

Лекция 6:

Градиент по Стратегии

Артём Сорокин | 07 Декабря

Целевая Функция в RL

Что мы хотим получить при помощи обучения с подкреплением:

$$\theta^* = \operatorname{argmax}_{\theta} \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_t \gamma^t r_t \right]$$

where:

- θ - параметры стратегии
- $p_{\theta}(\tau)$ - вероятностное распределение по траекториям в среде, которые генерирует стратегия π_{θ}
- $[\sum_t \gamma^t r_t]$ - кумулятивная дисконтированная награда за эпизод / доход с первого шага.

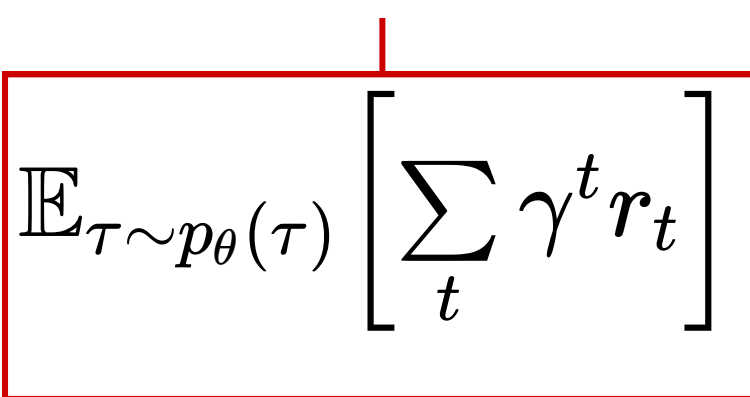
Целевая Функция в RL

$$\theta^* = \operatorname{argmax}_{\theta} \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_t \gamma^t r_t \right]$$

Целевая Функция в RL

$$\theta^* = \operatorname{argmax}_{\theta} \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_t \gamma^t r_t \right]$$

$J(\theta)$



Целевая Функция в RL

Целевая функция

→ $J(\theta)$

$$\theta^* = \operatorname{argmax}_{\theta} \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_t \gamma^t r_t \right]$$

Целевая Функция в RL

Целевая функция

→ $J(\theta)$

$$\theta^* = \operatorname{argmax}_{\theta} \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_t \gamma^t r_t \right]$$

← $r(\tau)$

Целевая Функция в RL

Целевая функция

$J(\theta)$

$\theta^* = \operatorname{argmax}_{\theta}$

$$\mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_t \gamma^t r_t \right]$$

$r(\tau)$

кумулятивная
награда за
траекторию

Целевая Функция в RL

Целевая функция

$J(\theta)$

$\theta^* = \operatorname{argmax}_{\theta}$

$$\mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_t \gamma^t r_t \right]$$

$r(\tau)$

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)] = \int p_{\theta}(\tau) r(\tau) d\tau$$

кумулятивная
награда за
траекторию

Целевая Функция в RL

Целевая функция

$J(\theta)$

$$\theta^* = \operatorname{argmax}_{\theta} \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_t \gamma^t r_t \right]$$

$r(\tau)$

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)] = \int p_{\theta}(\tau) r(\tau) d\tau$$

кумулятивная
награда за
траекторию

Задача:

Мы бы хотели найти градиент нашей целевой функции по параметрам стратегии π_{θ} , которая генерирует траектории

Градиент по Стратегии

Чтобы максимизировать средний ожидаемый доход:

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)}[r(\tau)] = \int p_{\theta}(\tau)r(\tau)d\tau$$

Найдем:

$$\nabla_{\theta}J(\theta) = \int \nabla_{\theta}p_{\theta}(\tau)r(\tau)d\tau$$

Градиент по Стратегии

Чтобы максимизировать средний ожидаемый доход:

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)] = \int p_{\theta}(\tau) r(\tau) d\tau$$

Найдем:

$$\nabla_{\theta} J(\theta) = \int \nabla_{\theta} p_{\theta}(\tau) r(\tau) d\tau$$

Log-derivative trick:

$$\nabla_{\theta} p_{\theta}(\tau) = p_{\theta}(\tau) \frac{\nabla_{\theta} p_{\theta}(\tau)}{p_{\theta}(\tau)} = p_{\theta}(\tau) \nabla_{\theta} \log p_{\theta}(\tau)$$

Градиент по Стратегии

Чтобы максимизировать средний ожидаемый доход:

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)] = \int p_{\theta}(\tau) r(\tau) d\tau$$

Найдем:

$$\begin{aligned} \nabla_{\theta} J(\theta) &= \int \nabla_{\theta} p_{\theta}(\tau) r(\tau) d\tau \\ &= \int p_{\theta}(\tau) \nabla_{\theta} \log p_{\theta}(\tau) r(\tau) d\tau = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right] \end{aligned}$$

Log-derivative trick:

$$\nabla_{\theta} p_{\theta}(\tau) = p_{\theta}(\tau) \frac{\nabla_{\theta} p_{\theta}(\tau)}{p_{\theta}(\tau)} = p_{\theta}(\tau) \nabla_{\theta} \log p_{\theta}(\tau)$$

Policy Gradients

Максимизируем средний ожидаемый доход:

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)]$$

Градиент по θ :

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

Policy Gradients

Максимизируем средний ожидаемый доход:

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)]$$

Градиент по θ :

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

Распишем $p_{\theta}(\tau)$:

$$p_{\theta}(\tau) = p_{\theta}(s_0, a_0, \dots, s_T, a_T) = p(s_0) \prod_{t=0}^T \pi_{\theta}(a_t | s_t) p(s_{t+1} | a_t, s_t)$$

Policy Gradients

Максимизируем средний ожидаемый доход:

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)]$$

Градиент по θ :

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

Распишем $p_{\theta}(\tau)$:

$$p_{\theta}(\tau) = p_{\theta}(s_0, a_0, \dots, s_T, a_T) = p(s_0) \prod_{t=0}^T \pi_{\theta}(a_t | s_t) p(s_{t+1} | a_t, s_t)$$

Возьмем логарифм:

$$\log p_{\theta}(\tau) = \log p(s_0) + \sum_{t=0}^T [\log \pi_{\theta}(a_t | s_t) + \log p(s_{t+1} | a_t, s_t)]$$


Policy Gradients

Максимизируем средний ожидаемый доход:

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)]$$

Градиент по θ :

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$


$$\log p(s_0) + \sum_{t=0}^T [\log \pi_{\theta}(a_t | s_t) + \log p(s_{t+1} | a_t, s_t)]$$


Policy Gradients

Максимизируем средний ожидаемый доход:

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)]$$

Градиент по θ :

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$


$$\nabla_{\theta} \left[\log p(s_0) + \sum_{t=0}^T [\log \pi_{\theta}(a_t | s_t) + \log p(s_{t+1} | a_t, s_t)] \right]$$

Policy Gradients

Максимизируем средний ожидаемый доход:

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)]$$

Градиент по θ :

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

$$\nabla_{\theta} \left[\cancel{\log p(s_0)} + \sum_{t=0}^T [\log \pi_{\theta}(a_t | s_t) + \cancel{\log p(s_{t+1} | a_t, s_t)}] \right]$$

Policy Gradients

Максимизируем средний ожидаемый доход:

$$J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)]$$

Градиент по θ :

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

$$\nabla_{\theta} \left[\cancel{\log p(s_0)} + \sum_{t=0}^T [\log \pi_{\theta}(a_t | s_t) + \cancel{\log p(s_{t+1} | a_t, s_t)}] \right]$$

Градиент по Стратегии:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) r(\tau) \right]$$

Оценка Градиента по Стратегии

Мы не знаем реального значения мат. ожидания

здесь:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

Оценка Градиента по Стратегии

Мы не знаем реального значения мат. ожидания

здесь:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

Как всегда можем оценить его используя сэмплирование:

Оценка Градиента по Стратегии

Мы не знаем реального значения мат. ожидания

здесь:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

Как всегда можем оценить его используя сэмплирование:

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) r(\tau_i) \right]$$

Оценка Градиента по Стратегии

Мы не знаем реального значения мат. ожидания

здесь:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

Как всегда можем оценить его используя сэмплирование:

$$\begin{aligned} \nabla_{\theta} J(\theta) &\approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) r(\tau_i) \right] \\ &= \frac{1}{N} \sum_{i=1}^N \left[\left(\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \right) \left(\sum_{t=0}^T r_{i,t} \right) \right] \end{aligned}$$

Алгоритм REINFORCE

Оцениваем Градиент по стратегии:

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\left(\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \right) \left(\sum_{t=0}^T r_{i,t} \right) \right]$$

Обновляем параметры стратегии:

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} J(\theta)$$

Псевдокод:

1. Сэмплируем i эпизодов $\{\tau^i\}$ стратегией π_{θ}
2. Оцениваем градиент по стратегии π_{θ} на эпизодах $\{\tau^i\}$
3. Обновляем параметры стратегии
4. Переходим к пункту 1

REINFORCE это on-policy алгоритм

REINFORCE оценивает градиент по стратегии (**P**olicy **G**radient):

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

Псевдокод:

1. Сэмплируем i эпизодов $\{\tau^i\}$ стратегией π_{θ}
2. Оцениваем градиент по стратегии π_{θ} на эпизодах $\{\tau^i\}$
3. Обновляем параметры стратегии
4. Переходим к пункту 1

REINFORCE это on-policy алгоритм

REINFORCE оценивает градиент по стратегии (**P**olicy **G**radient):

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

Для оценки градиента по параметрам θ нужно собирать сэмплы при помощи π_{θ} !

