ТюмГУ

# Обучение с подкреплением.

## Методы градиента политик. Актер-критик

Сергей Аксёнов,
Доцент отделения информационных технологий
Инженерной школы информационных технологий и робототехники
Томский политехнический университет

Тюмень-2024

# Цели ценностно-ориентированных и стратегических методов

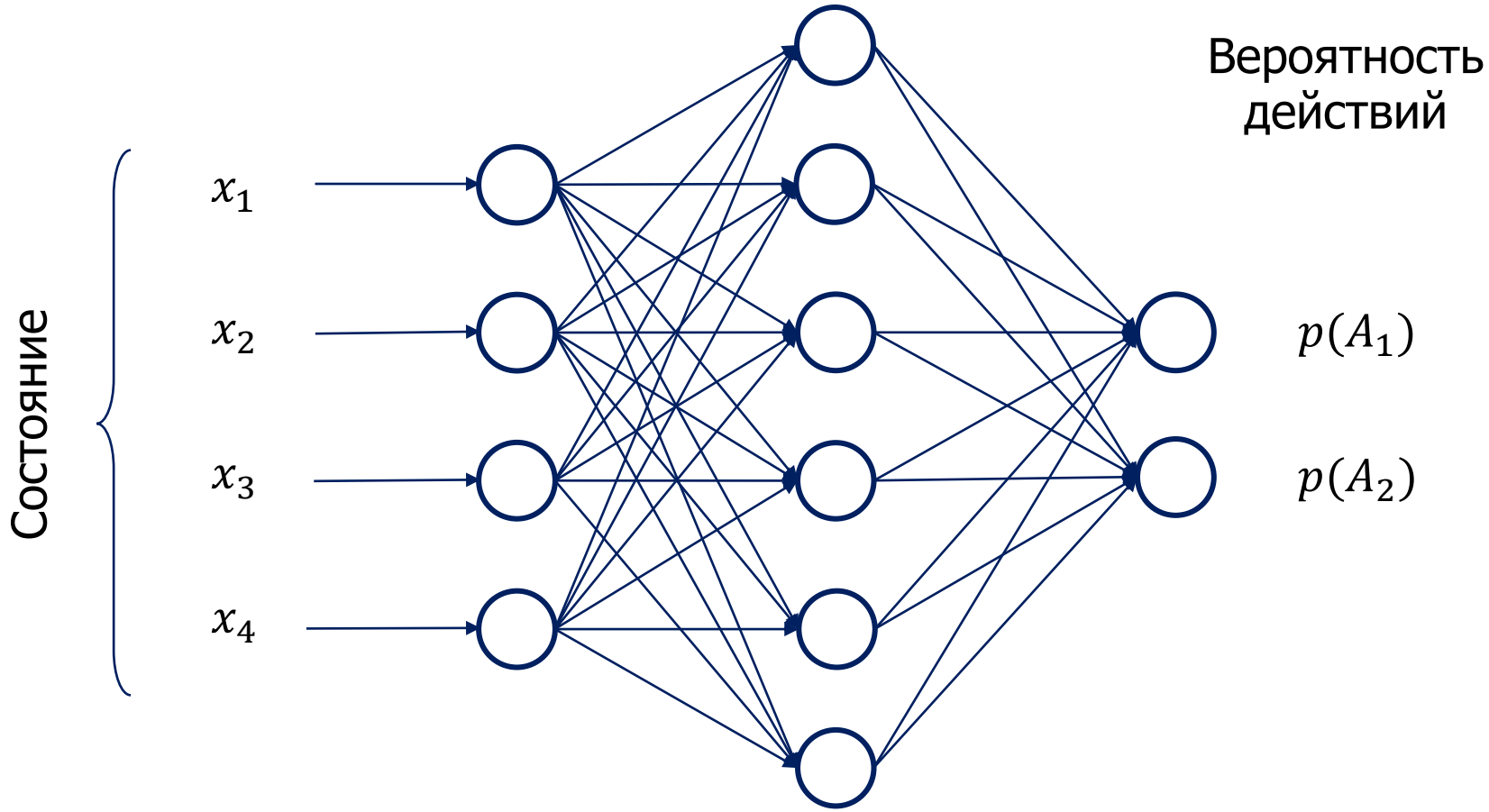
Цель ценностно-ориентированных методов:

$$L_i(\theta_i) = \mathbb{E}_{s,a}\left[\left(q_*(s,a) - Q(s,a;\theta_i)\right)^2\right]$$

Цель стратегических методов:

$$J(\theta) = \mathbb{E}_{s_0 \sim p_0}\left[v_{\pi_\theta}(s_0)\right]$$

# Сеть политики




Архитектура сети определяется особенностями данных

# Градиент политик - 1

Упрощенное уравнение цели: $J(\theta) = \mathbb{E}_{s_0 \sim p_0}\left[v_{\pi_\theta}(s_0)\right]$

Градиент функции: $\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{s_0 \sim p_0}\left[v_{\pi_\theta}(s_0)\right]$

Полная траектория: $\tau = S_0, A_0, R_1, S_1, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$

Выгода полной траектории: $G(\tau) = R_1 + \gamma R_2 + \cdots \gamma^{T-1} R_T$

Вероятность траектории:

$$p(\tau|\pi_\theta) = p_0(S_0)\pi(A_0|S_0;\theta)P(S_1, R_1|S_0, A_0) \ldots P(S_T, R_T|S_{T-1}, A_{T-1})$$

# Градиент политик - 2

Цель: $$\nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}[G(\tau)] = \nabla_\theta \mathbb{E}_{s_0\sim p_0}\left[v_{\pi_\theta}(s_0)\right]$$

Градиентная оценка функции вклада:

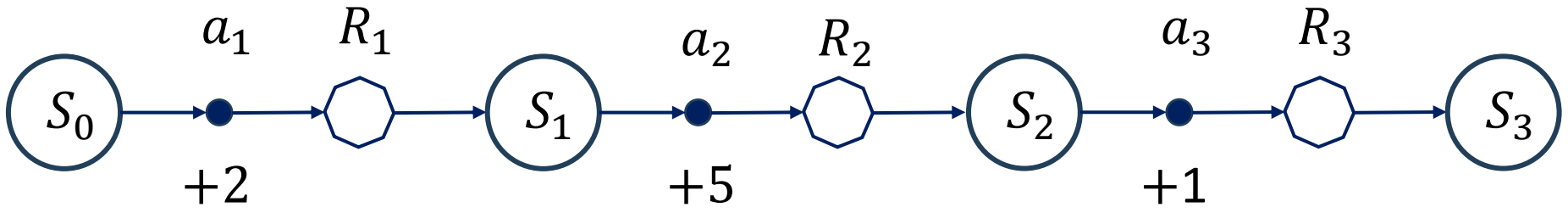
$$\nabla_\theta \mathbb{E}_x[f(x)] = \mathbb{E}_x[\nabla_\theta \log p\,(x|\theta) f(x)]$$

