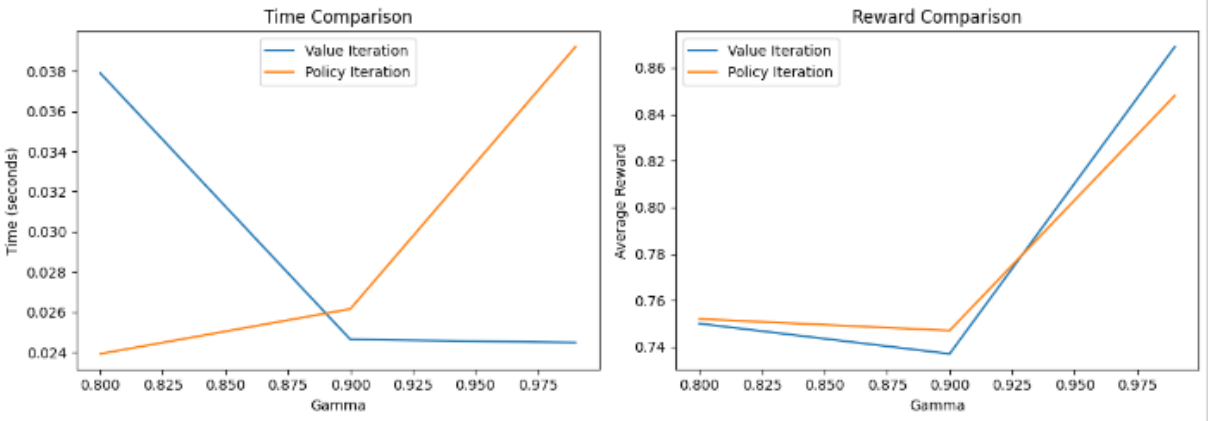
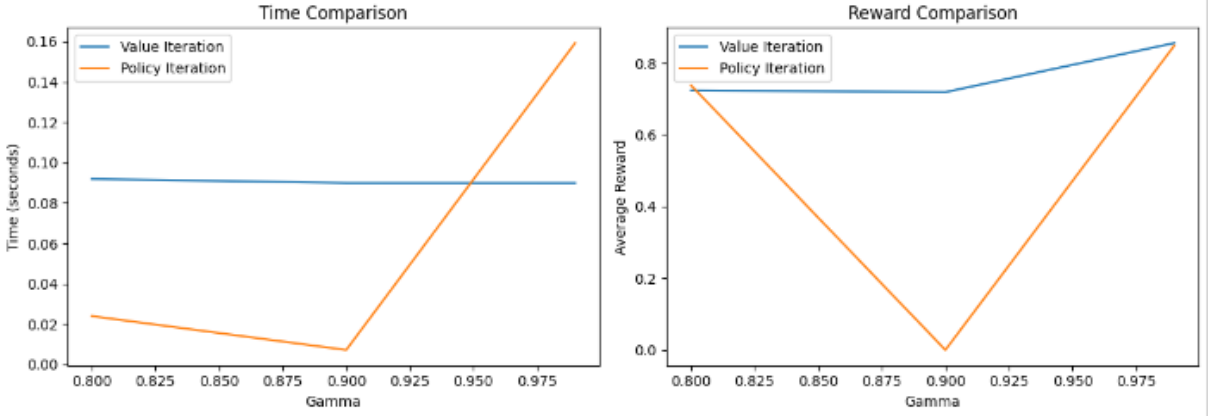
Отчет по Домашнему Заданию №3 для ODS RL’23

