Улучшение подхода Upside Down в обучении с подкреплением

Александр Никулин

Научный руководитель: А. А. Шпильман
Научный консультант: О. А. Свидченко

Задача обучения с подкреплением

Чего хотим мы

Получить агента способного решить конкретную задачу, например:

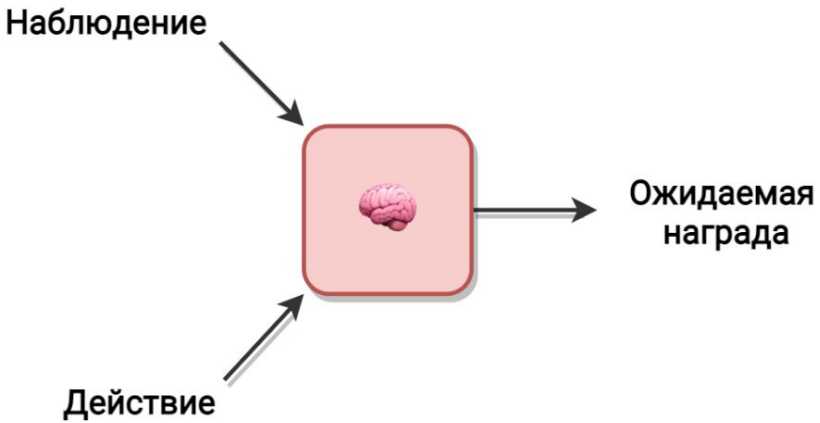
- посадить ракету
- научиться ходить
- выиграть в шахматы

Чего хочет агент

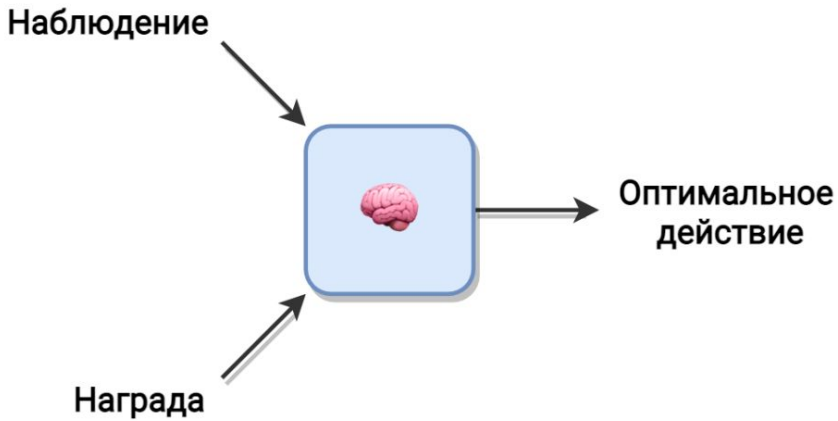
Выбирать действия так, чтобы получать максимальную награду

Обучение с подкреплением вверх ногами

Традиционное (RL)



Вверх ногами (UDRL)



Schmidhuber, J. (2020). Reinforcement Learning Upside Down: Don't Predict Rewards -- Just Map Them to Actions. ArXiv:1912.02875 [Cs]. http://arxiv.org/abs/1912.02875
Srivastava, R. K., Shyam, P., Mutz, F., Jaśkowski, W., & Schmidhuber, J. (2019). Training Agents using Upside-Down Reinforcement Learning. ArXiv:1912.02877 [Cs]. http://arxiv.org/abs/1912.02877

Обучение с подкреплением вверх ногами

В чем плюсы?