Обновление цели: $$\nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}[G(\tau)] = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log p\,(\tau|\pi_\theta) G(\tau)]$$

Окончательное выражение:

$$\nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}[G(\tau)] = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi\,(A_t|S_t) G(\tau)\right]$$

# Пример: Использование полученных наград




$\gamma = 0.99$

$G(\tau) = 2 + 0.99 \times 5 + 0.99^2 1 = 7.9301$

a) Использование выгоды всего эпизода:

$$7.9301 \times \nabla_\theta \log \pi_\theta (A_1|S_0) + 7.9301 \times \nabla_\theta \log \pi_\theta (A_2|S_1) + 7.9301 \times \nabla_\theta \log \pi_\theta (A_2|S_2)$$

b) Использование выгоды от каждого состояния:

$$7.9301 \times \nabla_\theta \log \pi_\theta (A_1|S_0) + 5.99 \times \nabla_\theta \log \pi_\theta (A_2|S_1) + 1 \times \nabla_\theta \log \pi_\theta (A_2|S_2)$$

# REINFORCE: Обучение политике, основанное на выгоде

1. Инициализировать веса сети $\theta$ случайными значениями

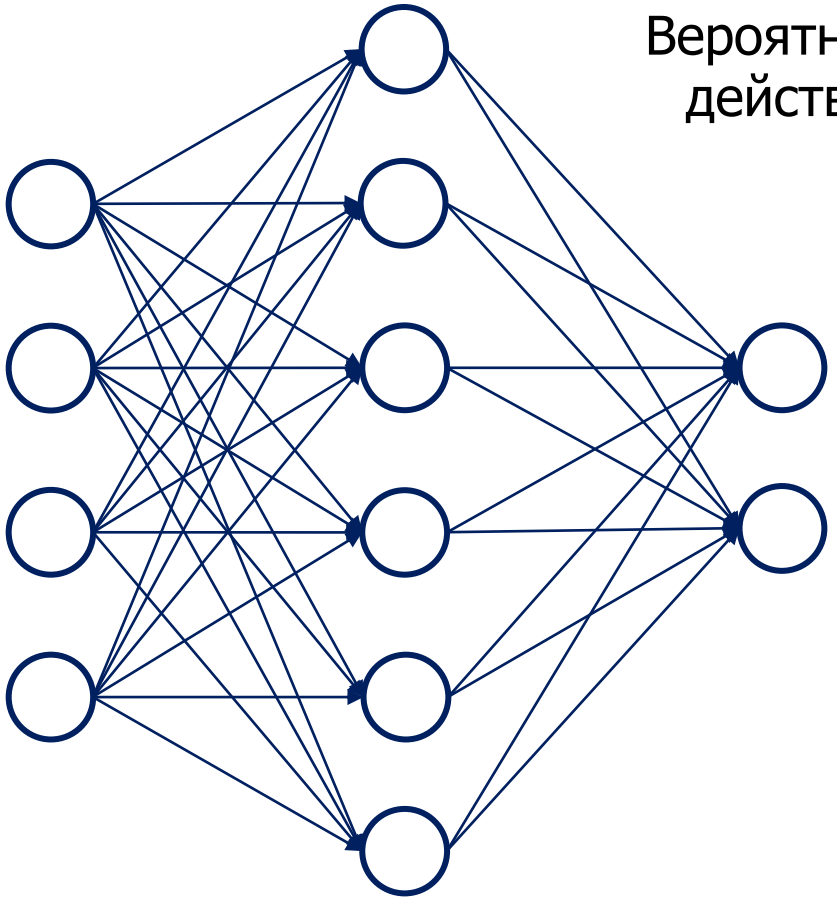
2. Сгенерировать $N$ траекторий $\{\tau^i\}_{i=1}^{N}$ согласно политике $\pi_\theta$
3. Рассчитать выгоду траектории $G(\tau)$
4. Вычислить градиент

$$\nabla_\theta J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t|s_t)G(\tau)\right]$$

5. Выполнить настройку параметров сети $\theta = \theta + \alpha\nabla_\theta J(\theta)$
6. Повторять шаги 2. - 5. до получения приемлемых результатов

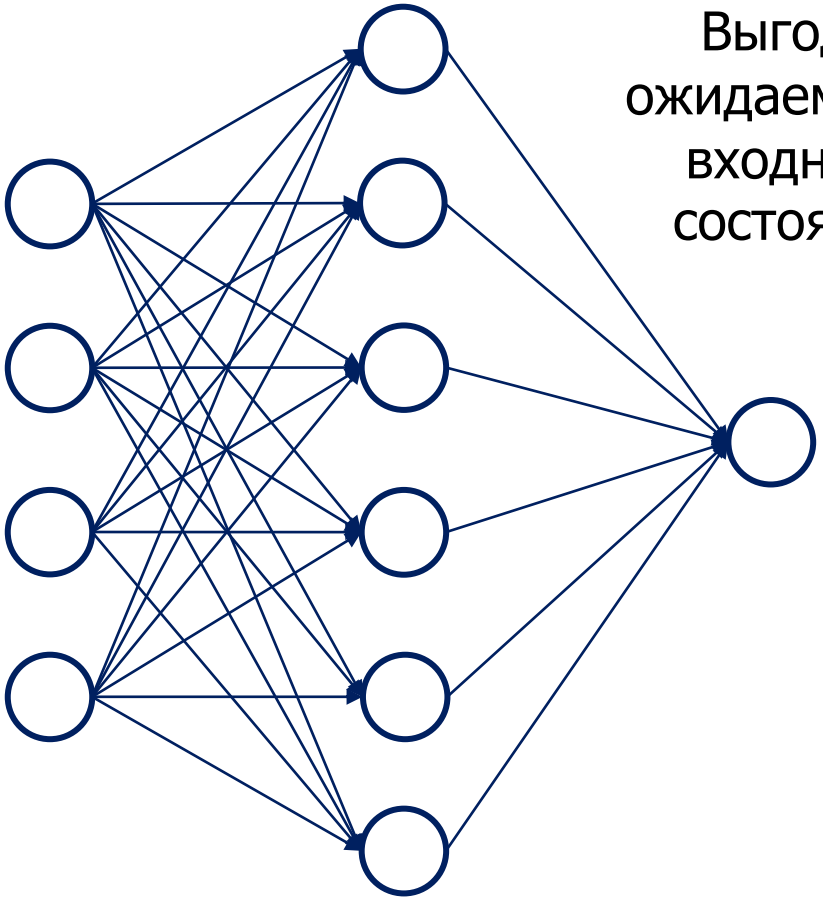
# Базовый градиент политик (VPG – REINFORCE with Baseline)

Сеть политики



Вероятность действий

Сеть функции ценности



Выгода, ожидаемая от входного состояния

# Базовый градиент политик (Vanilla Policy Gradient)

✓ Сокращение дисперсии функции потерь

Потери для функции ценности:

