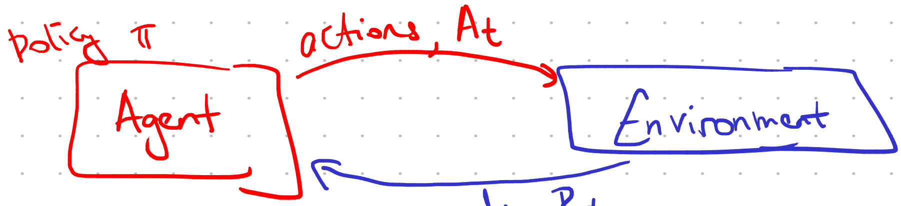
① Markov decision process (MDP)



Шахматы: / episodic task

- A_t – ход
- R_t – результат партии
- S_t – позиция / ~~доска~~ / расклад фигур / предыд. 50 ходов

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \ldots S_t, A_t, R_{t+1}, S_{t+1}, \ldots$$

(π: $S_t \to A_t$; p: $(S_t, A_t) \to (R_{t+1}, S_{t+1})$)

$\pi : S \to A$ – deterministic policy

π^* – optimal strategy

$\pi : S \to Prob(A)$ – stochastic policy

$$\pi(a|s) = p(A_t = a \mid S_t = s)$$

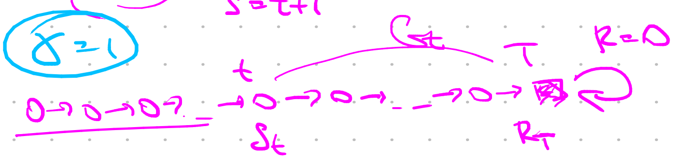
$$p(r, s' \mid s, a) = p(R_{t+1} = r, S_{t+1} = s' \mid S_t = s, A_t = a)$$

Episodic vs continuous tasks

Episodic: $t = 1, 2, \ldots, T$

$$G_t = \sum_{s=t+1}^{T} R_s \xrightarrow{\pi} \max$$

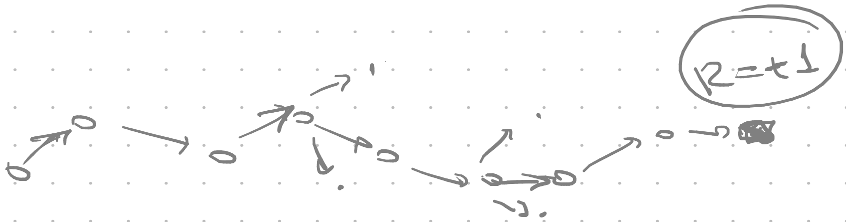
$\gamma = 1$



Continuous: $t = 1, 2, \ldots, T, \ldots$

$$G_t = \sum_{s=t+1}^{\infty} \gamma^{s-t-1} \cdot R_s \xrightarrow{\pi} \max$$

$\gamma < 1$



Problems

1) ~~Credit assignment~~
2) Exploration vs. exploitation

② Multiarmed bandits

$|S| = 1$, $A = a_1, a_2, \ldots, a_k$

Env – $p(r|a)$

$$a_* = \underset{a}{\operatorname{argmax}}\ \mathbb{E}_{p(r|a)}[r]$$

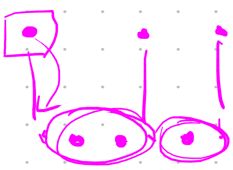
1) A/B testing: дёрнуть N раз за $\forall a$, выбрать max.

