$\pi(a|s,\bar\theta)$ $\quad J(\bar\theta)$

$$\bar\theta_{t+1} = \bar\theta_t + \alpha \nabla_{\bar\theta} J(\bar\theta)\Big|_{\bar\theta=\bar\theta_t}$$

$\bar\theta^T \bar x(s,a)$

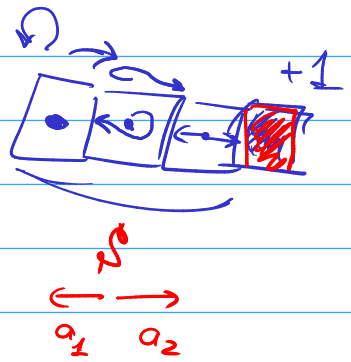
$=$ $h(s,a;\bar\theta)$ — "score"

$$\pi(a|s,\bar\theta) = \text{softmax}(h) = \frac{e^{h(s,a,\bar\theta)}}{\sum_{a'} e^{h(s,a',\bar\theta)}}$$

$\rightsquigarrow$ greedy

Можем

1) Сход. к детерм. страт.
2) π м.б. проще опт., чем Q
3) Можем сход. к стохаст. стратегии



Policy gradient theorem

$$J(\theta) = V_{\pi_\theta}^{\text{опт.}}(s_0)$$

$\gamma = 1$

$$\nabla_\theta V_\pi(s) = \nabla_\theta \left[\sum_a \pi_\theta(a|s) Q_{\pi_\theta}(s,a)\right] =$$

$$= \sum_a \left[\nabla_\theta \pi_\theta(a|s)\cdot Q_{\pi_\theta}(s,a) + \pi_\theta(a|s)\cdot \nabla_\theta Q_{\pi_\theta}(s,a)\right] =$$

$$= \sum_a \left[\nabla_\theta \pi_\theta(a|s)\cdot Q_{\pi_\theta}(s,a) + \pi_\theta(a|s)\cdot \nabla_\theta \sum_{r,s'} p(s',r|s,a)\left(r + V_\pi(s')\right)\right]$$

$$= \sum_a \left[-''- + \pi_\theta(a|s)\cdot \nabla_\theta \sum_{s'} p(s'|s,a)\cdot V_\pi(s')\right]$$

$$= \sum_a \left[\nabla_\theta \pi_\theta(a|s)\cdot Q_{\pi_\theta}(s,a) + \pi_\theta(a|s)\cdot \sum_{s'} p(s'|s,a)\cdot \nabla_\theta V_\pi(s')\right]$$

$\sum_a \pi(a|s) \sum_{s'} p(s'|s,a)$

$\sum_{a'} \left[\nabla \pi(a'|s') \cdot \ldots\right]$

$$= \sum_{s'} \left(\sum_{k=0}^{\infty} \Pr_{\pi}[s \to s' \text{ за } k \text{ шагов}] \right) \sum_a \nabla_\theta \pi_\theta(a|s') \cdot Q_\pi(s', a)$$

$\propto \mu(s)$

$$\nabla_\theta J(\theta) = \nabla_\theta V_\pi(s_0) = \sum_s \left(\sum_{k=0}^{\infty} \Pr[s_0 \to s \text{ за } k \text{ ш.}] \right) \left(\sum_a \nabla_\theta \pi(a|s) Q(s,a) \right)$$

сумм. врем. пров. в s

$\mu_\pi(s)$ = относ.-доля врем. в сост. s

Policy grad thm

$$\nabla_\theta J(\theta) \propto \sum_s \mu_\pi(s) \cdot \sum_a Q_{\pi_\theta}(s,a) \cdot \nabla_\theta \pi_\theta(a|s) =$$

$$= \mathbb{E}_\pi \left[\sum_a Q_\pi(S_t, a) \nabla_\theta \pi_\theta(a|S_t) \right]$$

1) „All-actions" алгоритм $\pi: S_0, A_0, S_1, A_1$ —

$$\bar{\theta}_{t+1} = \bar{\theta}_t + \alpha \sum_a \hat{Q}(S_t, a, \bar{w}) \cdot \nabla_\theta \pi_\theta(a|S_t)$$

2) REINFORCE (1992)

