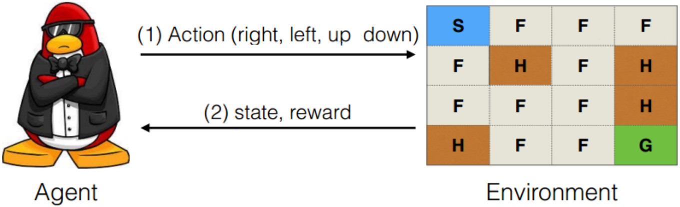
Лекция 4: Model-Free Reinforcement Learning

Антон Романович Плаксин

Пример: Frozen Lake

Frozen Lake World (OpenAI GYM)



Пример: Atari Games

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

Markov Decision Process

Markov Property

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

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

Markov Decision Process

Markov Property

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

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

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

- $\mathcal{S}$ — конечное ($|\mathcal{S}| = n$) пространство состояний
- $\mathcal{A}$ — конечное ($|\mathcal{A}| = m$) пространство действий
- $\mathcal{P}$ — неизвестная функция (тензор) вероятностей переходов между состояниями

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

- $\mathcal{R}$ — неизвестная функция (матрица) 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

Model-Free Algorithms

- Monte-Carlo Algorithm
- SARSA Algorithm
- Q-Learning Algorithm

Policy Iteration

Пусть инициализирована π^0 и заданы числа $L, K \in \mathbb{N}$.
Для каждого $k \in \overline{0, K}$ делаем

- (Policy evaluation) Iterative Policy Evaluation:

$$v^{l+1}(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v^l(s')\Big), \; l \in \overline{0, L-1}.$$

  Получаем $v^L \approx v_{\pi^k}$. По $v^L(s)$ построить $q^L(s,a) \approx q_{\pi^k}$.

- (Policy improvement) Greedy Policy Improvement:

$$\pi^{k+1}(a|s) = \begin{cases} 1, & \text{если } a \in \operatorname{argmax}_{a' \in \mathcal{A}} q^L(s,a') \\ 0, & \text{иначе} \end{cases}$$

Policy Iteration

Пусть инициализирована π^0 и заданы числа $L, K \in \mathbb{N}$.
Для каждого $k \in \overline{0, K}$ делаем

- (Policy evaluation) Iterative Policy Evaluation:

$$v^{l+1}(s) = \sum_a \pi(a|s)\Big(\mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v^l(s')\Big), \; l \in \overline{0, L-1}.$$

  Получаем $v^L \approx v_{\pi^k}$. По $v^L(s)$ построить $q^L(s,a) \approx q_{\pi^k}$.

- (Policy improvement) Greedy Policy Improvement:

$$\pi^{k+1}(a|s) = \begin{cases} 1, & \text{если } a \in \operatorname{argmax}_{a' \in \mathcal{A}} q^L(s,a') \\ 0, & \text{иначе} \end{cases}$$

$$q^L(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) v^L(s')$$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$
- получает награду R_0 и переходит в следующее состояние S_1

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- $\dots$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- $\ldots$
- совершает действие $A_{T-2} = \pi(S_{T-2})$,

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- $\ldots$
- совершает действие $A_{T-2} = \pi(S_{T-2})$,
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- ...
- совершает действие $A_{T-2} = \pi(S_{T-2})$,
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}
- совершает действие $A_{T-1} = \pi(S_{T-1})$,

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- $\ldots$
- совершает действие $A_{T-2} = \pi(S_{T-2})$,
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}
- совершает действие $A_{T-1} = \pi(S_{T-1})$,
- получает награду R_{T-1} и переходит в терминальное состояние S_T

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- $\ldots$
- совершает действие $A_{T-2} = \pi(S_{T-2})$,
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}
- совершает действие $A_{T-1} = \pi(S_{T-1})$,
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \ldots, S_T\}, \quad G(\tau) = \sum_{t=0}^{T-1} \gamma^t R_t,$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- $\ldots$
- совершает действие $A_{T-2} = \pi(S_{T-2})$,
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}
- совершает действие $A_{T-1} = \pi(S_{T-1})$,
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \ldots, S_T\}, \quad G(\tau) = \sum_{t=0}^{T-1} \gamma^t R_t,$

Задача

$$\text{Найти} \quad q_\pi(s,a) = \mathbb{E}_\pi[G \mid S_0 = s, A_0 = a]$$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$ $\quad q_\pi(S_0, A_0) =???$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- ...
- совершает действие $A_{T-2} = \pi(S_{T-2})$,
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}
- совершает действие $A_{T-1} = \pi(S_{T-1})$,
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \ldots, S_T\}, \quad G(\tau) = \sum_{t=0}^{T-1} \gamma^t R_t,$