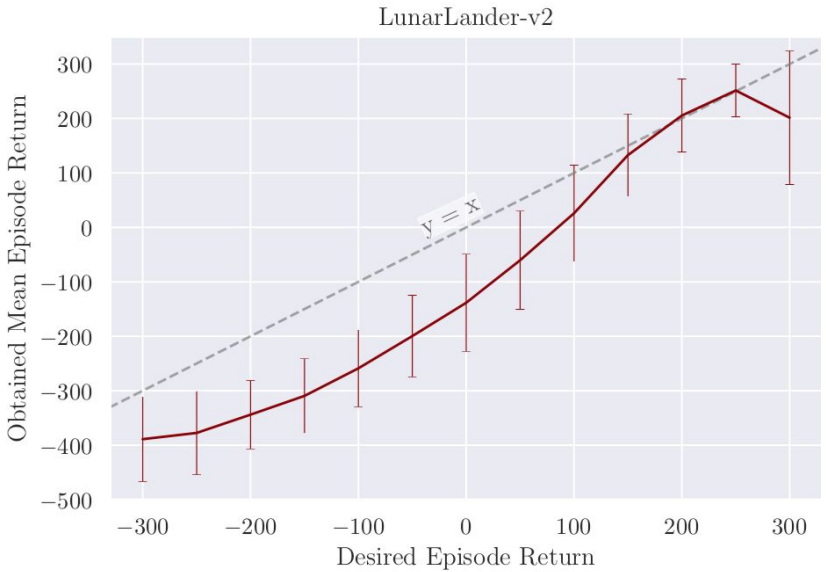
- легко представить как задачу обучения с учителем
- как следствие, стабильность и множество проверенных временем методов
- обученный **агент способен получать весь спектр возможных наград**
- гораздо больше данных для обучения в сравнении с традиционными методами
- в теории, нет проблем с разряженной наградой

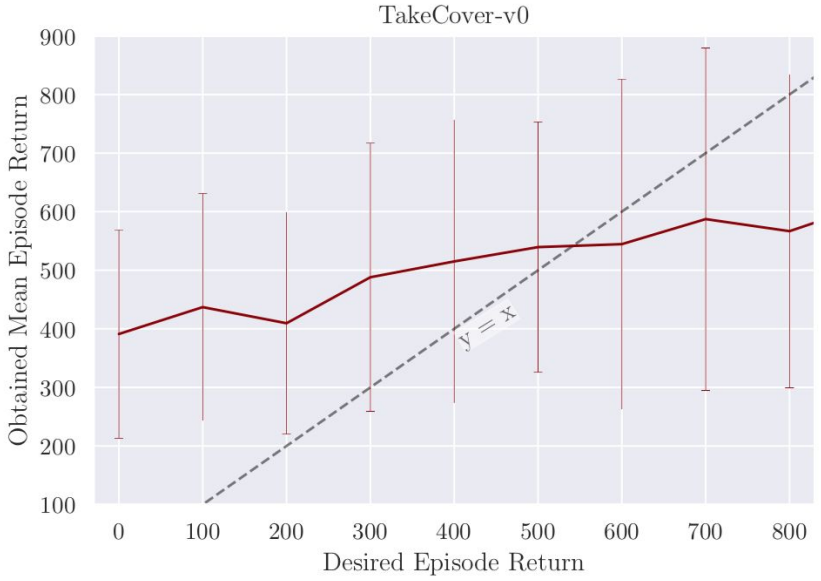
В чем минусы?

- теряет смысл, если награда не имеет осмысленного спектра, например: выиграл, проиграл

Schmidhuber, J. (2020). Reinforcement Learning Upside Down: Don't Predict Rewards -- Just Map Them to Actions. ArXiv:1912.02875 [Cs]. http://arxiv.org/abs/1912.02875
Srivastava, R. K., Shyam, P., Mutz, F., Jaśkowski, W., & Schmidhuber, J. (2019). Training Agents using Upside-Down Reinforcement Learning. ArXiv:1912.02877 [Cs]. http://arxiv.org/abs/1912.02877