2) Greedy: – дёрнуть 1 раз за каждую
– for $t = 1, 2, \ldots$

$$A_t = \underset{i}{\operatorname{argmax}}\ \hat{\mu}_i(t)$$

$$\hat{\mu}_i(t) = \frac{1}{n_i} \sum_{t : A_t = i} R_t$$

Binary bandits $r \in \{0, 1\}$

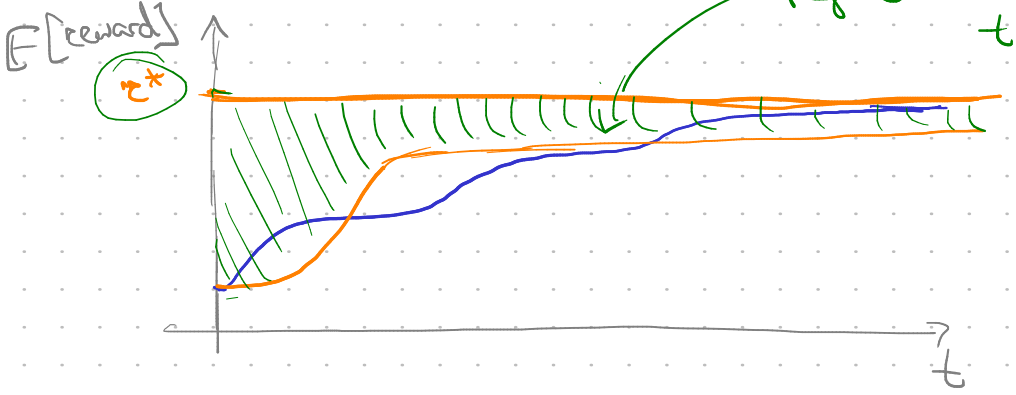


3) ε-Greedy

ε-soft strategies

$$A_t = \begin{cases} \text{argmax}_i \hat{\mu}_i(t), & \text{с вер. } 1-\varepsilon \\ \text{Unif}(\{1..K\}), & \text{с вер. } \varepsilon. \end{cases}$$

Regret / цена обучения



$$\text{regret} = \sum_{t=1}^{\infty} \mathbb{E}_{\text{learning}} [r^* - r_t]$$

$$\hat{\mu}(n) = \frac{1}{n} \sum_{i=1}^{n} r_i \qquad r_1, r_2, \dots, r_n, r_{n+1}$$

$$\hat{\mu}(n+1) = \frac{1}{n+1} \sum_{i=1}^{n+1} r_i = \frac{n}{n+1} \cdot \hat{\mu}(n) + \frac{1}{n+1} \cdot r_{n+1} = \hat{\mu}(n) + \frac{1}{n+1} \left(r_{n+1} - \hat{\mu}(n) \right)$$

Новая оценка = Старая оценка + α_n · (Мгновенное значение − Старая оценка)

$$\alpha_n = \frac{1}{n} \Rightarrow \hat{\mu} = \text{Avg}(r_i - r_n)$$

$$\sum_n \alpha_n = \infty, \quad \sum_n \alpha_n^2 < \infty \Rightarrow \text{оценка сходится}$$

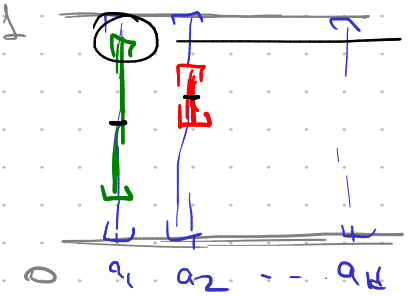
Nonstationary bandits

$\alpha_n = \alpha = \text{Const}$: EMA exp-mov.-avg

$$\hat{\mu}(n+1) = \hat{\mu}(n) + \alpha (r_{n+1} - \hat{\mu}(n)) = \alpha \cdot r_{n+1} + (1-\alpha) \hat{\mu}(n)$$

$$= \alpha \cdot r_{n+1} + \alpha (1-\alpha) r_n + \alpha (1-\alpha)^2 r_{n-1} + \dots$$

(3) Optimism under uncertainty



верх. грант. доверит. инт.