Задача

$$\text{Найти} \quad q_\pi(s, a) = \mathbb{E}_\pi[G \mid S_0 = s, A_0 = a]$$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$ $\quad q_\pi(S_0, A_0) = G(\tau)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- ...
- совершает действие $A_{T-2} = \pi(S_{T-2})$,
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}
- совершает действие $A_{T-1} = \pi(S_{T-1})$,
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \dots, S_T\}, \quad G(\tau) = \sum_{t=0}^{T-1} \gamma^t R_t,$

Задача

$$\text{Найти} \quad q_\pi(s, a) = \mathbb{E}_\pi[G \mid S_0 = s, A_0 = a]$$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$ $\quad q_\pi(S_0, A_0) = G(\tau)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- $\ldots$
- совершает действие $A_{T-2} = \pi(S_{T-2})$,
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}
- совершает действие $A_{T-1} = \pi(S_{T-1})$, $\quad q_\pi(S_{T-1}, A_{T-1}) = ???$
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \ldots, S_T\}$, $\quad G(\tau) = \sum\limits_{t=0}^{T-1} \gamma^t R_t$,

Задача

$$\text{Найти} \quad q_\pi(s, a) = \mathbb{E}_\pi[G \mid S_0 = s, A_0 = a]$$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$ $\quad q_\pi(S_0, A_0) = G(\tau)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- $\ldots$
- совершает действие $A_{T-2} = \pi(S_{T-2})$,
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}
- совершает действие $A_{T-1} = \pi(S_{T-1})$, $\quad q_\pi(S_{T-1}, A_{T-1}) = R_{T-1}$
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \ldots, S_T\}, \quad G(\tau) = \sum_{t=0}^{T-1} \gamma^t R_t,$

Задача

$$\text{Найти} \quad q_\pi(s, a) = \mathbb{E}_\pi[G \mid S_0 = s, A_0 = a]$$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$ $\quad q_\pi(S_0, A_0) = G(\tau)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- $\ldots$
- совершает действие $A_{T-2} = \pi(S_{T-2})$, $\quad q_\pi(S_{T-2}, A_{T-2}) = ???$
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}
- совершает действие $A_{T-1} = \pi(S_{T-1})$, $\quad q_\pi(S_{T-1}, A_{T-1}) = R_{T-1}$
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \ldots, S_T\}, \quad G(\tau) = \sum_{t=0}^{T-1} \gamma^t R_t,$

Задача

$$\text{Найти} \quad q_\pi(s, a) = \mathbb{E}_\pi[G \mid S_0 = s, A_0 = a]$$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$ $\quad q_\pi(S_0, A_0) = G(\tau)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- $\ldots$
- совершает действие $A_{T-2} = \pi(S_{T-2})$, $\quad q_\pi(S_{T-2}, A_{T-2}) = R_{T-2} + \gamma R_{T-1}$
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}
- совершает действие $A_{T-1} = \pi(S_{T-1})$, $\quad q_\pi(S_{T-1}, A_{T-1}) = R_{T-1}$
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \ldots, S_T\}, \quad G(\tau) = \sum\limits_{t=0}^{T-1} \gamma^t R_t,$

Задача

$$\text{Найти} \quad q_\pi(s, a) = \mathbb{E}_\pi[G \,|\, S_0 = s, A_0 = a]$$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$ $\quad q_\pi(S_0, A_0) = G(\tau)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$
- ...
- совершает действие $A_{T-2} = \pi(S_{T-2})$, $\quad q_\pi(S_{T-2}, A_{T-2}) = R_{T-2} + \gamma R_{T-1}$
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}
- совершает действие $A_{T-1} = \pi(S_{T-1})$, $\quad q_\pi(S_{T-1}, A_{T-1}) = R_{T-1}$
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \ldots, S_T\}$, $\quad G(\tau) = \sum_{t=0}^{T-1} \gamma^t R_t$, $\quad G_t = \sum_{k=t}^{T-1} \gamma^{k-t} R_t$

Задача

Найти $\quad q_\pi(s, a) = \mathbb{E}_\pi[G \mid S_0 = s, A_0 = a]$

Сессия. Детерминированный случай

- Мы задаем $\pi(s)$,
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 = \pi(S_0)$ $\quad q_\pi(S_0, A_0) = G_0$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 = \pi(S_1)$ $\quad q_\pi(S_1, A_1) = G_1$
- $\dots$
- совершает действие $A_{T-2} = \pi(S_{T-2})$, $\quad q_\pi(S_{T-2}, A_{T-2}) = G_{T-2}$
- получает награду R_{T-2} и переходит в следующее состояние S_{T-1}
- совершает действие $A_{T-1} = \pi(S_{T-1})$, $\quad q_\pi(S_{T-1}, A_{T-1}) = G_{T-1}$
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \dots, S_T\}$, $\quad G(\tau) = \sum_{t=0}^{T-1} \gamma^t R_t$, $\quad G_t = \sum_{k=t}^{T-1} \gamma^{k-t} R_t$

