Assignment 1

Due March 11

1 Theoretical assignment: TD(λ)

Данное теоретическое задание посвящено алгоритму TD(λ).

Задание TD.1 Доказать, что если вычислять обновления весов на каждом шаге (без их применения), то суммарное обновление онлайнового и оффлайного варианта одинаково.

Под оффлайновым обновлением подразумевается оффлайновый алгоритм λ-отдачи. Под онлайновым обновлением подразумевается алгоритм TD(λ). Подсказки можно найти в заданиях в конце главы 12.1 книги Саттона и Барто.

Задание TD.2 Доказать равнозначность оффлайного алгоритма λ-отдачи и истинно онлайнового алгоритма TD(λ).

См. соответствующую главу 12.5 книги Саттона и Барто и статью van Seijen et al., 2016.

2 Practical assignment: Imitation learning

Цель задания - реализовать и поэкспериментировать с алгоритмами имитационного обучения. По имеющимся демонстрациям предлагается обучить два варианта стратегии: с помощью алгоритма клонирования поведения и алгоритма DAgger (см. лекция по имитации поведения, статья по DAgger), а затем сравнить их качество в средах на базе Mujoco. Данное задание основано на первом задании курса по Deep RL Университета Беркли.

Код задания расположен по адресу

https://github.com/pkuderov/mipt-rl-hw-2022

Есть возможность выполнять задание как локально, так и в Google Colab. Подробности, в том числе по установке зависимостей, вы найдете в README в репозитории.

2.1 Behavioral Cloning

Требуется заполнить пропуски в реализации алгоритма клонирования поведения. По адресу `expert_data` содержатся *pickled* данные экспертных стратегий. Рекомендуется начать ознакомление с точки запуска `scripts/run_hw1.py` и затем продвигаться вглубь последовательно:

- `infrastructure/rl_trainer.py`,
- `agents/bc_agent.py`,

- `policies/MLP_policy.py`
- и далее, заполняя пропуски, отмеченные `TODO`.

Задание BC.1 Получите результаты вашей реализации в двух средах: Ant и любой другой. В Ant качество должно быть не хуже 25% от качества эксперта. Пример запуска найдете в README. Требуются следующие данные: среднее значение и стандартное отклонение отдачи по набору тестовых эпизодов. Эти данные логируются как `Eval_AverageReturn` и `Eval_StdReturn` соответственно.

Обратите внимание, чтобы собрать данные по нескольким эпизодам, потребуется при запуске указать параметр суммарного количества шагов ***`eval_batch_size`*** *в несколько раз больше параметра максимальной длины эпизода* ***`ep_len`****. Собственно, их отношение и задает минимальное число эпизодов, которые войдут в выборку.*

Выключение генерации видео при логировании достигается с использованием аргумента запуска `--video_log_freq -1`*. Удалите его при необходимости, но помните о возможном существенном замедлении выполнения.*

Задание BC.2 Поэкспериментируйте с разными наборами гиперпараметров (например, число шагов обучения, объем предоставленных экспертных данных и т.п.). Для одного фиксированного гиперпараметра постройте график изменения качества работы агента [на одной из сред из первого задания] и поясните ваш выбор этого параметра.

2.2 DAgger

Заполнив все пропуски `TODO`, вы также сможете запустить DAgger (см. аргументы запуска в README).

Задание DA.1 Проведите запуски алгоритма на выбранных в задании 1 средах. Постройте график обучения DAgger'а—число итераций алгоритма vs. средняя отдача за эпизод (с указанием стандартного отклонения). Добавьте в этот график горизонтальными линиями результаты эксперта и клонирования поведения. Укажи использованные гиперпараметры.

3 Формат сдачи

Сдача предполагается в виде предложений изменения кода (pull request) в репозитории ссылке в начале практического задания.

Ожидается, что предложение будет содержать непосредственно код заполненных вами недостающих частей выданного решения и логи финальных запусков (для каждого задания и каждой из использованных сред).

Оригинально все логи лежат в папке `data`. Логи финальных запусков скопируйте из `data` в отдельную папку `run_logs` и отправьте вместе с вашим решением.

Также в сообщении к предложению необходимо добавить результаты/описание/решение по каждому из пунктов задания (в соответствии с тем, что оно требует). Разметка (markdown) позволяет и вставку картинок, и оформление табличек. Опционально, вы можете оформить результаты в виде отдельного файла .doc или .pdf и добавить их в посылку (commit), а в сообщении сослаться на этот файл. Не забудьте также добавить к каждому пункту задания код запуска, чтобы можно было воспроизвести ваши результаты.