Псевдокод:

1. Сэмплируем i эпизодов $\{\tau^i\}$ стратегией π_{θ}
2. Оцениваем градиент по стратегии π_{θ} на эпизодах $\{\tau^i\}$
3. Обновляем параметры стратегии
4. Переходим к пункту 1

REINFORCE это on-policy алгоритм

REINFORCE оценивает градиент по стратегии (**P**olicy **G**radient):

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

Для оценки градиента по параметрам θ нужно собирать сэмплы при помощи π_{θ} !

On-policy алгоритмы:

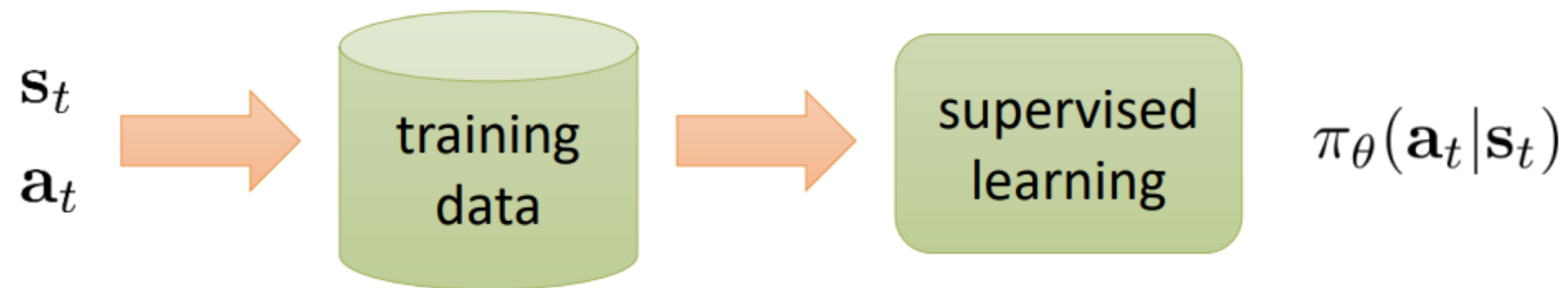
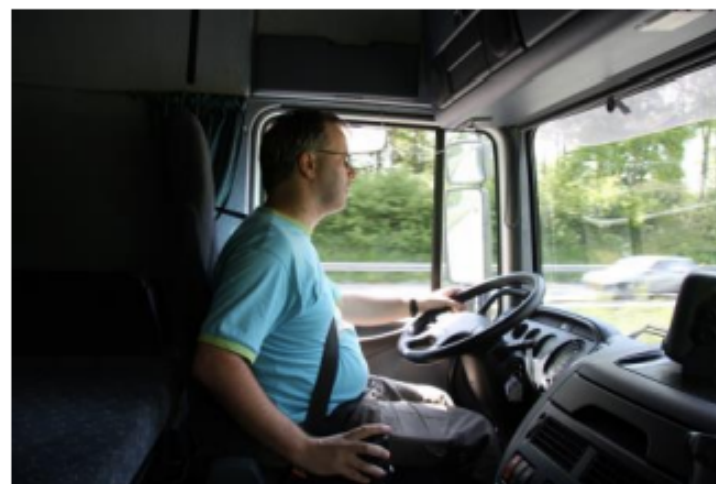
- После обновления параметров сэмплы собранные со старыми параметрами становятся бесполезны.
- Алгоритмы на основе PG требуют много сэмплов!

Псевдокод:

1. Сэмплируем i эпизодов $\{\tau^i\}$ стратегией π_{θ}
2. Оцениваем градиент по стратегии π_{θ} на эпизодах $\{\tau^i\}$
3. Обновляем параметры стратегии
4. Переходим к пункту 1

Основная идея Градиента по Стратегии

Представим, что учим стратегию по экспертным траекториями при помощи обучения с учителем:

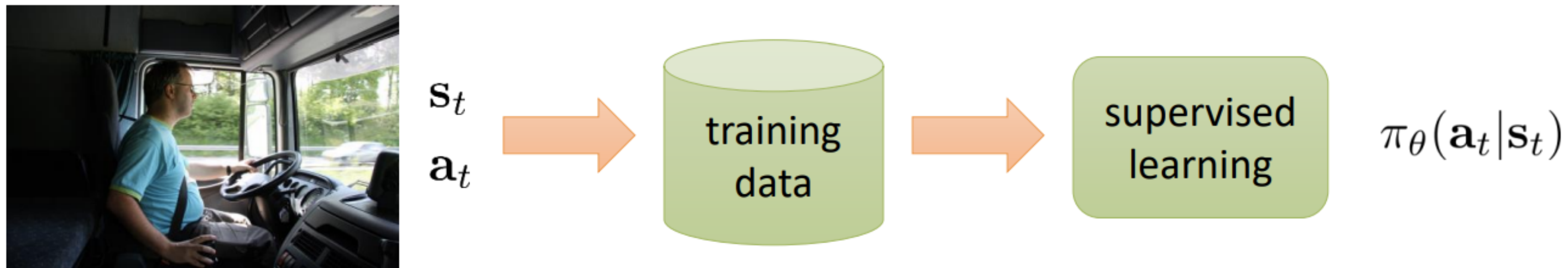


Стратегия в s_t : $\pi_{\theta}(* | s_t) = \bar{y} = \begin{bmatrix} 0.2 \\ 0.7 \\ 0.1 \end{bmatrix}$

Ground Truth из датасета в s_t : $y = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

Основная идея Градиента по Стратегии

Представим, что учим стратегию по экспертным траекториями при помощи обучения с учителем:



Используем Cross Entropy-loss для каждого перехода (s_t, a_t, s_{t+1}) в датасете :

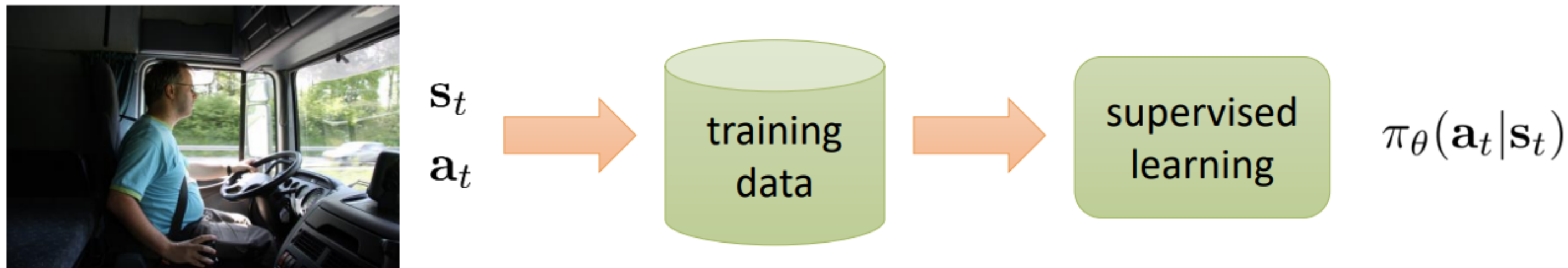
$$H(\bar{y}, y_t) = \frac{1}{|C|} \sum_j^{|C|} -y_j \log \bar{y}_j = -\log \bar{y}_{a_t} \frac{1}{|C|}$$

Стратегия в s_t : $\pi_{\theta}(*|s_t) = \bar{y} = \begin{bmatrix} 0.2 \\ 0.7 \\ 0.1 \end{bmatrix}$

