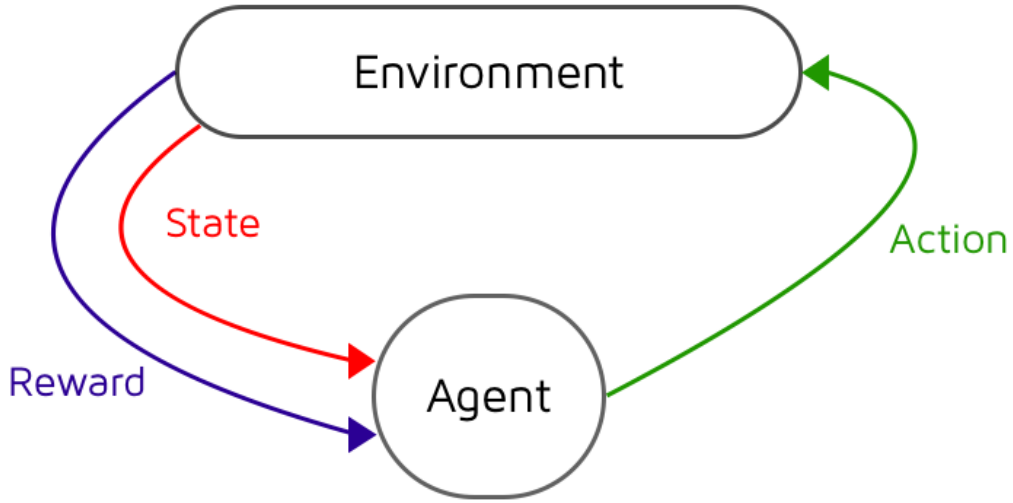
Нижние границы на regret в обучении с подкреплением

Сергей Володин

МФТИ

Обучение с подкреплением

Агент взаимодействует со средой:



Определения

Definition

(Марковский процесс принятия решений)
ММПР — это кортеж $(\mathcal{S}, \mathcal{A}, R, P)$, где

1. $\mathcal{S} = \{1, ..., S\}$ — множество состояний среды
2. $\mathcal{A} = \{1, ..., A\}$ — множество действий, доступных агенту
3. $R(s, a)$ — функция награды. Для данных $s \in \mathcal{S}$ и $a \in \mathcal{A}$ случайная величина $R(s, a) \in [0, 1]$ — награда за действие
4. $P(s, a)$ — функция переходов. Для данных $s \in \mathcal{S}$ и $a \in \mathcal{A}$ случайная величина $P(s, a) \in \mathcal{S}$ — новое состояние среды

Definition

(Взаимодействие агента со средой)
Имеется ММПР $(\mathcal{S}, \mathcal{A}, R, P)$. Вводится время $t \in \mathbb{N}$. Для каждого момента времени **??**:

1. Агент получает состояние $s_t \in \mathcal{S}$
2. [illegible] $\in \mathcal{A}$