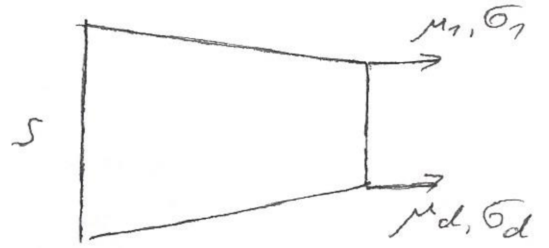
07. 12. 18 dl

Policy gradients methods, $\pi(a|s) = \pi(a|s,\theta)$, $a \in \mathbb{R}^d$ $\swarrow$



$$\pi(a|s,\theta) = \prod_{j=1}^{d} N(a_j | \mu_j(s,\theta), \sigma_j^2(s,\theta))$$

$$a \in \{1 \dots k\},\ \pi(a|s,\theta) = sm(output(s,\theta))$$

$$F(\theta) = \mathbb{E}_{p(s)} \mathbb{E}_{\pi(a|s,\theta)} Q^{\pi_\theta}(s,a) \rightarrow \max_\theta$$

$$p(\tau|\theta) = p(s_0) \prod_{j=0}^{\infty} p(s_{j+1}|s_j,a_j)\pi(a_j|s_j,\theta)$$

$$\nabla_\theta F(\theta) = \nabla_\theta \overset{?}{\curvearrowright} \mathbb{E}_{p(\tau|\theta)} f(\tau),\ \text{log-der-trick}$$

$$\nabla_\theta F(\theta) = \nabla_\theta \int p(\tau|\theta) f(\tau) d\tau = \int \nabla_\theta p(\tau|\theta) f(\tau) d\tau =$$

$$= \left\{ \nabla_\theta \log p(\tau|\theta) = \frac{1}{p(\tau|\theta)} \nabla_\theta p(\tau|\theta) \right\} = \int p(\tau|\theta) \nabla_\theta \log p(\tau|\theta) f(\tau) d\tau =$$

$$= \mathbb{E}_{p(\tau|\theta)} \nabla_\theta \log p(\tau|\theta) f(\tau)$$

$? \pi(s_t,\theta) \leftarrow$ дет.

$//$ $\swarrow$ случ.

$$\nabla_\theta \log p(\tau|\theta) = \nabla_\theta \log \left(p(s_0) \prod_{t=0}^{\cancel{\infty} \nearrow T} p(s_{t+1}|s_t,a_t)\pi(a_t|s_t,\theta)\right) =$$

$$= \sum_{t=0}^{\cancel{\infty} \nearrow T} \nabla_\theta \log \pi(a_t|s_t,\theta)$$

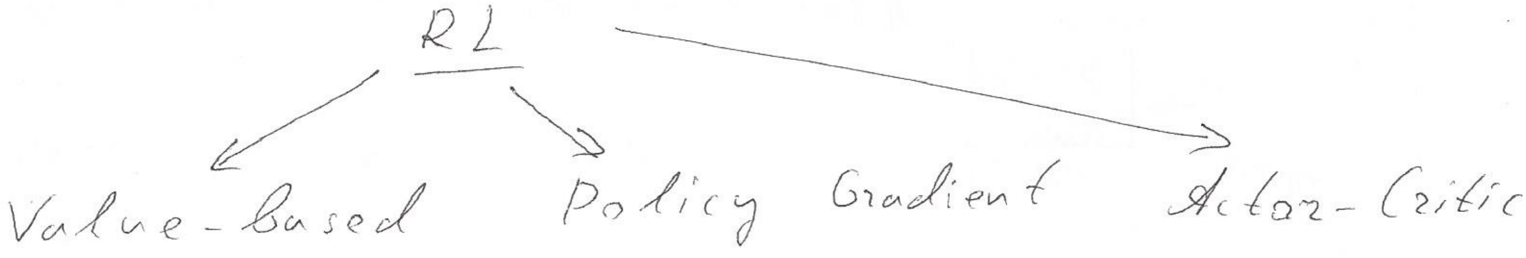
Алгоритм REINFORCE

```
Иниц. θ
повторять
    $\tau = \{s_0, a_0, s_1, a_1, \dots, s_T\}$
    $\nabla F(\theta) = \left(\sum_{t=0}^{T} \nabla_\theta \log \pi(a_t, s_t|\theta)\right)\left(\sum_{t=0}^{T} \gamma^t r(s_t,a_t)\right)$
    $\theta \leftarrow \theta + \alpha \nabla_\theta F(\theta)$
```