Базовый уровень

$$L_v(\phi) = \frac{1}{N}\sum_{n=0}^{N}\left[\left(G_t - V(S_t;\phi)\right)^2\right]$$

Потери политики:

Логарифм вероятности выбранного действия

$$L_\pi(\theta) = -\frac{1}{N}\sum_{n=0}^{N}\left[\left(G_t - V(S_t;\phi)\right)\log\pi(A_t|S_t;\theta) + \beta H\left(\pi(S_t;\theta)\right)\right]$$

Прогнозируемое преимущество

Взвешенная энтропия

# REINFORCE с направляющей (VPG)

1. Инициализировать параметры сети политики $\theta$ и параметры сети ценности $\phi$
2. Сгенерировать $N$ траекторий $\{\tau^i\}_{i=1}^{N}$ согласно политике $\pi_\theta$
3. Рассчитать выгоду $G_t$
4. Вычислить градиент

$$\nabla_\theta J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t|s_t)(G_t - V_\phi(s_t))\right]$$

5. Обновить параметры сети политик $\theta$ с использованием градиентного подъёма

$$\theta = \theta + \alpha\nabla_\theta J(\theta)$$

6. Рассчитать MSE для сети ценности

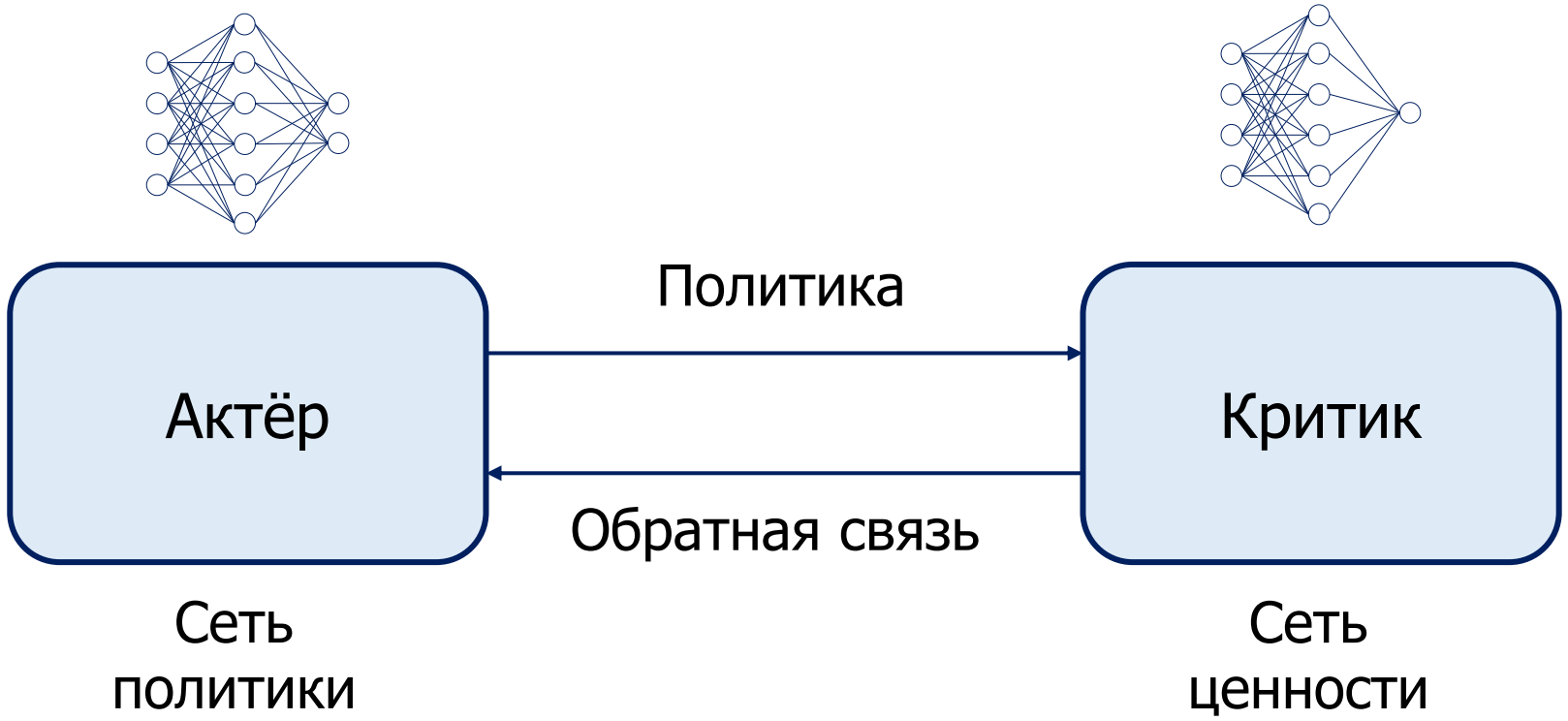
$$J(\phi) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\left(G_t - V_\phi(s_t)\right)^2$$

7. Обновить параметры сети ценности с помощью градиентного спуска

$$\phi = \phi - \alpha\nabla_\phi J(\phi)$$

8. Повторять шаги 2.-7. до получения приемлемых результатов

# «Актёр-критик»




- ✓ Бутстрэппинг, аналогичный тому, который применяется в методе TD
- ✓ Обновление параметров сетей политики и ценности на каждом переходе эпизода

# Алгоритм «Актёр-критик»

1. Инициализировать параметры сети актёра $\theta$ и сети критика $\phi$ случайными значениями.
2. Выполнить $N$ эпизодов, каждый из которых состоит из шага 3.
3. Для каждого шага в эпизоде $t = 0, \dots, T-1$
    a. Выбрать действие согласно политике $a_t \sim \pi_\theta(s_t)$
    b. Получить кортеж опыта $(s_t, a_t, r_t, s'_t)$
    c. Вычислить градиент политик:

$$\nabla_\theta J(\theta) = \nabla_\theta \log \pi_\theta(a_t|s_t)\left(r + \gamma V_\phi(s'_t) - V_\phi(s_t)\right)$$

    d. Обновить параметры сети актёра $\theta$ с помощью градиентного подъёма:

$$\theta = \theta + \alpha \nabla_\theta J(\theta)$$

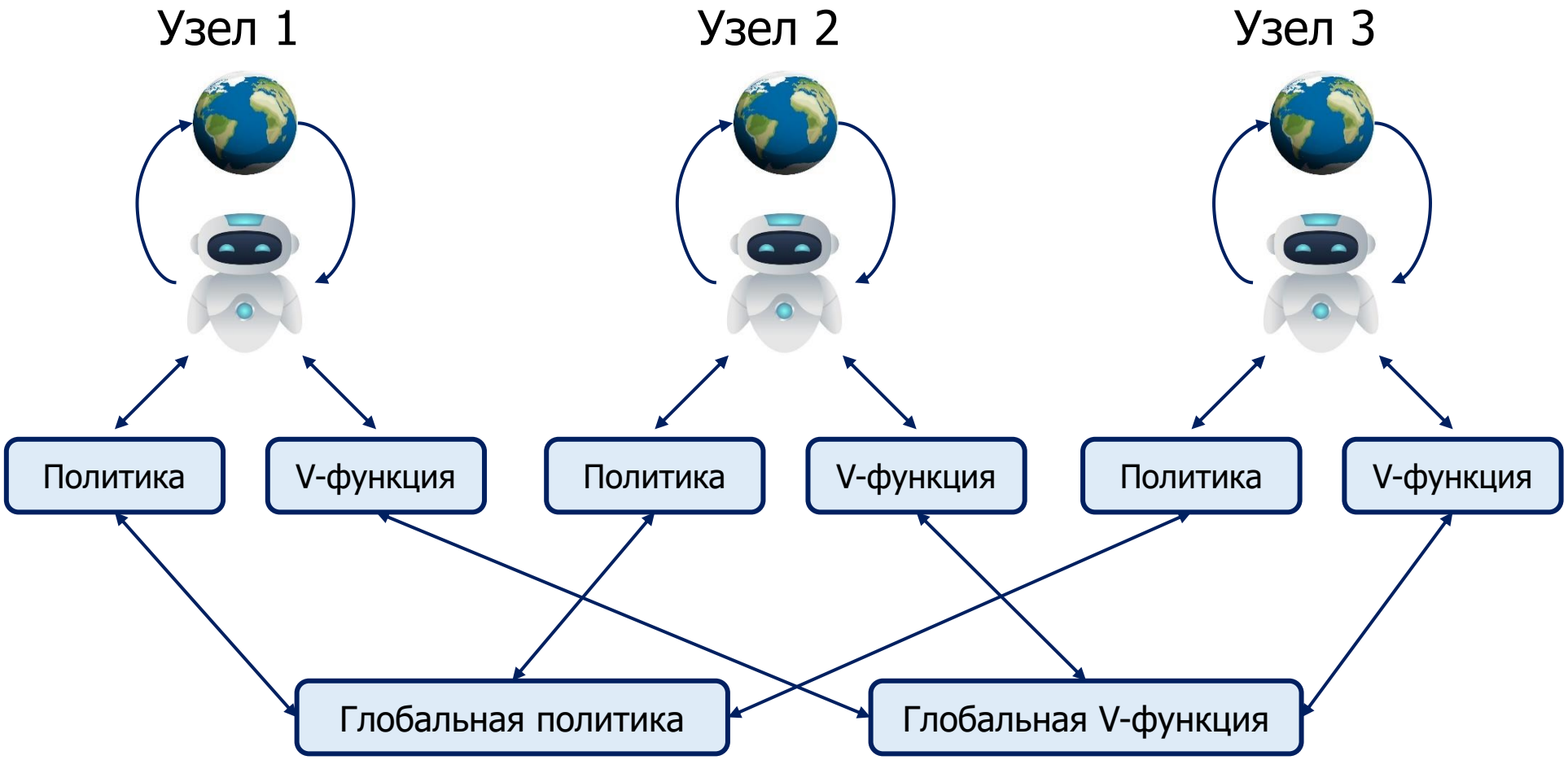
    e. Вычислить потери сети критика:

$$J(\phi) = r + \gamma V_\phi(s'_t) - V_\phi(s_t)$$

    f. Вычислить градиенты $\nabla_\phi J(\theta)$ и обновить параметра сети критика $\phi$ с помощью градиентного спуска:

$$\phi = \phi - \alpha \nabla_\phi J(\theta)$$

# Асинхронный алгоритм «актёр-критик» с преимуществом (A3C)




*Использованы изображения из freepik*

# Обобщённое программирование преимущества (GAE)

$$A^1(S_t, A_t; \phi) = R_t + \gamma V(S_{t+1}; \phi) - V(S_t; \phi)$$

$$A^2(S_t, A_t; \phi) = R_t + \gamma R_{t+1} + \gamma^2 V(S_{t+2}; \phi) - V(S_t; \phi)$$

$$A^3(S_t, A_t; \phi) = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \gamma^3 V(S_{t+3}; \phi) - V(S_t; \phi)$$

$$A^n(S_t, A_t; \phi) = R_t + \gamma R_{t+1} + \cdots + \gamma^n R_{t+n} + \gamma^{n+1} V(S_{t+n+1}; \phi) - V(S_t; \phi)$$

Аналог TD(λ) для преимуществ: $A^{GAE(\gamma,\lambda)}(S_t, A_t; \phi) = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}$

$$A^{GAE(\gamma,0)}(S_t, A_t; \phi) = R_t + \gamma V(S_{t+1}; \phi) - V(S_t; \phi)$$

$$A^{GAE(\gamma,1)}(S_t, A_t; \phi) = \sum_{l=0}^{\infty} \gamma^l R_{t+l} - V(S_t; \phi)$$

# Использование оценок на основе n-шагового бутстрэппинга

Функция преимущества: $A(S_t, A_t; \phi) = G_t - V(S_t; \phi)$

Функция преимущества с n-шаговой выгодой с бутстрэппом:

$$A(S_t, A_t; \phi) = R_t + \gamma R_{t+1} + \cdots + \gamma^n R_{t+n} + \gamma^{n+1} V(S_{t+n+1}; \phi) - V(S_t; \phi)$$

Функция потерь политики с прогнозом n-шаговой выгоды:

$$L_\pi(\theta) = -\frac{1}{N} \sum_{n=0}^{N} \left[ A(S_t, A_t; \phi) \log \pi(A_t | S_t; \theta) + \beta H\big(\pi(S_t; \theta)\big) \right]$$

Функция потерь ценности:

$$L_v(\phi) = \frac{1}{N} \sum_{n=0}^{N} \left[ \left( R_t + \gamma R_{t+1} + \cdots + \gamma^n R_{t+n} + \gamma^{n+1} V(S_{t+n+1}; \phi) - V(S_t; \phi) \right)^2 \right]$$

# Потенциальные способы оценки градиента политик

Анализируемый градиент: $$g = \mathbb{E}\left[\sum_{t=0}^{\infty} \Psi_t \nabla_\theta \log \pi(A_t|S_t;\theta)\right]$$