Ground Truth из датасета в s_t : $y = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

Основная идея Градиента по Стратегии

Представим, что учим стратегию по экспертным траекториями при помощи обучения с учителем:



Используем Cross Entropy-loss для каждого перехода (s_t, a_t, s_{t+1}) в датасете :

$$H(\bar{y}, y_t) = \frac{1}{|C|} \sum_j^{|C|} -y_j \log \bar{y}_j = -\log \bar{y}_{a_t} \frac{1}{|C|}$$

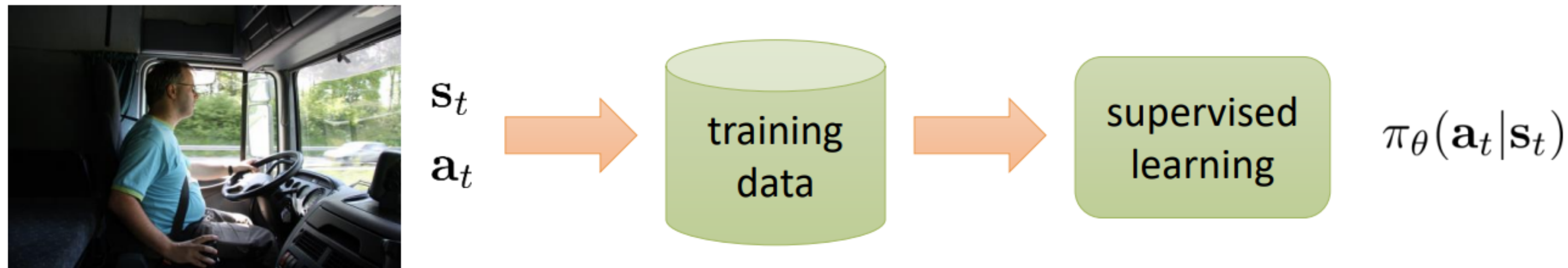
Стратегия в s_t : $\pi_{\theta}(*|s_t) = \bar{y} = \begin{bmatrix} 0.2 \\ 0.7 \\ 0.1 \end{bmatrix}$

a_t

Ground Truth из датасета в s_t : $y = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

Основная идея Градиента по Стратегии

Представим, что учим стратегию по экспертным траекториями при помощи обучения с учителем:



Используем Cross Entropy-loss для каждого перехода (s_t, a_t, s_{t+1}) в датасете :

$$\begin{aligned} H(\bar{y}, y_t) &= \frac{1}{|C|} \sum_j^{|C|} -y_j \log \bar{y}_j = -\log \bar{y}_{a_t} \frac{1}{|C|} \\ &= -\log \pi_{\theta}(a_t | s_t) \text{ } c \end{aligned}$$

Стратегия в s_t : $\pi_{\theta}(* | s_t) = \bar{y} = \begin{bmatrix} 0.2 \\ 0.7 \\ 0.1 \end{bmatrix}$

a_t

Ground Truth из датасета в s_t : $y = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

Основная идея Градиента по Стратегии

Градиент при клонировании поведения (Behaviour Clonning):

$$\nabla_{\theta} J_{BC}(\theta) = \mathbb{E}_{\tau \sim D} \left[\sum_{t=0}^T \nabla_{\theta} -\log \pi_{\theta}(a_t | s_t) c \right] \quad \boxed{\text{Цель минимизировать } J_{BC}(\theta)}$$

Основная идея Градиента по Стратегии

Градиент при клонировании поведения (Behaviour Clonning):

$$\nabla_{\theta} J_{BC}(\theta) = \mathbb{E}_{\tau \sim D} \left[\sum_{t=0}^T \nabla_{\theta} -\log \pi_{\theta}(a_t | s_t) c \right] \quad \text{Цель минимизировать } J_{BC}(\theta)$$

Градиент по стратегии:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) r(\tau) \right] \quad \text{Цель максимизировать } J(\theta)$$

Основная идея Градиента по Стратегии

Градиент при клонировании поведения (Behaviour Clonning):

$$\nabla_{\theta} J_{BC}(\theta) = \mathbb{E}_{\tau \sim D} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) c \right] \quad \text{Цель максимизировать } -J_{BC}(\theta)$$

Градиент по стратегии:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) r(\tau) \right] \quad \text{Цель максимизировать } J(\theta)$$

Основная идея Градиента по Стратегии

Градиент при клонировании поведения (Behaviour Clonning):

$$\nabla_{\theta} J_{BC}(\theta) = \mathbb{E}_{\tau \sim D} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) c \right]$$

Цель максимизировать $-J_{BC}(\theta)$

BC учит модель выбирать те же действия что и эксперт!

Градиент по стратегии:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) r(\tau) \right]$$

Цель максимизировать $J(\theta)$

Основная идея Градиента по Стратегии

Градиент при клонировании поведения (Behaviour Clonning):

$$\nabla_{\theta} J_{BC}(\theta) = \mathbb{E}_{\tau \sim D} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) c \right] \quad \text{Цель максимизировать } -J_{BC}(\theta)$$

BC учит модель выбирать те же действия что и эксперт!

Градиент по стратегии:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) r(\tau) \right] \quad \text{Цель максимизировать } J(\theta)$$

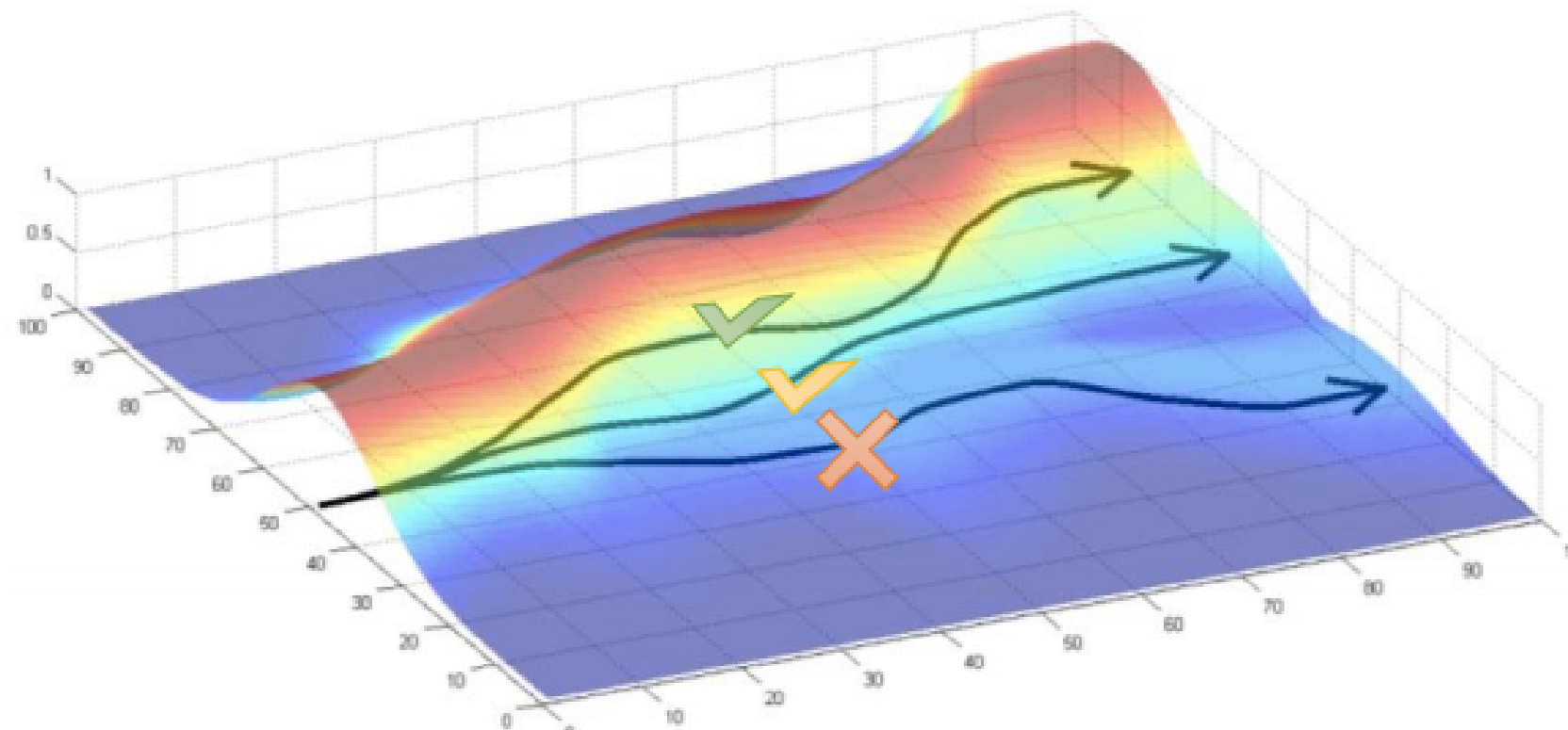
PG учит модель выбирать действия ведущие к высоким наградам за эпизод!

