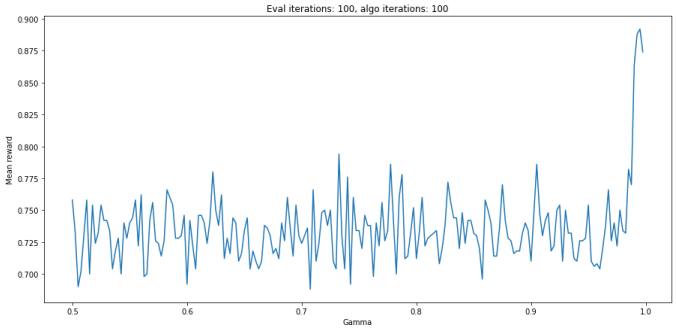
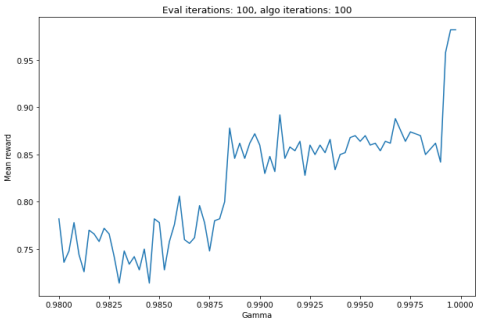
Задание 1. Исследование гиперпараметра gamma.

Эксперимент: зависимость награды от параметра gamma.



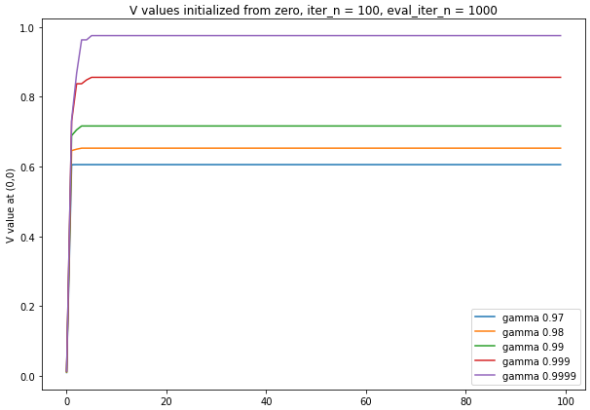


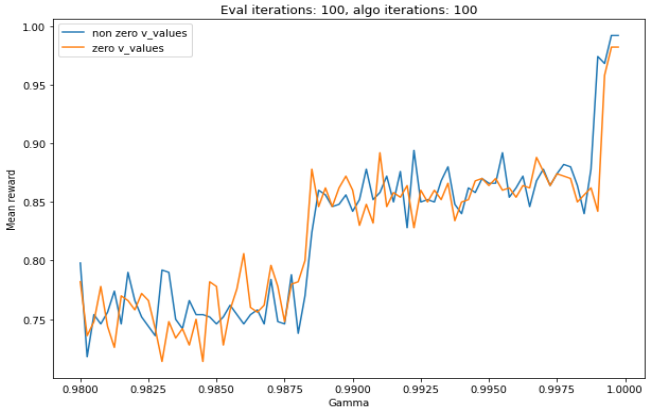
Вывод.

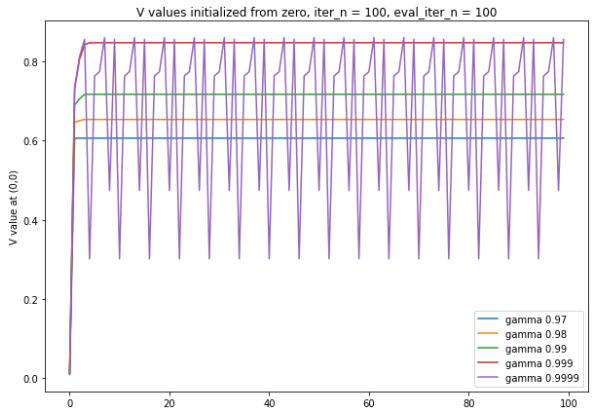
Средняя награда существенно вырастает только при gamma > 0.99, если уменьшить шаг приращения до 0.00025, то при gamma = 0.99999 средняя награда будет около 0.98.

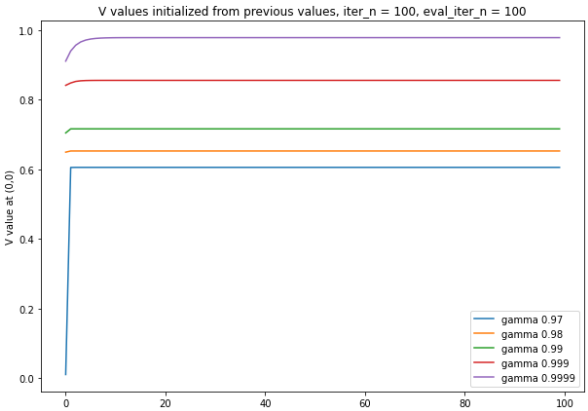
Задание 2. Не нулевые v_values

Эксперимент: сравнение нулевых v_values с v_values обученных на предыдущем шаге.







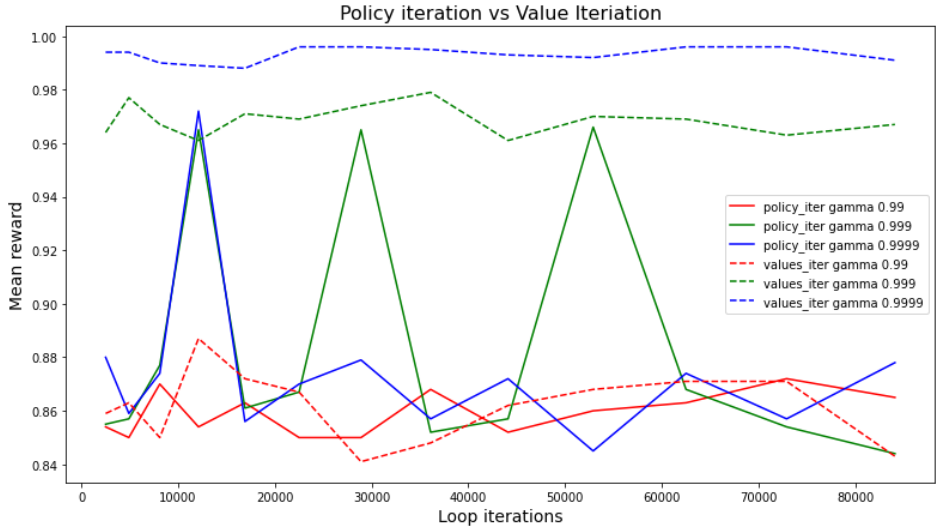


Вывод.

Предобученные v_values на высоких значениях gamma сходятся быстрее. При значении gamma 0.99 v_values сходятся после 4 итераций и скорость сходимости падает по мере увеличения gamma. При нулевой инициализации высокое значение gamma не дает v_values сойтись за 100 итераций и их число необходимо поднять до 1000.

Задание 3. Value Iteriation

Эксперимент: сравнение награды от количества итераций в циклах и параметра gamma.



Вывод.

Для корректного сравнения я взял диапазон количества итераций от 50 до 290. В случае policy iteration оба цикла iter_n и eval_iter_n имели одинаковую длину равной параметру из диапазона и сумма итераций на каждом параметре была равна iter_n * eval_iter_n. В случае values iteration параметр из диапазона возводился в квадрат и столько итераций совершалось в цикле. Общее количество итераций для каждого шага видно на оси X.

Value iteration работает более стабильно и не имеет резких скачков в средней награде.