огромная дисперсия из-за log-der-trick

$$\mathbb{E}_{p(\tau|\theta)} \nabla_\theta \log p(\tau|\theta) = \int p(\tau|\theta) \frac{1}{p(\tau|\theta)} \nabla_\theta p(\tau|\theta)\, d\tau = 0$$

Baseline B



Actor: $\pi(a|s,\theta)$

Critic: $Q(s,a|w) \approx Q^{\pi}(s,a)$

$$F(\theta) = \mathbb{E}_{p(s)} \mathbb{E}_{\pi(a|s,\theta)} Q(s,a|w) \to \max_{\theta}$$

$$Q^{\pi}(s,a) = r(s,a) + \gamma\, \mathbb{E}_{p(s'|s,a)} \mathbb{E}_{\pi(a'|s')} Q^{\pi}(s',a')$$

$$G(w) = \frac{1}{|S||A|} \sum_{s,a} \left(Q(s,a|w) - \left[r(s,a) + \gamma\, \mathbb{E}_{p(s'|s,a)} \mathbb{E}_{\pi(a'|s',\theta)} Q(s',a'|w) \right] \right)^2 \to \min_{w}$$

Алгоритм QAC

```
Иниц. $\theta, w$                                   1) memory replay
повторять                                           2) parallel learning
    сэмпл $(s, a, r, s')$ $\leftarrow$
    $a' \sim \pi(a'|s', \theta)$
    $y = r(s,a) + \gamma Q(s', a'|w)$
    $\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi(a|s,\theta) Q(s,a|w)$
    $w \leftarrow w - \beta \cdot 2 (Q(s,a|w) - y) \nabla_w Q(s,a|w)$
```

Baseline: $\nabla_\theta F(\theta) = \nabla_\theta \mathbb{E}_{p(s)} \mathbb{E}_{\pi(a|s,\theta)} (Q(a,s|w) - B(s))$

$$\nabla_\theta \mathbb{E}_{p(s)} \mathbb{E}_{\pi(a|s,\theta)} B(s) = \nabla_\theta \int p(s) \int \pi(a|s,\theta) B(s)\, da\, ds =$$

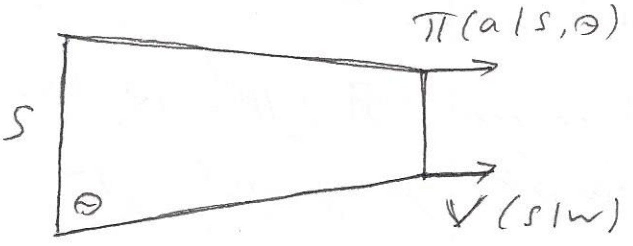
$$= \int p(s) B(s) \Big(\nabla_\theta \underbrace{\int \pi(a|s,\theta)\, da}_{1} \Big) ds = 0$$

$$\underset{\text{advantage}}{A^{\pi}}(s,a) = Q^{\pi}(s,a) - \underbrace{V^{\pi}(s)}_{B(s)}$$

$$Q^{\pi}(s,a) = r(s,a) + \gamma\, \mathbb{E}_{p(s'|s,a)} V^{\pi}(s'')$$

Схема A2C

```
Инициализация $\theta, w$
повторять
  сэмплы $(s, a, r, s')$
  $y = r(s,a) + \gamma V(s'|w)$
  $A(s,a) = y - V(s|w)$
  $\theta \leftarrow \theta - \alpha \nabla_{\theta} \log \pi(a|s,\theta) A(s,a)$
  $w \leftarrow w - \beta 2 (V(s|w) - y) \nabla_{w} V(s|w)$
```



Процессы Дирихле

$$\begin{cases} f(x) \sim GP(0, k(\cdot,\cdot)) \\ y_i | f(x_i) \sim p(y|f) \end{cases} \Rightarrow p(f|y,x)$$

$$p(y_{test} | x_{test}) = \int p(y_{test} | f(x_{test})) \cdot p(f|y,x)\, df$$

непараметрическая модель, зависит от числа объектов

$$\begin{cases} \pi \sim Dir(\pi|\alpha) \\ \theta_1, \dots \theta_K \sim p(\theta) \\ z_1 \dots z_N \sim Discrete(z|\pi) \\ x_i | z_i, \theta \sim p(x|\theta_{z_i}) \end{cases}$$

$$p(X, Z, \theta, \pi) = \underbrace{p(\theta) p(\pi)}_{\text{непараметрическая модель ?}} p(Z|\pi) p(X|Z,\theta)$$

(здесь нет, но нужно сделать)