Лекция 1: Введение в обучение с подкреплением. Метод Cross-Entropy.

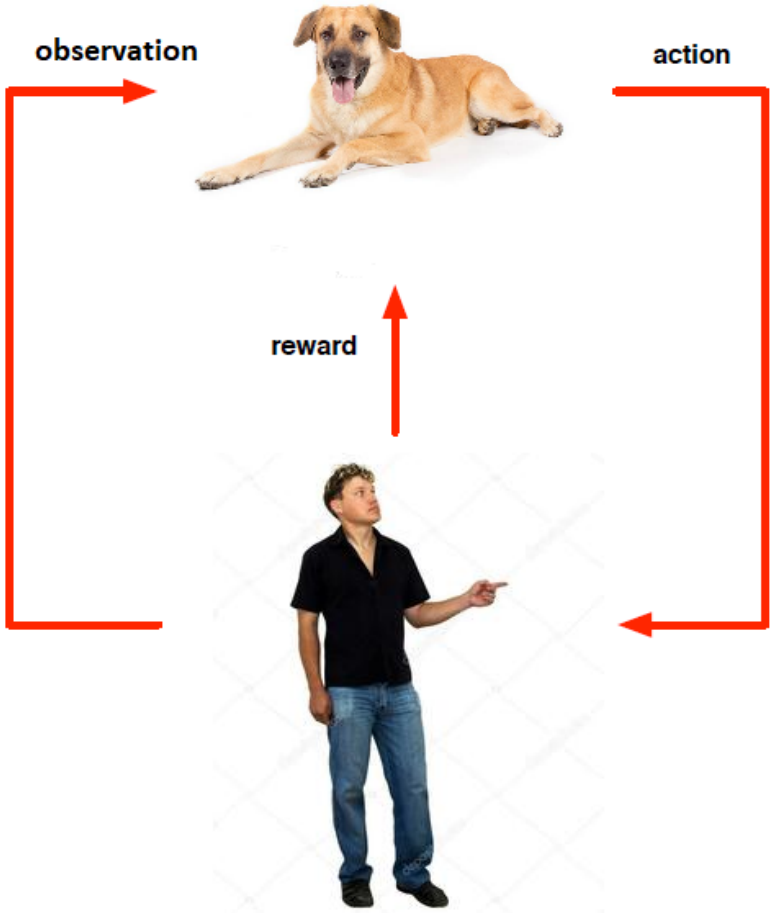
Антон Романович Плаксин

Организационные вопросы

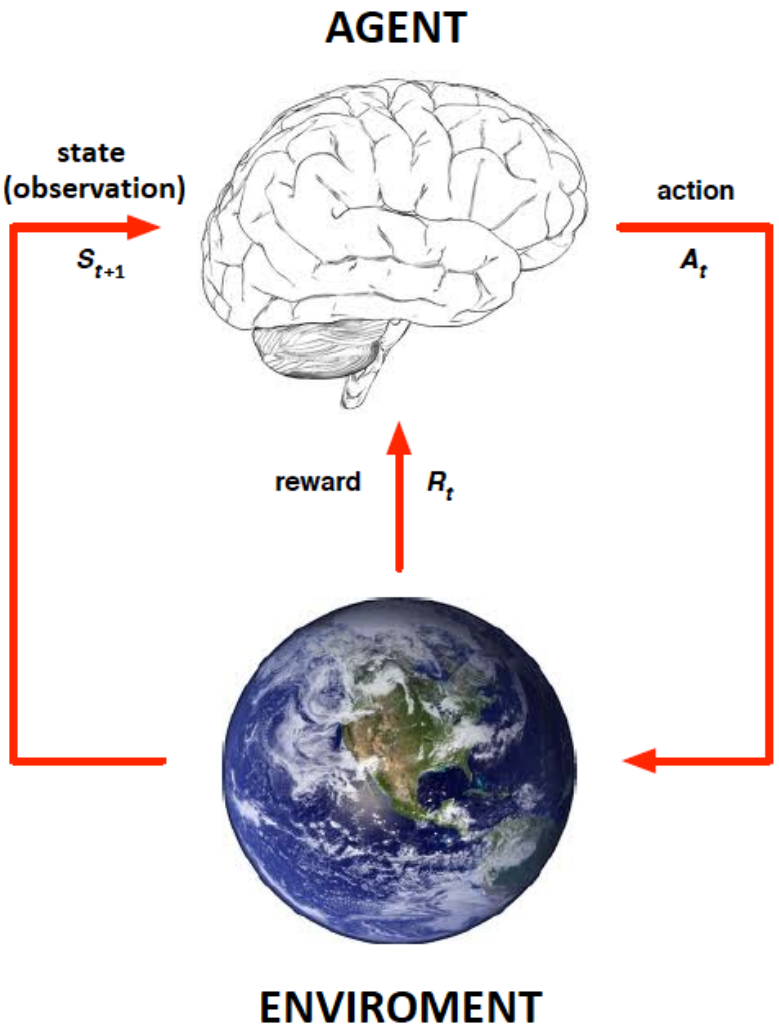
- Четверг, 16:10, онлайн
- Лекции и практики
- Отчетность: домашние работы
- Слайды: https://github.com/imm-rl-lab/UrFU_course
- E-mail для связи: a.r.plaksin@gmail.com
- Вопросу по ходу можно и нужно!