Общая выгода: $$\Psi_t = \sum_{t=0}^{T} \gamma^t R_t$$

Выгода с текущего шага: $$\Psi_t = \sum_{t'=t}^{T} \gamma^{t'-t} R_{t'}$$

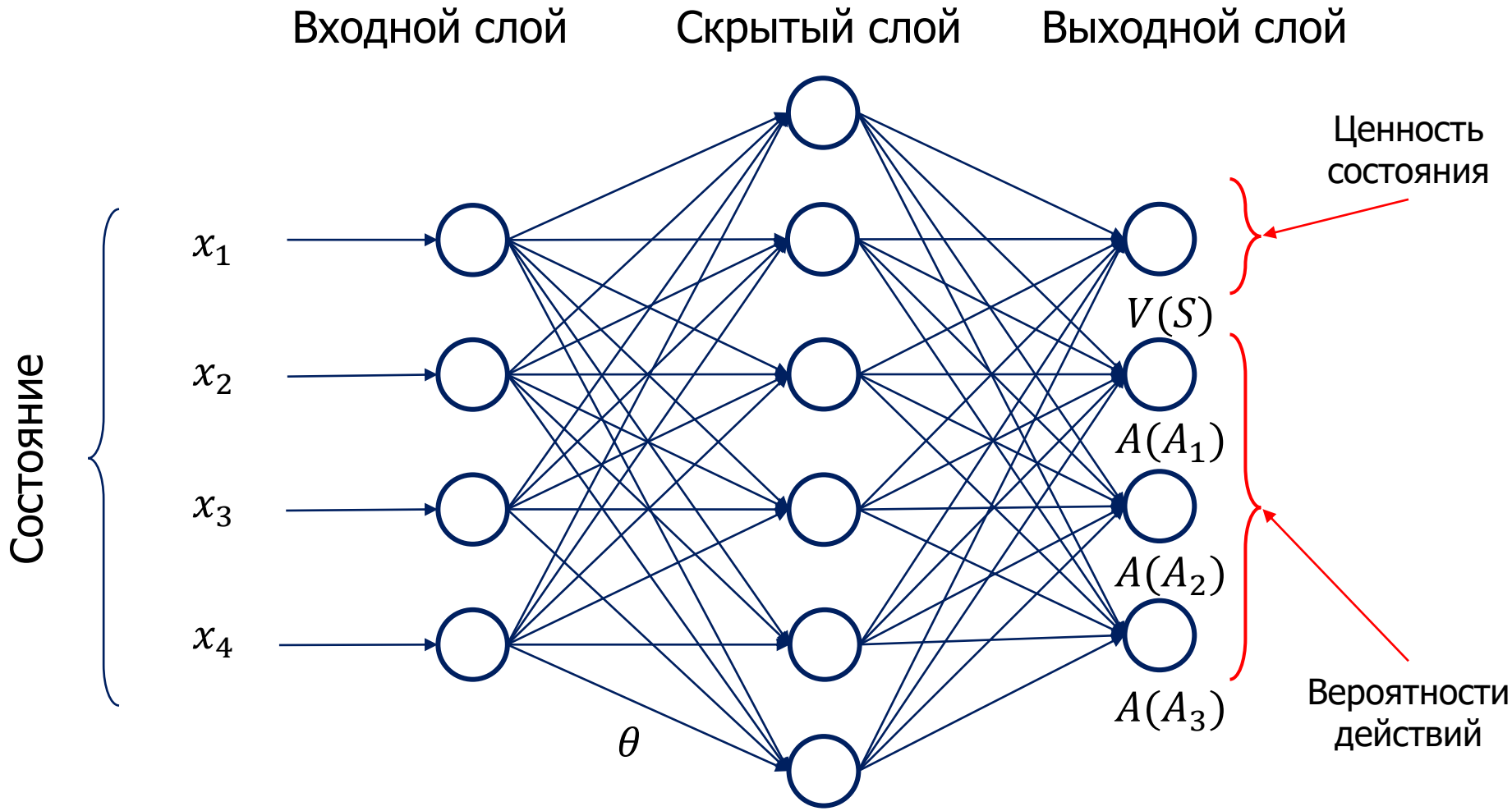
Использование направляющей: $$\Psi_t = \sum_{t'=t}^{T} \gamma^{t'-t} R_{t'} - b(S_t)$$

N-шаговая оценка преимущества: $$\Psi_t = a_\pi(S_t, A_t)$$

Функция ценности действий: $$\Psi_t = q_\pi(S_t, A_t)$$

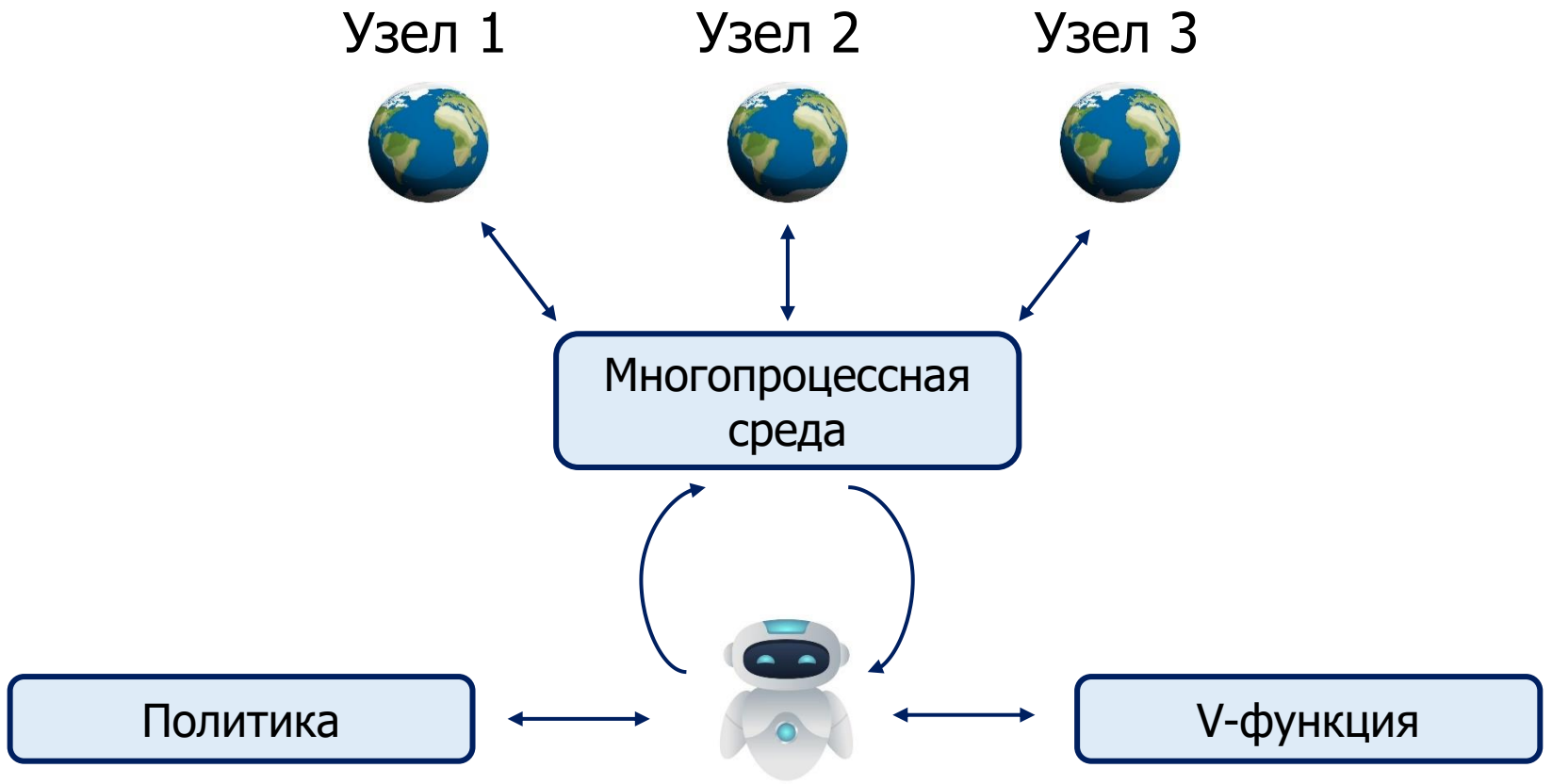
Погрешность TD: $$\Psi_t = R_t + v_\pi(S_{t+1}) - v_\pi(S_t)$$

