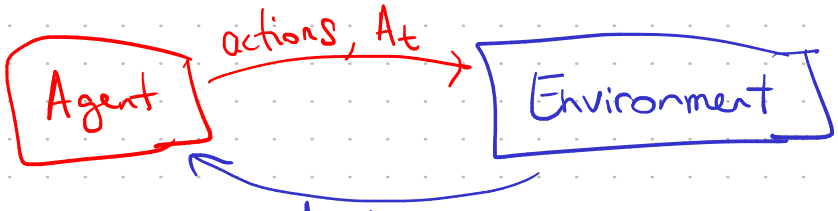
1) Machine Learning

- Supervised learning: $D = \{(\bar{x}, y)\}$, $f: \bar{x} \mapsto y$
- Unsupervised learning: $D = \{\bar{x}\}$, $p(\bar{x})$
- → Reinforcement Learning

2) Markov decision process (MDP)



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

$\pi: S \to Prob(A(S))$

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

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

Шахматы: action — ход; reward — 0 / 1/2 / 1 в конце партии; state — позиция + счётчики

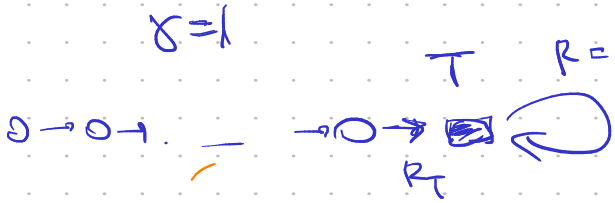
} episodic tasks

Episodic vs continuous

$t = 1, 2, \dots, T$

$G_t = \sum_{t=1}^{T} R_t \xrightarrow{\pi} \max$

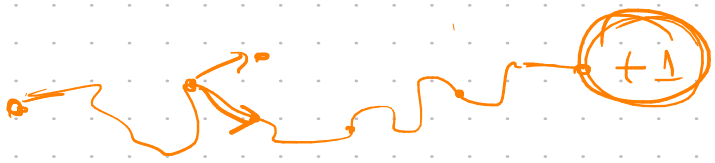
$\gamma = 1$



$t = 1, 2, \dots, T, \dots$

$G_t = \sum_{t=1}^{\infty} \gamma^{t-1} R_t \xrightarrow{\pi} \max$

$\gamma < 1$



1) ~~Credit assignment~~

2) Exploration vs. exploitation

(3) Multiarmed bandits

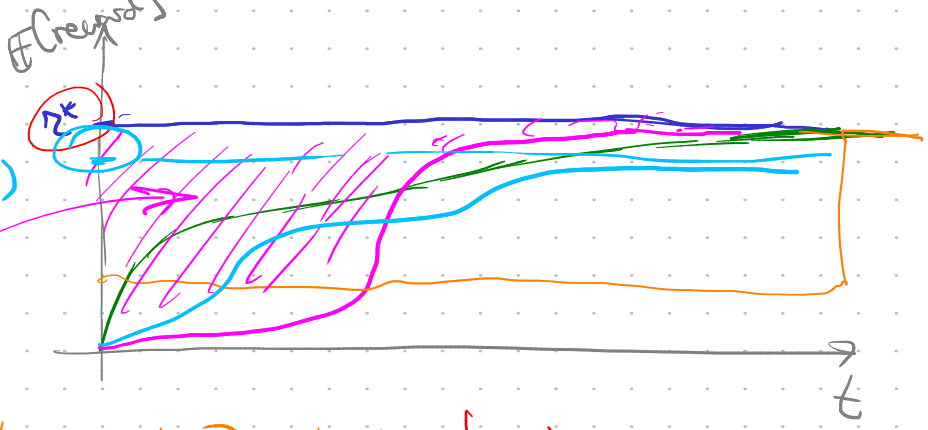
$|S| = 1$, $A = a_1, a_2, \dots, a_k$

$R_{t+1} \sim p(R_{t+1} = r \mid A_t = a)$ $\quad p(r|a)$

$$a_* = \arg\max_a \mathbb{E}_{p(r|a)}[r]$$

$R_1 - R_t$

$\hat{\mu}_i = \frac{1}{t} \sum r_j$ — regret



Greedy:
- делаем 1 раз за каждую ручку
- for $t = 1, \dots$

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

$$A_t = \arg\max_i \hat{\mu}_i(t)$$

Thompson sampling

