Why Don't Combine Policy Iteration and Value Iteration

Kayzwer

Abstract

In the book Reinforcement Learning: An Introduction, Value Iteration, and Policy Iteration are used to calculate the optimal policy by using Dynamic Programming. A new algorithm can be created to combine both Value Iteration and Policy Iteration. The comparisons of Value Iteration, Policy Iteration, and the new algorithm are shown in the paper to see what situation that new algorithm can outperform Value Iteration and Policy Iteration.

1 SHIN Value Iteration

SHIN value iteration is a variant of value iteration that will run n sweeps each iteration instead of 1. Below shows the algorithm of SHIN value iteration.

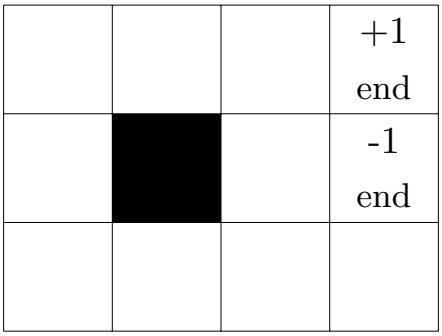
Algorithm 1 SHIN Value Iteration, for estimating $\pi \approx \pi_*$

Parameters: $n > 0$, $\gamma \in (0, 1]$, $\theta > 0$
Initialize $V(s)$, for all $s \in S^+$, $V(terminal) = 0$, $\pi(s)$, for all s, except $s \in terminal$
while *true* **do**
  for 1 to n **do**
    $\Delta \leftarrow 0$
    for all s $\in S^+$ **do**
      $v_{old} \leftarrow V(s)$
      $V(s) \leftarrow \mathbb{E}_\pi[r + \gamma V(s')|s]$
      $\Delta \leftarrow \max(\Delta, |v_{old} - V(s)|)$
    end for
    if $\Delta < \theta$ **then**
      break
    end if
  end for
  for all s $\in S^+$ **do**
    $\pi_{new}(s) \leftarrow \arg\max_a \mathbb{E}_\pi[r + \gamma V(s')|s, a]$
  end for
  if $\pi_{new} = \pi$ **then**
    break
  else
    $\pi \leftarrow \pi_{new}$
  end if
end while

2 Comparisons

Three of the algorithms will be applied to different types of grid world. The total step each algorithm takes to make policy converge will be the metric. One sweep of Policy Evaluation is counted as 1 step and Policy Improvement is also counted as 1 step. The initial policy π for all states is uniformly distributed. State with "end" represent the terminal state.

2.1 Grid World I



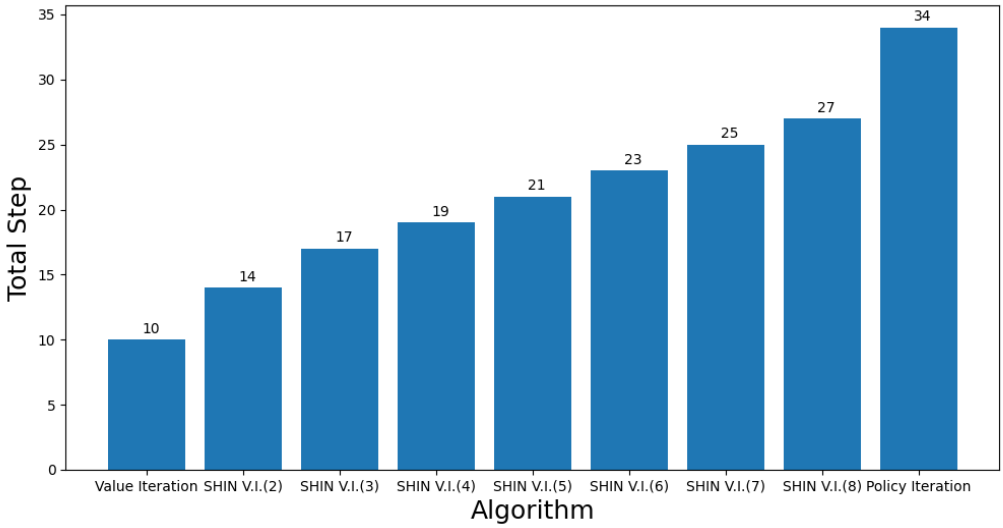
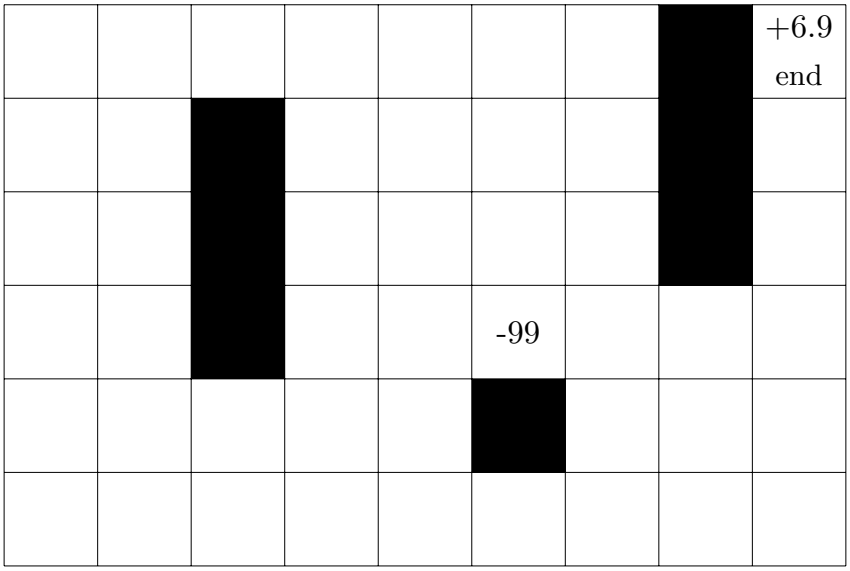