Задача

$$\text{Найти} \quad q_\pi(s, a) = \mathbb{E}_\pi[G \mid S_0 = s, A_0 = a]$$

Сессия. Общий случай

- Мы задаем $\pi(a|s)$. Инициализируем $W(s,a) = 0$ и $N(s,a) = 0$
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 \sim \pi(\cdot|S_0)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 \sim \pi(\cdot|S_1)$
- $\ldots$
- совершает действие $A_{T-1} \sim \pi(\cdot|S_{T-1})$,
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \ldots, S_T\}$, $\quad G(\tau) = \sum_{t=0}^{T-1} \gamma^t R_t$, $\quad G_t = \sum_{k=t}^{T-1} \gamma^{k-t} R_t$

Задача

$$\text{Найти} \quad q_\pi(s,a) = \mathbb{E}_\pi[G \mid S_0 = s, A_0 = a]$$

Сессия. Общий случай

- Мы задаем $\pi(a|s)$. Инициализируем $W(s,a)=0$ и $N(s,a)=0$
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 \sim \pi(\cdot|S_0)$
  $W(S_0,A_0) \leftarrow W(S_0,A_0)+G_0,\ \ N(S_0,A_0) \leftarrow N(S_0,A_0)+1$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 \sim \pi(\cdot|S_1)$
  $W(S_1,A_1) \leftarrow W(S_1,A_1)+G_1,\ \ N(S_1,A_1) \leftarrow N(S_1,A_1)+1$
- $\ldots$
- совершает действие $A_{T-1} \sim \pi(\cdot|S_{T-1})$,
  $W(S_{T-1},A_{T-1}) \leftarrow W(S_{T-1},A_{T-1})+G_{T-1}$,
  $N(S_{T-1},A_{T-1}) \leftarrow N(S_{T-1},A_{T-1})+1$
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau=\{S_0,A_0,S_1,A_1,\ldots,S_T\}, \quad G(\tau)=\sum\limits_{t=0}^{T-1}\gamma^t R_t, \quad G_t=\sum\limits_{k=t}^{T-1}\gamma^{k-t}R_t$

Задача

$$\text{Найти} \quad q_\pi(s,a)=\mathbb{E}_\pi[G \,|\, S_0=s, A_0=a]$$

Сессия. Общий случай

- Мы задаем $\pi(a|s)$. Инициализируем $W(s,a)=0$ и $N(s,a)=0$
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 \sim \pi(\cdot|S_0)$
  $W(S_0,A_0) \leftarrow W(S_0,A_0)+G_0,\;\; N(S_0,A_0) \leftarrow N(S_0,A_0)+1$
  $Q(S_0,A_0) \leftarrow W(S_0,A_0)/N(S_0,A_0)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 \sim \pi(\cdot|S_1)$
  $W(S_1,A_1) \leftarrow W(S_1,A_1)+G_1,\;\; N(S_1,A_1) \leftarrow N(S_1,A_1)+1$
  $Q(S_1,A_1) \leftarrow W(S_1,A_1)/N(S_1,A_1)$
- $\ldots$
- совершает действие $A_{T-1} \sim \pi(\cdot|S_{T-1})$,
  $W(S_{T-1},A_{T-1}) \leftarrow W(S_{T-1},A_{T-1})+G_{T-1}$,
  $N(S_{T-1},A_{T-1}) \leftarrow N(S_{T-1},A_{T-1})+1$
  $Q(S_{T-1},A_{T-1}) \leftarrow W(S_{T-1},A_{T-1})/N(S_{T-1},A_{T-1})$
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau=\{S_0,A_0,S_1,A_1,\ldots,S_T\}, \quad G(\tau)=\sum\limits_{t=0}^{T-1}\gamma^t R_t, \quad G_t=\sum\limits_{k=t}^{T-1}\gamma^{k-t}R_t$

Задача

$$\text{Найти} \quad q_\pi(s,a)=\mathbb{E}_\pi[G \mid S_0=s, A_0=a] \approx Q(s,a)$$

Формула для упрощения

$$Q_N = \frac{1}{N}\sum_{i=1}^{N} w_i$$

Формула для упрощения

$$Q_N = \frac{1}{N}\sum_{i=1}^{N} w_i$$

Тогда

$$Q_{N+1} = \frac{1}{N+1}\sum_{i=1}^{N+1} w_i = \frac{1}{N+1}\left(\sum_{i=1}^{N} w_i + w_{N+1}\right)$$