Основная Идея Градиента по Стратегии

Градиент по стратегии:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) r(\tau) \right]$$

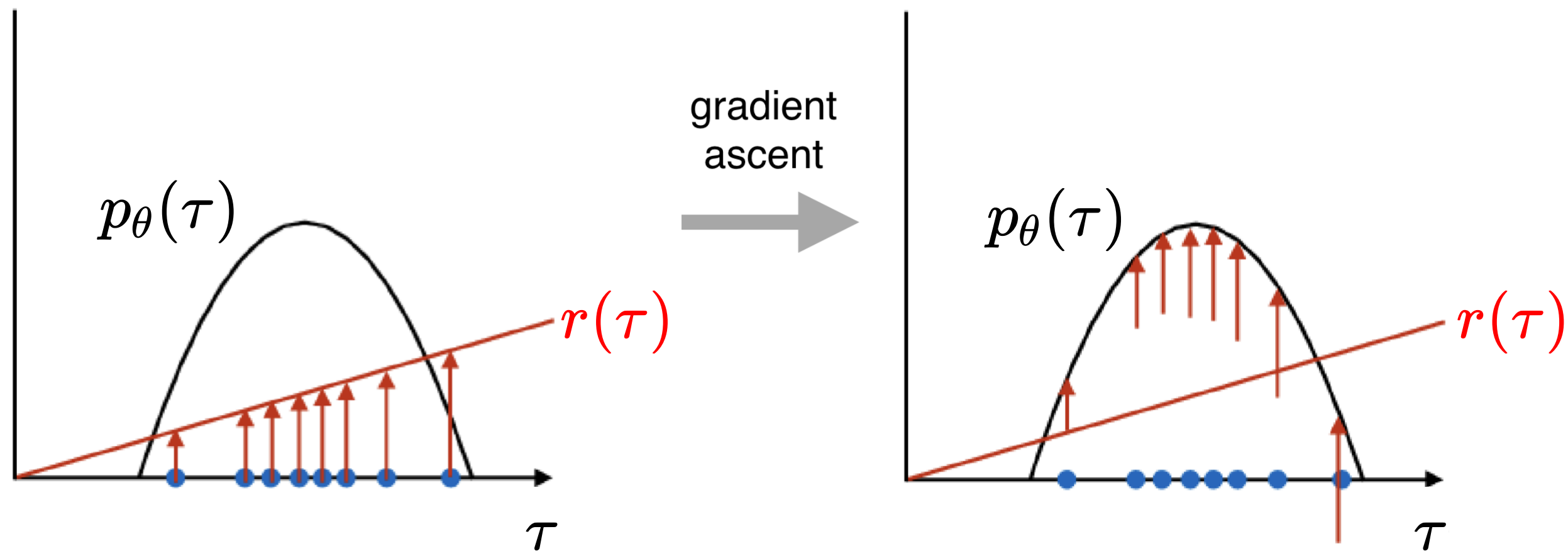
PG учит стратегию выбирать действия ведущие к высокому доходу за эпизод!



Проблемы Градиенты по Стратегии

Проблема: **высокая дисперсия!**

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) r(\tau) \right]$$



Уменьшаем Дисперсию: Причинность

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=0}^T \gamma^{t'} r_{i,t'} \right) \right] \leftarrow \text{выглядит подозрительно!}$$

Уменьшаем Дисперсию: Причинность

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=0}^T \gamma^{t'} r_{i,t'} \right) \right] \leftarrow \text{выглядит подозрительно!}$$

Принцип причинности: действие на шаге t не может повлиять на награду за шаг t' если $t' < t$

Уменьшаем Дисперсию: Причинность

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=0}^T \gamma^{t'} r_{i,t'} \right) \right] \leftarrow \boxed{\text{выглядит подозрительно!}}$$

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=t}^T \gamma^{t'} r_{i,t'} \right) \right]$$

Уменьшаем Дисперсию: Причинность

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=0}^T \gamma^{t'} r_{i,t'} \right) \right] \leftarrow \text{выглядит подозрительно!}$$

Принцип причинности: действие на шаге t не может повлиять на награду за шаг t' если $t' < t$

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=t}^T \gamma^{t'} r_{i,t'} \right) \right]$$

Последние действия становятся менее значимыми!

Уменьшаем Дисперсию: Причинность

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=0}^T \gamma^{t'} r_{i,t'} \right) \right] \leftarrow \text{выглядит подозрительно!}$$

Принцип причинности: действие на шаге t не может повлиять на награду за шаг t' если $t' < t$

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\gamma^t \sum_{t'=t}^T \gamma^{t'-t} r_{i,t'} \right) \right]$$

Последние действия становятся менее значимыми!

Уменьшаем Дисперсию: Причинность

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=0}^T \gamma^{t'} r_{i,t'} \right) \right] \leftarrow \text{выглядит подозрительно!}$$

Принцип причинности: действие на шаге t не может повлиять на награду за шаг t' если $t' < t$

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\gamma^t \sum_{t'=t}^T \gamma^{t'-t} r_{i,t'} \right) \right]$$

Последние действия становятся менее значимыми!

Финальная Версия:

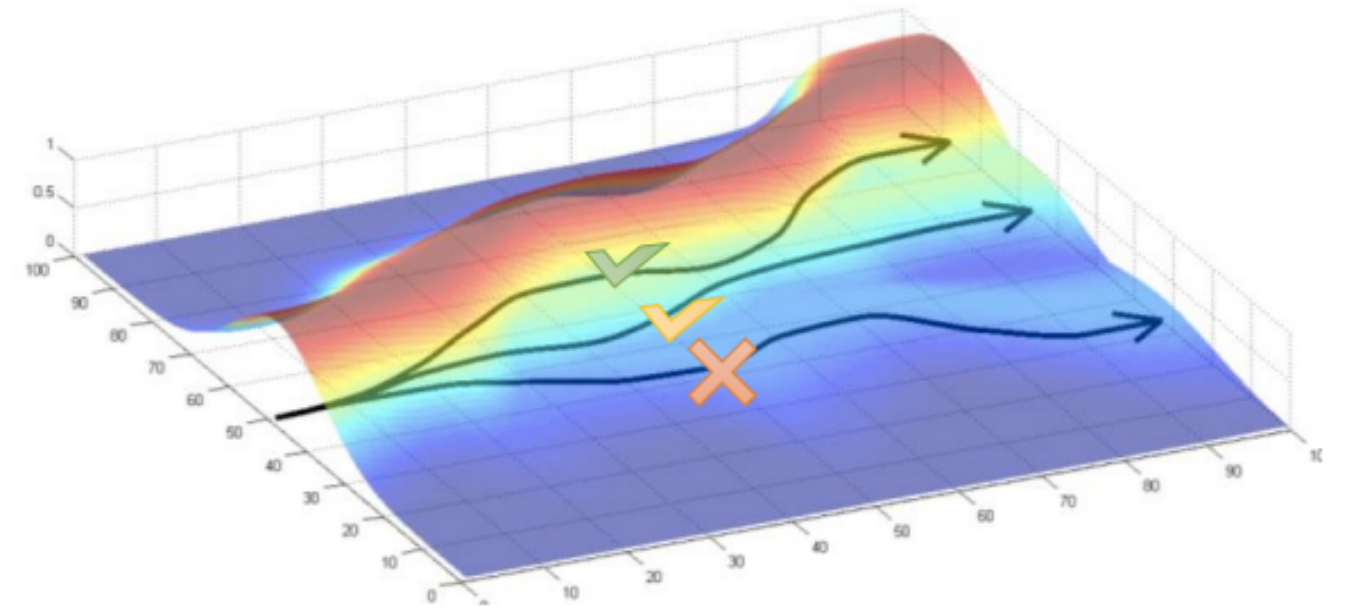
$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=t}^T \gamma^{t'-t} r_{i,t'} \right) \right]$$