Наказненко Павел, 2023

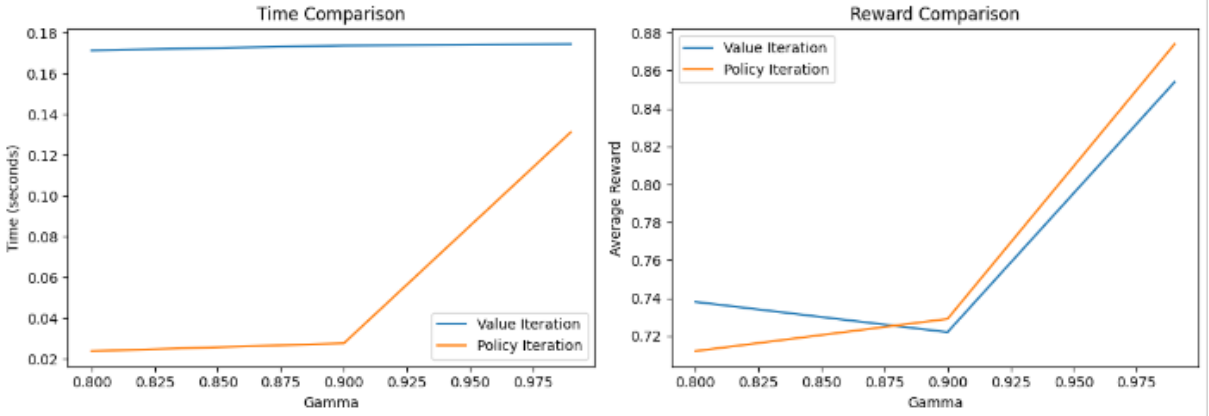
Iteration: 100



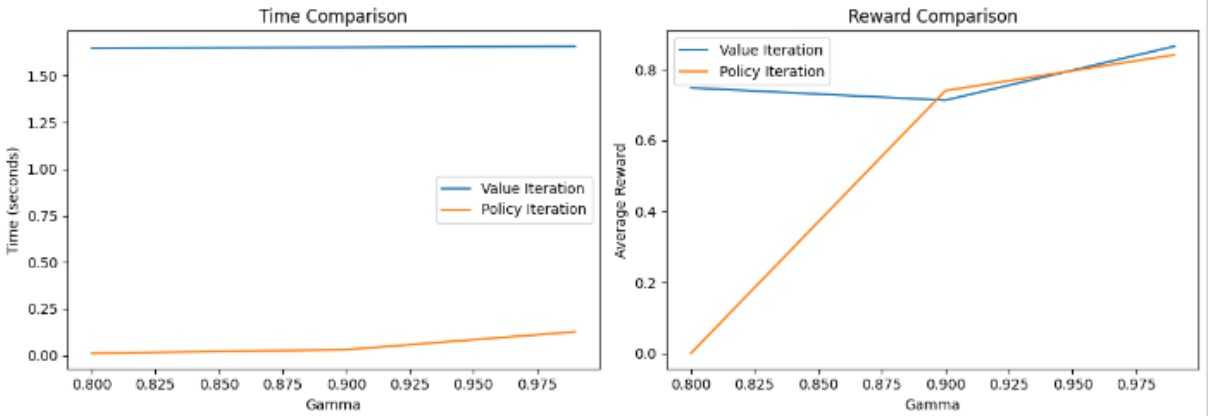
Iteration: 500



Iteration: 1000



Iteration: 10000



1. Выбор гиперпараметра gamma для Policy Iteration

Оптимальное значение `gamma` зависит от конкретных целей и условий задачи. Выбор `gamma` влияет на взвешивание немедленных и отложенных вознаграждений. С увеличением `gamma` модель будет больше стремиться к получению отложенного вознаграждения. На представленных графиках видно, что с увеличением `gamma` возрастает и средний `total_reward`, что указывает на потенциальную выгоду от учёта будущих вознаграждений. Однако стоит отметить, что `gamma` больше 1 может приводить к нестабильности и переоценке будущих наград, что не всегда является желательным и может привести к плохой сходимости алгоритма. Исходя из графиков, `gamma` следует выбирать в интервале от 0.8 до 0.99.

2. Использование предыдущих значений в Policy Evaluation

При переходе к следующему шагу Policy Evaluation с сохранением ранее вычисленных `values`, алгоритм должен продолжать работать и, в теории, сходиться быстрее, так как он будет использовать уже частично оптимизированную оценку состояний. Это позволяет алгоритму избегать пересчёта значений с нуля на каждой итерации, что может ускорить сходимость и улучшить эффективность алгоритма.

3. Сравнение Value Iteration и Policy Iteration

При сравнении этих двух алгоритмов важно учитывать не только их результативность (например, среднее вознаграждение), но и вычислительную сложность. Политика, полученная через Value Iteration, часто сходится быстрее, так как это более прямой метод оптимизации. Однако Policy Iteration может быть более эффективным при условии использования эффективной стратегии для Policy Evaluation. Адекватным сравнением алгоритмов будет учёт не только итераций алгоритма, но и общего количества обращений к среде, что отражает реальную вычислительную нагрузку. На представленных графиках видно, что с увеличением `gamma` и количества итераций разница во времени выполнения между алгоритмами увеличивается. Но также видно, что Policy Iteration достигает более высокой награды на некоторых значениях `gamma` при большем числе итераций, что указывает на его потенциальное преимущество в долгосрочной перспективе.

Таким образом, выбор конкретного алгоритма и настройка его параметров должны быть обоснованы спецификой задачи, требуемой эффективностью и доступными вычислительными ресурсами.