Figure 1: Parameters: $\gamma = 0.99, \theta = 0.01$

2.2 Grid World II



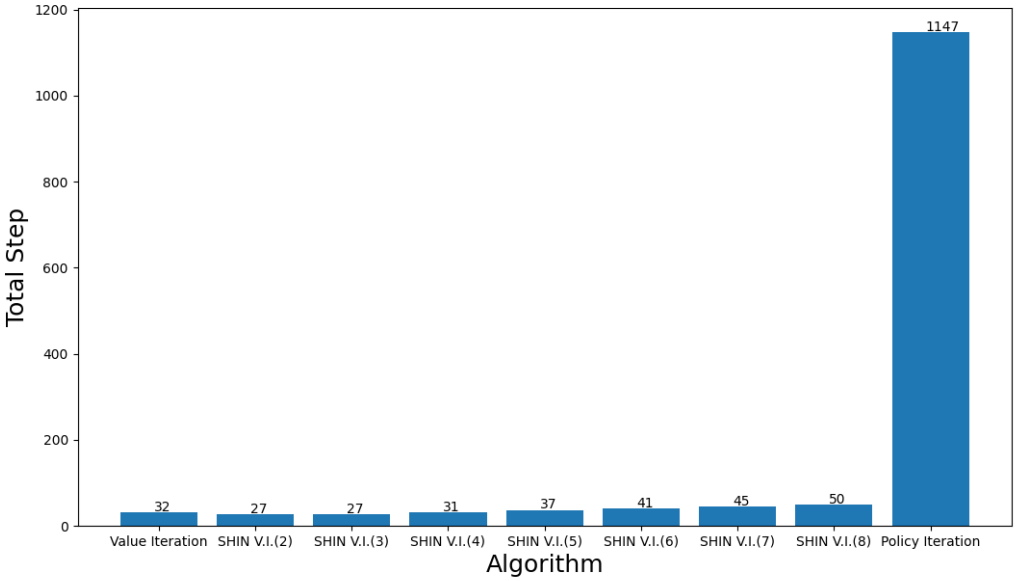