$n_i \to \infty$

$$\text{Priority}_i(t) = \hat{\mu}_i(t) + \text{Exploration bonus}$$

UCB – upper confidence bound

доверительная

$$UCB_1: \text{Priority}_i(t) = \hat{\mu}_i(t) + c \cdot \sqrt{\frac{\log t}{n_i}}$$

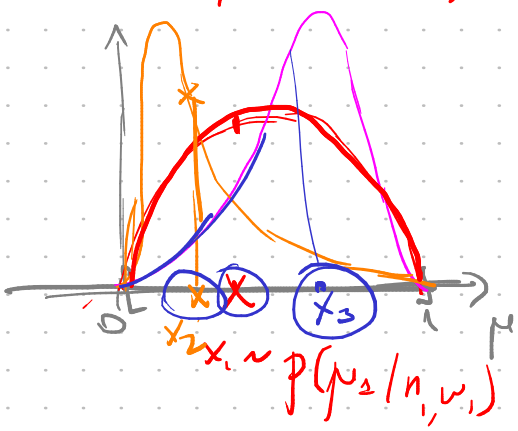
для i

Thompson sampling

$p(\mu_i) = \text{Beta}(1,1)$

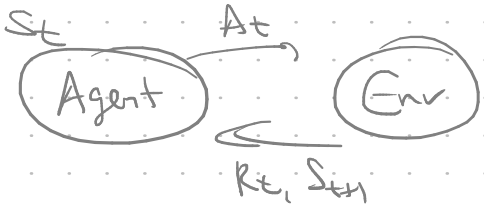
n эксп., w - wins,

$p(\mu_i | n, w) = \text{Beta}(w+1, n-w+1)$



- $x_i \sim p(\mu_i | D)$
- $A_t = \underset{i}{\text{argmax}}\, x_i$

(4) Value functions



Reward: $R_t, R_{t+1}, R_{t+2}, \dots$

Return: $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$

$G_t = R_{t+1} + \gamma \cdot G_{t+1}$

(State) value function:

$$V_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right]$$

$$V_*(s) = \max_\pi V_\pi(s) = V_{\pi_*}(s)$$

$\mathbb{E}$ по $R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, \dots$

State-action value function:

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_k \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right]$$

$$Q_*(s,a) = \max_\pi Q_\pi(s,a) \qquad \pi_*(s) = \underset{a}{\text{argmax}}\, Q_*(s,a)$$

$$Q_\pi(s,a) = \mathbb{E}_\pi[R_{t+1} + \gamma \cdot G_{t+1} | S_t = s, A_t = a] =$$

$$= \sum_{r,s'} p(r,s' | s,a) \cdot \mathbb{E}_\pi[r + \gamma \cdot G_{t+1} | S_{t+1} = s']$$

$$\boxed{Q_\pi(s,a) = \sum_{r,s'} p(r,s' | s,a)\left(r + \gamma \cdot V_\pi(s')\right)}$$

$$V_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s] = \mathbb{E}_{A_t}[\mathbb{E}_\pi[\dots]] =$$

$\pi(a|s)$ $\quad A(s) = A$

$$= \sum_a \pi(a|s) \cdot \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right]$$

$$\boxed{V_\pi(s) = \sum_a \pi(a|s)\, Q(s,a)}$$

$$\boxed{V_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(r,s'|s,a)\left(r + \gamma \cdot V_\pi(s')\right)}$$

Bellman equations

$y = f(x)$

$$\boxed{Q_\pi(s,a) = \sum_{r,s'} p(r,s'|s,a)\left(r + \gamma \cdot \sum_{a'} \pi(a'|s')\, Q(s',a')\right)}$$

Алгоритм: - для каждой π:

- решаем ур-я Беллмана, найти $Q_\pi(s,a)\ \forall s,a$
- $\pi_*(s) = \arg\max_a \left[\max_\pi Q_\pi(s,a)\right]$

Проблемы: 1) π слишком много

2) уравнений слишком много $|S|,\ |S| \times |A|$

3) Часто мы не знаем $p(r,s'|s,a)$

1) $$V_*(s) = \max_\pi V_\pi(s) = \max_\pi \sum_a \pi(a|s) \sum_{r,s'} p(r,s'|s,a)\left(r + \gamma V_\pi(s')\right) =$$

$$= \max_a \sum_{r,s'} p(r,s'|s,a)\left(r + \gamma \max_\pi V_\pi(s')\right)$$

$$\boxed{V_*(s) = \max_a \sum_{r,s'} p(r,s'|s,a)\left(r + \gamma \cdot V_*(s')\right)}$$