Что такое Reinforcement Learning?

Что такое Reinforcement Learning?

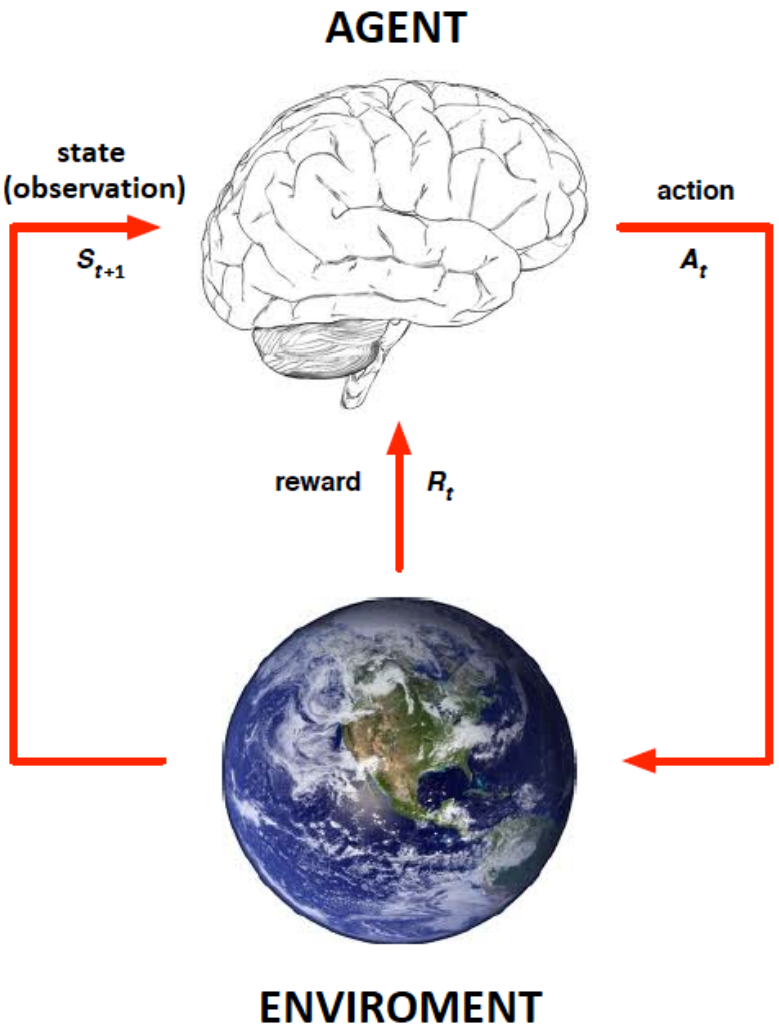


Что такое Reinforcement Learning?



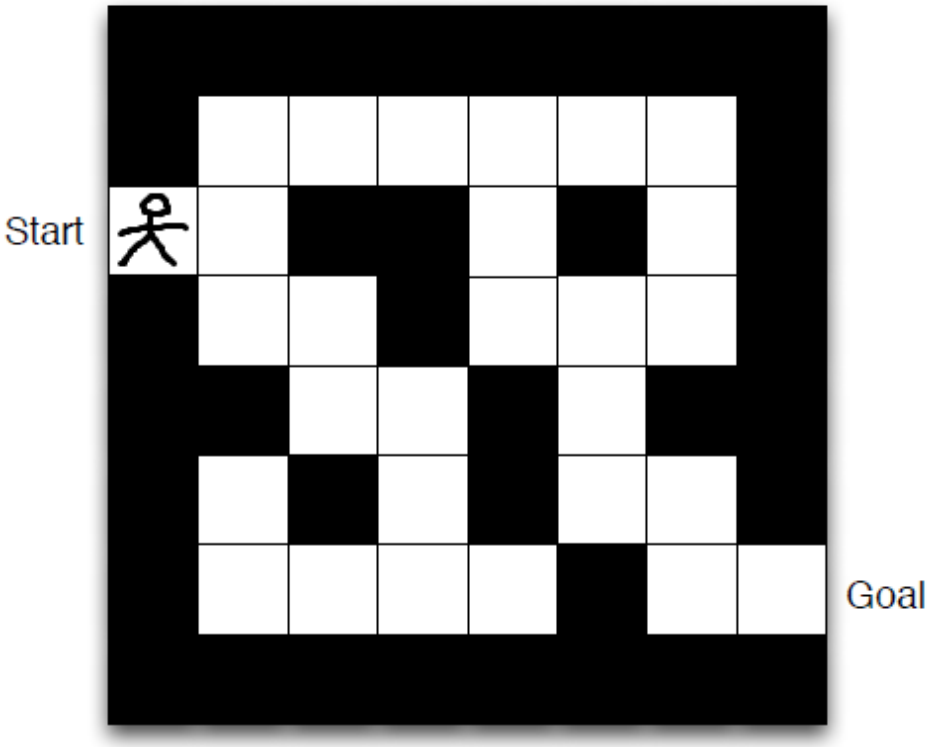
Цель агента - ???

