Машинное обучение с подкреплением. МТИИ 2021.
Домашнее задание №1.

Сроки выполнения с 17 февраля по 28 февраля, 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) необходимо оформлять в виде одной Jupyter тетрадки со всеми необходимыми пояснениями и комментариями. Название тетрадки должно совпадать с вашей фамилией (например, Петров.ipynb). Тетрадка должно выполняться в google colab. Загружать тетрадки нужно через Dropbox по следующему адресу: `https://www.dropbox.com/request/aFmPvtPvdiwwAaTwX3DC`. Датой отправки считается дата, значащаяся в Dropbox.

Задание 1. Теоретическая часть: Смещенность оценок при оценке стратегии. (50 баллов)

Оценка стратегии – это один из ключевых методов в обучении с подкреплением, который является составной частью обобщенной итерации по стратегиям. Существует два основных алгоритма оценки стратегии: метод Монте-Карло (MC, две вариации: с первым посещением $MC(1)$ и с каждым посещением $MC(all)$ и метод временных различий $TD(0)$. Методы оценки стратегии обладают двумя важными свойствами: смещенностью и дисперсией. В этой части вам предлагается изучить эти свойства в методах $MC(1)$, $MC(all)$ и $TD(0)$.

- Дайте определение свойств смещенности и дисперсии для оценки случайной величины. **(4 балла)**
- Является ли оценка отдачи по алгоритму $MC(1)$ смещенной? Попробуйте это доказать формально. **(10 баллов)**
- Является ли оценка отдачи по алгоритму $MC(all)$ смещенной? Попробуйте это доказать формально. **(10 баллов)**
- Попробуйте сравнить аналитически дисперсии оценок отдачи по $MC(1)$ и $MC(all)$. **(6 баллов)**
- Является ли оценка отдачи по алгоритму TD(0) смещенной? Попробуйте это доказать формально. **(20 баллов)**

Задание 2. Практическая часть: $Q(\lambda)$ (50 баллов)

Вашей задачей будет реализация N-шагового обучения (или n-step bootstrapping) подхода вместе с алгоритмом Q-обучения. Подробности смотрите в книге Саттона и Барто: `https://yadi.sk/i/XHwgifNrMBvHiA`. Q-обучение было разобрано на семинарских занятиях.

- Реализуйте N-step вариант Q-обучения и проверьте его работу для различных N. **(30 баллов)**
- Реализуйте тот же вариант для алгоритма SARSA: N-step SARSA. Сравните его работу с N-step Q-обучением. **(20 баллов)**

Проверить сходимость алгоритмов нужно на следующих средах: **Taxi-v3, CliffWalking-v0, NChain-v0**. Для каждого набора параметров алгоритма нужно провести несколько экспериментов, а затем усреднить их, чтобы результаты были статистически значимы.