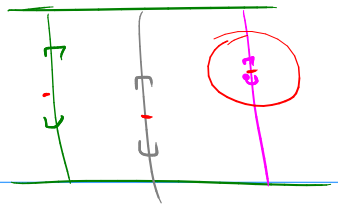
$a_1 \dots a_n$




regret

A/B testing

Contextual bandits

+ context $c \in C$

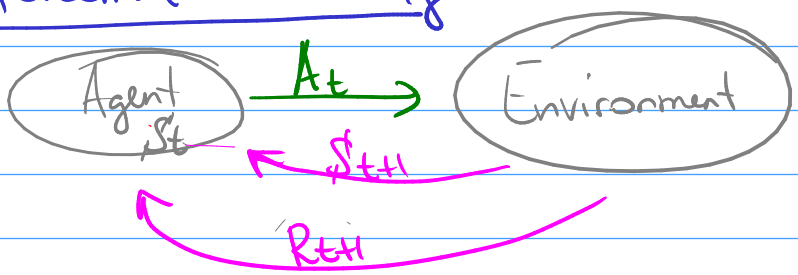
$$\mathbb{E}_p[f] = \int p(x) f(x)\, dx = \int q(x) \left[\frac{p(x)}{q(x)} f(x)\right] dx =$$

$$= \mathbb{E}_q\left[f \cdot \frac{p}{q}\right]$$

importance weights

---

## Reinforcement learning




MDP - Markov decision process

$\mathcal{S}$, $\mathcal{R}$

Trajectory: $S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \dots$ (env / agent)

$\mathcal{A}(s)$

Strategy $\pi(a|s)$ $\quad \pi: \mathcal{S} \to \text{Probs}[\mathcal{A}(s)]$

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

$p(s'|s,a)$

Return $G_t = R_{t+1} + R_{t+2} + \dots + R_T$ $\quad \mathbb{E}[G_t] \to \max$

$R_{t+1}, R_{t+2}, \dots, R \dots$ $\quad \gamma < 1$ (discount)

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

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

$$V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_k \gamma^k R_{t+k+1} \mid S_t = s\right]$$

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

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

$$V_\pi(s) = \mathbb{E}_{a\sim\pi(s)}[Q_\pi(s,a)]$$

$$Q_\pi(s,a) = r(s,a) + \gamma \cdot \mathbb{E}_{s'\sim p(s'|s,a)}[V_\pi(s')]$$

## Bellman equations

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

$$= \mathbb{E}_{a\sim\pi(a|s),\, r,s'\sim p(s',r|a,s)}[R_{t+1} + \gamma \mathbb{E}_\pi[G_{t+1}] \mid S_t = s] =$$

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

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

$$Q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] =$$

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

$$= \mathbb{E}_{a'\sim\pi(a'|s')}[\mathbb{E}_\pi[\ldots]]$$

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

RL? $\pi^* = \arg\max_\pi V_\pi(s)$

$V_\pi(s)$, sc $\S$ $\quad \overline{V_\pi(s)} = A \cdot \overline{V_\pi(s)} + \bar{b}$

$\bar{x} = f(\bar{x}) \qquad \bar{x}^{(k)} = f(\bar{x}^{(k-1)})$

---

① Нельзя перб. $\pi$, когда $V_*, Q_*$

② Не знаем $p(s', r | s, a)$

③ Даже $|S|$ ассимптотический, если записать ур. Беллмана будет

---

optimal value function

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

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

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

PLANNING

$V_*(s) = \max_\pi V_\pi(s) = \max_\pi \mathbb{E}_\pi[G_t | S_t = s] =$

$= \max_a \mathbb{E}\left[R_{t+1} + \gamma \max_\pi V_\pi(S_{t+1})\right]$

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

Bellman optimality equation

$Q_*(s,a) = \max_\pi \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a\right] =$

$= \mathbb{E}\left[R_{t+1} + \gamma \max_\pi V_\pi(S_{t+1}) \mid S_t = s, A_t = a\right]$

$$\max_{a',\pi} Q_\pi(S_{t+1}, a') = \max_{a'} Q_*(S_{t+1}, a')$$

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

$\bar{a} = \pi(\bar{Q})$

## Policy improvement

$\pi \rightsquigarrow \pi'$

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

## Policy improvement theorem

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

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

$\pi \quad \pi'(s) = \arg\max_a Q_\pi(s,a)$

$Q_\pi(s,a)$

Доказ:

$\pi^{(0)} \to \pi^{(1)} \to \pi^{(2)} \to \dots \to \pi_*$

$Q \quad Q \quad \dots$

$$V_\pi(s) \leq Q_\pi(s, \pi'(s)) =$$

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

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

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

$$= \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma \mathbb{E}\left[R_{t+2} + \gamma V_\pi(S_{t+2}) \mid S_{t+1}, A_{t+1} = \pi'(S_{t+1})\right] \mid S_t = s\right]$$

$$= \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 V_\pi(S_{t+2}) \mid S_t = s\right] \leq \dots$$

$$\leq \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s\right] = V_{\pi'}(s)$$