Что такое Reinforcement Learning?

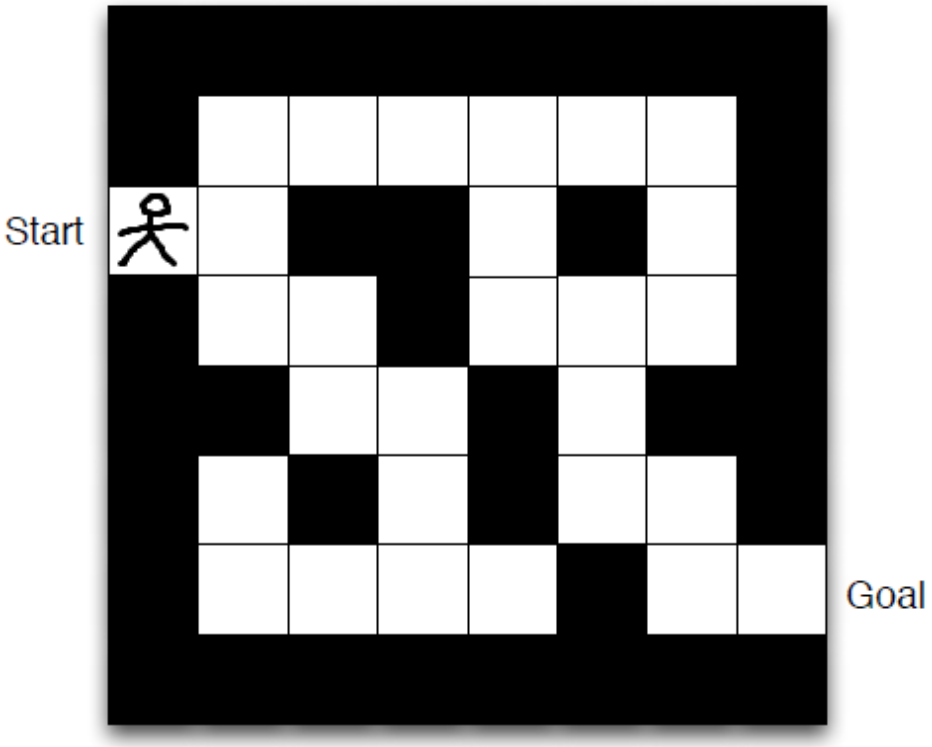


Цель агента - максимизировать $G = \sum_{t=0}^{\infty} \gamma^t R_t, \quad \gamma \in [0, 1].$

