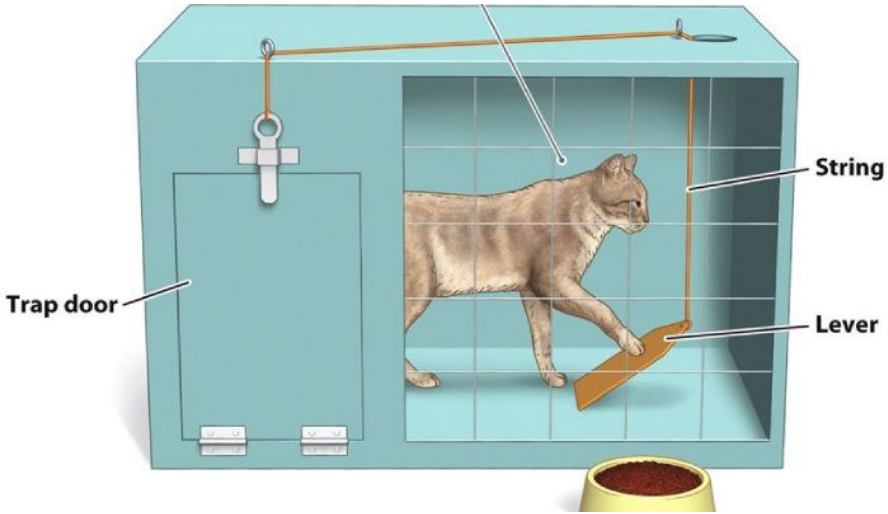
Faculty of Computer Science

Reinforcement Learning 1

Обучение с подкреплением 1

Karim Aitkhodjaev, 201



План

- Постановка задачи
- Описание величин и формул
- Методы и их классификация
- Примеры работы на openai gym

Постановка задачи

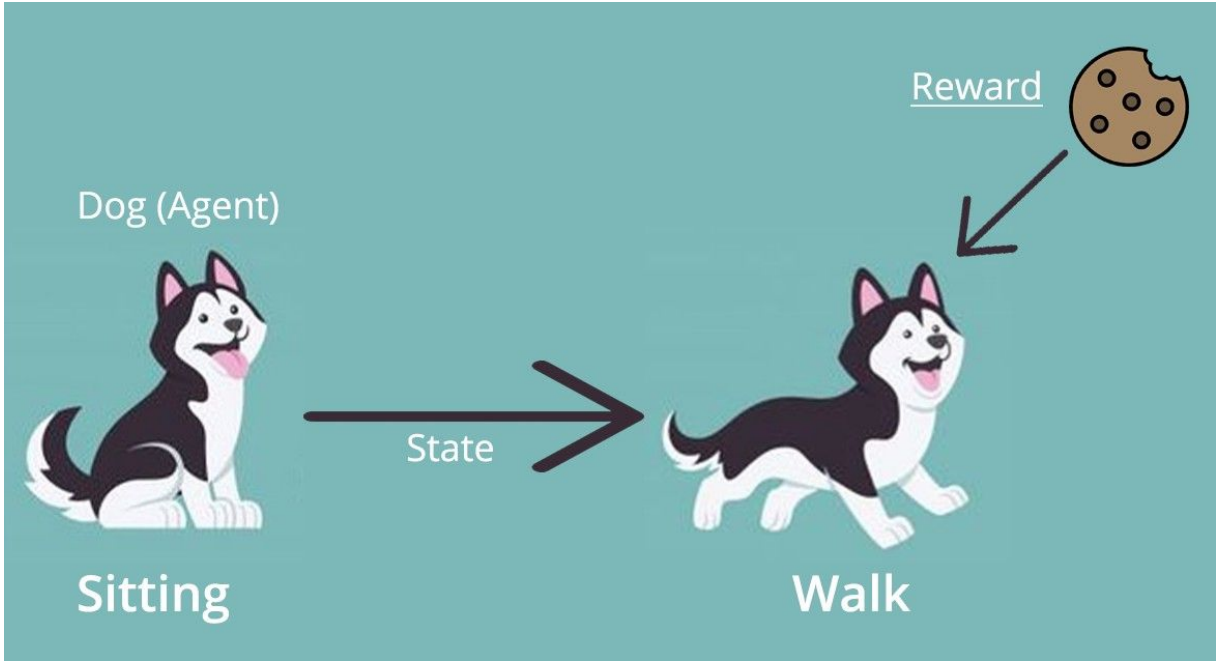
Что такое RL?