Обучение с подкреплением вверх ногами





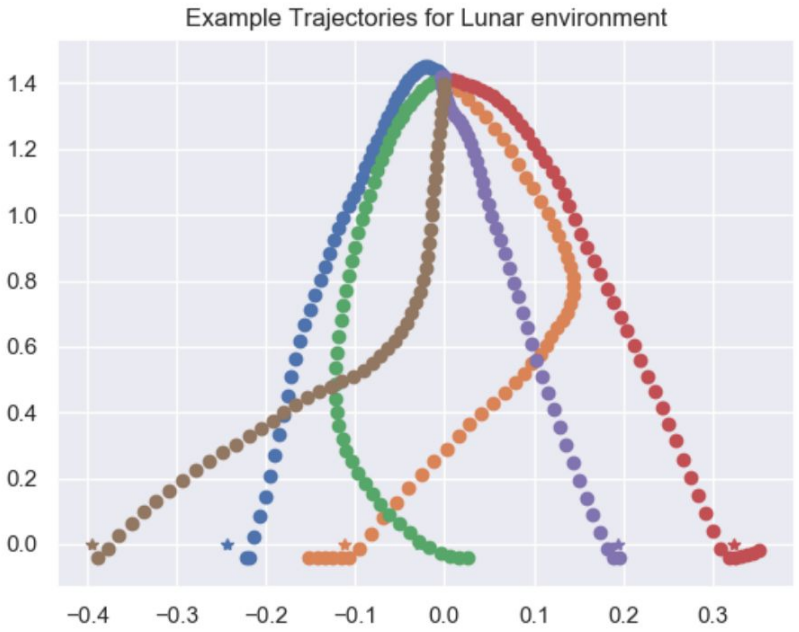
Schmidhuber, J. (2020). Reinforcement Learning Upside Down: Don't Predict Rewards -- Just Map Them to Actions. ArXiv:1912.02875 [Cs]. http://arxiv.org/abs/1912.02875
Srivastava, R. K., Shyam, P., Mutz, F., Jaśkowski, W., & Schmidhuber, J. (2019). Training Agents using Upside-Down Reinforcement Learning. ArXiv:1912.02877 [Cs]. http://arxiv.org/abs/1912.02877

Обзор схожих методов

Goal-conditioned RL

- в качестве цели - состояние, а не награда
- награда - достижение цели
- популярен в робототехнике
- обучение с помощью традиционных для RL алгоритмов

Однако, метод предложенный (Ghosh et al., 2020) почти полностью повторяет метод (Srivastava et al., 2019)



Ghosh, D., Gupta, A., Reddy, A., Fu, J., Devin, C., Eysenbach, B., & Levine, S. (2020). Learning to Reach Goals via Iterated Supervised Learning. ArXiv:1912.06088 [Cs, Stat]. http://arxiv.org/abs/1912.06088
Srivastava, R. K., Shyam, P., Mutz, F., Jaśkowski, W., & Schmidhuber, J. (2019). Training Agents using Upside-Down Reinforcement Learning. ArXiv:1912.02877 [Cs]. http://arxiv.org/abs/1912.02877

Цель и задачи

Существующие подходы либо требуют датасета с экспертными траекториями, либо получают агентов, которые так и не научаются точно выполнять команды.

Цель: Разработка нового алгоритма **UDRL**, исправляющего недостатки существующих подходов.

Задачи:

- разработать алгоритм, способный обучать агента **UDRL**
- разработать синтетическое окружение, подходящее для тестирования свойств полученного агента
- проанализировать свойства полученного агента на основе экспериментов в существующих и разработанном окружениях

Архитектура агента

Из-за схожести задач в основу агента был взят алгоритм Soft Actor-Critic

Критик

Учится предсказывать условное распределение исходов для актора:

$$o \sim p(\cdot|s, a, c)$$

Вероятность выполнить команду:

$$Q(s, a, c) = p(o = c|s, a, c)$$

Условное распределение представляется в виде смеси Гауссиан.

Актор

Учится выбирать действия, которые приведут к выполнению команды:

$$a \sim \pi(\cdot|s, c)$$

Для этого максимизируется функция ценности:

$$V(s, c) = \mathbb{E}_{a\sim\pi(\cdot|s,c)}[Q(s, a, c)] + \alpha \cdot H(\pi(\cdot|s, c))$$

Учится максимизировать вероятность выполнить команду.

Окружения: SumGame

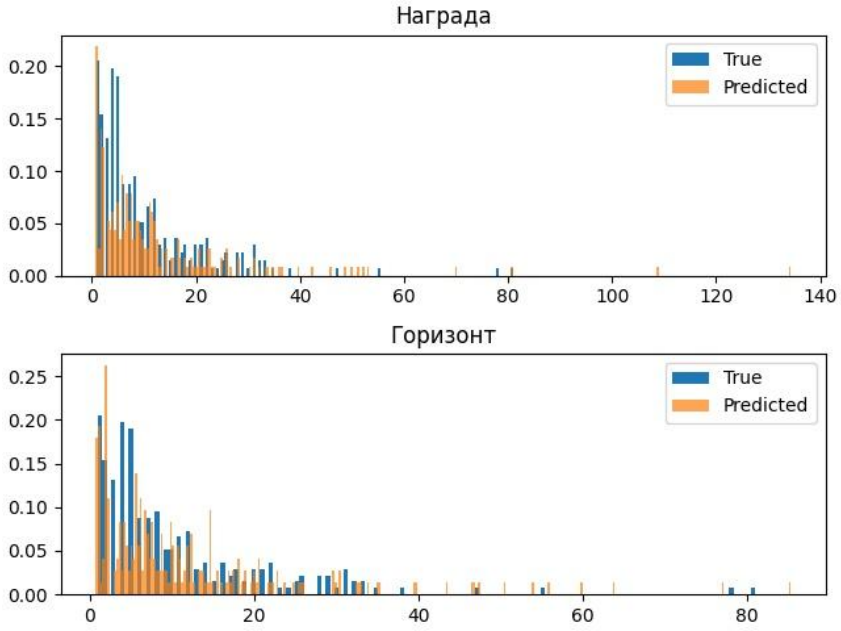
Отладочная среда для тестирования.