# Разделение весов между выходными политиками и ценностью




Архитектура сети определяется особенностями данных

# A2C: Синхронная версия A3C




*Использованы изображения из freepik*

# Глубокий градиент по детерминированным политикам (DDPG)

Функция потерь DQN для Q-функции:

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{U}(\mathcal{D})}\left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta_i)\right)^2\right]$$

Функция потерь с argmax:

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{U}(\mathcal{D})}\left[\left(r + \gamma Q\left(s', \operatorname{argmax}_{a'} Q(s', a', \theta^-); \theta^-\right) - Q(s', a'; \theta_i)\right)^2\right]$$

Функция потерь DDPG с сетью стратегий $\mu$:

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{U}(\mathcal{D})}\left[\left(r + \gamma Q(s', \mu(s', \phi^-); \theta^-) - Q(s', a'; \theta_i)\right)^2\right]$$

Функция потерь детерминированной политики в DDPG:

$$J_i(\phi_i) = \mathbb{E}_{s\sim\mathcal{U}(\mathcal{D})}[Q(s, \mu(s, \phi); \theta)]$$

# Двойное обучение в DDPG (TD3)

$$J_i(\theta_i^a) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{U}(\mathcal{D})}\left[\left(\mathcal{TWIN}^{target} - Q(s,a;\theta_i^a)\right)^2\right]$$

$$J_i(\theta_i^b) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{U}(\mathcal{D})}\left[\left(\mathcal{TWIN}^{target} - Q(s,a;\theta_i^b)\right)^2\right]$$

$$\mathcal{TWIN}^{target} = r + \gamma \min_n Q(s';\mu(s';\phi^-);\theta^{n,-})$$

# Глубокий детерминированный градиент политик (DDPG)

1. Инициализировать веса главных сетей критика $\theta$ и актёра $\phi$
2. Инициализировать параметры целевой сети критика $\theta'$ копированием параметров главной сети критика $\theta$
3. Инициализировать параметры целевой сети актёра $\phi'$ копированием параметров главной сети актёра $\phi$
4. Инициализировать буфер воспроизведения $\mathcal{D}$
5. Для $N$ эпизодов повторять шаги 6 и 7
6. Инициализировать $\mathcal{N}$ - случайный процесс Орнштейна-Уленбека для исследования пространства действий
7. Для каждого шага в эпизоде $t = 0, \dots, T-1$
    a. Выбрать действие $\mathrm{a}$ согласно политике $\mu_\phi(s)$ и разведочного шума:
    $$a = \mu_\phi(s) + \mathcal{N}$$
    b. Получить кортеж опыта $(s_t, a_t, r_t, s'_t)$ и сохранить информацию о переходе в буфер воспроизведения $\mathcal{D}$

# Алгоритм DDPG. Окончание

c. Случайно отобрать пакет $K$ кортежей опыта из буфера воспроизведения $\mathcal{D}$

d. Рассчитать целевое значение критика: $y_i = r_i + \gamma Q_{\theta'}\left(s_i', \mu_{\phi'}\left(s_i'\right)\right)$

e. Вычислить потери сети критика:

$$J(\theta) = \frac{1}{K}\sum_i \left(y_i - Q_\theta(s_i, a_i)\right)^2$$

f. Вычислить градиент потерь критика $\nabla_\theta J(\theta)$ и обновить параметры сети критика с использованием градиентного спуска

$$\theta = \theta - \alpha\, \nabla_\theta J(\theta)$$

g. Вычислить градиент потерь актёра $\nabla_\phi J(\phi)$ и обновить параметры сети актёра с использованием градиентного подъёма

$$\phi = \phi + \alpha\, \nabla_\phi J(\phi)$$

h. Обновить параметры сетей актёра и критика:

$$\theta' = \tau\theta + (1-\tau)\theta' \text{ для } j = 1, 2 \text{ and } \phi' = \tau\phi + (1-\tau)\phi'$$

# Сглаживание целей

Значение $x$ в пределах $l$ и $h$:

$$clamp(x, l, h) = \max(\min(x, h), l)$$

Сглаживание действия за счет добавления ограниченного гауссового шума:

Гауссовый шум

$$a'^{,smooth} = clamp\big(\mu(s'; \phi') + clamp\big(\varepsilon, \varepsilon_{low}, \varepsilon_{high}\big), a_{low}, a_{high}\big)$$

$\varepsilon_{low}$ - Мин. значение для $\varepsilon$

$\varepsilon_{high}$ - Макс. значение для $\varepsilon$

$a_{low}$ - Мин. значение для $a$

$a_{high}$ - Макс. значение для $a$

Целевая функция $TD3$ :

$$\mathcal{TD}3^{target} = r + \gamma \min_n Q\big(s'; a'^{,smooth}; \theta^{n,-}\big)$$

# Максимизация ожидаемой выгоды и энтропии (SAC)

Добавление энтропии в уравнение Беллмана:

$$q_\pi(s,a) = \mathbb{E}_{r,s' \sim P(s,a), a' \sim \pi(s')}\left[r + \gamma\left(q_\pi(s',a') + \alpha\mathcal{H}\left(\pi(\cdot\,|s')\right)\right)\right]$$

Цель функции ценности действий:

$$\mathcal{SAC}^{target} = r + \gamma\left[\min_n Q(s';\hat{a}';\theta^{n,-}) - \alpha\log\pi(\hat{a}'|s';\phi)\right]$$

Функция цели коэффициента энтропии $\alpha$:

$$J(\alpha) = \mathbb{E}_{s\sim\mathcal{U}(\mathcal{D}),\hat{a}\sim\pi}[\alpha(\mathcal{H} + \log\pi(\hat{a}|s;\phi))]$$