Улучшаем РС: Бейзлайн

Обновляем стратегию пропорционально доходу $\tau(r)$:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) r(\tau) \right]$$

где: $b = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)]$

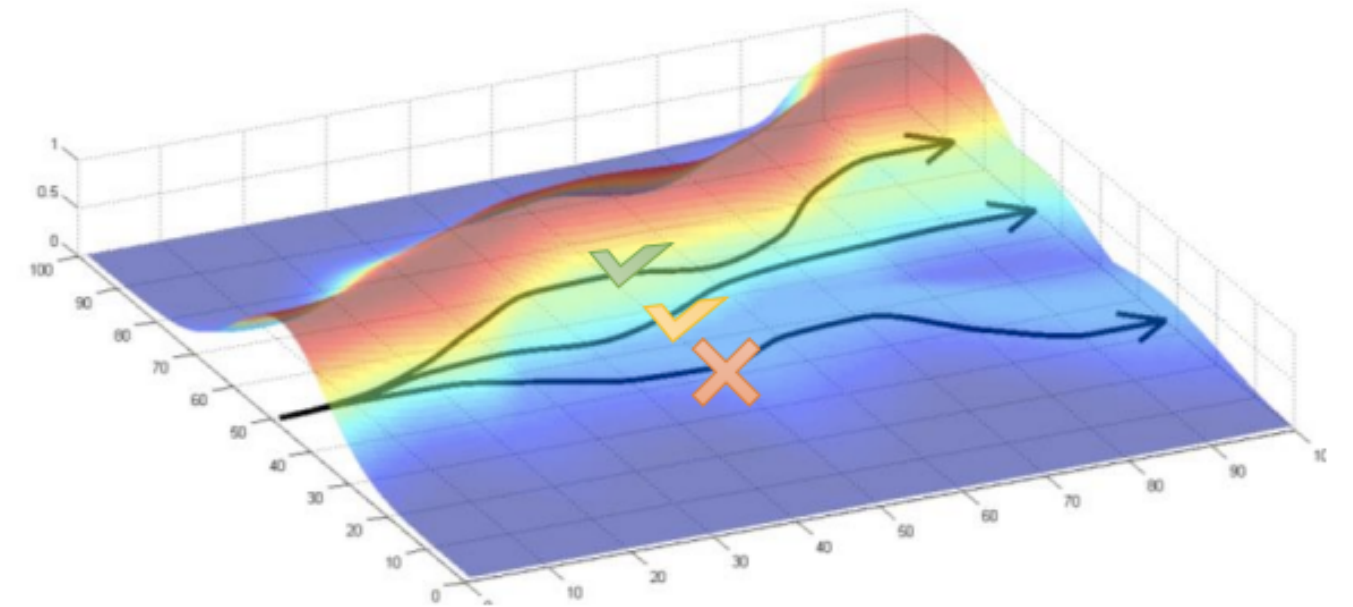


Улучшаем РС: Бейзлайн

Обновляем стратегию пропорционально тому на сколько $\tau(r)$ лучше чем средний доход:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) (r(\tau) - b) \right]$$

где: $b = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)]$

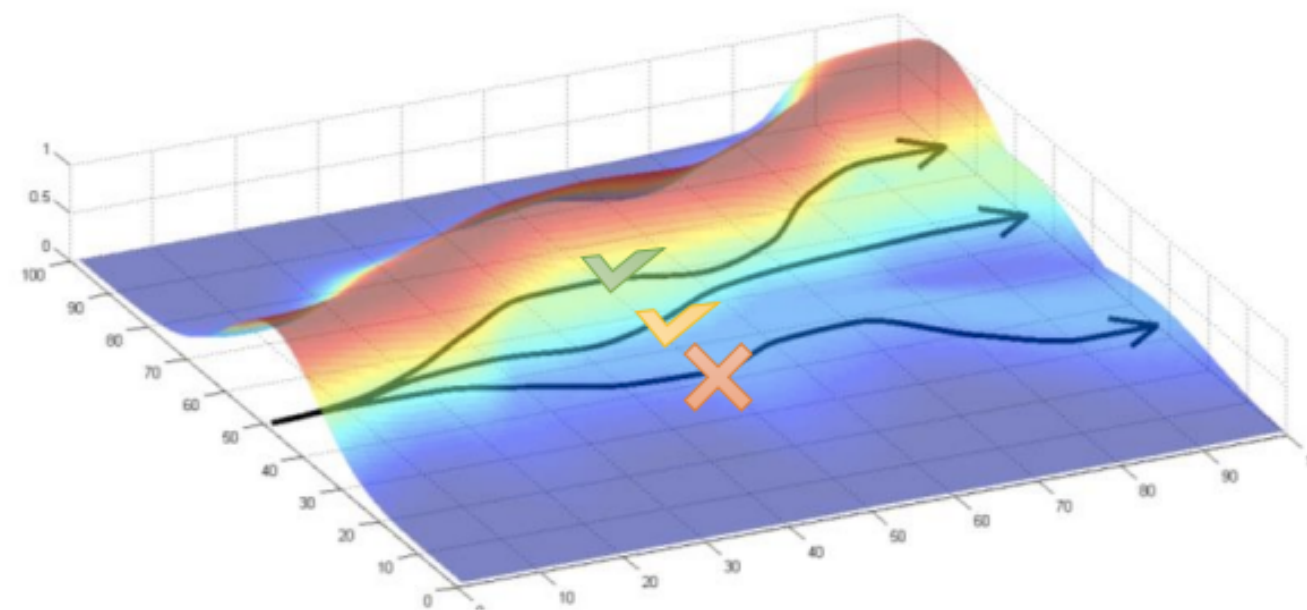


Улучшаем РС: Бейзлайн

Обновляем стратегию пропорционально тому на сколько $\tau(r)$ лучше чем средний доход:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) (r(\tau) - b) \right]$$

где: $b = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)]$



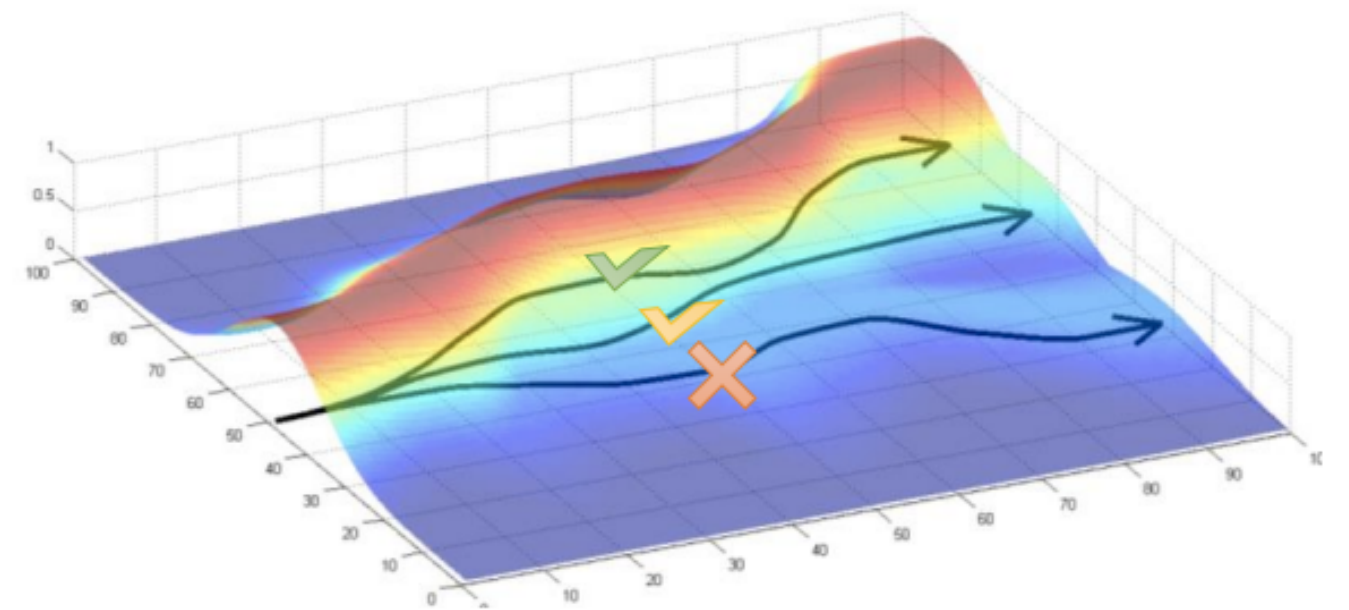
Вычитание бейзлайна дает **несмещенную оценку** (и часто работает лучше):