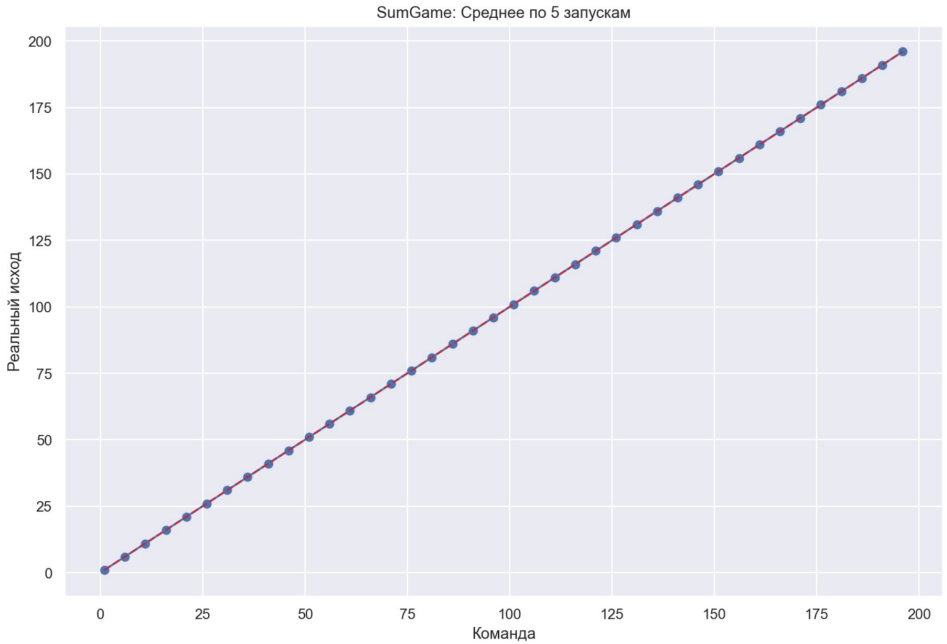
Имеет только одно состояние.
На каждом шаге агенту доступно два действия:

- получить 1 награды и продолжить игру
- получить 1 награды и закончить игру.

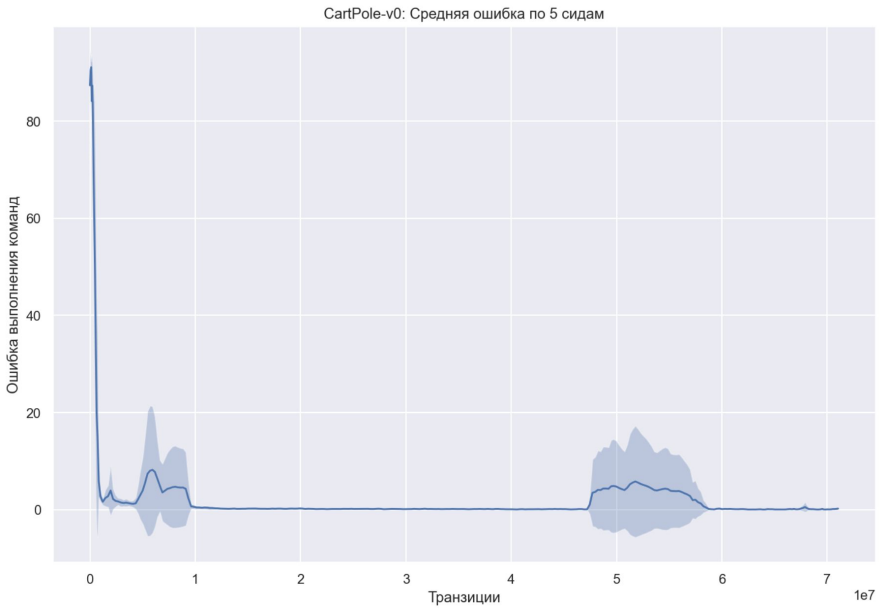
Игра ограничена 200 шагами. Награда всегда равна длине траектории, что удобно для визуализации и тестирования.

