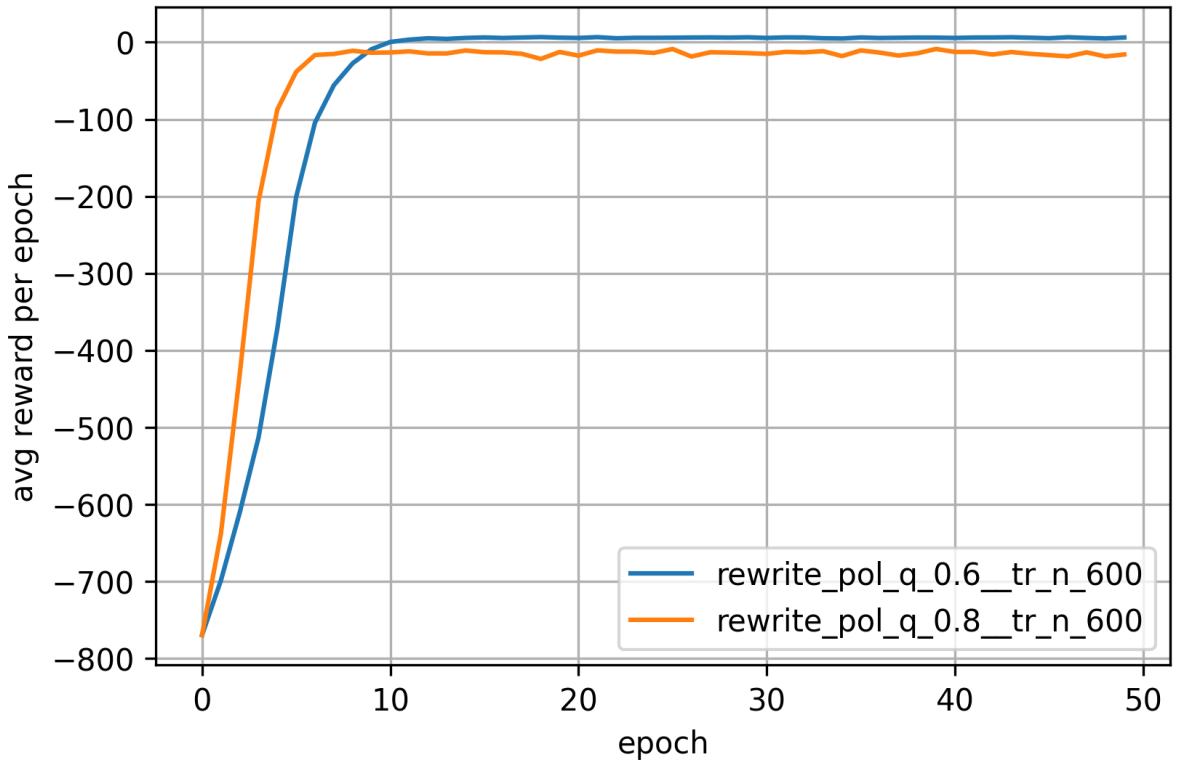
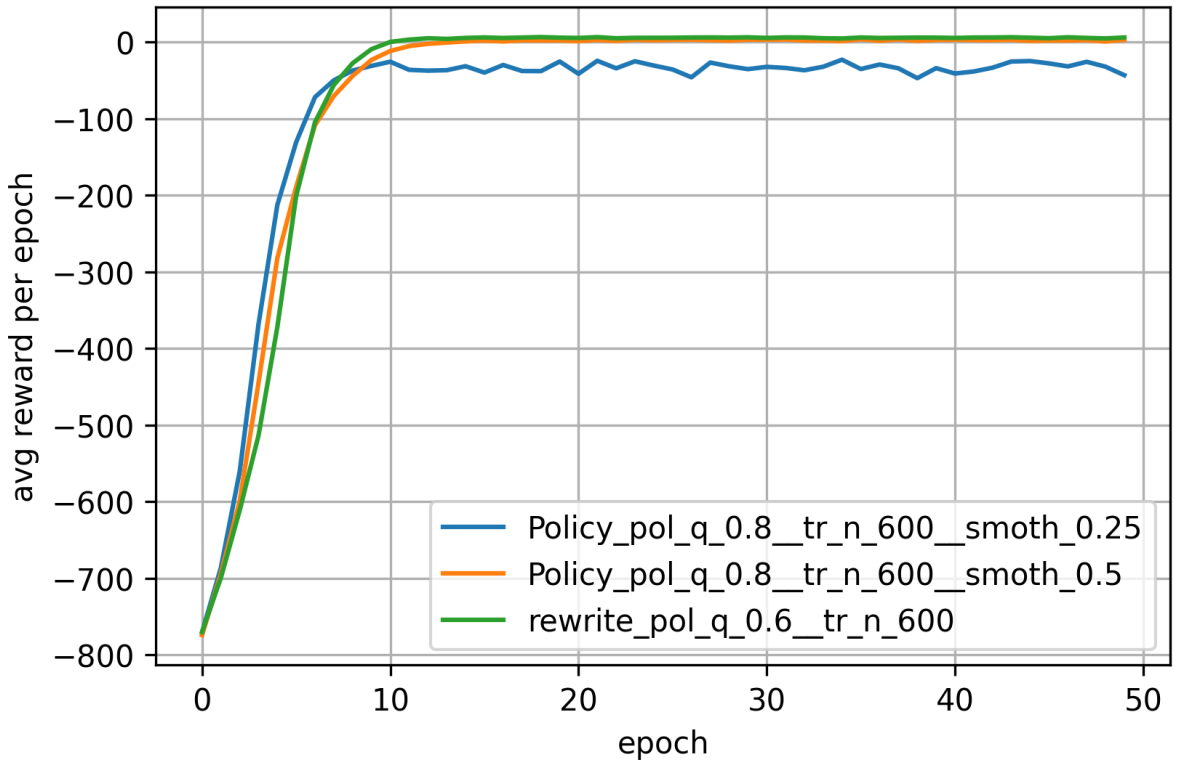
1. При подборе параметров для первой задачи, я увеличил количество траекторий по сравнению с задачей gym-maze, тк поле имеет ту же размерность, но количество состояний в 20 раз больше. Такое большое количество состояний уменьшает вероятность успешного завершения симуляции. После этого начал экспериментировать с квантилем. Как и ожидалось, при увеличении квантиля и достаточном количестве траекторий, скорость сходимости — выше, но сходится алгоритм не к глобальному оптимуму из-за недостаточного исследования.