Figure 2: Parameters: $\gamma = 0.99, \theta = 0.01$

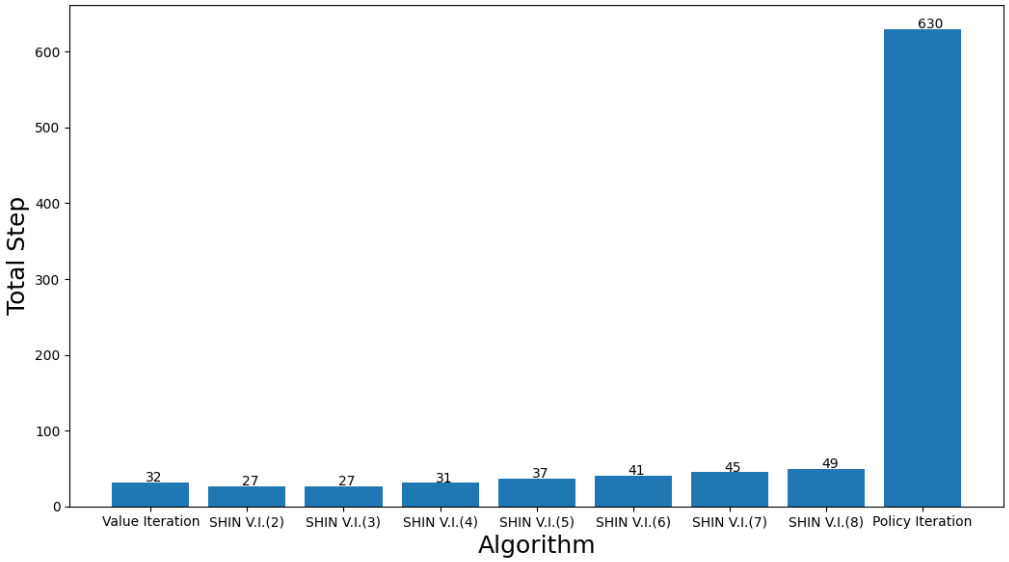
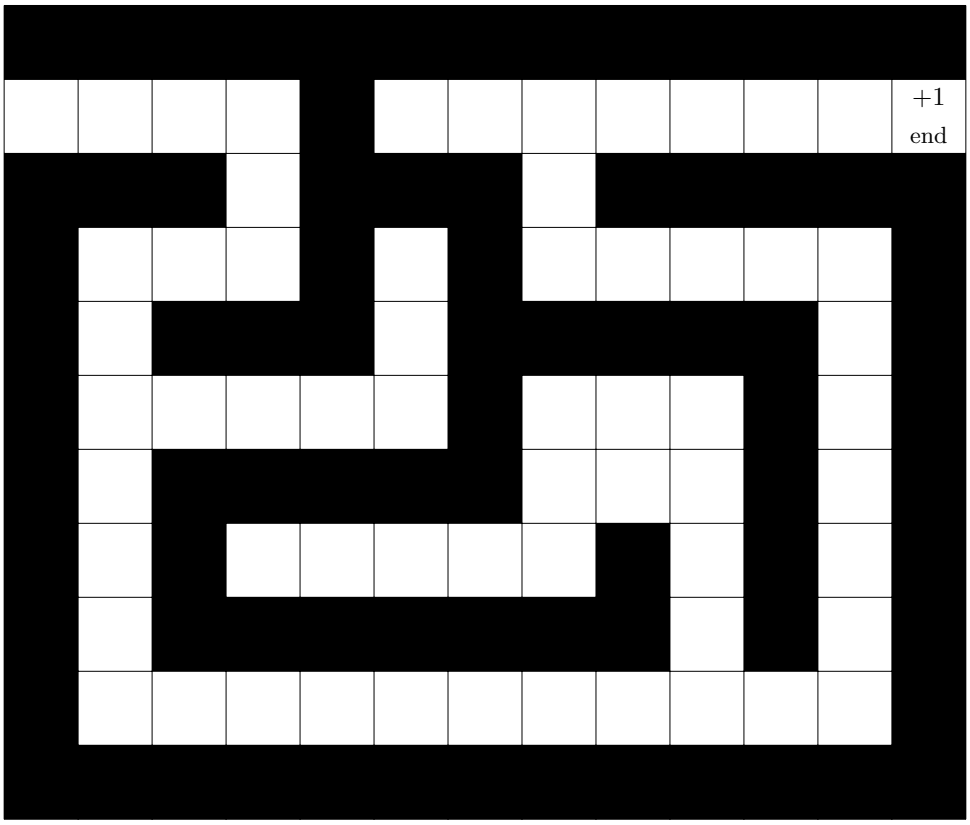