Чем он отличается от других типов задач МО?

Где применяется?

Постановка задачи

Пример

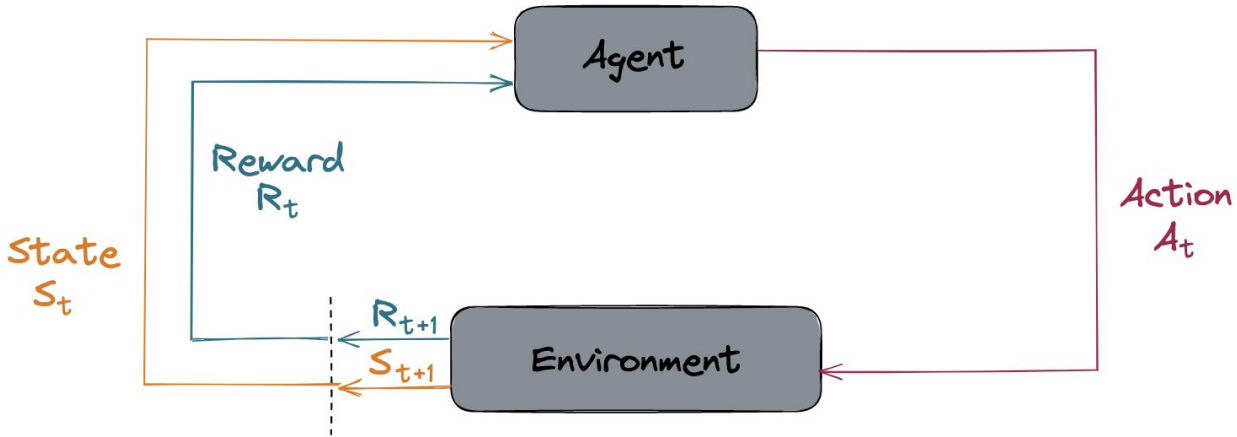


Постановка задачи

Марковский процесс принятия решений(Markov Decision Process)

Понятия и величины

- Среда
- Агент
- Действие
- Состояние
- Награда



Описание величин и формул

- Предположим, что у нас конечное число состояний, действий и наград
- На каждом промежутке времени ***t=0, 1, 2…*** агент получает представление состояния ***S_t***, относительно которого принимает действие ***A_t***
- Получаем пару (***S_t, A_t***)
- Увеличиваем ***t -> t+1*** , по паре (***S_t, A_t***) получаем ***R_{t+1}***
- Здесь уже знакомимся с функцией награды ***f:***(***S_t, A_t***) -> ***R_{t+1}***
- Весь процесс выглядит так

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \dots$$

Описание величин и формул

Что вообще хотим получить?

- Найти оптимальную политику \pi по которой получим максимальную суммарную награду

Как и что оптимизируем?

- Суммарную награду

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T,$$

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \\ &= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}. \end{aligned}$$

Описание величин и формул

Вероятности перехода(Transition probabilities)

$$p\left(s', r \mid s, a\right) = \Pr\left\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\right\}.$$

Хотим узнать с какой вероятностью агент перейдет в то или иное состояние

$$\sum_{r \in \mathcal{R}} p(s', r | s, a).$$

Описание величин и формул

v и q функции

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\}$$

- v функция, aka state-value function
- Нужна чтобы оценить как хорошо находиться в любом данном состоянии ***s*** для агента выполняющего политику \pi

Описание величин и формул

v и q функции

$$Q^{\pi}(s,a) = E_{\pi}\{R_t | s_t = s, a_t = a\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\}$$

- q функция, aka action-value function
- Нужна чтобы оценить как хорошо находиться в любом данном состоянии ***s*** и при любом данном действии ***a*** для агента выполняющего политику \pi

Методы

Cross entropy method

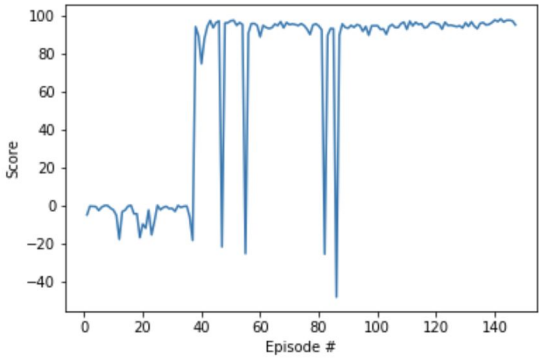
1. Create a Gaussian distribution $N(\mu,\sigma)$ that describes the weights θ of the neural network.
2. Sample ***N*** batch samples of θ from the Gaussian.
3. Evaluate all ***N*** samples of θ using the value function, e.g. running trials.
4. Select the top % of the samples of θ and compute the new μ and σ to parameterize the new Gaussian distribution.
5. Repeat steps 1-4 until convergence.