Формула для упрощения

$$Q_N = \frac{1}{N}\sum_{i=1}^{N} w_i$$

Тогда

$$Q_{N+1} = \frac{1}{N+1}\sum_{i=1}^{N+1} w_i = \frac{1}{N+1}\left(\sum_{i=1}^{N} w_i + w_{N+1}\right)$$

$$= \frac{1}{N+1}(NQ_N + w_{N+1}) = Q_N + \frac{1}{N+1}(w_{N+1} - Q_N)$$

Формула для упрощения

$$Q_N = \frac{1}{N}\sum_{i=1}^{N} w_i$$

Тогда

$$Q_{N+1} = \frac{1}{N+1}\sum_{i=1}^{N+1} w_i = \frac{1}{N+1}\left(\sum_{i=1}^{N} w_i + w_{N+1}\right)$$

$$= \frac{1}{N+1}(NQ_N + w_{N+1}) = Q_N + \frac{1}{N+1}(w_{N+1} - Q_N)$$

$$Q_{N+1} = Q_N + \frac{1}{N+1}(w_{N+1} - Q_N)$$

Сессия. Общий случай

- Мы задаем $\pi(a|s)$. Инициализируем $Q(s,a) = 0$ и $N(s,a) = 0$
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 \sim \pi(\cdot|S_0)$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 \sim \pi(\cdot|S_1)$
- ...
- совершает действие $A_{T-1} \sim \pi(\cdot|S_{T-1})$,
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \dots, S_T\}, \quad G(\tau) = \sum_{t=0}^{T-1} \gamma^t R_t, \quad G_t = \sum_{k=t}^{T-1} \gamma^{k-t} R_t$

Задача

$$\text{Найти} \quad q_\pi(s,a) = \mathbb{E}_\pi[G \mid S_0 = s, A_0 = a] \approx Q(s,a)$$

Сессия. Общий случай

- Мы задаем $\pi(a|s)$. Инициализируем $Q(s,a) = 0$ и $N(s,a) = 0$
- АГЕНТ находится в начальном состоянии S_0,
- совершает действие $A_0 \sim \pi(\cdot|S_0)$
  $Q(S_0, A_0) \leftarrow Q(S_0, A_0) + \frac{1}{N(S_0,A_0)+1}(G_0 - Q(S_0, A_0))$,
  $N(S_0, A_0) \leftarrow N(S_0, A_0) + 1$
- получает награду R_0 и переходит в следующее состояние S_1
- совершает действие $A_1 \sim \pi(\cdot|S_1)$
  $Q(S_1, A_1) \leftarrow Q(S_1, A_1) + \frac{1}{N(S_1,A_1)+1}(G_1 - Q(S_1, A_1))$,
  $N(S_1, A_1) \leftarrow N(S_1, A_1) + 1$
- ...
- совершает действие $A_{T-1} \sim \pi(\cdot|S_{T-1})$,
  $Q(S_{T-1}, A_{T-1}) \leftarrow Q(S_{T-1}, A_{T-1}) + \frac{1}{N(S_{T-1},A_{T-1})+1}(G_{T-1} - Q(S_{T-1}, A_{T-1}))$,
  $N(S_{T-1}, A_{T-1}) \leftarrow N(S_{T-1}, A_{T-1}) + 1$
- получает награду R_{T-1} и переходит в терминальное состояние S_T
- $\tau = \{S_0, A_0, S_1, A_1, \ldots, S_T\}, \quad G(\tau) = \sum_{t=0}^{T-1} \gamma^t R_t, \quad G_t = \sum_{k=t}^{T-1} \gamma^{k-t} R_t$

Задача

Найти $\quad q_\pi(s,a) = \mathbb{E}_\pi[G \mid S_0 = s, A_0 = a] \approx Q(s,a)$

Monte-Carlo Policy Evaluation

Пусть выбрана π. Пусть $Q(s,a)=0$ и $N(s,a)=0$.
Для каждого эпизода $k \in \overline{1,K}$ делаем

- В согласии с π получаем траекторию $\tau = (S_0, A_0, \ldots, S_T)$ и награды $(R_0, \ldots, R_{T-1})$. По ним определяем $(G_0, \ldots, G_{T-1})$.

Monte-Carlo Policy Evaluation

Пусть выбрана π. Пусть $Q(s, a) = 0$ и $N(s, a) = 0$.
Для каждого эпизода $k \in \overline{1, K}$ делаем

- В согласии с π получаем траекторию $\tau = (S_0, A_0, \ldots, S_T)$ и награды $(R_0, \ldots, R_{T-1})$. По ним определяем $(G_0, \ldots, G_{T-1})$.
- Для каждого $t \in \overline{0, T-1}$ обновляем Q и N:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t) + 1}\big(G_t - Q(S_t, A_t)\big),$$

$$N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$$

Monte-Carlo Policy Evaluation

Пусть выбрана π. Пусть $Q(s,a) = 0$ и $N(s,a) = 0$.
Для каждого эпизода $k \in \overline{1, K}$ делаем

