Машинное обучение с подкреплением
Assignment III

Выполнил: студент 5 курса M05-205
Аментес Артем Важаевич

1 Задание TD

1.1 Упражнение 10.6 книги Саттона и Барто

При использовании Марковского процесса без скидок (discount factor = 1) и с единичным вознаграждением за достижение состояния A, дифференциалы для полезностей во всех трех состояниях будут равны 1. То есть, $U'(A) = U'(B) = U'(C) = 1$

1.2 Упражнение 10.6 книги Саттона и Барто

Если оценка уже принимала значение 1/3 и осталась на этом уровне, то последовательность $Rt - R\hat{}t$ будет равна 0 во всех случаях (так как истинное среднее вознаграждение равно 1/3, а оценка уже принимала это значение).
Последовательность δt будет также равна 0 во всех случаях (так как ошибка будет равна разности между текущей оценкой и текущим вознаграждением, которые равны).
Обновление $R\hat{}t+1$ с использованием ошибки δt дает более устойчивую оценку среднего вознаграждения, потому что ошибка учитывает не только текущее вознаграждение, но и текущую оценку. Это позволяет алгоритму быстрее корректировать оценку, если она начинает отклоняться от истинного значения.

2 Q-обучение

3 Actor Critic

3.1 Базовый эксперимент CartPole

Частота обучения критика и обновления целевых функций влияет на то, как быстро и точно алгоритм сходится к оптимальному решению.
Если частота обучения критика слишком низкая, то агент может не успеть учиться на новых данных и оценки будут устаревать. Это может привести к тому, что агент будет действовать неоптимально и неэффективно.
С другой стороны, если частота обучения критика слишком высокая, то алгоритм может начать переобучаться на имеющихся данных и не сможет обобщать на новые ситуации. Это может привести к тому, что агент будет действовать субоптимально и неэффективно.
Частота обновления целевых функций также влияет на скорость сходимости алгоритма. Если частота обновления целевых функций слишком низкая, то алгоритм может не успеть корректировать оценки и не сможет достичь оптимального решения. С другой стороны, если частота обновления целевых функций слишком высокая, то алгоритм может начать переобучаться и не сможет обобщать на новые ситуации.
Таким образом, оптимальная частота обучения критика и обновления целевых функций зависит от конкретной задачи и требует экспериментального подбора.

Полученные результаты в ходе эксперимента: Обновление градиентов актора и критика в задаче машинного обучения с подкреплением actor-critic должно быть одинаковым для того, чтобы достичь лучшего показателя, потому что они взаимодействуют друг с другом в процессе обучения. Критик оценивает, насколько хорошо актор выполняет задачу, и предоставляет обратную связь для корректировки стратегии актора. Если обновление градиентов актора и критика различается, то может возникнуть дисбаланс между обучением актора и критика, что может привести к неэффективному

обучению.
Частота обновления градиентов актора и критика должна быть равной 10 для того, чтобы достичь лучшего показателя, так как это является оптимальным значением, полученным в результате экспериментального подбора.

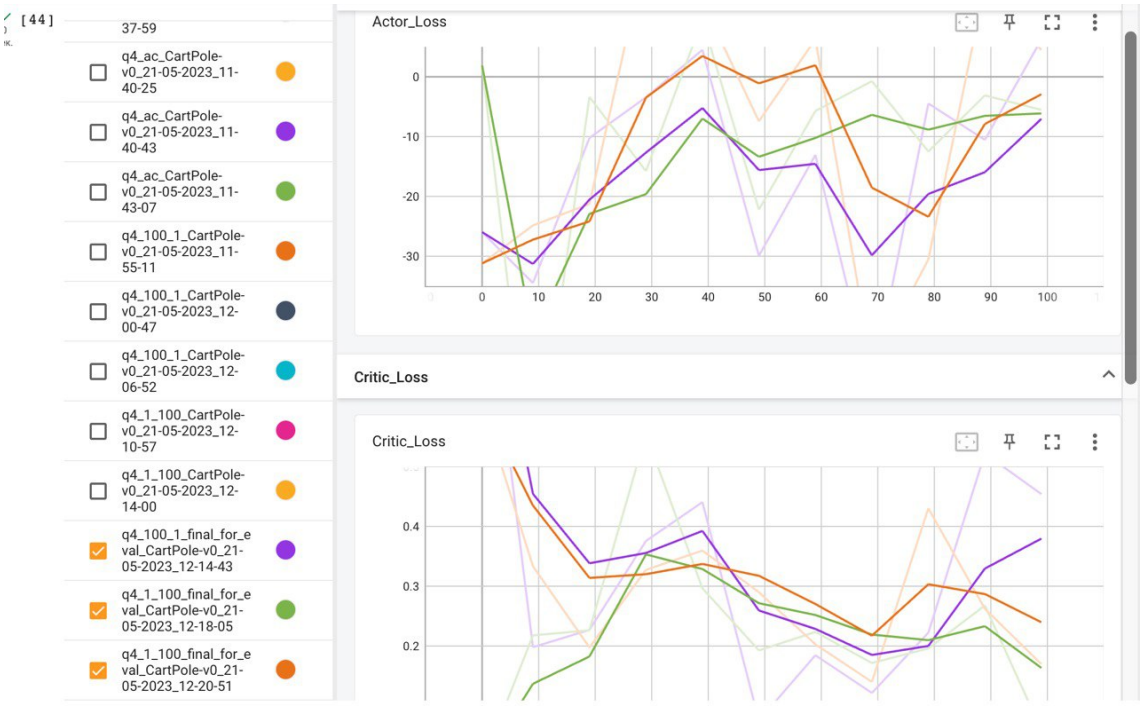

Рис. 3.1: Сравнение показателей AC(100-1, 1-100, 10-10).