Улучшаем РС: Бейзлайн

Обновляем стратегию пропорционально тому на сколько $\tau(r)$ лучше чем средний доход:

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} \left[\nabla_{\theta} \log p_{\theta}(\tau) (r(\tau) - b) \right]$$

где: $b = \mathbb{E}_{\tau \sim p_{\theta}(\tau)} [r(\tau)]$



Вычитание бейзлайна дает **несмещенную оценку** (и часто работает лучше):

$$\begin{aligned} \mathbb{E}[\nabla_{\theta} \log p_{\theta}(\tau) b] &= \int p_{\theta}(\tau) \nabla_{\theta} \log p_{\theta}(\tau) b d\tau = \int b \nabla_{\theta} p_{\theta}(\tau) d\tau = \\ &= b \nabla_{\theta} \int p_{\theta}(\tau) d\tau = b \nabla_{\theta} 1 = 0 \end{aligned}$$

Улучшаем РС: Регуляризация энтропией

В методах на основе функций ценности (DQN, Q-learning, SARSA, и тд.) мы использовали ϵ -жданую стратегию, чтобы агент исследовал новые варианты в среде

Улучшаем РС: Регуляризация энтропией

В методах на основе функций ценности (DQN, Q-learning, SARSA, и тд.) мы использовали ϵ -жданую стратегию, чтобы агент исследовал новые варианты в среде

В методах с явным представлением стратегии агента, можно использовать более гибкий вариант:

Улучшаем РС: Регуляризация энтропией

В методах на основе функций ценности (DQN, Q-learning, SARSA, и тд.) мы использовали ϵ -жданую стратегию, чтобы агент исследовал новые варианты в среде

В методах с явным представлением стратегии агента, можно использовать более гибкий вариант:

Регуляризация энтропии стратегии агента:

$$H(\pi_{\theta}(\cdot|s_t)) = - \sum_{a \in A} \pi_{\theta}(a|s_t) \log \pi_{\theta}(a|s_t)$$

Улучшаем РС: Регуляризация энтропией

В методах на основе функций ценности (DQN, Q-learning, SARSA, и тд.) мы использовали ϵ -жадную стратегию, чтобы агент исследовал новые варианты в среде

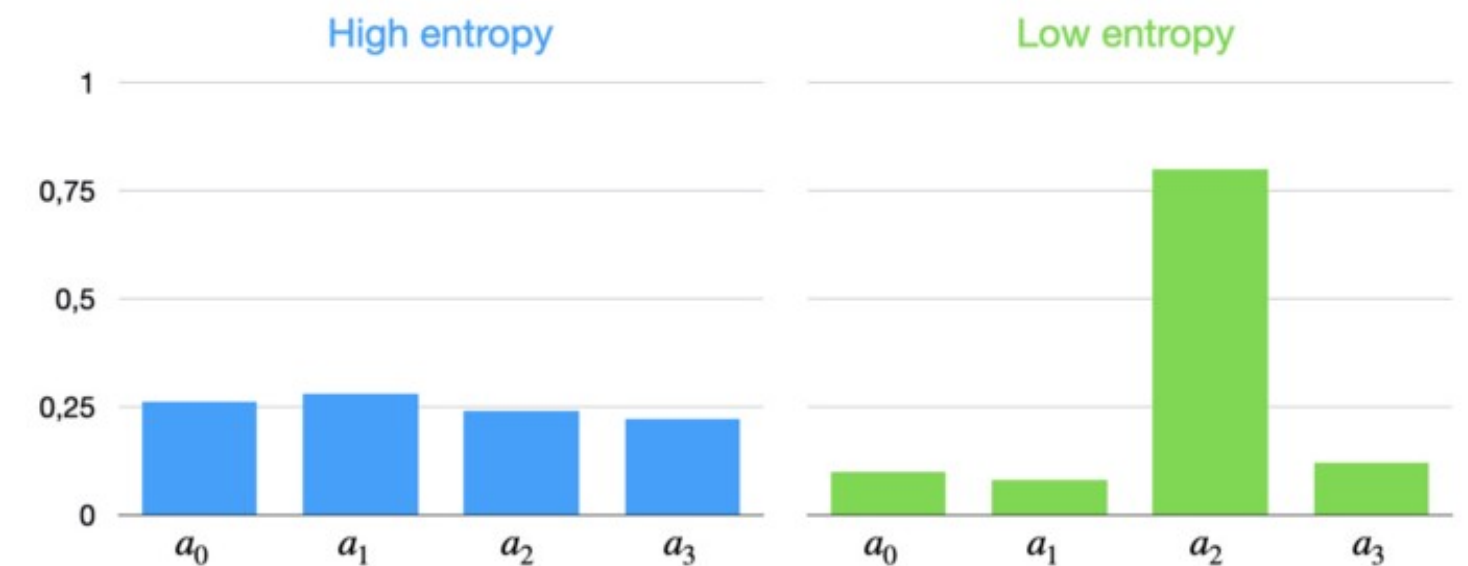
В методах с явным представлением стратегии агента, можно использовать более гибкий вариант:

Регуляризация энтропии стратегии агента:

$$H(\pi_{\theta}(\cdot|s_t)) = - \sum_{a \in A} \pi_{\theta}(a|s_t) \log \pi_{\theta}(a|s_t)$$

Добавление к функции потерь $-H(\pi_{\theta})$:

- поощряет агента действовать более случайно
- накладывает менее строгие ограничения чем ϵ -жадная стратегия



Алгоритмы Исполнитель-Критик

Финальная версия REINFORCE с "учетом причинности" и бейзлайном:

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=t}^T \gamma^{t'-t} r_{i,t'} - b \right) \right]$$

Алгоритмы Исполнитель-Критик

Финальная версия REINFORCE с "учетом причинности" и бейзлайном:

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=t}^T \gamma^{t'-t} r_{i,t'} - b \right) \right]$$

Что это?

Алгоритмы Исполнитель-Критик

Финальная версия REINFORCE с "учетом причинности" и бейзлайном:

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=t}^T \gamma^{t'-t} r_{i,t'} - b \right) \right]$$

Что это?

Вспоминаем Функции ценности:

$$Q_{\pi}(s, a) = \mathbb{E}_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid S_t = s, A_t = a \right]$$

$$V_{\pi}(s) = \mathbb{E}_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid S_t = s \right]$$

Алгоритмы Исполнитель-Критик

Финальная версия REINFORCE с "учетом причинности" и бейзлайном:

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(\sum_{t'=t}^T \gamma^{t'-t} r_{i,t'} - b \right) \right]$$

Оценка $Q_{\pi_{\theta}}(s_{i,t}, a_{i,t})$ по одному сэмплу

Вспоминаем Функции ценности:

$$Q_{\pi}(s, a) = \mathbb{E}_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid S_t = s, A_t = a \right]$$

$$V_{\pi}(s) = \mathbb{E}_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid S_t = s \right]$$

Алгоритмы Исполнитель-Критик

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(Q_{\pi_{\theta}}(s_{i,t}, a_{i,t}) - b \right) \right]$$

Алгоритмы Исполнитель-Критик

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(Q_{\pi_{\theta}}(s_{i,t}, a_{i,t}) - b \right) \right]$$

дисперсия меньше чем
у оценки по одному
сэмплу

Алгоритмы Исполнитель-Критик

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(Q_{\pi_{\theta}}(s_{i,t}, a_{i,t}) - b \right) \right]$$

дисперсия меньше чем
у оценки по одному
сэмплу

Вспомним про бейзлайн:

Алгоритмы Исполнитель-Критик

Совместим *Градиент по стратегии* и *Функции Ценности*!

дисперсия меньше чем
у оценки по одному
сэмплу

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(Q_{\pi_{\theta}}(s_{i,t}, a_{i,t}) - b \right) \right]$$

Вспомним про бейзлайн:

$$b = \mathbb{E}_{\tau \sim \pi_{\theta}} [r(\tau)] =$$

Алгоритмы Исполнитель-Критик

Совместим *Градиент по стратегии* и *Функции Ценности*!

дисперсия меньше чем
у оценки по одному
сэмплу

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(Q_{\pi_{\theta}}(s_{i,t}, a_{i,t}) - b \right) \right]$$

Вспомним про бейзлайн:

$$b = \mathbb{E}_{\tau \sim \pi_{\theta}} [r(\tau)] =$$

Тут тоже стоит учесть причинность....