Пример: Maze

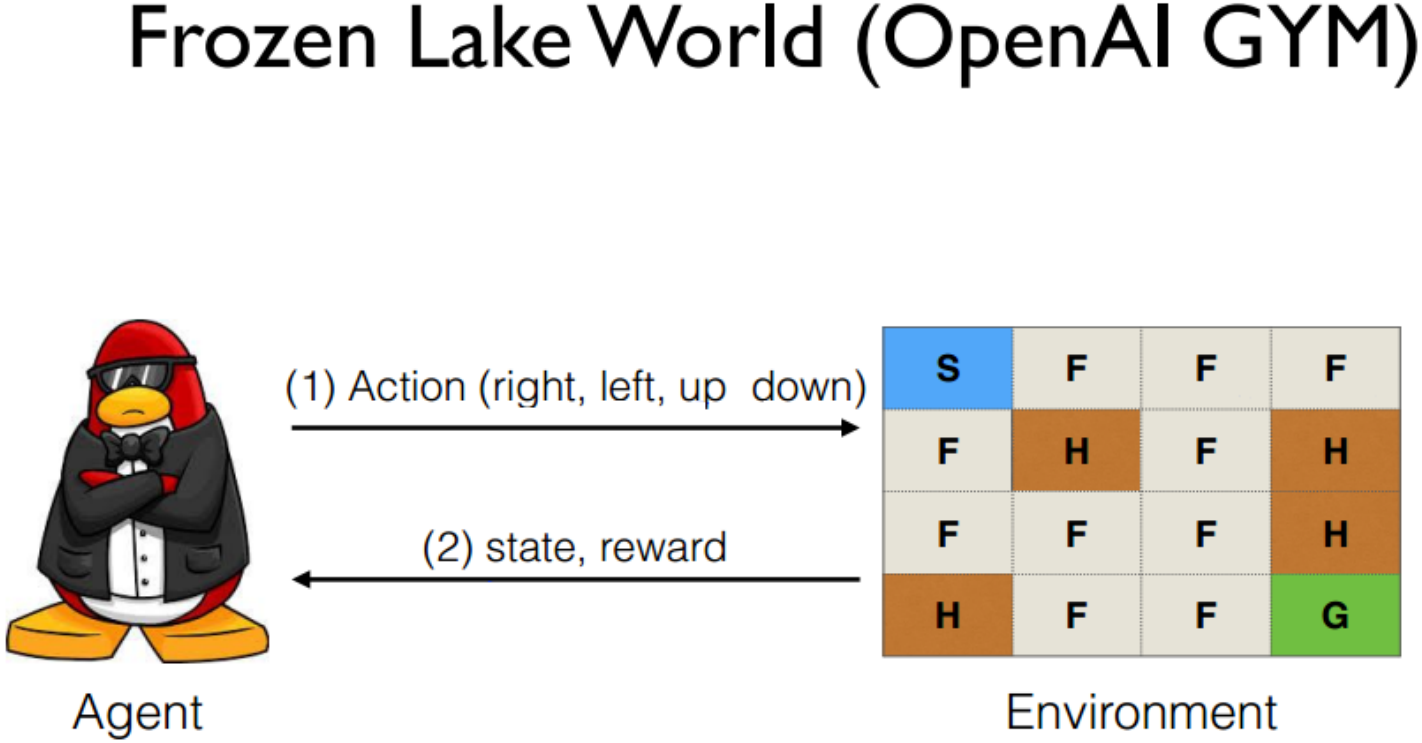


Пример: Maze



- Состояния: белые клетки
- Действия: $\uparrow$, $\rightarrow$, $\downarrow$, $\leftarrow$
- Награда: -1 на каждом шаге,
  0 если стоять в Goal

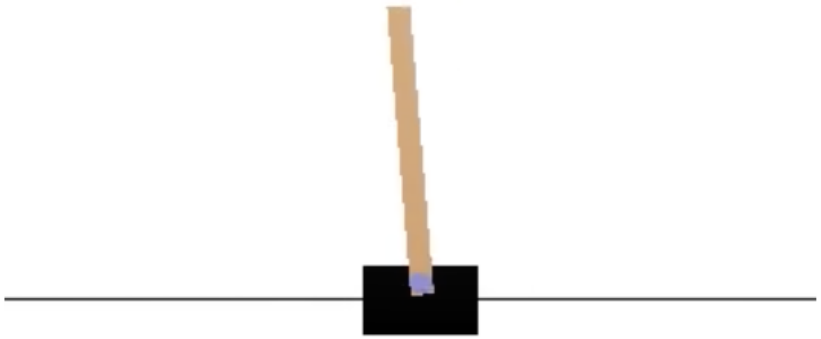
Пример: Frozen Lake



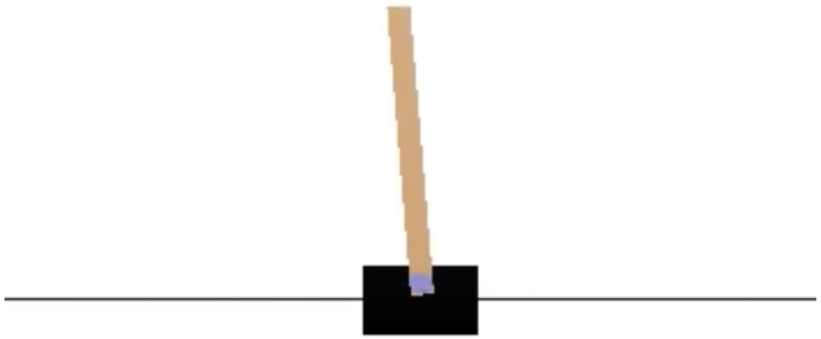
Пример: Atari Games

- Состояния: пиксели с экрана
- Действия: $\rightarrow$, $\leftarrow$, «0»
- Награда: очки в игре