Ниже приведен график отображающий данное поведение на 50 эпохах при квантили 0.6 и 0.8 соответственно.

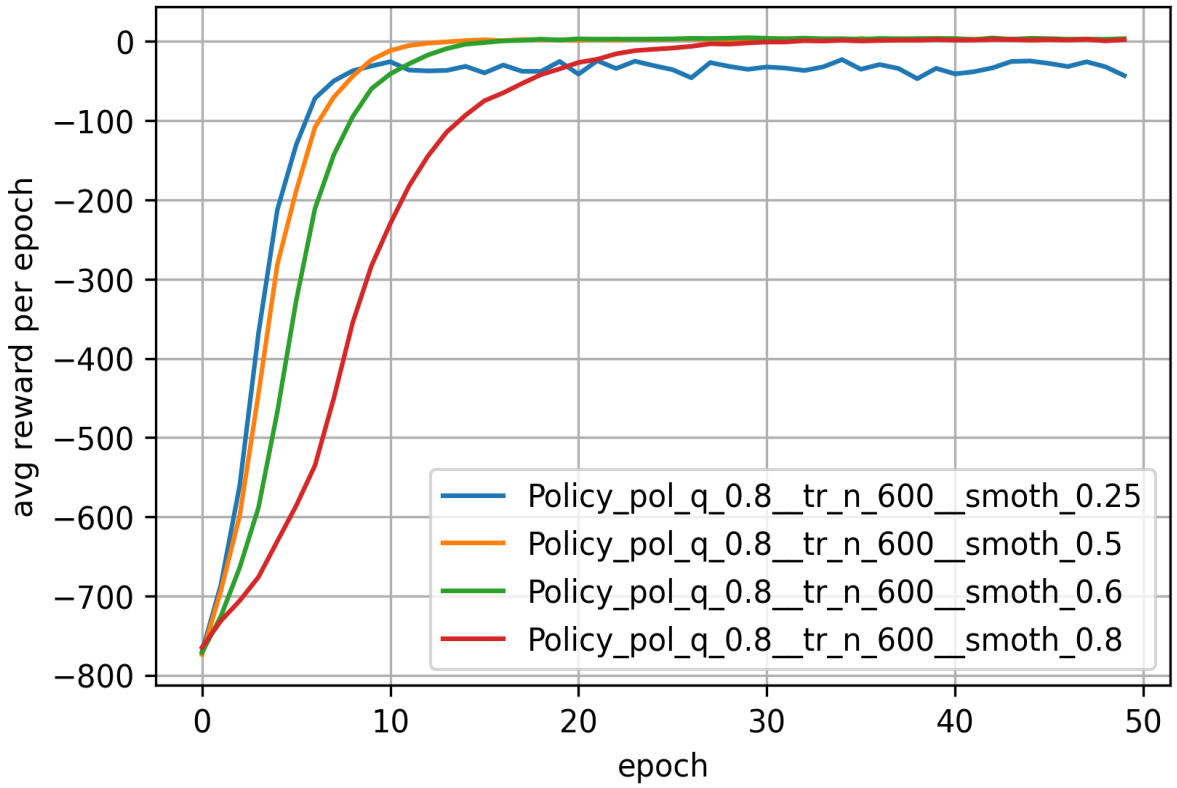


2. Сглаживание должно уберечь нас от попадания в локальный оптимум. Изначально я предполагал, что обычное сглаживание политики будет почти эквивалентно выбору меньшего параметра q на начальных эпохах (тк при меньшем q мы получаем набор семплов согласно предыдущей политике тем самым частично сохраняя ее), и не оказывать сильного влияния на поздних эпохах тк геометрическая прогрессия образованная q^epoch убывает достаточно быстро. Такое поведение подтвердилось. Правда я ожидал увидеть похожие графики наград для обучения без сглаживания и q=0.6 и для обучения при наличии сглаживания и q=0.8, λ = 0.25 (тк 0*0.25 + 0.8*(1-0.25) = 0.6). Однако при λ = 0.5 результат похож на желаемый больше