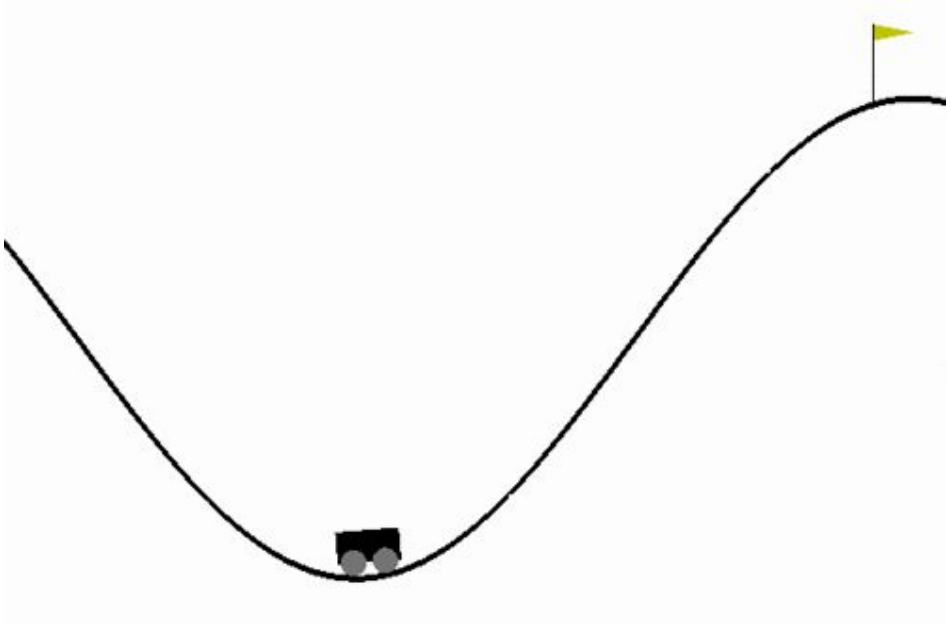
Методы

Cross entropy method

```
Episode 10      Average Score: -1.44
Episode 20      Average Score: -3.98
Episode 30      Average Score: -4.18
Episode 40      Average Score: 2.57
Episode 50      Average Score: 18.74
Episode 60      Average Score: 29.35
Episode 70      Average Score: 38.69
Episode 80      Average Score: 45.65
Episode 90      Average Score: 47.98
Episode 100     Average Score: 52.56
Episode 110     Average Score: 62.09
Episode 120     Average Score: 72.28
Episode 130     Average Score: 82.21
Episode 140     Average Score: 89.48

Environment solved in 47 iterations!    Average Score: 90.83
```





Методы

Cross entropy method

+

- Не требует градиентов
- Для разных seed’ов дает почти одинаковые результаты
- Хорошо параллелизуем

–

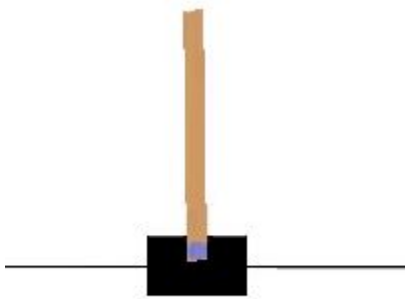
- В стохастических средах может быть нестабилен, т.е. не оптимален
- Работает хорошо только для коротких по сценарию задач
- Слабый Exploration, т.е. смотрит на траекторию эпизодов а не отдельных действий