Пример: Cartpole

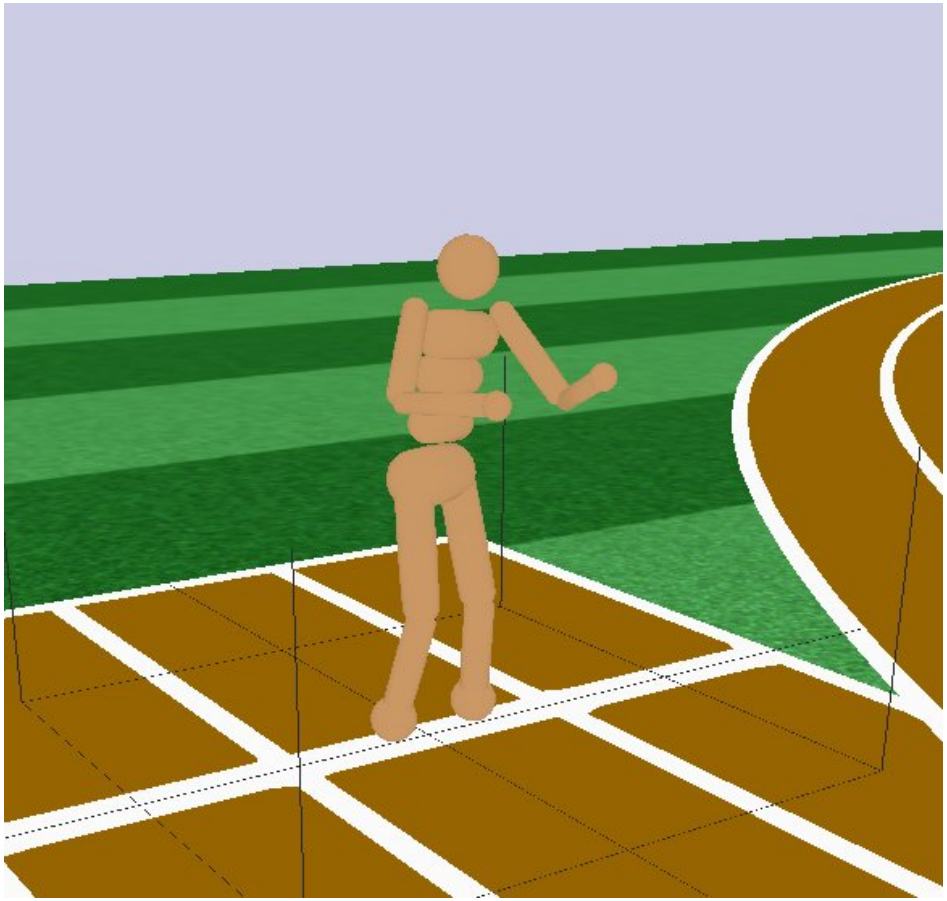


Пример: Cartpole



- Состояния: $\mathbb{R}^4$
  или пиксели с экрана
- Действия: $\rightarrow$, $\leftarrow$, «0»
- Награда: $+1$ на каждом шаге, $-\infty$ если маятник падает

Пример



- Состояния: $\mathbb{R}^{26}$
- Действия: $\mathbb{R}^{6}$
- Награда: $+1$ в каждый момент времени

Markov Decision Process

Markov Property

$$\mathbb{P}[S_{t+1}|S_t, A_t] = \mathbb{P}[S_{t+1}|S_1, A_1, S_2, A_2 \ldots, S_t, A_t]$$

$$\mathbb{P}[R_t|S_t, A_t] = \mathbb{P}[R_t|S_1, A_1, S_2, A_2 \ldots, S_t, A_t] = 1$$

Markov Decision Process

Markov Property

$$\mathbb{P}[S_{t+1}|S_t, A_t] = \mathbb{P}[S_{t+1}|S_1, A_1, S_2, A_2 \ldots, S_t, A_t]$$

$$\mathbb{P}[R_t|S_t, A_t] = \mathbb{P}[R_t|S_1, A_1, S_2, A_2 \ldots, S_t, A_t] = 1$$

Markov Decision Process $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

- $\mathcal{S}$ — пространство состояний
- $\mathcal{A}$ — пространство действий
- $\mathcal{P}$ — функция (матрица) вероятностей переходов между состояниями