Binary bandits

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

ε-Greedy — — — —
- for $t = 1 \dots$

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

ε-Soft strategies

$$\hat{\mu}(n) = \frac{1}{n} \sum_{i=1}^{n} r_i \quad \in r_{n+1}$$

$$\hat{\mu}(n+1) = \frac{(n \cdot \hat{\mu}(n) + r_{n+1})}{n+1} = \hat{\mu}(n) + \underbrace{\frac{1}{n+1}}_{=\alpha_n} (r_{n+1} - \hat{\mu}(n))$$

Новая оценка = Старая оценка + α_n (Новое значение − старая оценка)

$\alpha_n \sim \frac{1}{n}$

$$\sum_{n=1}^{\infty} \alpha_n = \infty$$

$$\sum \alpha_n^2 < \infty$$

$\alpha_n = \alpha = \text{const}$

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

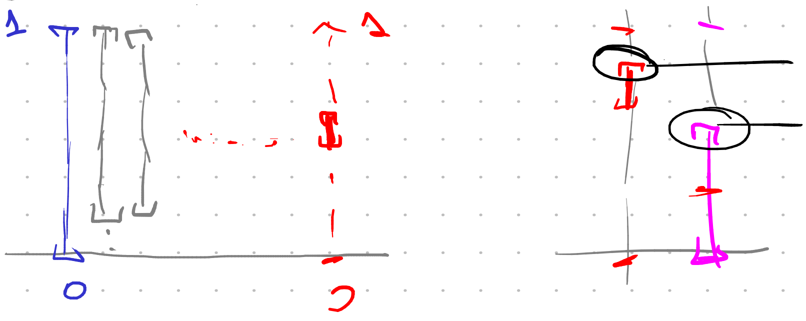
$$= \alpha r_{n+1} + (1-\alpha)(\alpha r_n + (1-\alpha) \hat{\mu}_{n-1}) = \dots$$

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

exponential moving average

→ nonstationary bandits

④ Optimism under uncertainty



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

UCB – Upper Confidence Bound

UCB1: $\text{Priority}_i(t) = \hat{\mu}_i(t) + c \cdot \sqrt{\dfrac{\log t}{n_i}}$

(всеки такт)

(бр. дърп. с i-та ръчка)

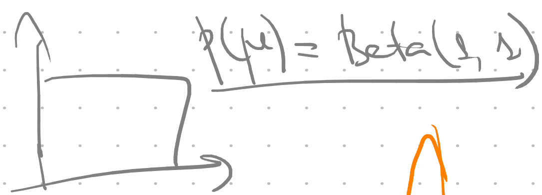
Thm. $c = \sqrt{2} \Rightarrow$

$\text{Regret} = O(\sqrt{K \cdot T \cdot \log T})$, и

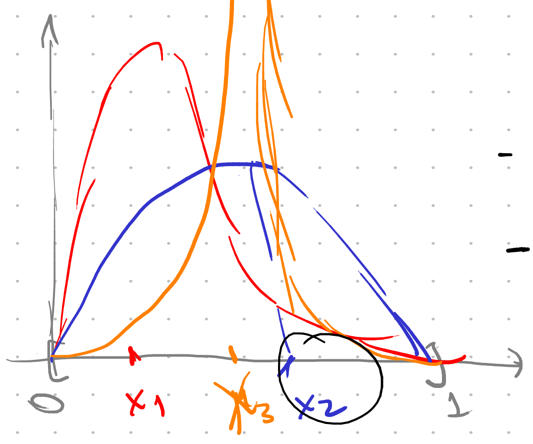
за зависим.-от задачата дефиниции $O(\log T)$ [illegible]