$y = f(x,t)$

$$\boxed{Q_*(s,a) = \sum_{r,s'} p(r,s'|s,a)\left(r + \gamma \cdot \max_{a'} Q_*(s',a')\right)}$$

Bellman equations

2) Tabular RL $\longrightarrow$ Approximate RL

S	V(s)
s_1	$V(s_1)$
s_2	$V(s_2)$

$s \to [\theta] \to V(s)$

$s \to [\theta] \to Q(s,a)$

(5) Policy improvement theorem

Thm. Если для двух детерм. π, π'

$$\forall s \quad Q_\pi(s, \pi'(s)) \geq V_\pi(s) = Q_\pi(s, \pi(s)),$$

то $\forall s \quad V_{\pi'}(s) \geq V_\pi(s)$.

Док-во:

$$V_\pi(s) \leq Q_\pi(s, \pi'(s)) = \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = \pi'(s)\right]$$

$$= \mathbb{E}_{R_{t+1}, S_{t+1}}\left[R_{t+1} + \gamma \cdot \underbrace{\mathbb{E}_\pi\left[G_{t+1} \mid S_{t+1}\right]}_{V_\pi(S_{t+1}) \leq Q_\pi(S_{t+1}, \pi'(S_{t+1}))} \;\middle|\; \begin{matrix} S_t = s, \\ A_t = \pi'(s) \end{matrix}\right] \leq$$

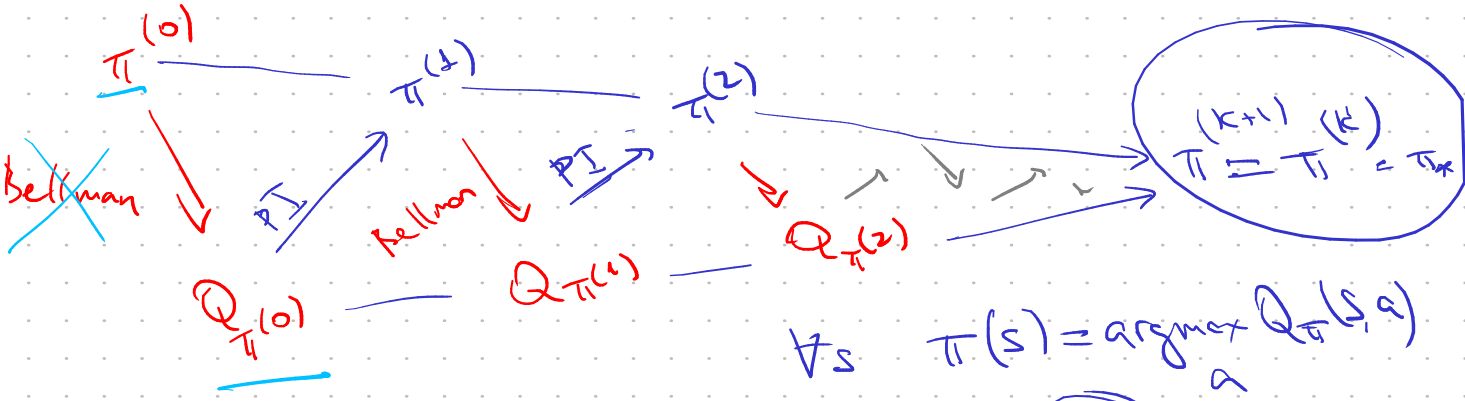
$$\leq \mathbb{E}_{R_{t+1}, S_{t+1}}\left[R_{t+1} + \gamma \cdot \mathbb{E}_{R_{t+2}, S_{t+2}}\left[R_{t+2} + \gamma \cdot \mathbb{E}_t\left[G_{t+2} \mid S_{t+2}\right] \;\middle|\; \begin{matrix} S_{t+1}, \\ A_{t+1} = \pi'(S_{t+1}) \end{matrix}\right]\right.$$

$$\left. \;\middle|\; S_t = s, A_t = \pi'(s)\right] \leq \cdots$$

$$\leq \cdots \leq V_{\pi'}(s) \qquad \text{QED}$$

$\pi \rightsquigarrow \pi'$ — лучше?