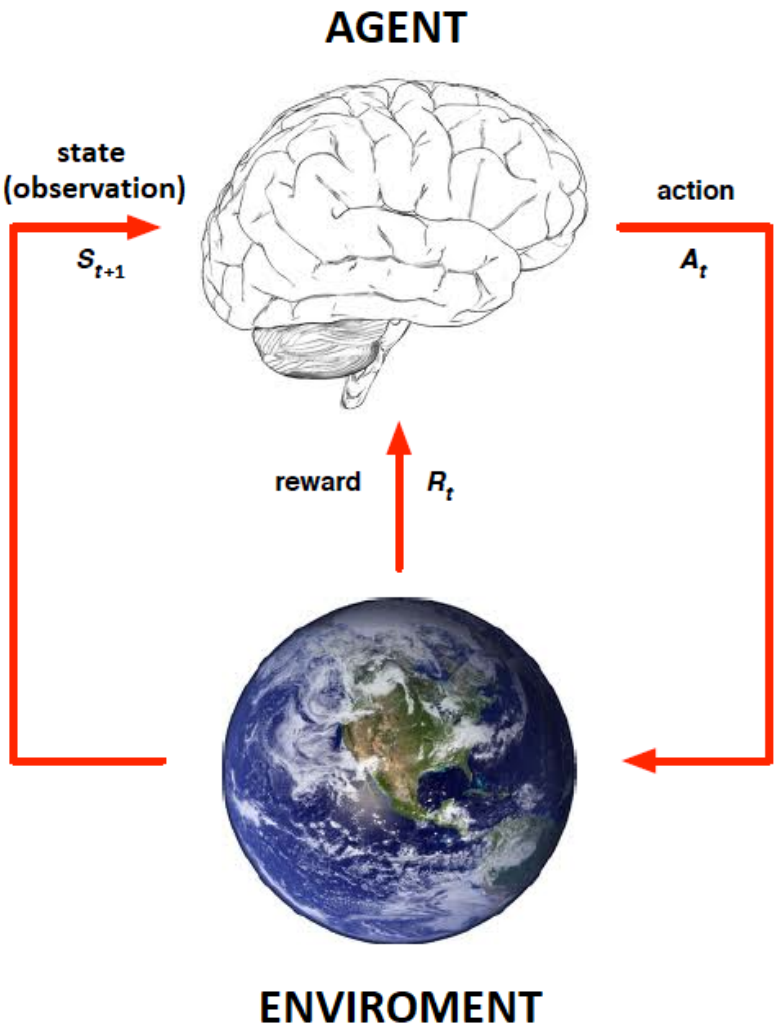
$$\mathcal{P}(s'|s, a) = \mathbb{P}[S_{t+1} = s'|S_t = s, A_t = a]$$

- $\mathcal{R}$ — функция (вектор) вознаграждений

$$\mathcal{R}(s, a) = R_t \quad \Leftrightarrow \quad \mathbb{P}[R_t|S_t = s, A_t = a] = 1$$

- $\gamma \in [0, 1]$ — коэффициент дисконтирования

Наша задача. Что мы хотим?



Policy

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

Policy

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

- Мы задаем π

Policy

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

- Мы задаем π
- АГЕНТ находится в начальном состоянии $S_0 \in \mathcal{S}$

Policy

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

- Мы задаем π
- АГЕНТ находится в начальном состоянии $S_0 \in \mathcal{S}$
- совершает действие $A_0 = \pi(S_0)$

Policy

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

- Мы задаем π
- АГЕНТ находится в начальном состоянии $S_0 \in \mathcal{S}$
- совершает действие $A_0 = \pi(S_0)$
- получает награду $R_0 = \mathcal{R}(S_0, A_0)$ и переходит в следующее состояние $S_1 \sim \mathcal{P}(\cdot|S_0, A_0)$

Policy

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

- Мы задаем π
- АГЕНТ находится в начальном состоянии $S_0 \in \mathcal{S}$
- совершает действие $A_0 = \pi(S_0)$
- получает награду $R_0 = \mathcal{R}(S_0, A_0)$ и переходит в следующее состояние $S_1 \sim \mathcal{P}(\cdot|S_0, A_0)$
- совершает действие $A_1 = \pi(S_1)$

Policy

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

- Мы задаем π
- АГЕНТ находится в начальном состоянии $S_0 \in \mathcal{S}$
- совершает действие $A_0 = \pi(S_0)$
- получает награду $R_0 = \mathcal{R}(S_0, A_0)$ и переходит в следующее состояние $S_1 \sim \mathcal{P}(\cdot|S_0, A_0)$
- совершает действие $A_1 = \pi(S_1)$
- получает награду $R_1 = \mathcal{R}(S_1, A_1)$ и переходит в следующее состояние $S_2 \sim \mathcal{P}(\cdot|S_1, A_1)$

Policy

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

- Мы задаем π
- АГЕНТ находится в начальном состоянии $S_0 \in \mathcal{S}$
- совершает действие $A_0 = \pi(S_0)$
- получает награду $R_0 = \mathcal{R}(S_0, A_0)$ и переходит в следующее состояние $S_1 \sim \mathcal{P}(\cdot|S_0, A_0)$
- совершает действие $A_1 = \pi(S_1)$
- получает награду $R_1 = \mathcal{R}(S_1, A_1)$ и переходит в следующее состояние $S_2 \sim \mathcal{P}(\cdot|S_1, A_1)$
- $\ldots$
- $\tau = \{S_0, A_0, S_1, A_1, S_2, A_2, \ldots\}, \quad G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t)$