- В согласии с π получаем траекторию $\tau = (S_0, A_0, \dots, S_T)$ и награды $(R_0, \dots, R_{T-1})$. По ним определяем $(G_0, \dots, G_{T-1})$.
- Для каждого $t \in \overline{0, T-1}$ обновляем Q и N:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t) + 1}\big(G_t - Q(S_t, A_t)\big),$$

$$N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$$

$$Q(s,a) \approx q_\pi(s,a)$$

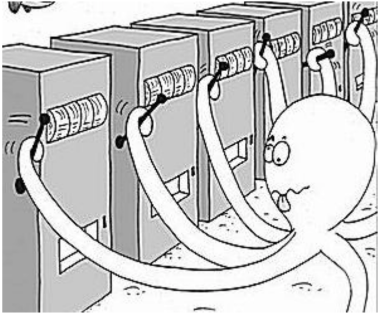
Будет ли это работать?

Пусть инициализированы π^0 и заданы числа $K > 0$.
Для каждой итерации $k \in \overline{1, K}$ делаем

- (Policy evaluation) Monte-Carlo Policy Evaluation.
  Получаем $Q^k(s, a) \approx q_{\pi^k}(s, a)$
- (Policy improvement) Greedy Policy Improvement:

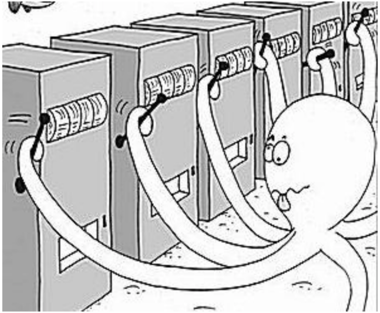
$$\pi^{k+1}(a|s) = \begin{cases} 1, & \text{если } a \in \operatorname{argmax}_{a' \in \mathcal{A}} Q^k(s, a') \\ 0, & \text{иначе} \end{cases}$$

Многорукий бандит



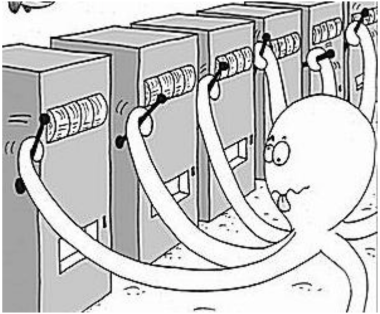
- Состояния: начальное и терминальное
- Действия: $\rightarrow$, $\leftarrow$
- Награда:
  $R(S_0, \leftarrow) = 1$,
  $R(S_0, \rightarrow) = 2$
- Начальная Policy:
  $\pi^0(S_0, \leftarrow) = 1$,
  $\pi^0(S_0, \rightarrow) = 0$
- Q^0 =???

Многорукий бандит



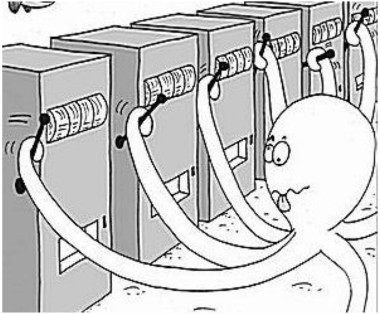
- Состояния: начальное и терминальное
- Действия: $\rightarrow$, $\leftarrow$
- Награда:
  $R(S_0, \leftarrow) = 1$,
  $R(S_0, \rightarrow) = 2$
- Начальная Policy:
  $\pi^0(S_0, \leftarrow) = 1$,
  $\pi^0(S_0, \rightarrow) = 0$
- $Q^0(S_0, \leftarrow) = 1$,
  $Q^0(S_0, \rightarrow) = 0$

Многорукий бандит



- Состояния: начальное и терминальное
- Действия: $\rightarrow$, $\leftarrow$
- Награда:
  $R(S_0, \leftarrow) = 1$,
  $R(S_0, \rightarrow) = 2$
- Начальная Policy:
  $\pi^0(S_0, \leftarrow) = 1$,
  $\pi^0(S_0, \rightarrow) = 0$
- $Q^0(S_0, \leftarrow) = 1$,
  $Q^0(S_0, \rightarrow) = 0$
- $\pi^1 = \pi^0$,

Многорукий бандит



- Состояния: начальное и терминальное
- Действия: $\rightarrow$, $\leftarrow$
- Награда:
  $R(S_0, \leftarrow) = 1$,
  $R(S_0, \rightarrow) = 2$
- Начальная Policy:
  $\pi^0(S_0, \leftarrow) = 1$,
  $\pi^0(S_0, \rightarrow) = 0$
- $Q^0(S_0, \leftarrow) = 1$,
  $Q^0(S_0, \rightarrow) = 0$
- $\pi^1 = \pi^0$,
- $Q^1 = Q^0$,