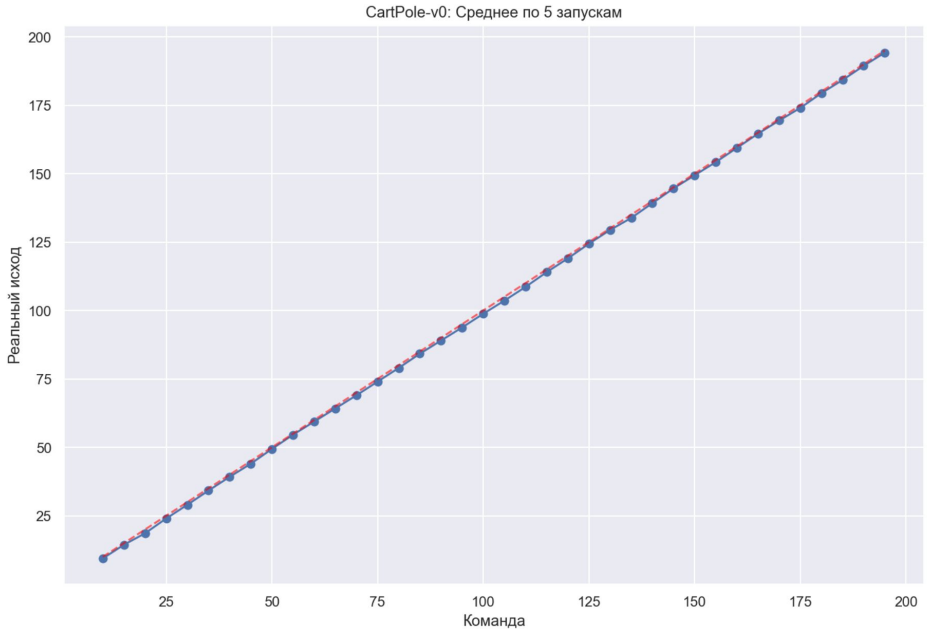
Эксперименты: SumGame



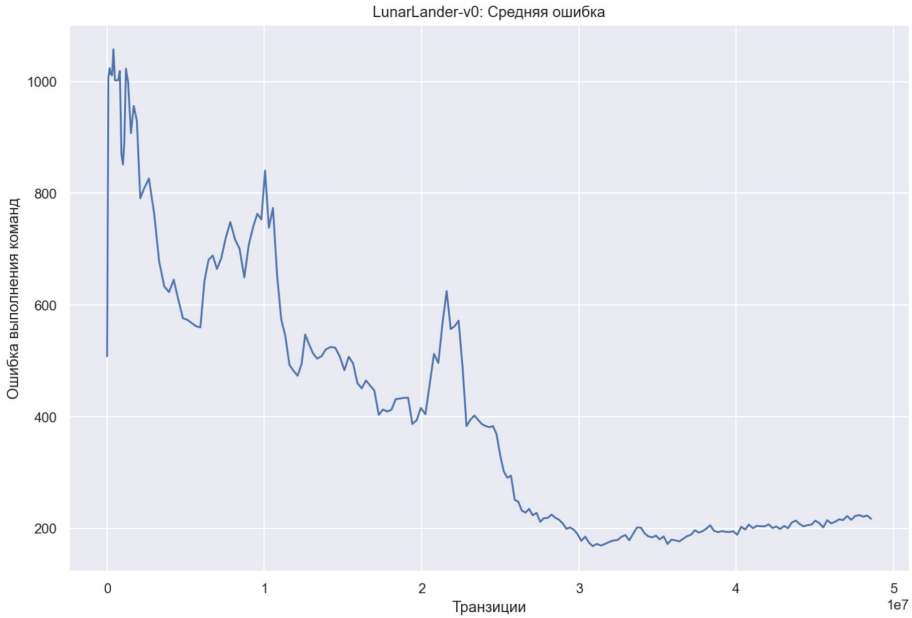


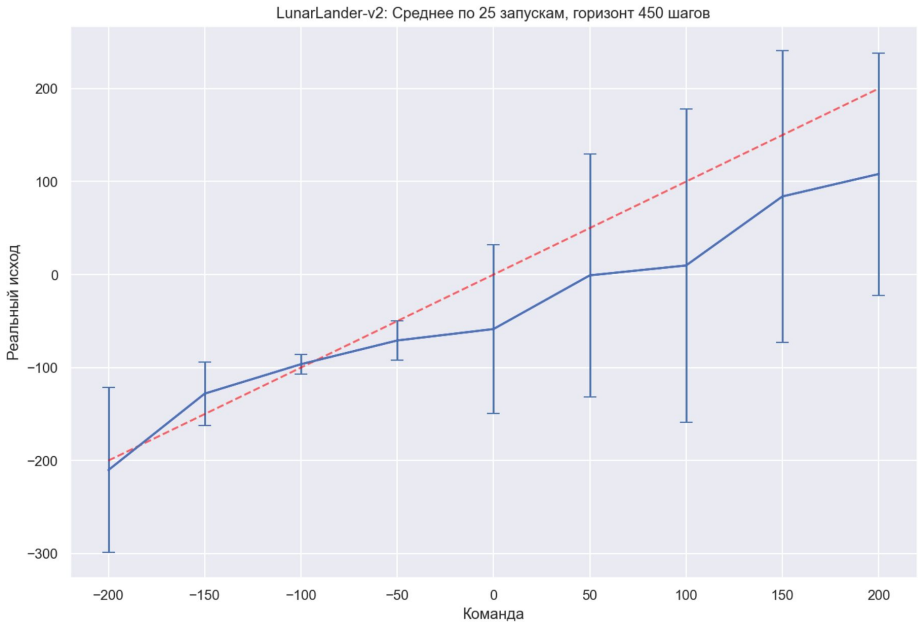
Эксперименты: CartPole





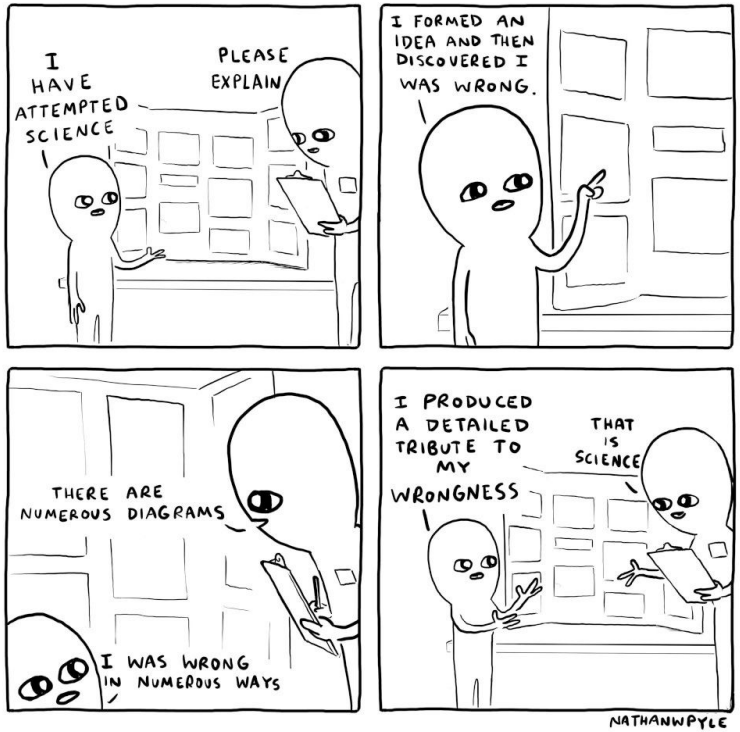
Эксперименты: LunarLander





Результаты

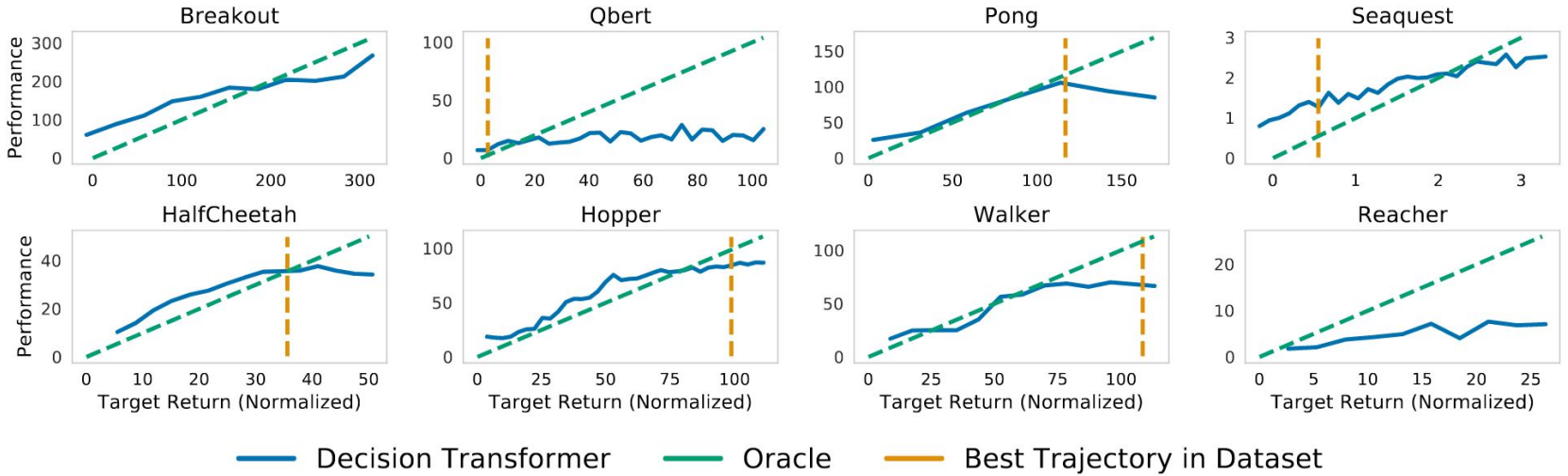
- Реализован алгоритм, способный обучать агента UDRL на основе SAC
- Реализована отладочная среда SumGame
- Алгоритм протестирован на отладочной и на двух реальных средах