Policy

$$\pi : \mathcal{S} \mapsto \mathcal{A}$$

- Мы задаем π
- АГЕНТ находится в начальном состоянии $S_0 \in \mathcal{S}$
- совершает действие $A_0 = \pi(S_0)$
- получает награду $R_0 = \mathcal{R}(S_0, A_0)$ и переходит в следующее состояние $S_1 \sim \mathcal{P}(\cdot|S_0, A_0)$
- совершает действие $A_1 = \pi(S_1)$
- получает награду $R_1 = \mathcal{R}(S_1, A_1)$ и переходит в следующее состояние $S_2 \sim \mathcal{P}(\cdot|S_1, A_1)$
- $\ldots$
- $\tau = \{S_0, A_0, S_1, A_1, S_2, A_2, \ldots\}, \quad G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t)$

Наша задача

$$\mathbb{E}_\pi[G] \quad \longrightarrow \quad \max_\pi$$

Stochastic policy

$$\pi(a|s) \in [0, 1], \quad a \in \mathcal{A}, \quad s \in \mathcal{S}$$

- Мы задаем π
- АГЕНТ находится в начальном состоянии $S_0 \in \mathcal{S}$
- совершает действие $A_0 \sim \pi(\cdot|S_0)$
- получает награду $R_0 = \mathcal{R}(S_0, A_0)$ и переходит в следующее состояние $S_1 \sim \mathcal{P}(\cdot|S_0, A_0)$
- совершает действие $A_1 \sim \pi(\cdot|S_1)$
- получает награду $R_1 = \mathcal{R}(S_1, A_1)$ и переходит в следующее состояние $S_2 \sim \mathcal{P}(\cdot|S_1, A_1)$
- $\ldots$
- $\tau = \{S_0, A_0, S_1, A_1, S_2, A_2, \ldots\}, \quad G(\tau) = \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(S_t, A_t)$

Наша задача

$$\mathbb{E}_\pi[G] \quad \longrightarrow \quad \max_\pi$$

Что такое $\mathbb{E}_\pi[G]$

$$\tau = \{S_0, A_0, S_1, A_1, S_2, A_2, \ldots\}$$

Что такое $\mathbb{E}_\pi[G]$

$$\tau = \{S_0, A_0, S_1, A_1, S_2, A_2, \ldots\}$$

$$\mathbb{P}(\tau) = ?$$

Что такое $\mathbb{E}_\pi[G]$

$$\tau = \{S_0, A_0, S_1, A_1, S_2, A_2, \ldots\}$$

$$\begin{aligned} \mathbb{P}(\tau) &= \mathbb{P}(A_0|S_0)\mathbb{P}(S_1|S_0, A_0) \\ &\times \mathbb{P}(A_1|S_1)\mathbb{P}(S_2|S_1, A_1) \\ &\times \ldots \end{aligned}$$

Что такое $\mathbb{E}_\pi[G]$

$$\tau = \{S_0, A_0, S_1, A_1, S_2, A_2, \ldots\}$$

$$\begin{aligned} \mathbb{P}(\tau) &= \mathbb{P}(A_0|S_0)\mathbb{P}(S_1|S_0, A_0) \\ &\times \mathbb{P}(A_1|S_1)\mathbb{P}(S_2|S_1, A_1) \\ &\times \ldots \end{aligned}$$

$$\mathbb{P}(\tau|\pi) := \prod_{t=0}^{\infty} \pi(A_t|S_t)\mathcal{P}(S_{t+1}|S_t, A_t)$$

Что такое $\mathbb{E}_\pi[G]$

$$\tau = \{S_0, A_0, S_1, A_1, S_2, A_2, \ldots\}$$

$$\begin{aligned}\mathbb{P}(\tau) &= \mathbb{P}(A_0|S_0)\mathbb{P}(S_1|S_0, A_0) \\ &\times \mathbb{P}(A_1|S_1)\mathbb{P}(S_2|S_1, A_1) \\ &\times \ldots\end{aligned}$$

$$\mathbb{P}(\tau|\pi) := \prod_{t=0}^{\infty} \pi(A_t|S_t)\mathcal{P}(S_{t+1}|S_t, A_t)$$

$$\mathbb{E}_\pi[G] = \int_\tau G(\tau)\mathbb{P}(\mathrm{d}\tau|\pi)$$

Общая схема Cross-Entropy Method

На каждой итерации:

- Policy evaluation
- Policy improvement

Policy evaluation

Policy evaluation

Закон больших чисел

Если X_k, $k \in \overline{1, K}$ — независимые случайные величины с одним распределением. Тогда

$$\frac{1}{K} \sum_{k=1}^{K} X_k \to E[X] \text{ по вероятности}$$

Policy evaluation

Закон больших чисел

Если X_k, $k \in \overline{1, K}$ — независимые случайные величины с одним распределением. Тогда

$$\frac{1}{K} \sum_{k=1}^{K} X_k \to E[X] \text{ по вероятности}$$

Подход Монте-Карло

$$E_\pi[G] \approx \frac{1}{K} \sum_{k=1}^{K} G(\tau_k)$$

Квантиль (Перцентиль)