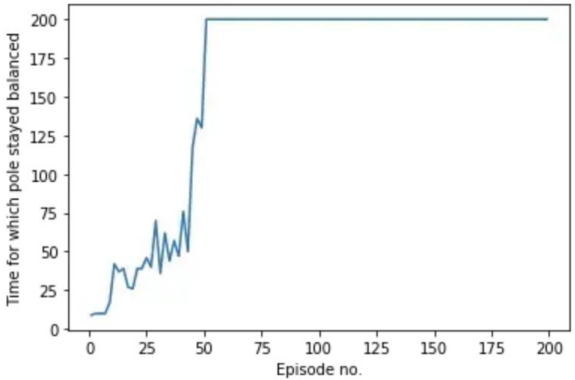
Методы

Табличные и Проксимальные методы

- В проксимальном методе используется параметризованная нелинейная функция
- Она нужна чтобы вычислить функцию значений (value function)
- Табличный метод используется когда число состояний, действий и наград конечно
- Чтобы вычислить эту нелинейную функцию можно использовать ANN
- Чтобы вычислить value function в табличном случае можно использовать динамическое программирование

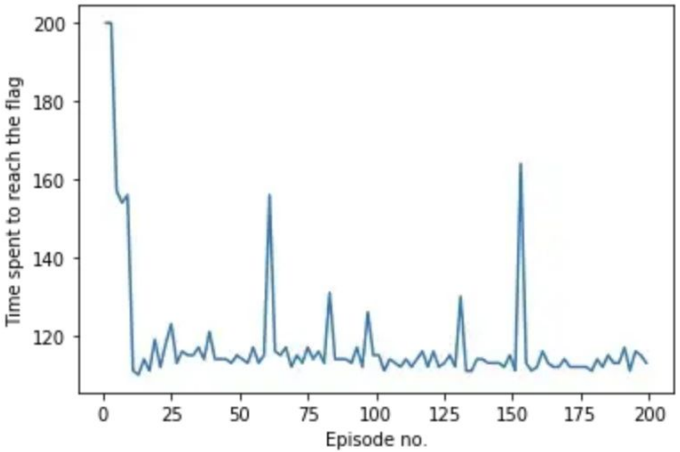
Примеры





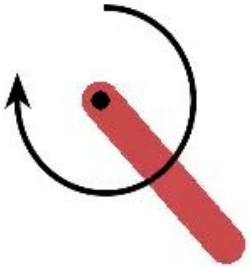
Хотим удерживать столб в вертикальном положении, за это агент награждается +1

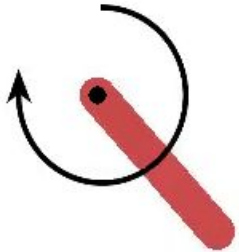
Примеры

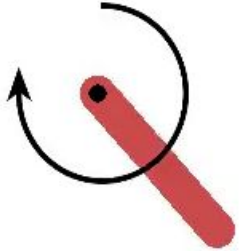


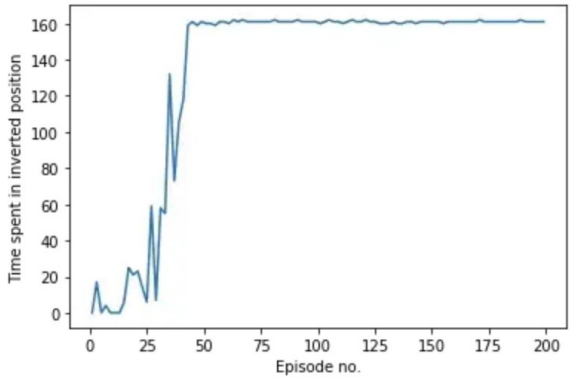
Хотим дойти до флажка, здесь уже требуется своя функция награды, в данном случае подойдет при увеличении механической энергии

Примеры









Хотим сбалансировать пендулум используя момент по или против часовой стрелки, здесь уже нужно чтобы механическая энергия увеличивалась до потенциальной обратного момента

Источники

Материалы

1. http://incompleteideas.net/book/RLbook2020.pdf
2. https://www.cs.upc.edu/~mmartin/Ag4-4x.pdf
3. https://en.wikipedia.org/wiki/Reinforcement_learning

Картинки и формулы

1. https://deeplizard.com/learn/video/a-SnJtmBtyA
2. https://towardsdatascience.com/open-ai-gym-classic-control-problems-rl-dqn-reward-functions-16a1bc2b007
3. https://jetnew.io/blog/2021/cem/

Код

1. https://github.com/yandexdataschool/Practical_RL/blob/master/week01_intro
2. https://github.com/udacity/deep-reinforcement-learning/blob/master/cross-entropy

Faculty of Computer science