Figure 3: Parameters: $\gamma = 0.99, \theta = 0.1$

2.3 Grid World III



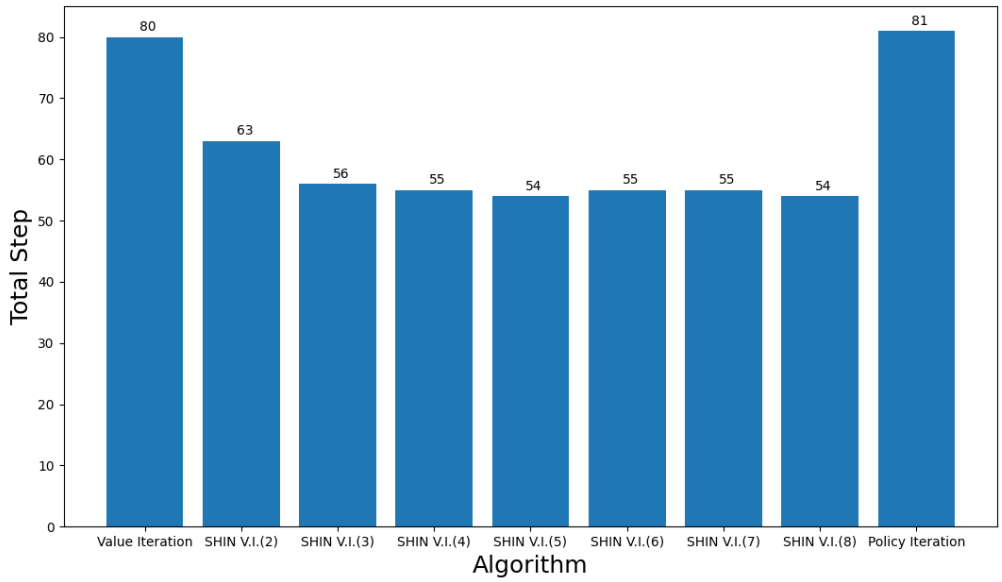


Figure 4: Parameters: $\gamma = 0.99, \theta = 0.01$

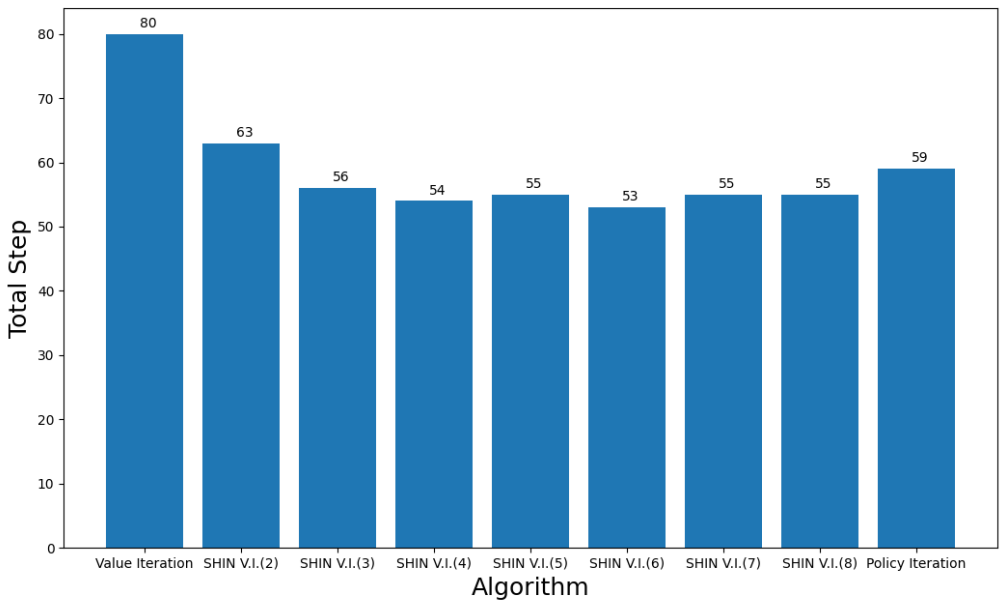


Figure 5: Parameters: $\gamma = 0.99, \theta = 0.1$

3 Conclusion

From Figure 2 and Figure 4, SHIN Value Iteration outperformed Value Iteration and Policy Iteration in Maze Grid World and Grid World that is more open. However, a proper n have to choose in order to make the convergence faster.

Reference

- Andrew Barto and Richard S.(2018) SuttonReinforcement Learning: An Introduction. Chapter 4 Dynamic Programming
- Grid World I `https://towardsdatascience.com/reinforcement-learning-implement-grid-world-from-scratch-c5963765ebff`
- Grid World II `https://arxiv.org/pdf/1405.5459.pdf`
- Grid World III `https://github.com/kenndanielso/mlrefined`