- На основе логов были изучены свойства итогового алгоритма
- **Итоговый алгоритм имеет ряд недостатков**:
  - чувствительность к параметру энтропии
  - резкие скачки расстояния Кульбака — Лейблера в акторе
  - нестабильность градиентов критика



Обзор схожих методов

RL как предсказание последовательностей

Основная идея: награда в траектории как сумма до конца эпизода



Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision Transformer: Reinforcement Learning via Sequence Modeling. ArXiv:2106.01345 [Cs].

http://arxiv.org/abs/2106.01345

Представление и обучение критика

Для представления критика используется архитектура **MixtureDensityNetwork**[1], с помощью которой возможно выучивать условные распределения как смеси:

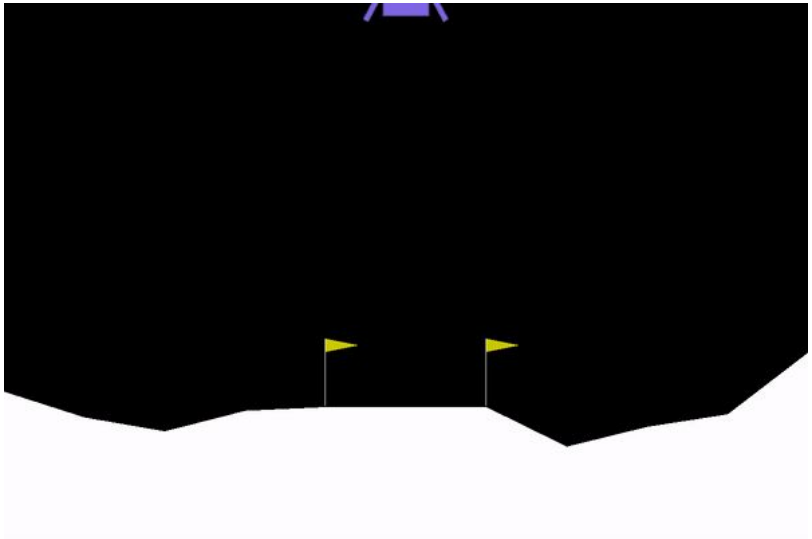
$$p(y|x) = \sum_k \alpha_k(x)\mathcal{N}(y|\mu_k(x), \sigma_k^2(x))$$