$\sum_a \pi(a|S_t) \cdot (\ldots)$

$$\nabla J(\theta) \propto \mathbb{E}_\pi \left[\sum_a Q_\pi(S_t, a) \nabla_\theta \pi_\theta(a|S_t) \cdot \frac{\pi_\theta(a|S_t)}{\pi_\theta(a|S_t)} \right] =$$

$$= \mathbb{E}_\pi \left[Q_\pi(S_t, A_t) \cdot \frac{\nabla_\theta \pi_\theta(A_t|S_t)}{\pi_\theta(A_t|S_t)} \right] =$$

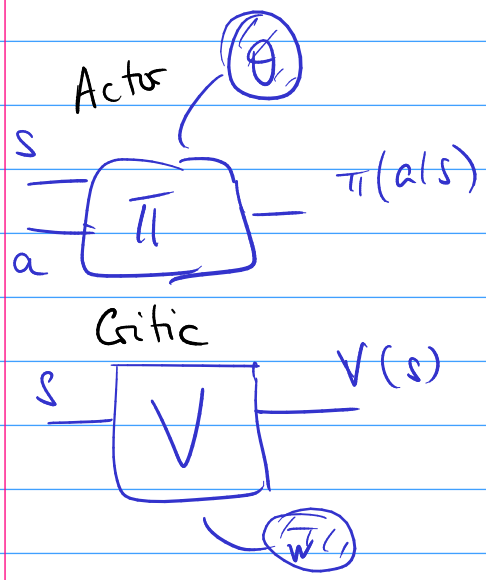
" $\mathbb{E}_\pi[G_t]$ " $\nabla_\theta \ln \pi_\theta(A_t|S_t)$

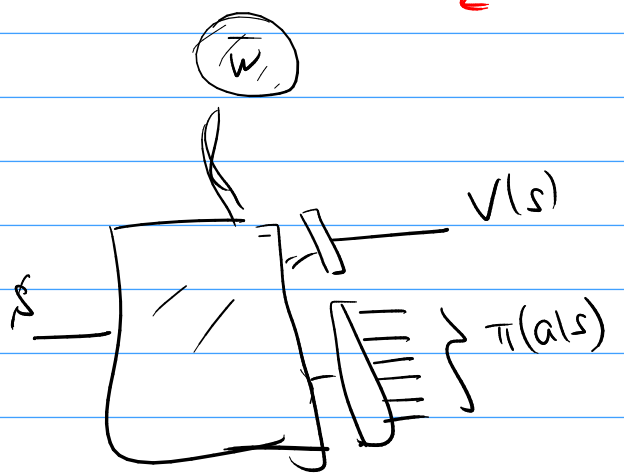
$$\boxed{\nabla J(\theta) \propto \mathbb{E}_\pi \left[G_t \cdot \nabla_\theta [\ln \pi_\theta(A_t|S_t)] \right]}$$

REINFORCE: $\theta_{t+1} = \theta_t + \alpha \cdot G_t \cdot \nabla_\theta \ln \pi_\theta(A_t|S_t) =$

$\pi \sim \dots S_t, A_t, G_t, -$

$= \theta_t + \alpha \cdot \frac{G_t}{\pi(A_t|S_t)} \cdot \nabla_\theta \pi(A_t|S_t)$

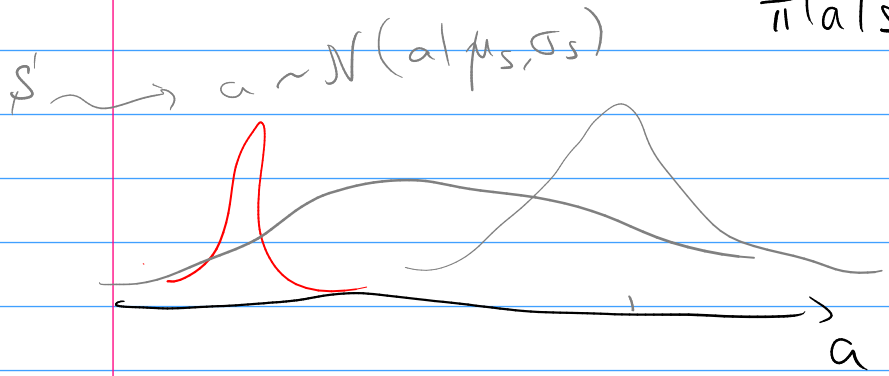




$$\pi(a|s,\theta) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{1}{2\sigma^2}(a-\mu)^2}$$

$\mu(s,\theta)$

$\sigma(s,\theta)$



$F(\theta) \to \max$

$g(\theta) \le \beta$

$F(\theta) - \lambda g(\theta) \to \max$