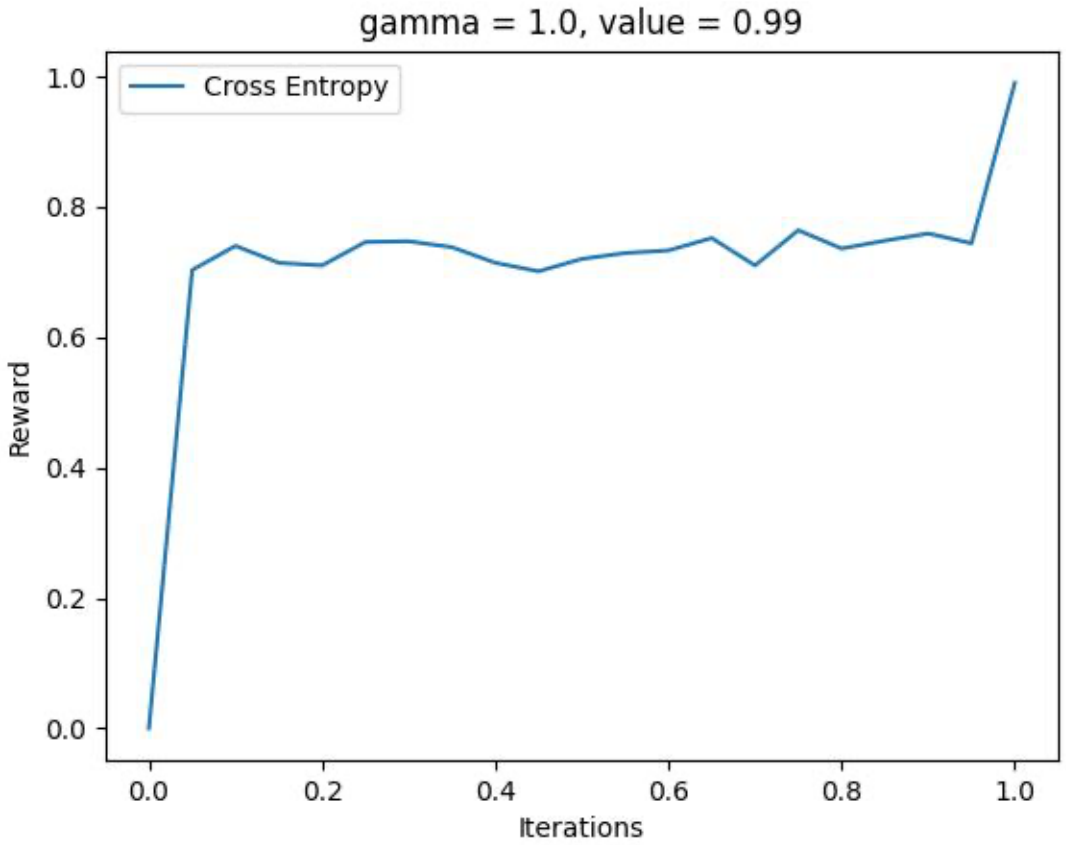
Отчет по домашнему заданию.

1. В алгоритме Policy Iteration важным гиперпараметром является gamma. Требуется ответить на вопрос, какой gamma лучше выбирать. Качество обученной политики можно оценивать например запуская среду 1000 раз и взяв после этого средний total_reward.

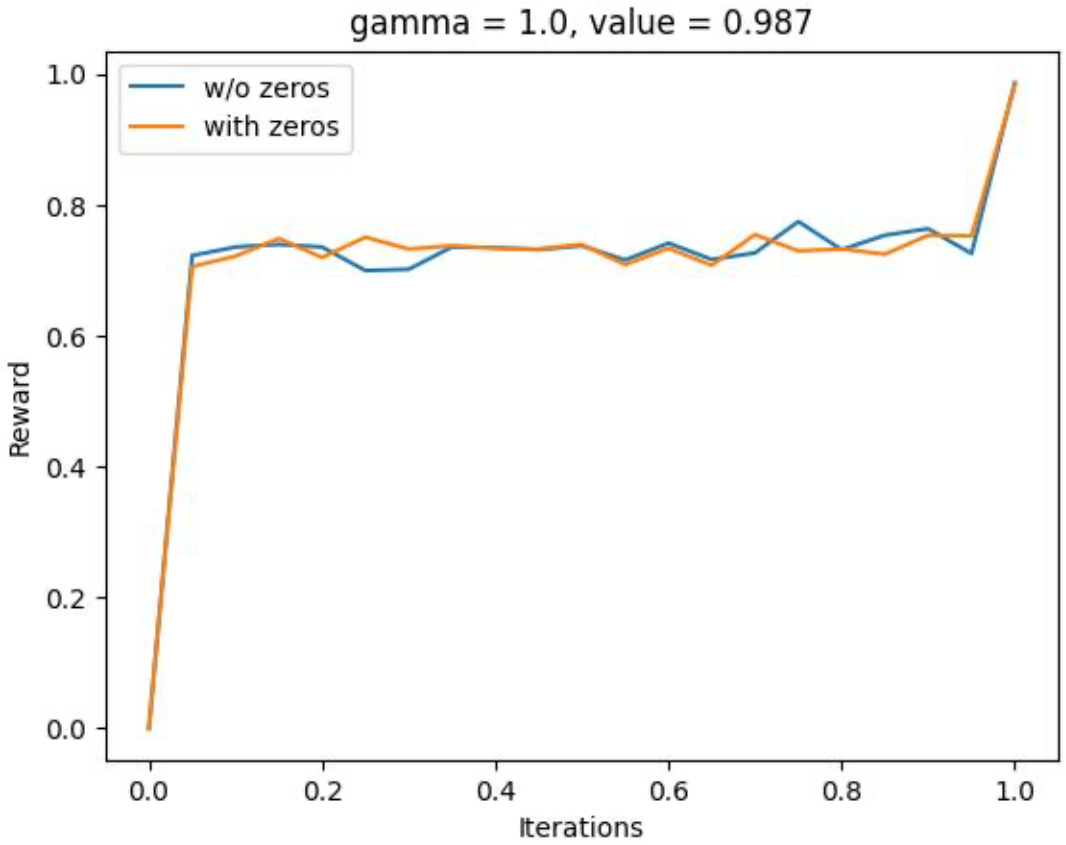
   В первом задании нужно перебрать параметр gamma. Я запустил цикл, который перебирает параметры гамма [0,1]. Результаты:



   Итог: Максимальное значение достигается при gamma=1, mean_total_reward = 1.

2. На шаге Policy Evaluation мы каждый раз начинаем с нулевых values. А что будет если вместо этого начинать с values обученных на предыдущем шаге? Будет ли алгоритм работать? Если да, то будет ли он работать лучше?

Я оставил инициализацию нулями, а также передавал в функцию «policy_evaluation» для дальнейшего обновления значений. Далее провел эксперимент с одинаковыми параметрами, но с разными подходами инициализации. Результаты:



Итог: Большой разницы в оценке награды не заметно, но можно сказать, что подход итеративного улучшения значения v_values работает немного лучше.