$\Delta_i = |\mu_* - \mu_i|$

Thompson sampling

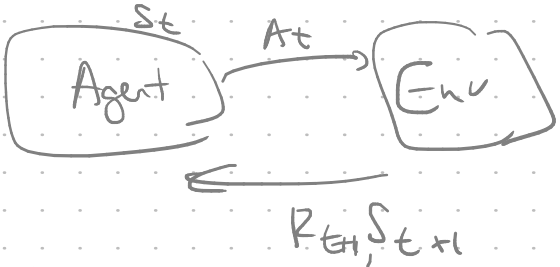


$$\times\ \mu^{w}(1-\mu)^{n-w} \propto p(\mu \mid D) = \text{Beta}(w+1,\ n-w+1)$$



- $x_i \sim p(\mu_i \mid D)$
- $A_t = \arg\max x_i$

(5) Value functions & Bellman equations



$\pi(a|s)$ – стратегия $\quad \pi^*$?

$p(r, s'|s, a)$ – MDP dynamics

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

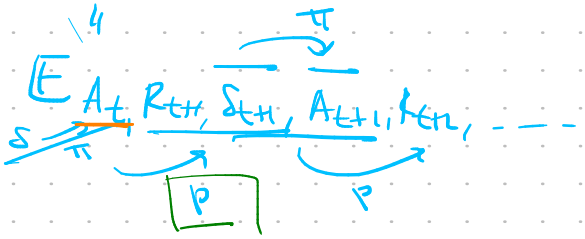
Return – $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

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

(State) value function:

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

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

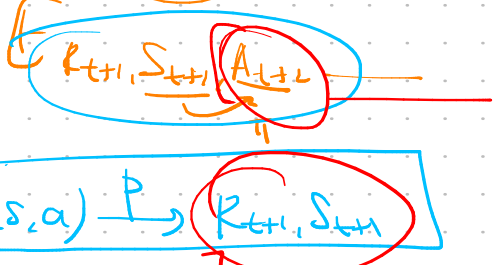


State-action value function:

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

$$Q_*(s, a) = \max_\pi Q_\pi(s, a)$$

$$\pi_*(s) = \arg\max_a Q_*(s, a)$$



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

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

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

$a \in \mathcal{A}(s) \qquad a \in \mathcal{A}$

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

$$= \sum_a \pi(a|s) \cdot Q_\pi(s,a)$$

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

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

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

Алгоритм:

- для каждой $\pi: S \to A$
  - решить $\gamma < 1$, найти $Q_\pi(s,a)$
- выбрать $\pi_*(s) = \arg\max_a \max_\pi Q_\pi(s,a)$

Проблемы:

1) π слишком много
2) $|S|$ слишком много, $|S| \cdot |A|$
3) Часто мы не знаем $p(r,s'|s,a)$

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

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

$|S| \cdot |A|$

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

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

$$\pi_*(s) = \arg\max_a Q_*(s,a)$$

Tabular RL

Approximate RL

$s, a \to$ [ML, θ] $\to Q$

s	V
s_1	$V(s_1)$
s_2	
⋮	
s_N	$V(s_N)$

$V(s_0) = \{ \dots - V(s_i) - \dots \}$

$V(s_1) = \max_{a_1, a_2, a_3} \{ \dots \}$

$V(s_2) = \max \{ - - - \}$

(6) Policy improvement theorem

Если для π, π'

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

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

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

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

$\leq \mathbb{E} [R_{t+1} + \gamma \cdot Q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s, A = \pi'(s)] \leq$

$\leq \dots \mathbb{E} [R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \dots \mid S_t = s, A_t = \pi'(s), A_{t+1} = \pi'(S_{t+1}), \dots]$

$= V_{\pi'}(s)$

Policy iteration: $\pi \longrightarrow Q_\pi \longrightarrow \boxed{\pi'(s) = \arg\max_a Q_\pi(s, a)}$