ε-Greedy Policy Improvement

$\pi = \varepsilon$-greedy(Q)

$$\pi'(a|s) = \begin{cases} 1 - \varepsilon + \varepsilon/m, & \text{если } a \in \operatorname{argmax}_{a' \in \mathcal{A}} Q(s, a'), \\ \varepsilon/m, & \text{иначе} \end{cases}$$

ε-Greedy Policy Improvement

$\pi = \varepsilon\text{-greedy}(Q)$

$$\pi'(a|s) = \begin{cases} 1 - \varepsilon + \varepsilon/m, & \text{если } a \in \operatorname{argmax}_{a' \in \mathcal{A}} Q(s, a'), \\ \varepsilon/m, & \text{иначе} \end{cases}$$

Policy Improvement Theorem

Пусть $Q(s, a)$ — некоторая функция.
Пусть $\pi = \varepsilon\text{-greedy}(Q)$ и $\pi' = \varepsilon\text{-greedy}(q_\pi)$.
Тогда $\pi' \geq \pi$ (т.е. $v_{\pi'}(s) \geq v_\pi(s)$, $\forall s$)

Learning with Monte-Carlo Policy Evaluation

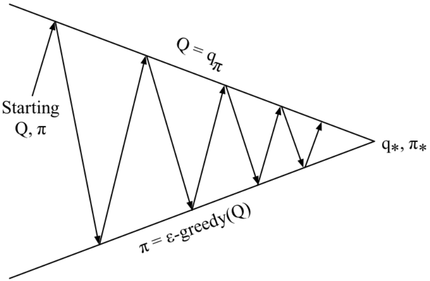
Пусть инициализированы π^0 и заданы числа $K > 0$ и $\varepsilon = 1$.
Для каждой итерации $k \in \overline{1, K}$ делаем

- (Policy evaluation) Monte-Carlo Policy Evaluation - получаем $Q^k(s, a) \approx q_{\pi^k}(s, a)$
- (Policy improvement) ε-Greedy Policy Improvement - получаем π^{k+1} по Q^k. Определяем $\varepsilon = 1/k$

Learning with Monte-Carlo Policy Evaluation

Пусть инициализированы π^0 и заданы числа $K > 0$ и $\varepsilon = 1$.
Для каждой итерации $k \in \overline{1, K}$ делаем

- (Policy evaluation) Monte-Carlo Policy Evaluation - получаем $Q^k(s, a) \approx q_{\pi^k}(s, a)$
- (Policy improvement) ε-Greedy Policy Improvement - получаем π^{k+1} по Q^k. Определяем $\varepsilon = 1/k$



Learning with Monte-Carlo Policy Evaluation

Пусть инициализированы π^0 и заданы числа $K > 0$ и $\varepsilon = 1$.
Для каждой итерации $k \in \overline{1, K}$ делаем

- (Policy evaluation) Monte-Carlo Policy Evaluation - получаем $Q^k(s,a) \approx q_{\pi^k}(s,a)$
- (Policy improvement) ε-Greedy Policy Improvement - получаем π^{k+1} по Q^k. Определяем $\varepsilon = 1/k$

Теорема

Алгоритм сходится, то есть $Q^k \to q_*$ и $\pi^k \to \pi_*$ при $k \to \infty$.

Monte-Carlo Algorithm

Пусть $Q(s, a) = 0$, $N(s, a) = 0$ и $\varepsilon = 1$.
Для каждого эпизода $k \in \overline{1, K}$ делаем:

- Согласно $\pi = \varepsilon\text{-greedy}(Q)$ получаем траекторию $\tau = (S_0, A_0, \dots, S_T)$ и награды $(R_0, \dots, R_{T-1})$. По ним определяем $(G_0, \dots, G_{T-1})$.
- Для каждого $t \in \overline{0, T-1}$ обновляем Q и N:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t) + 1}\big(G_t - Q(S_t, A_t)\big),$$

$$N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$$

Определяем $\varepsilon = 1/k$

Monte-Carlo Algorithm

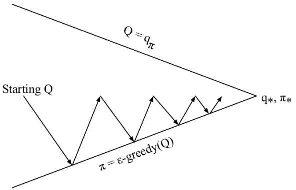
Пусть $Q(s,a) = 0$, $N(s,a) = 0$ и $\varepsilon = 1$.
Для каждого эпизода $k \in \overline{1, K}$ делаем:

- Согласно $\pi = \varepsilon\text{-greedy}(Q)$ получаем траекторию $\tau = (S_0, A_0, \ldots, S_T)$ и награды $(R_0, \ldots, R_{T-1})$. По ним определяем $(G_0, \ldots, G_{T-1})$.
- Для каждого $t \in \overline{0, T-1}$ обновляем Q и N:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t) + 1}\big(G_t - Q(S_t, A_t)\big),$$

$$N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$$

Определяем $\varepsilon = 1/k$



Monte-Carlo Algorithm

Пусть $Q(s,a) = 0$, $N(s,a) = 0$ и $\varepsilon = 1$.
Для каждого эпизода $k \in \overline{1, K}$ делаем:

