Нижние границы на сожаление в обучении с подкреплением

Сергей Володин

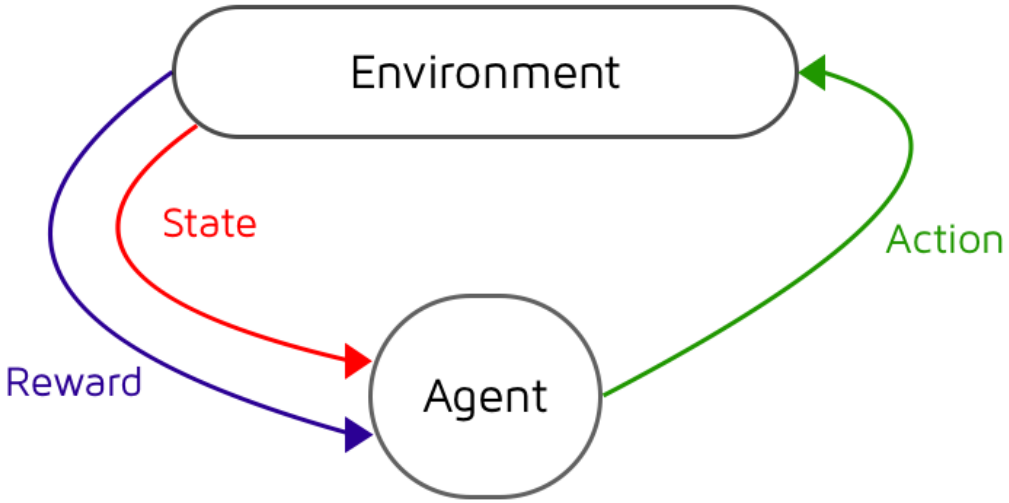
МФТИ

По статье

Ian Osband, Benjamin Van Roy. On Lower Bounds for Regret in Reinforcement Learning

Обучение с подкреплением

Агент взаимодействует со средой:



Определения

Definition

ММПР — кортеж $(\mathcal{S}, \mathcal{A}, R, P)$, где

1. $\mathcal{S} = \{1, ..., S\}$ — состояния
2. $\mathcal{A} = \{1, ..., A\}$ — действия
3. $R(s, a)$ — функция награды.
4. $P(s, a)$ — функция переходов.

Definition

$t \in \mathbb{N}$ — время.

1. Агент получает состояние $s_t \in \mathcal{S}$
2. Агент выбирает действие $a_t \in \mathcal{A}$
3. Агент получает награду $r_t \sim R(s_t, a_t) \in [0, 1]$
4. Среда переходит в новое состояние $s_{t+1} \sim P(s_t, a_t)$

Определения

Definition

Политика μ — отображение $\mu\colon \mathcal{S} \to \mathcal{A}$.

Definition

Средняя награда:

$$\lambda_\mu^M(s) = \lim_{T\to\infty} \mathbb{E}_{M,\mu}\left[\frac{1}{T}\sum_{t=1}^{T} \overline{r}(s_t, a_t)\Big| s_1 = s\right]$$

$$\lambda_*^M(s) = \lambda_{\mu^M}^M(s)$$

где $\overline{r}(s,a) = \mathbb{E}R(s,a)$ и $\mu^M \in \arg\max_\mu \lambda_\mu^M(s)$

Определения

Definition

$\mathcal{H}_t = (s_1, a_1, r_1, ..., s_{t-1}, a_{t-1}, r_{t-1})$ — история для t.

Definition

$\pi = \{\pi_t \big| t \in \mathbb{N}\}$ — алгоритм RL, если π_t — функция, сопоставляющая истории $\mathcal{H}_t$ распределение над политиками

Definition

Сожаление:

$$\text{Regret}(T, \pi, M, s) = T\lambda_*^M(s) - \sum_{t=1}^{T} r_t$$

Определения

Definition

Многорукий бандит M — ММПР с $S = 1$

Theorem

Для любого π существует функция награды R, такая что

$$\mathbb{E}Regret(T, \pi, M) \geqslant \frac{1}{24}\sqrt{AT}$$