# Мягкий «Актёр-критик» (SAC). Начало

1. Инициализировать параметры сети критика $\psi$, параметры Q - сети $\theta_1$ и $\theta_2$, а также параметры сети актёра $\phi$
2. Инициализировать параметра целевой сети критика $\psi'$ копированием параметра главной сети критика $\psi$
3. Инициализировать буфер воспроизведения $\mathcal{D}$
4. Для $N$ эпизодов выполнять шаг 5
5. Для каждого перехода в эпизоде $t = 0, \dots, T-1$
    a. Выбрать действие $a$ на основе политики $\pi_\phi$: $a = \pi_\phi(s)$
    b. Выполнить действие $a$, перейти в состояние $s'$, получить награду $r$, сохранить кортеж опыта в буфере воспроизведения $\mathcal{D}$
    c. Случайно отобрать пакет $K$ кортежей опыта из буфера воспроизведения $\mathcal{D}$
    d. Вычислить значение целевого состояния

$$y_{v_i} = \min_{j=1,2} Q_{\theta_j}(s_i, a_i) - \alpha \log \pi_\phi(s_i, a_i)$$

# SAC. Окончание

e. Вычислить потери сети ценности

$$J_v\ (\psi) = \frac{1}{K}\sum_i \left(y_{v_i} - V_\psi(s_i)\right)^2$$

и обновить $\psi$ параметр с использованием градиентного спуска

$$\psi = \psi - \alpha\ \nabla_\psi J(\psi)$$

f. Вычислить целевое значение ценности действия Q: $y_{q_i} = r_i + \gamma V_{\psi'}(s_i')$

g. Рассчитать потери для сети ценности действия:

$$J_Q\left(\theta_j\right) = \frac{1}{K}\sum_i \left(y_{q_i} - Q_{\theta_j}(s_i, a_i)\right)^2, j = 1{,}2$$

и обновить параметры с помощью градиентного спуска

$$\theta_j = \theta_j - \lambda\nabla_{\theta_j} J\left(\theta_j\right), j = 1{,}2$$

h. Вычислить градиенты целевой функции актёра $\nabla_\phi J(\phi)$ и обновить параметры актёра с помощью градиентного подъёма, $\phi = \phi + \lambda\nabla_\phi J(\phi)$

i. Обновить параметр целевой сети ценности $\psi' = \tau\psi + (1-\tau)\psi'$

# Ограничение обновлений политики (PPO)

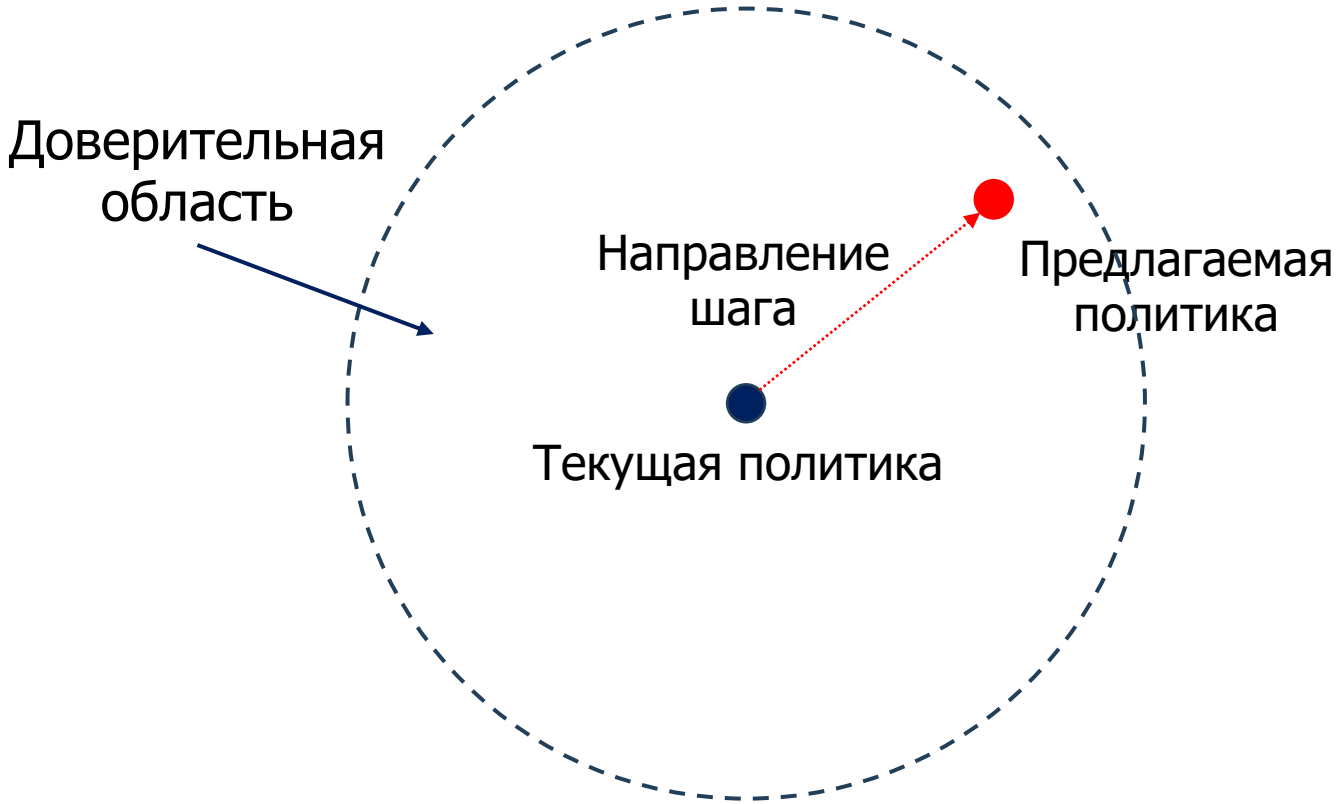
Ограниченная цель политики:

$$J(\phi, \phi') = \mathbb{E}_{(s,a,A^{GAE})\sim\mathcal{U}\left(\mathcal{D}(\phi')\right)}\left\{min\left[\frac{\pi(a|s;\phi)}{\pi(a|s;\phi')}A^{GAE}, clamp\left(\frac{\pi(a|s;\phi)}{\pi(a|s;\phi')}, 1-\varepsilon, 1+\varepsilon\right)A^{GAE}\right]\right\}$$

Потери функции ценности:

$$L(\theta, \theta') = \mathbb{E}_{(s,a,G,V)\sim\mathcal{U}\left(\mathcal{D}(\theta')\right)}\{max[G - V(s;\theta), G - (V(s;\theta) - V, -\delta, \delta)]\}$$

# Концепция методов доверительной области



# Оптимизация политики доверительной области (TRPO)

1. Инициализировать параметры сети политик $\theta$ и сети ценности $\phi$
2. Сгенерировать $N$ траекторий $\{\tau^i\}_{i=1}^{N}$ согласно политике $\pi_\theta$
3. Вычислить выгоду $G_t$
4. Рассчитать значение преимущества $A_t$
5. Вычислить градиент политики:

$$g = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t|s_t)A_t\right]$$

6. Рассчитать $s = H^{-1}g$ с использованием метода сопряженных градиентов
7. Обновить параметры $\theta$ сети политик:

$$\theta = \theta_k + \alpha^j\sqrt{\frac{2\delta}{s^T H s}}s$$

8. Вычислить MSE для сети ценности:

$$J(\phi) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\left(G_t - V_\phi(s_t)\right)^2$$

