Машинное обучение с подкреплением. МТИИ 2021. Домашнее задание №2.

Сроки выполнения с 24 марта по 4 апреля, до 23:59 по Москве. За каждый день просрочки **-1 балл** к итоговой оценке по всему домашенму заданию по 10-балльной шкале. Решения(и теоретическую и практическую части) рекомендуется оформлять в виде одной Jupyter тетрадки со всеми необходимыми пояснениями и комментариями. Название тетрадки должно совпадать с вашей фамилией (например, Петров.ipynb). Загружать тетрадки нужно через Dropbox по следующему адресу: `https://www.dropbox.com/request/XFQTfyrYJZwDUGokl2HQ`. Датой отправки считается дата, значащаяся в Dropbox.

Задание 1. Теоретическая часть: решение оптимизационной задачи. (50 баллов)

Мы с вами рассмотрели на лекции, что задача оптимизации градиента стратегииможет быть сведена к оптимизационной задаче с ограничениями. Основнойчастью этой задачи является суррогатная целевая функция и ограничение нарасстояние между стратегиями. На лекции мы привели теорему о нижнейгранице оценки на полезность любой стохастической стратегии. В этом заданиипредлагается разобраться с этим немного подробнее.

- Найдите или постарайтесь самостоятельно привести доказательство теоремы о нижней границе, в которой был задан коэффициент при расстоянии между стратегиями равный $\frac{4\epsilon\gamma}{(1-\gamma)^2}$ **(24 балла)**
- Сформулируйте оптимизационные задачи со штрафом и согранчением на расстояние между стратегиями. **(10 баллов)**
- Выпишите полное решение оптимизационной задачи соштрафом методом множителей Лагранжа при использованииаппроксимации по Тейлору. **(16 баллов)**

Задание 2. Практическая часть: Vanilla Policy Gradient. (50 баллов)

Вашей задачей будет реализация алгоритма Vanilla Policy Gradient (VPG). Вам нужно поработать с двумя классами: задающим самого агента *MLPPolicyPG* и его стратегию – *MLP_policy*. Заготовка кода доступна по ссылке: `https://colab.research.google.com/drive/1lTpcxmswU7XyhuPR_L-gAIkpuA1MtEMo`. Вам необходимо вписать недостающий код в пропуски и провести эксперименты.

- Необходимо реализовать базовый алгоритм VPG. Для оценки траектории используется сумма дисконтированных вознаграждений. **(20 баллов)**
- Проведите эксперименты каждого из вариантов с окружением *CartPole-v0* и *InvertedPendulum-v2* с разными вариантами по размеру батча (batchsize $\in [1000, 5000]$). Изобразите результаты на графике прокомментируйте результаты. **(10 баллов)**
- Заполните пропуски в методе *estimate_advantage*. **(10 баллов)**
- Проведите эксперименты со средой *LunarLanderContinuous-v2*. Попробуйте подобрать гиперпараметры, чтобы увеличить скорость сходимости алгоритма. Постройте графики и прокомментируйте результаты. **(20 баллов)**

Графики и ваши выводы рекомендуется оформить в отдельном разделе тетрадки (отделить от кода).