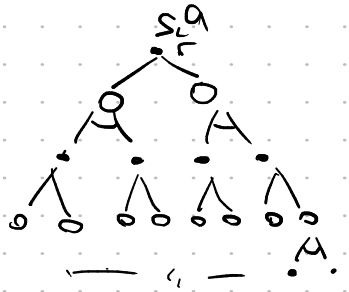
Policy iteration: $\pi \longrightarrow Q_\pi \longrightarrow \pi'(s) = \operatorname{argmax}_a Q_\pi(s, a)$

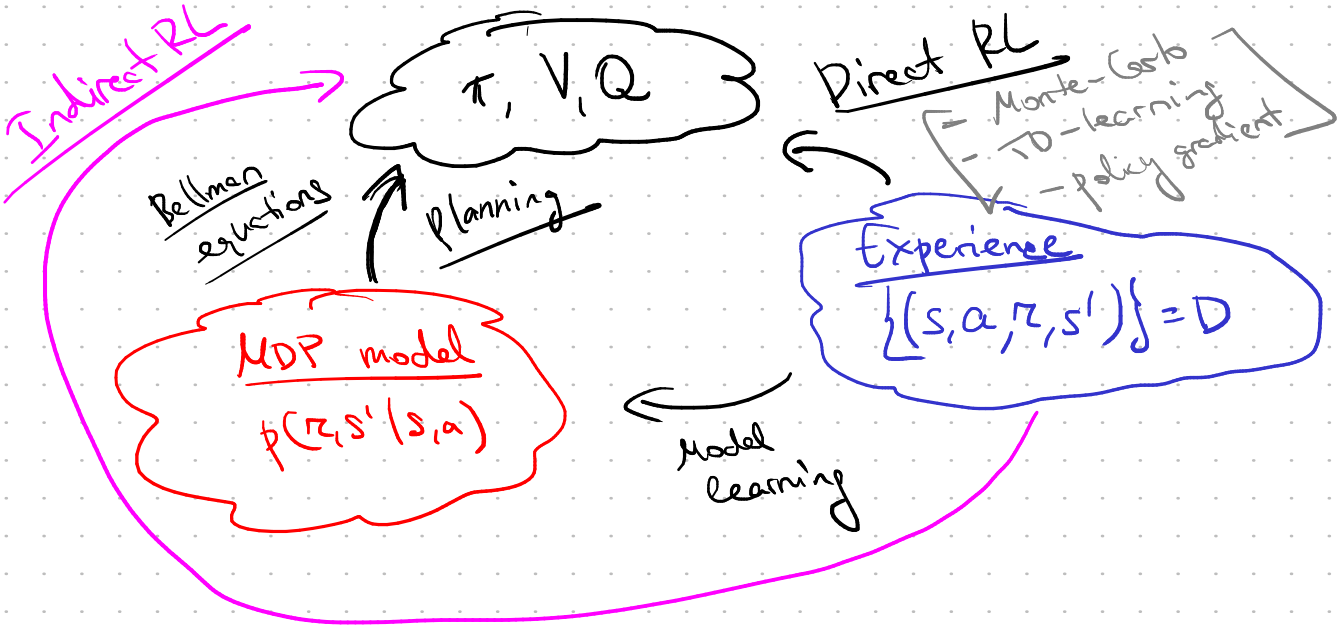


$\forall s \quad \pi(s) = \operatorname{argmax}_a Q_\pi(s, a)$

$\Rightarrow \pi = \pi_*$

$\pi \rightsquigarrow$ Env $\rightsquigarrow$ Experience $\rightsquigarrow Q_\pi$




Indirect RL
π, V, Q
Direct RL
- Monte-Carlo
- TD-learning
- policy gradient
Bellman equations
planning
Experience
{(s, a, r, s')} = D
MDP model
p(r, s' | s, a)
model learning