- Согласно $\pi = \varepsilon\text{-greedy}(Q)$ получаем траекторию $\tau = (S_0, A_0, \ldots, S_T)$ и награды $(R_0, \ldots, R_{T-1})$. По ним определяем $(G_0, \ldots, G_{T-1})$.
- Для каждого $t \in \overline{0, T-1}$ обновляем Q и N:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t) + 1}\big(G_t - Q(S_t, A_t)\big),$$

$$N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$$

Определяем $\varepsilon = 1/k$

Теорема

Алгоритм сходится, то есть $Q^k \to q_*$ и $\pi^k \to \pi_*$ при $k \to \infty$.

Использование Bellman Equation

Bellman Expectation Equation для q_π

$$q_\pi(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \sum_{a'} \pi(a'|s') q_\pi(s',a')$$

$$\Downarrow$$

$$q_\pi(s,a) = \mathbb{E}_\pi[R_t + \gamma q_\pi(S_{t+1}, A_{t+1}) \,|\, S_t = s,\; A_t = a]$$

$$\Downarrow$$

Temporal-Difference

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_t + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

Temporal-Difference Policy Evaluation

Пусть выбрана π. Пусть $Q(s, a) = 0$.
Для каждого эпизода $k \in \overline{1, K}$ делаем

- В согласии с π получаем траекторию $\tau = (S_0, A_0, \ldots, S_T)$ и награды $(R_0, \ldots, R_{T-1})$.

Temporal-Difference Policy Evaluation

Пусть выбрана π. Пусть $Q(s, a) = 0$.
Для каждого эпизода $k \in \overline{1, K}$ делаем

- В согласии с π получаем траекторию $\tau = (S_0, A_0, \ldots, S_T)$ и награды $(R_0, \ldots, R_{T-1})$.
- Для каждого $t \in \overline{0, T-2}$ обновляем значения

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_t + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

Temporal-Difference Policy Evaluation

Пусть выбрана π. Пусть $Q(s, a) = 0$.
Для каждого эпизода $k \in \overline{1, K}$ делаем

- В согласии с π получаем траекторию $\tau = (S_0, A_0, \ldots, S_T)$ и награды $(R_0, \ldots, R_{T-1})$.
- Для каждого $t \in \overline{0, T-2}$ обновляем значения

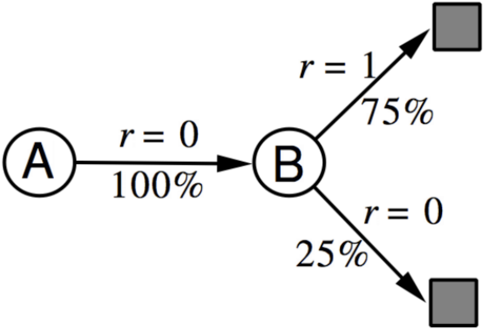
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_t + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

$$Q(s, a) \approx q_\pi(s, a)$$

Сравнение MC и TD Policy Evaluation

Two states A, B; no discounting; 8 episodes of experience

A, 0, B, 0
B, 1
B, 1
B, 1
B, 1
B, 1
B, 1
B, 0



What is $V(A)$, $V(B)$?

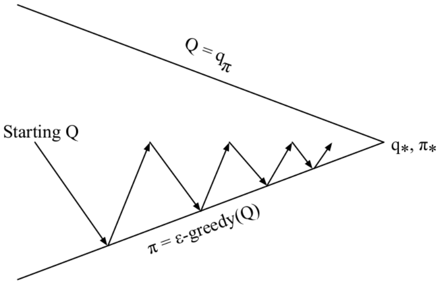
Learning with Temporal-Difference Policy Evaluation

Пусть $Q(s,a) = 0$ и $\varepsilon = 1$.
Для каждого эпизода $k \in \overline{1,K}$ делаем:

- Согласно $\pi = \varepsilon\text{-greedy}(Q)$ получаем траекторию $\tau = (S_0, A_0, \dots, S_T)$ и награды $(R_0, \dots, R_{T-1})$.
- Для каждого $t \in \overline{0, T-2}$ обновляем Q:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_t + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

Определяем $\varepsilon = 1/k$



SARSA Algorithm

Пусть $Q(s, a) = 0$ и $\varepsilon = 1$.
Для каждого эпизода k делаем:
Пока эпизод не закончен делаем:

- Находясь в состоянии S_t совершаем действие $A_t \sim \pi(\cdot|S_t)$, где $\pi = \varepsilon$-greedy(Q), получаем награду R_t, переходим в состояние S_{t+1}, совершаем действие $A_{t+1} \sim \pi(\cdot|S_{t+1})$
- По $(S_t, A_t, R_t, S_{t+1}, A_{t+1})$ обновляем Q:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_t + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