Сложилось впечатление, что подбирать параметр q такой, чтобы не сойтись в локальный минимум может быть сложнее, чем выбрать

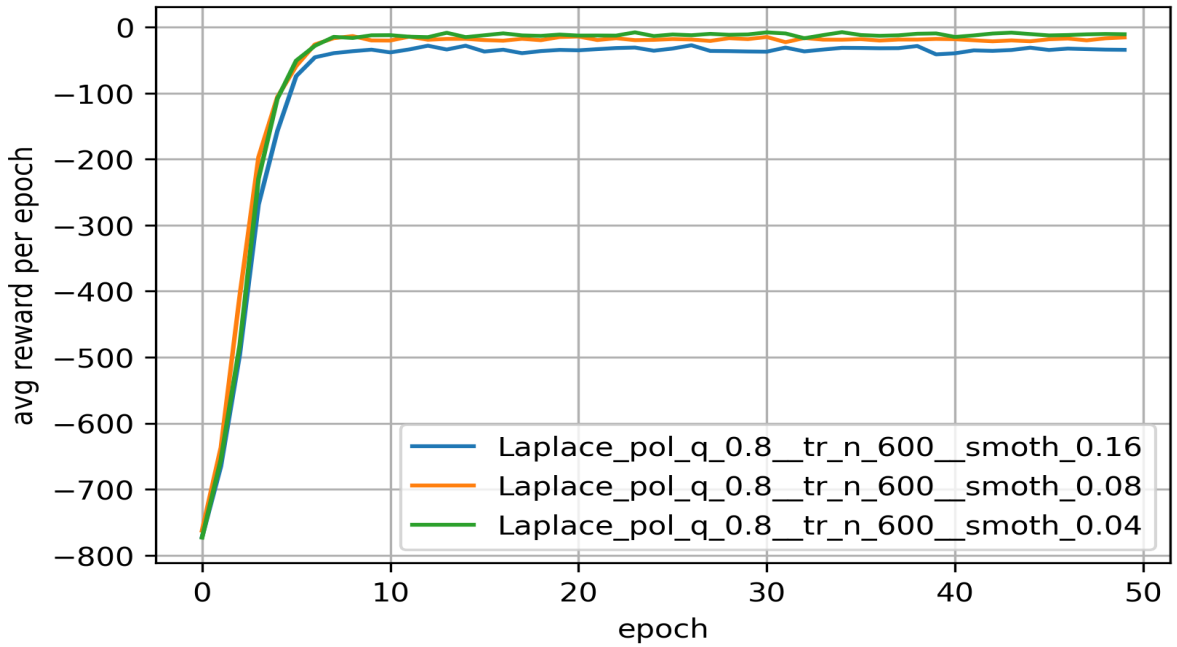
сглаживание побольше для большего q.



Сглаживание лапласа для данной задачи будто подходит хуже. Мне кажется, что подобрать хороший диапазон параметра довольно сложно. В отличии от обычного сглаживания, сглаживание Лапласа, с ростом количества эпох, не угасает для редких состояний. Например, если мы уже неплохо обучились и знаем, что в клетке с пассажиром нам надо его только подбирать, то мы побываем в этом состоянии единожды, а сглаживания лапласа все остальные действия вероятности оставит равновероятными и равными

$$(0 + \lambda)/(1 + \lambda|A|).$$

Таким образом модель не стремится к оптимуму при наличии данного вида сглаживания. (На графике ниже при различных параметрах средняя награда не была никогда положительна)



3. Хоть в данной задаче среда является детерминированной, начальные условия даже при фиксированной политике — разные. Если бы начальные условия были постоянными, то увеличив количество траекторий в эпоху кратно количеству траекторий для усреднения при фиксированной политике, мы получили бы результат идентичный обучению без усреднения по политике. Но поскольку в элитные траектории теперь входят не только элитные траектории, но и те которые попали в одну политику с хорошими траекториями то скорость обучения замедляется (и из-за возросшего количества траекторий и из-за попадания в обучающую выборку более плохих реализаций траекторий). По поведению обучение с осреднение похоже на обучение с обычным сглаживанием, нам удается легко достигнуть оптимума, при этом жертвуя скоростью схождения.