## Policy Iteration

$\pi^{(0)} \to V^{(0)} \to \pi^{(1)} \to V^{(1)} \to \dots$

## Value Iteration

$$V^{(k+1)}(s) = \sum_a \pi^{(k+1)}(a|s) \sum_{s',r} p(s',r|s,a)\left(r + \gamma V^{(k)}(s')\right) =$$

$$= \max_a \sum_{s',r} p(s',r|s,a)\left(r + \gamma V^{(k)}(s')\right)$$

$$\pi^{(k+1)}(s) = \arg\max_a Q_{\pi^{(k)}}(s,a) =$$

$$= \arg\max_a \sum_{s',r} p(s',r|s,a)\left(r + \gamma V^{(k)}(s')\right)$$

$$V^{(k+1)}(s) = \max_a \sum_{s',r} p(s',r|s,a)\left(r + \gamma V^{(k)}(s')\right)$$

Vs

$$\pi(s) = \arg\max_a Q_\pi(s,a)$$

$$V(s) = \max_a \sum p(s',r|s,a)\left(r + \gamma V(s')\right)$$

$$\Rightarrow V = V_*$$

---

(2) $p(s',r|s,a) = ?$ Traj $S_0, A_0, R_1, S_1, A_1, R_2, S_2, \ldots$

Monte Carlo — $V, Q$ оценивать из траекторий

input: $\pi$ $Q(s,a)$ $Ret(s,a)$ output: $V_\pi(s)$

MC prediction
- init $V(s)$, $Returns(s)$
- loop:
  - generate episode $S_0, A_0, R_1, S_1, A_1 \ldots R_T$ по $\pi$
  - compute $G_T = 0$, $G_t = \gamma G_{t+1} + R_{t+1}$
  - append $G_t$ to $Returns(S_t)$ $Ret(S_t, A_t)$
  - $V(S_t) = Avg(Returns(S_t))$ $Ret(S_t, A_t)$

first-visit MC / every-visit MC

Exploring starts

$\pi^{(0)} \to V_{\pi^{(0)}}, Q_{\pi^{(0)}} \to \pi^{(1)} = Greedy(Q_{\pi^{(0)}}) \to Q_{\pi^{(1)}} \to \pi^{(2)} \to \ldots$

loop:
- $\pi(s) := \arg\max_a Q(s,a)$
- $Q(s,a) := MCPred(\pi)$

prediction $\pi \to V_\pi, Q_\pi$

control $\pi$

## MC control

- init $\pi$, $Q$, Ret
- loop!
  - gen. ep. by $\pi$: $S_0, A_0, R_1, \dots, S_{T-1}, A_{T-1}, R_T$
  - compute $G_t$, $t = 0 \dots T-1$
  - ∀t: append $G_t$ to $Ret(S_t, A_t)$

$$Q(S_t, A_t) := Avg(Ret(S_t, A_t))$$

$$\pi(S_t) := \underset{a}{\operatorname{argmax}}\, Q(S_t, a)$$

## Soft policy

### $\varepsilon$-greedy

$$\pi(a|s) = \begin{cases} 1 - \varepsilon + \varepsilon/|A(s)|, & a = a_* \\ \varepsilon/|A(s)|, & \forall\, a \neq a_* \end{cases}$$

$\pi'$ – $\varepsilon$-greedy

$\pi$ – $\varepsilon$-soft  $\pi(a|s) \geq \varepsilon/|A(s)|$ $\forall a$

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

$$= \frac{\varepsilon}{|A(s)|} \cdot \sum_{a \neq a^*} Q_\pi(s,a) + \left(1 - \varepsilon + \frac{\varepsilon}{|A(s)|}\right) Q_\pi(s, a_*) =$$

$$= \max_{a'} Q_\pi(s, a')$$

$$= \frac{\varepsilon}{|A(s)|} \sum_a Q_\pi(s,a) + (1-\varepsilon) Q_\pi(s, a_*) =$$

$$= \frac{\varepsilon}{|A(s)|} \sum_a Q_\pi(s,a) + (1-\varepsilon) \max_{a'} Q_\pi(s,a')$$

$$\sum_a \left(\pi(a|s) - \frac{\varepsilon}{|A(s)|}\right) = 1 - \varepsilon$$

$$V_\pi(s) = Q_\pi(s, \pi|s) = \frac{\varepsilon}{|A(s)|} \sum_a Q_\pi(s,a) + \sum_a \left(\pi(a|s) - \frac{\varepsilon}{|A(s)|}\right) \cdot Q_\pi(s,a)$$

$$(1-\varepsilon) \cdot \sum_a \left(\frac{\pi(a|s) - \varepsilon/|A(s)|}{1-\varepsilon}\right) \cdot Q_\pi(s,a)$$

$$Q_\pi(s, \pi'(s)) \geq V_\pi(s)$$