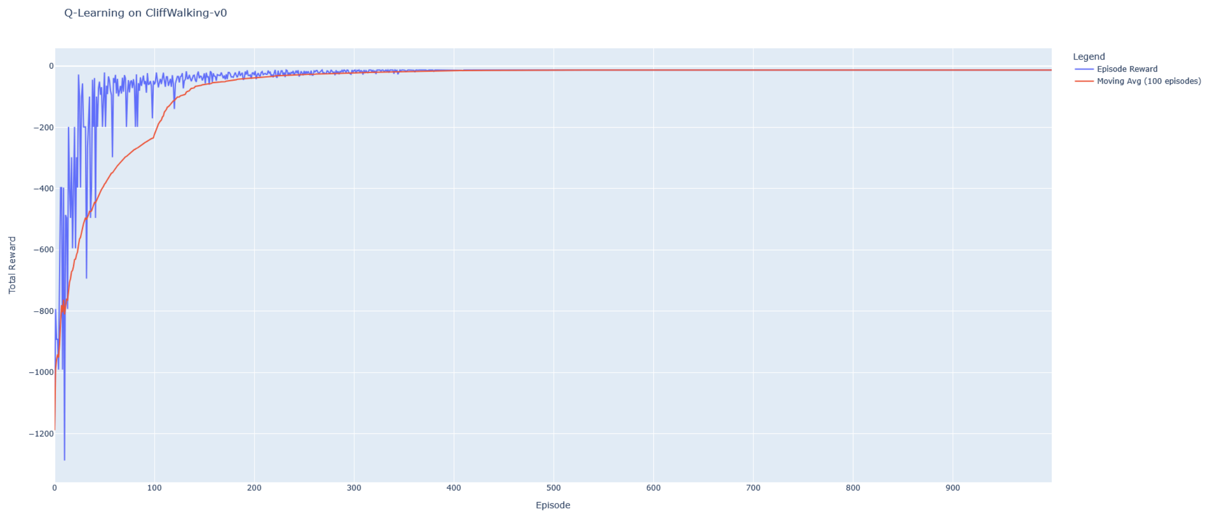
Three values of max steps (13, 100, 1000) were tested. All other parameters were kept at their baseline values.

Steps = 13



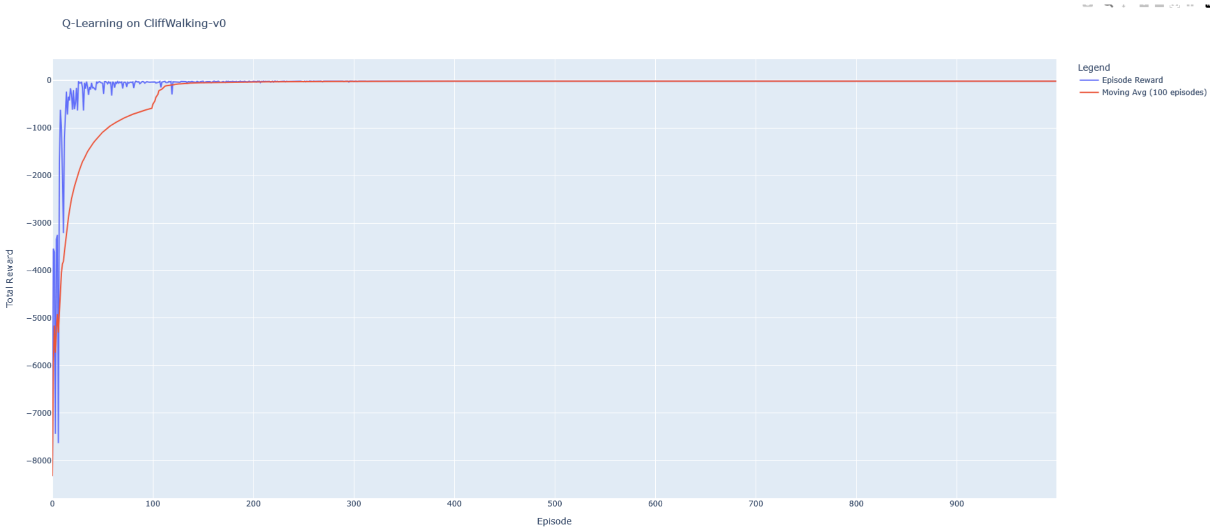
Too restrictive. Many episodes terminated early, limiting learning.

Steps = 100



Adequate steps for reaching the goal and optimising the path.

Steps = 1000



No significant improvement over 100 steps. Led to longer episode times without added value.

Conclusion

Alpha:

Alpha = 0.5 seemed to converge the fastest, followed by Alpha = 1.0 and Alpha = 0.1. A low value of alpha means that the agent makes very small updates to its Q-values at each step. The ‘new information’ only slightly adjusts the existing estimate, which leads to slow but stable convergence. A higher value of alpha means that the agent relies more on ‘new information’, which allows the agent to react very quickly to recent rewards and experiences.

However, this can also lead to instability in convergence and a moderate value like 0.5 seems to provide the best of both worlds.

Gamma:

Gamma controls how far into the future the agent considers the reward.

When gamma is high, the agent aims for long-term benefits, but this can make the training more sensitive to the reward structure. When it is too low, the agent becomes shortsighted and fails to develop meaningful strategies.

The results suggest that long-term planning is essential in cliff-like environments, where short-term gains often lead to punishment.

Epsilon:

The exploration rate plays a key role in discovering the optimal policy.

Too much exploration results in the agent wandering without improving, while too little prevents it from finding better paths.

The results show that a moderate amount of exploration early in training, gradually replaced by exploitation, leads to the most effective learning process.

When Epsilon = 0, the policy corresponds to a simple greedy policy, which led to suboptimal convergence due to a lack of initial exploration.

Epsilon Decay:

Gradually reducing exploration ensures that the agent can both learn about the environment and later consolidate its learning into an optimal strategy.

A well-chosen decay rate helps the agent balance these phases smoothly.

Too slow decay prevents convergence; too fast decay cuts learning short.

Max Steps:

Allowing enough steps per episode ensures the agent has time to reach the goal and learn from the outcome.

However, more steps do not always equate to better learning and can cause unnecessary delays.

There is a point beyond which increasing the episode length does not lead to better results.