9. Обновить параметры сети ценности: $\phi = \phi - \alpha\nabla_\phi J(\phi)$
10. Repeat steps 2 to 9 for several iterations

# Оптимизация проксимальной политики с обрезкой (PPO-Clipped)

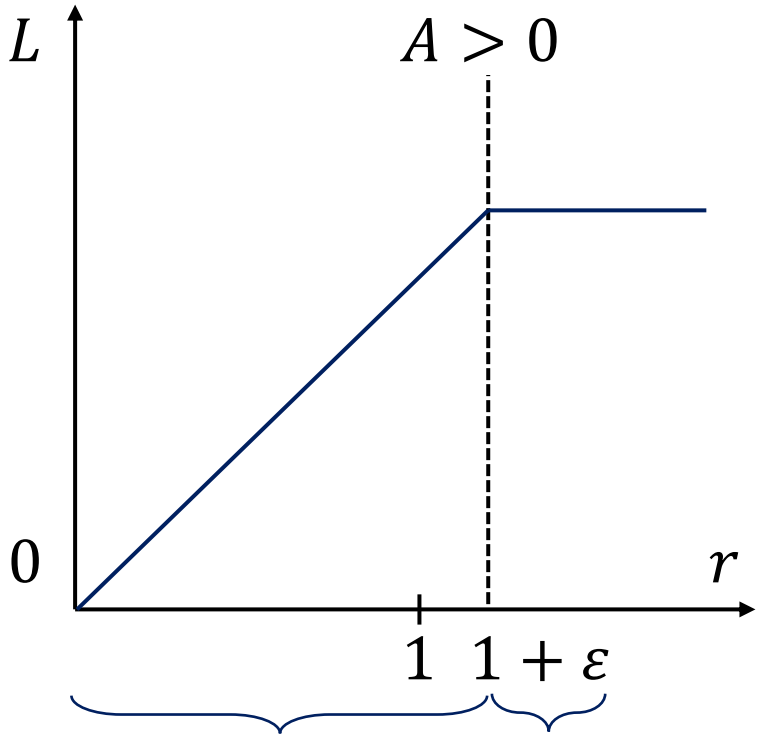
1. Инициализировать параметры сети политики $\theta$ и параметры сети ценности $\phi$
2. Собрать $N$ траекторий $\{\tau^i\}_{i=1}^N$ согласно политике $\pi_\theta$
3. Расчёт выгоды $G_t$
4. Вычислить целевую функцию: $L(\theta) = \mathbb{E}_t[min(r_t(\theta)A_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)]$

5. Рассчитать градиент целевой функции $\nabla_\theta L(\theta)$
6. Обновить параметры сети политики $\theta$ с использованием градиентного подъёма:
$$\theta = \theta + \alpha\nabla_\theta L(\theta)$$
7. Вычислить MSE для целевой сети:
$$J(\phi) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\left(G_t - V_\phi(s_t)\right)^2$$

8. Рассчитать градиент целевой сети $\nabla_\phi J(\phi)$
9. Обновить параметры сети ценности $\phi$ с использованием градиентного спуска:
$$\phi = \phi - \alpha\nabla_\phi J(\phi)$$
10. Повторять шаги 2. – 9. в течение нескольких итераций

# PPO-Обрезка

Величина $r_t(\theta)$ при положительном преимуществе

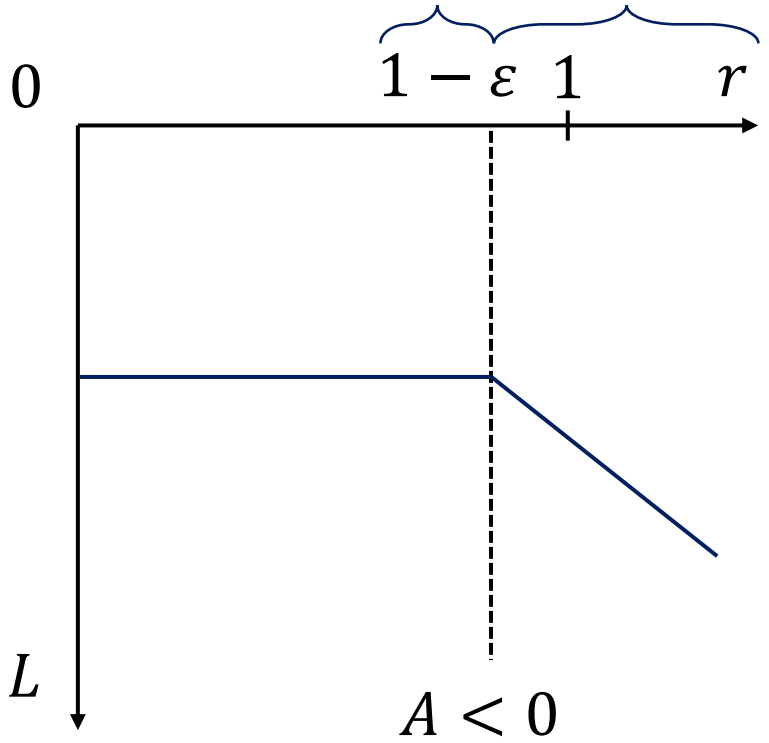



Увеличение $r$

Обрезка $r$ у значения $1+\varepsilon$

Обрезка $r$ у значения $1-\varepsilon$

Уменьшение $r$




Величина $r_t(\theta)$ при отрицательном преимуществе

# Оптимизация проксимальной политики со штрафом (PPO-penalty)

1. Инициализировать параметры сети политики $\theta$, сети ценности $\phi$, коэффициент штрафа $\beta_1$ и целевую дивергенцию Кульбака-Лейблера $\delta$
2. Выполнить итерации $i = 1,2, \dots, K$ :
   a. Собрать $N$ траекторий согласно политике $\pi_\theta$
   b. Вычислить выгоду $G_t$
   c. Рассчитать целевую функцию:

$$L(\theta) = \mathbb{E}_t \left[ \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t))} A_t - \beta_i KL\left[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)\right] \right]$$

   d. Найти градиент целевой функции $\nabla_\theta L(\theta)$
   e. Обновить параметры сети политики $\theta$: $\theta = \theta + \alpha \nabla_\theta L(\theta)$
   f. Если $d \geq 1.5\delta$, то обновляем $\beta_{i+1} = 2\beta_i$; если $d \leq \delta/1.5$ обновляем $\beta_{i+1} = \beta_i/2$
   g. Рассчитать MSE сети ценности:

$$J(\phi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \left( G_t - V_\phi(s_t) \right)^2$$

   h. Вычислить градиент сети ценности $\nabla_\phi J(\phi)$
   i. Update the value network parameter $\phi$ using gradient descent $\phi = \phi - \alpha \nabla_\phi J(\phi)$