Пусть $q \in [0, 1]$. q-квантилем чисел $G_1, G_2, \ldots, G_K$ называем, такое число γ_q, что

$$\frac{|\{G_k, k \in \overline{1, K} : G_k \leq \gamma_q\}|}{|\{G_k, k \in \overline{1, K}\}|} \geq q$$

$$\frac{|\{G_k, k \in \overline{1, K} : G_k > \gamma_q\}|}{|\{G_k, k \in \overline{1, K}\}|} \leq 1 - q$$

Пусть $p \in [0, 100]$. p-перцентиль — это $(p/100)$-квантиль

Cross-Entropy Method

Пусть π_0 — начальная (равномерная) policy, N — количество итераций алгоритма, $q \in (0, 1)$ — параметр для определения элитных траекторий. Для каждого $n \in \overline{0, N}$ делаем

Cross-Entropy Method

Пусть π_0 — начальная (равномерная) policy, N — количество итераций алгоритма, $q \in (0, 1)$ — параметр для определения элитных траекторий. Для каждого $n \in \overline{0, N}$ делаем

- (Policy evaluation) Действуя в согласии с текущей policy π_n реализуем K сессий, получаем траектории τ_k, $k \in \overline{1, K}$ и награды $G(\tau_k)$ для каждой из них. Оцениваем policy π_n:

$$\mathbb{E}_{\pi_n}[G] \approx V_{\pi_n} := \frac{1}{K} \sum_{k=1}^{K} G(\tau_k)$$

  Если $V_{\pi_n} \approx V_{\pi_{n-1}}$, то break с ответом π_n

Cross-Entropy Method

Пусть π_0 — начальная (равномерная) policy, N — количество итераций алгоритма, $q \in (0, 1)$ — параметр для определения элитных траекторий. Для каждого $n \in \overline{0, N}$ делаем

- (Policy evaluation) Действуя в согласии с текущей policy π_n реализуем K сессий, получаем траектории τ_k, $k \in \overline{1, K}$ и награды $G(\tau_k)$ для каждой из них. Оцениваем policy π_n:

$$\mathbb{E}_{\pi_n}[G] \approx V_{\pi_n} := \frac{1}{K} \sum_{k=1}^{K} G(\tau_k)$$

  Если $V_{\pi_n} \approx V_{\pi_{n-1}}$, то break с ответом π_n
- (Policy improvement) Выбираем элитные траектории $\mathcal{T}_n = \{\tau_k, k \in \overline{1, K} : G(\tau_k) > \gamma_q\}$ (γ_q — q-квантиль чисел $G(\tau_k)$, $k \in \overline{1, K}$). Если $\mathcal{T}_n \neq \emptyset$, то определяем следующую Policy

$$\pi_{n+1}(a|s) = \frac{\text{количество пар } (a|s) \text{ в траекториях из } \mathcal{T}_n}{\text{количество } s \text{ в траекториях из } \mathcal{T}_n}$$

В чем проблема алгоритма?

В чем проблема алгоритма?

- Необходима большое количество сессий
- Выбор policy сильно зависит от случайности
- Проблемы со стохастической средой
- Работает только с конечными $\mathcal{S}$ и $\mathcal{A}$

Проблема: выбор policy сильно зависит от случайности

РЕШЕНИЕ:

- Сглаживание по Лапласу

$$\pi_{n+1}(a|s) = \frac{|(a|s) \in \mathcal{T}_n| + \lambda}{|s \in \mathcal{T}_n| + \lambda|\mathcal{A}|}, \quad \lambda > 0$$

- Сглаживание по policy

$$\pi_{n+1}(a|s) \leftarrow \lambda\pi_{n+1}(a|s) + (1-\lambda)\pi_n(a|s), \quad \lambda \in (0, 1]$$

Проблемы со стохастической средой

РЕШЕНИЕ:

По стохастической policy π_n насэмплировать детерминированные policy $\pi_{n,m}$, $m \in \overline{1, M}$. В согласии с каждой из них реализовать K сессий и получить траектории $\tau_{m,k}$, $m \in \overline{1, M}$, $k \in \overline{1, K}$.
Определить величины

$$V_{\pi_{n,m}} = \frac{1}{K} \sum_{k=1}^{K} G(\tau_{m,k})$$

Выбираем элитные траектории
$\mathcal{T}_n = \{\tau_{m,k}, m \in \overline{1, M}, k \in \overline{1, K} : V_{\pi_{n,m}} > \gamma_q\}$ (γ_q — q-квантиль чисел $V_{\pi_{n,m}}$, $m \in \overline{1, M}$).

ВОПРОСЫ?