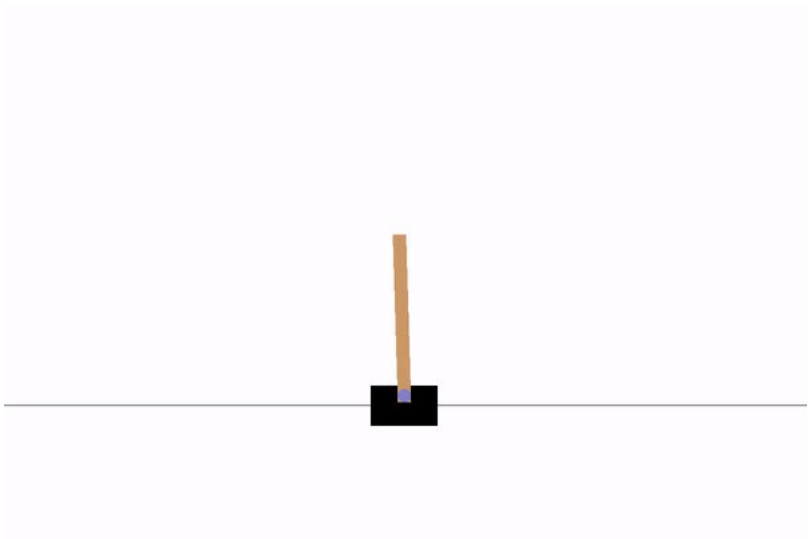
Обучение модели происходит с помощью отрицательной функции правдоподобия:

$$L(\theta) = -\frac{1}{N}\log p_\theta(\boldsymbol{Y}|\boldsymbol{X})$$

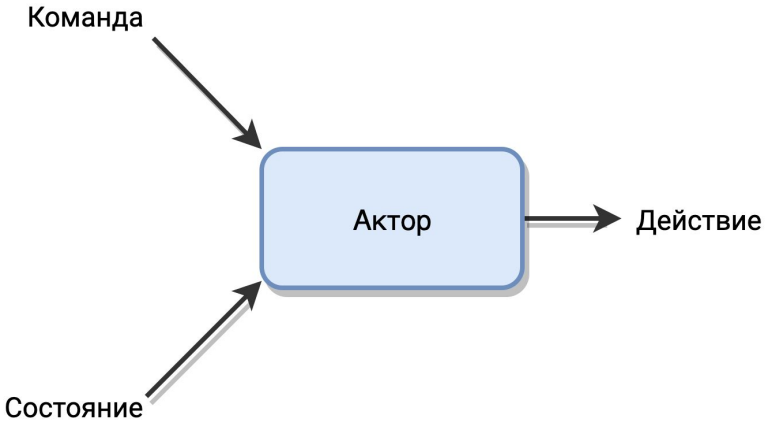
1. Bishop, Christopher M. (1994). Mixture density networks. Technical Report. Aston University, Birmingham. (Unpublished)

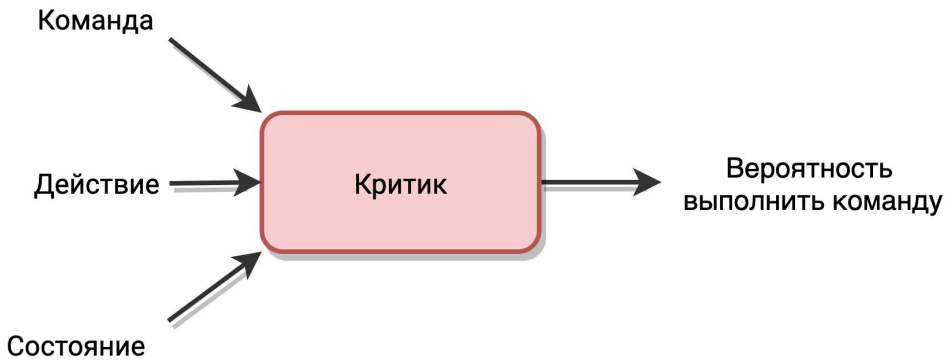
Окружения: LunarLander & CartPole





Актор и критик визуально





Визуализация траекторий