Policy and Value Iterations. Отчет

Содержание

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 ходе экспериментов был исследован гиперпараметр gamma для алгоритма Policy Iteration.

Результаты можно посмотреть в ClearML

Значения gamma выбирались из:
1) 0 до 0.9 с шагом ~0.25
2) От 0.901 до 1 с шагом 0.001

График среднего кол-ва обращений к средне на 1000 запусков среды в зависимости от разных gamma:

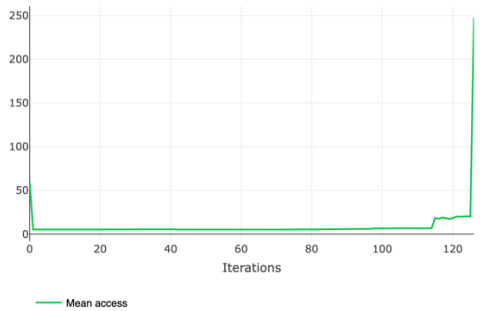


График среднего reward на 1000 запусков среды в зависимости от разных gamma:

Далее был выбран gamma с наилучшем reward и было запущено обучение с ним и тестовый запуск среды:

- Gamma = 1
- Reward = 1
- Кол-во обращений к среде = 421

Вывод:
Оптимальный gamma по reward является 1, но при таком значении происходит очень много обращений к среде. Gamma'ы от 0 до 0.99 не включительно имеют не большие значения среднего reward ~<0.8, но небольшое среднее кол-во обращений к среде ~5-6. Gamm'ы от 0.99 до 1 не включительно имеют средний reward ~>0.84, но среднее кол-во обращений к среде~17-20.
На мой взгляд, что **оптимальными являются gamm'ы от 0.99 до 1 не включительно,** так как

у них в среднем reward выше среднего и не такое большое кол-во обращений к среде по сравнению с gamma = 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?

В ходе экспериментов были исследованы использование values с предыдущего шага обучения, а также гиперпараметр gamma.

Результаты можно посмотреть в ClearML

Значения gamma выбирались из:
1) 0 до 0.9 с шагом ~0.25
2) От 0.901 до 1 с шагом 0.001

График среднего кол-ва обращений к средне на 1000 запусков среды в зависимости от разных gamma:

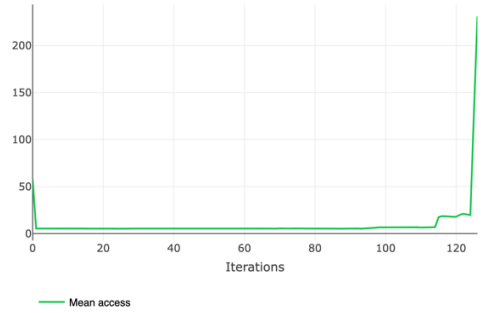
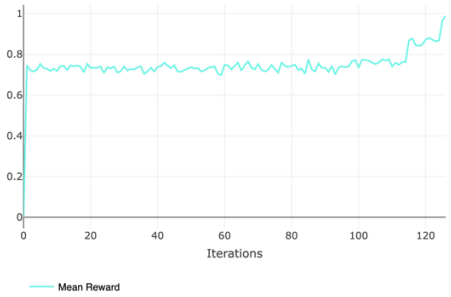


График среднего reward на 1000 запусков среды в зависимости от разных gamma:



Далее был выбран gamma с наилучшем reward и было запущено обучение с ним и тестовый запуск среды:

- Gamma = 1
- Reward = 1

Кол-во обращений к среде = 148

Вывод:
Результаты по gamma сравнимы с результатами задачи 1.
Можно заметить, что если использовать values с предыдущего шага обучения для получения q_values, а потом values инициализировать снова нулями, то **среднее обращение к среде уменьшается примерно на 25%**.

3. Написать Value Iteriation. Исследовать гиперпараметры (в том числе gamma). Сравнить с Policy Iteration. Поскольку в Policy Iteration есть еще внутренний цикл, то адекватным сравнением алгоритмов будет не графики их результативности относительно внешнего цикла, а графики относительно, например, количества обращения к среде.

Входе эксперимента были алгоритм Value Iteration и гиперпараметр gamma.

Результаты можно посмотреть в ClearML

Значения gamma выбирались из:

3) 0 до 0.9 с шагом ~0.25
4) От 0.901 до 1 с шагом 0.001

График среднего кол-ва обращений к средне на 1000 запусков среды в зависимости от разных gamma:

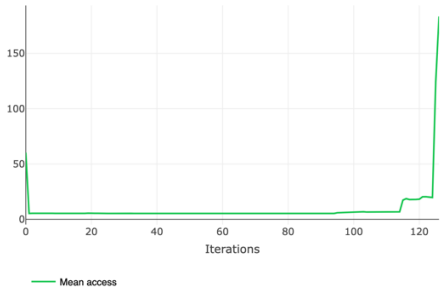
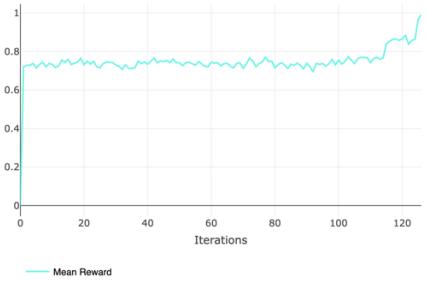


График среднего reward на 1000 запусков среды в зависимости от разных gamma:



Далее был выбран gamma с наилучшем reward и было запущено обучение с ним и тестовый запуск среды:

- Gamma = 1
- Reward = 1

Кол-во обращений к среде = 198

Далее был выбран gamma с наименьшим кол-вом обращений к среде и было запущено обучение с ним и тестовый запуск среды:

- Gamma = 1
- Reward = 0.966

Кол-во обращений к среде = 6

Вывод:

Результаты по gamma сравнимы с результатами задачи 1.

Среднее кол-во итераций заметно меньше, чем при использовании алгоритма Policy **Iteration ~2-5 раз.**
При чем если выбирать gamma по среднему кол-ву обращений к среде, то кол-во обращений становится очень маленьким ~6 вместо 421 (Задача 1), 148 (Задача 2) и 198 при выборе gamma по reward в Value Iteration без потери качества reward (на тестовом запуске).