Алгоритмы Исполнитель-Критик

Совместим *Градиент по стратегии* и *Функции Ценности*!

дисперсия меньше чем
у оценки по одному
сэмплу

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(Q_{\pi_{\theta}}(s_{i,t}, a_{i,t}) - b \right) \right]$$

Вспомним про бейзлайн:

$$b = \mathbb{E}_{\tau \sim \pi_{\theta}} [r(\tau)] = \mathbb{E}_{a \sim \pi_{\theta}(a|s)} [Q_{\pi_{\theta}}(s, a)] =$$

Тут тоже стоит учесть причинность....

Алгоритмы Исполнитель-Критик

Совместим *Градиент по стратегии* и *Функции Ценности*!

дисперсия меньше чем
у оценки по одному
сэмплу

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(Q_{\pi_{\theta}}(s_{i,t}, a_{i,t}) - b \right) \right]$$

Вспомним про бейзлайн:

$$b = \mathbb{E}_{\tau \sim \pi_{\theta}} [r(\tau)] = \mathbb{E}_{a \sim \pi_{\theta}(a|s)} [Q_{\pi_{\theta}}(s, a)] = V_{\pi_{\theta}}(s)$$

Тут тоже стоит учесть причинность....

Advantage Actor-Critic: A2C

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(Q_{\pi_{\theta}}(s_{i,t}, a_{i,t}) - V_{\pi_{\theta}}(s_{i,t}) \right) \right]$$

Advantage Actor-Critic: A2C

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(Q_{\pi_{\theta}}(s_{i,t}, a_{i,t}) - V_{\pi_{\theta}}(s_{i,t}) \right) \right]$$

Функция преимущества / Advantage Function:

$$A(a, s) = Q_{\pi_{\theta}}(s, a) - V_{\pi_{\theta}}(s)$$

Advantage Actor-Critic: A2C

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(Q_{\pi_{\theta}}(s_{i,t}, a_{i,t}) - V_{\pi_{\theta}}(s_{i,t}) \right) \right]$$

Функция преимущества / Advantage Function:

$$A(a, s) = Q_{\pi_{\theta}}(s, a) - V_{\pi_{\theta}}(s) \quad \text{на сколько } \mathbf{a}_t \text{ лучше чем обычное поведение стратегии}$$

Advantage Actor-Critic: A2C

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(A_{\pi_{\theta}}(s_{i,t}, a_{i,t}) \right) \right]$$

Функция преимущества / Advantage Function:

$$A(a, s) = Q_{\pi_{\theta}}(s, a) - V_{\pi_{\theta}}(s) \quad \text{на сколько } \mathbf{a}_t \text{ лучше чем обычное поведение стратегии}$$

Advantage Actor-Critic: A2C

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(A_{\pi_{\theta}}(s_{i,t}, a_{i,t}) \right) \right]$$

Функция преимущества / Advantage Function:

$$A(a, s) = Q_{\pi_{\theta}}(s, a) - V_{\pi_{\theta}}(s) \quad \text{на сколько } \mathbf{a}_t \text{ лучше чем обычное поведение стратегии}$$

Легче учить только одну функцию!

Advantage Actor-Critic: A2C

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(A_{\pi_{\theta}}(s_{i,t}, a_{i,t}) \right) \right]$$

Функция преимущества / Advantage Function:

$$A(a, s) = Q_{\pi_{\theta}}(s, a) - V_{\pi_{\theta}}(s) \quad \text{на сколько } \mathbf{a}_t \text{ лучше чем обычное поведение стратегии}$$

Легче учить только одну функцию! ...но можно сделать еще лучше:

Advantage Actor-Critic: A2C

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(A_{\pi_{\theta}}(s_{i,t}, a_{i,t}) \right) \right]$$

Функция преимущества / Advantage Function:

$$A(a, s) = Q_{\pi_{\theta}}(s, a) - V_{\pi_{\theta}}(s) \quad \text{на сколько } \mathbf{a_t} \text{ лучше чем обычное поведение стратегии}$$

Легче учить только одну функцию! ...но можно сделать еще лучше:

$$A(a, s) = \mathbb{E}_{s' \sim p(s' | a, s)} [r(s, a) + E_{a' \sim \pi_{\theta}(s' | s')} [Q_{\pi_{\theta}}(a', s')]] - V_{\pi_{\theta}}(s_t)$$

Advantage Actor-Critic: A2C

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(A_{\pi_{\theta}}(s_{i,t}, a_{i,t}) \right) \right]$$

Функция преимущества / Advantage Function:

$$A(a, s) = Q_{\pi_{\theta}}(s, a) - V_{\pi_{\theta}}(s) \quad \text{на сколько } \mathbf{a}_t \text{ лучше чем обычное поведение стратегии}$$

Легче учить только одну функцию! ...но можно сделать еще лучше:

$$\begin{aligned} A(a, s) &= \mathbb{E}_{s' \sim p(s' | a, s)} [r(s, a) + \mathbb{E}_{a' \sim \pi_{\theta}(s' | s')} [Q_{\pi_{\theta}}(a', s')]] - V_{\pi_{\theta}}(s_t) \\ &= r(s, a) + \mathbb{E}_{s' \sim p(s' | a, s)} [V_{\pi_{\theta}}(s')] - V_{\pi_{\theta}}(s) \end{aligned}$$

Advantage Actor-Critic: A2C

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(A_{\pi_{\theta}}(s_{i,t}, a_{i,t}) \right) \right]$$

Функция преимущества / Advantage Function:

$$A(a, s) = Q_{\pi_{\theta}}(s, a) - V_{\pi_{\theta}}(s) \quad \text{на сколько } \mathbf{a}_t \text{ лучше чем обычное поведение стратегии}$$

Легче учить только одну функцию! ...но можно сделать еще лучше:

$$\begin{aligned} A(a, s) &= \mathbb{E}_{s' \sim p(s' | a, s)} [r(s, a) + \mathbb{E}_{a' \sim \pi_{\theta}(s' | s')} [Q_{\pi_{\theta}}(a', s')]] - V_{\pi_{\theta}}(s_t) \\ &= r(s, a) + \mathbb{E}_{s' \sim p(s' | a, s)} [V_{\pi_{\theta}}(s')] - V_{\pi_{\theta}}(s) \end{aligned}$$

аппроксимируем это значение при помощи одного сэмпла

Advantage Actor-Critic: A2C

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(A_{\pi_{\theta}}(s_{i,t}, a_{i,t}) \right) \right]$$

Функция преимущества:

$$A_{\pi_{\theta}}(a_t, s_t) \approx r_t + V_{\pi_{\theta}}(s_{t+1}) - V_{\pi_{\theta}}(s_t)$$

Выучить V -функцию легче, т.к. она зависит от меньшего числа аргументов

Advantage Actor-Critic: A2C

Совместим *Градиент по стратегии* и *Функции Ценности*!

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_{i,t} | s_{i,t}) \left(A_{\pi_{\theta}}(s_{i,t}, a_{i,t}) \right) \right]$$

Функция преимущества:

на сколько a_t лучше чем обычное поведение стратегии

$$A_{\pi_{\theta}}(a_t, s_t) \approx r_t + V_{\pi_{\theta}}(s_{t+1}) - V_{\pi_{\theta}}(s_t)$$

