Отчет

По домашнему заданию №1

Pavel Nakaznenko

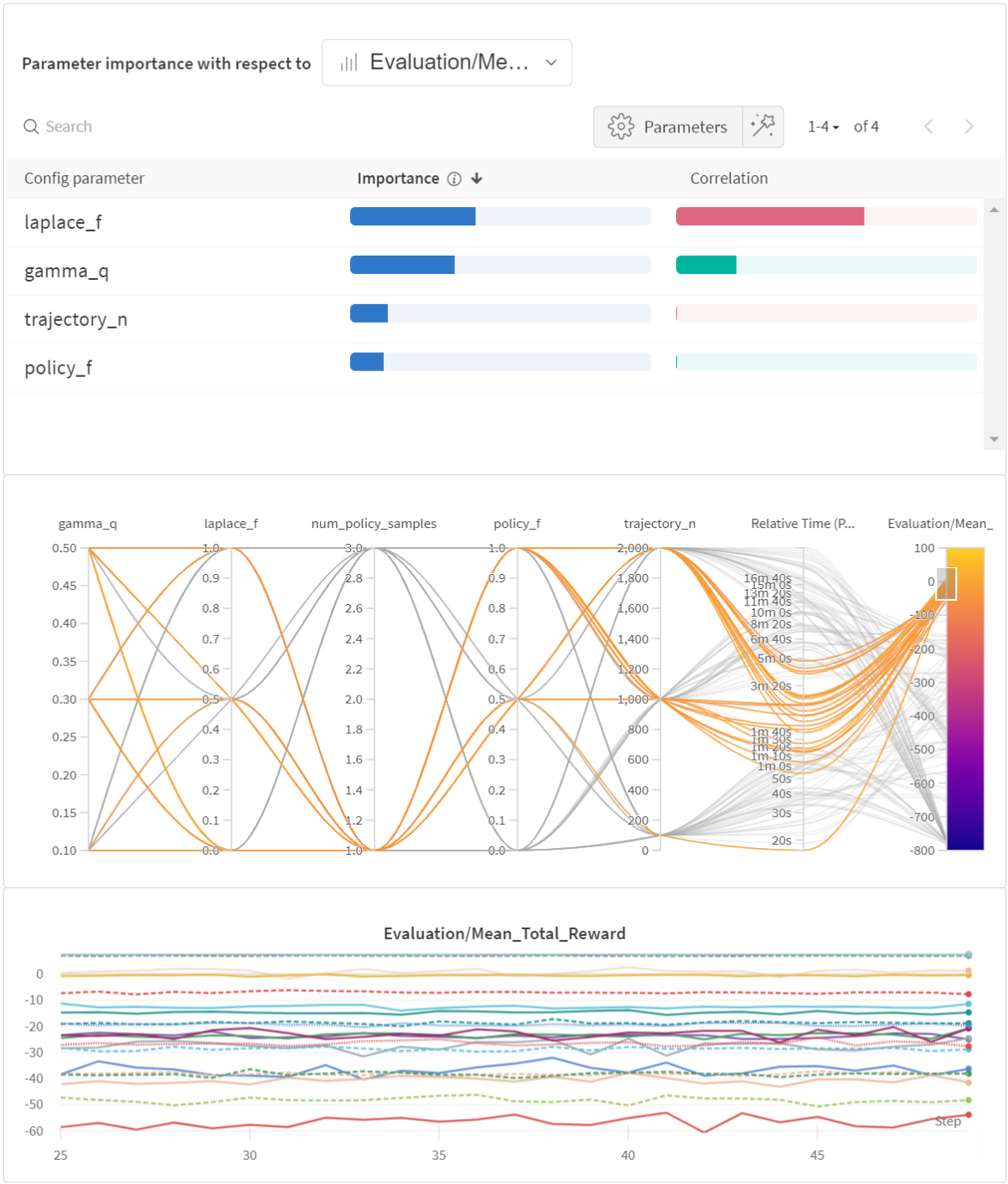
Гиперпараметры:

- gamma_q - гамма параметр квантиля
- laplace_f - фактор интерполяции по Лапласу
- policy_f - фактор линейной интерполяции от предыдущей политики к новой
- trajectory_n - количество траекторий в выборке
- iteration_n - (фиксированный в данном отчете, равен 25) количество итераций обучения
- max_trjectory_len - (фиксированный в данном отчете, равен 200) максимально допустимая длина траектории в выборке. Значение выбрано в соответствии с критерием останова симуляции.

Эксперименты:

1. Имплементация классического Cross Entropy Method
2. Имплементация сглаживания по Лапласу (laplace_f) и сглаживания политики (policy_f)
3. Имплементация множественного сэмплинга (num_policy_samples) стохастической среды (num_policy_samples > 1)
4. Определение теоретически возможной максимальной средней награды (эмпирически измеряем mean total reward для всех возможных изначальных значений state)
5. Интерполяция гиперпараметров trajectory_n и gamma_q от эпизода к эпизоду
6. grid search для определения влияния гиперпараметров

Результаты:


Parameter importance with respect to
Evaluation/Me…
Search
Parameters
1-4 of 4
Config parameter
Importance
Correlation
laplace_f
gamma_q
trajectory_n
policy_f
gamma_q
laplace_f
num_policy_samples
policy_f
trajectory_n
Relative Time (P...
Evaluation/Mean_
Evaluation/Mean_Total_Reward
Step

Наблюдения и выводы:

1. Для оптимальной производительности важна "насмотренность": такое сочетание гиперпараметров trajectory_n и gamma_q, при котором агент "наблюдает" достаточно различных последствий своих действий.
2. Эмпирически посчитанный максимально возможный mean total reward 7.93
3. Лучший результат обучения методом кросс энтропии находится в окрестности теоретически возможного на eval-сете с большим количеством эпизодов
4. Оптимальные гиперпараметры находятся в окрестности следующих значений: trajectory_n >=1000, gamma_q <= 0.3, iteration_n >= 100
5. Техники по улучшению результата обучения на стохастических средах в среднем дают ожидаемое ухудшение производительности на детерминированной среде, а так же увеличивают время сходимости
6. Интерполяция гиперпараметров trajectory_n и gamma_q от более обширного обхвата до более элитарного в среднем увеличивает скорость сходимости обучения, сохраняя баланс между результатом и скоростью обучения.