Полагаем, например, $\varepsilon = 1/k$

SARSA Algorithm

Пусть $Q(s, a) = 0$ и $\varepsilon = 1$.
Для каждого эпизода k делаем:
Пока эпизод не закончен делаем:

- Находясь в состоянии S_t совершаем действие $A_t \sim \pi(\cdot|S_t)$, где $\pi = \varepsilon\text{-greedy}(Q)$, получаем награду R_t, переходим в состояние S_{t+1}, совершаем действие $A_{t+1} \sim \pi(\cdot|S_{t+1})$
- По $(S_t, A_t, R_t, S_{t+1}, A_{t+1})$ обновляем Q:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_t + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t))$$

Полагаем, например, $\varepsilon = 1/k$

Теорема

Алгоритм сходится, то есть $Q^k \to q_*$ и $\pi^k \to \pi_*$ при $k \to \infty$.

Использование Bellman Optimality Equation

Bellman Optimality Equation для q_*

$$q_*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \max_{a'} q_*(s',a')$$

$$\Downarrow$$

$$q_*(s,a) = \mathbb{E}[R_t + \gamma \max_{a'} q_*(S_{t+1},a') \mid S_t = s,\ A_t = a]$$

$$\Downarrow$$

Q-Learning

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha(R_t + \gamma \max_{a'} Q(S_{t+1},a') - Q(S_t,A_t))$$

Q-Learning Algorithm

Пусть $Q(s,a) = 0$ и $\varepsilon = 1$.
Для каждого эпизода k делаем:
Пока эпизод не закончен делаем:

- Находясь в состоянии S_t совершаем действие $A_t \sim \pi(\cdot|S_t)$, где $\pi = \varepsilon\text{-greedy}(Q)$, получаем награду R_t переходим в состояние S_{t+1}.
- По (S_t, A_t, R_t, S_{t+1}) обновляем Q:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_t + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

Полагаем, например, $\varepsilon = 1/k$

Q-Learning Algorithm

Пусть $Q(s, a) = 0$ и $\varepsilon = 1$.
Для каждого эпизода k делаем:
Пока эпизод не закончен делаем:

- Находясь в состоянии S_t совершаем действие $A_t \sim \pi(\cdot|S_t)$, где $\pi = \varepsilon\text{-greedy}(Q)$, получаем награду R_t переходим в состояние S_{t+1}.
- По (S_t, A_t, R_t, S_{t+1}) обновляем Q:

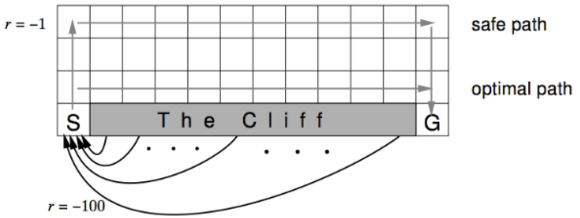
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_t + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t))$$

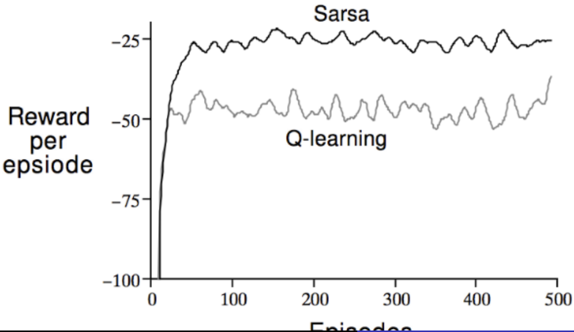
Полагаем, например, $\varepsilon = 1/k$

Теорема

Алгоритм сходится, то есть $Q^k \to q_*$ и $\pi^k \to \pi_*$ при $k \to \infty$.

Сравнение SARSA и Q-Learning





Model-based и Model-free

Q-Policy Iteration	Sarsa
$Q(s,a) \leftarrow \mathbb{E}\left[R + \gamma Q(S',A') \mid s,a\right]$	$Q(S,A) \overset{\alpha}{\leftarrow} R + \gamma Q(S',A')$
Q-Value Iteration	Q-Learning
$Q(s,a) \leftarrow \mathbb{E}\left[R + \gamma \max_{a' \in \mathcal{A}} Q(S',a') \mid s,a\right]$	$Q(S,A) \overset{\alpha}{\leftarrow} R + \gamma \max_{a' \in \mathcal{A}} Q(S',a')$

Организационные вопросы

- Пятница, 17:50, аудитория 622
- Отчетность: домашние работы
- Странчика курса: https://github.com/imm-rl-lab/UrFU_course
- E-mail для связи:

ВОПРОСЫ?