Выучить V -функцию легче, т.к. она зависит от меньшего числа аргументов

A2C: Обучение

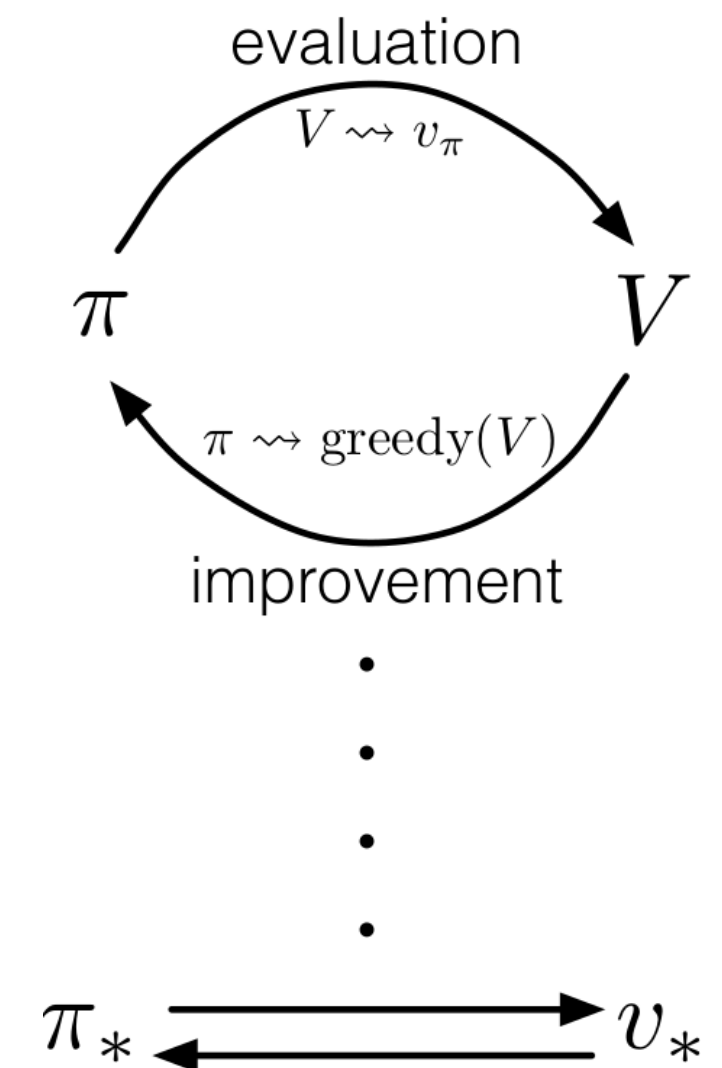
- Сэмплируем $\{\tau\}$ при помощи $\pi_\theta(a_t|s_t)$
- **Policy Improvement шаг:**
 - Учим исполнителя при помощи градиента по стратегии:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) A_{\pi_\theta}(s_{i,t}, a_{i,t}) \right]$$

- **Policy Evaluation шаг:**
 - Учим Критика через MSE (по аналогии с DQN)

$$\nabla_\phi L(\phi) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_\phi \|(r_t + \gamma V_{\hat{\phi}}(s_{t+1})) - V_\phi(s_t)\|^2 \right]$$

Policy Iteration
напоминка:



A2C: Обучение

- Сэмплируем $\{\tau\}$ при помощи $\pi_\theta(a_t|s_t)$
- **Policy Improvement шаг:**
 - Учим исполнителя при помощи градиента по стратегии:

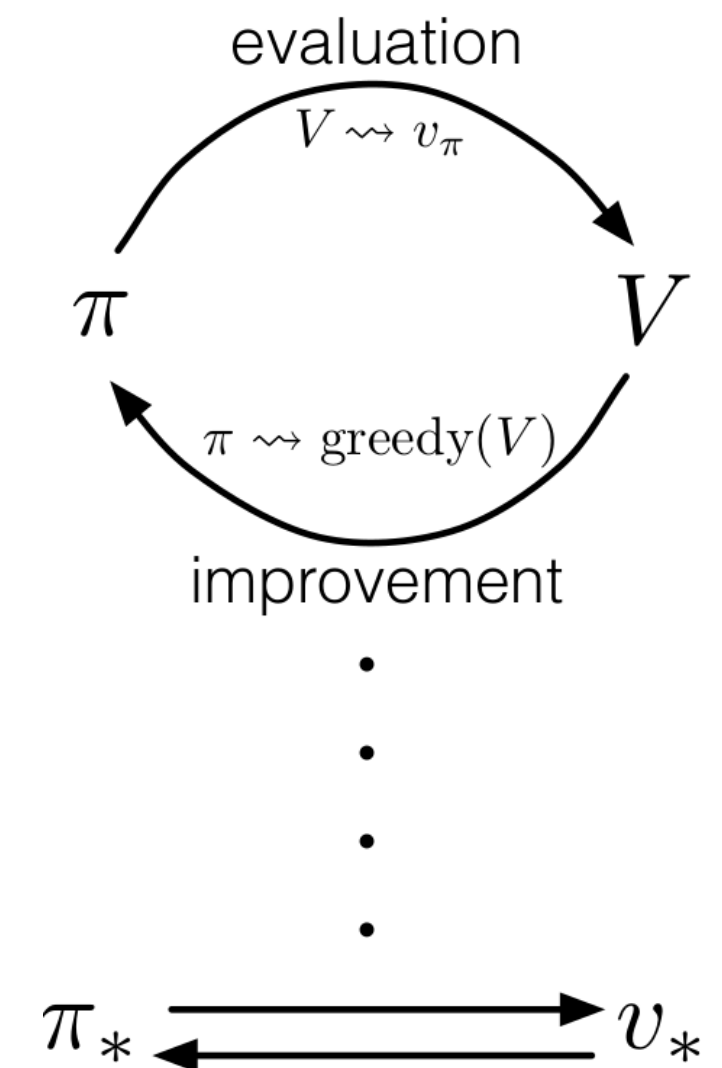
$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) A_{\pi_\theta}(s_{i,t}, a_{i,t}) \right]$$

- **Policy Evaluation шаг:**
 - Учим Критика через MSE (по аналогии с DQN)

$$\nabla_\phi L(\phi) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_\phi \|(r_t + \gamma V_{\hat{\phi}}(s_{t+1})) - V_\phi(s_t)\|^2 \right]$$

ϕ : свой набор параметров

Policy Iteration
напоминка:



A2C: Обучение

- Сэмплируем $\{\tau\}$ при помощи $\pi_\theta(a_t|s_t)$
- **Policy Improvement шаг:**
 - Учим исполнителя при помощи градиента по стратегии:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) A_{\pi_\theta}(s_{i,t}, a_{i,t}) \right]$$

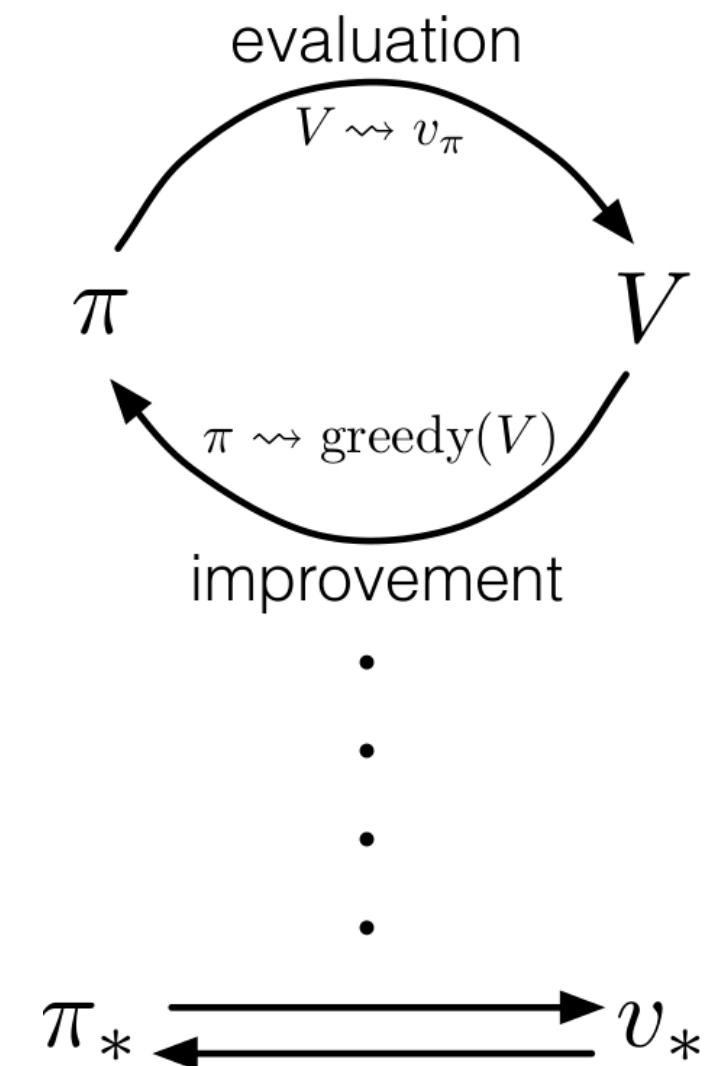
- **Policy Evaluation шаг:**
 - Учим Критика через MSE (по аналогии с DQN)

$$\nabla_\phi L(\phi) \approx \frac{1}{N} \sum_{i=1}^N \left[\sum_{t=0}^T \nabla_\phi \|(r_t + \gamma V_{\hat{\phi}}(s_{t+1})) - V_\phi(s_t)\|^2 \right]$$

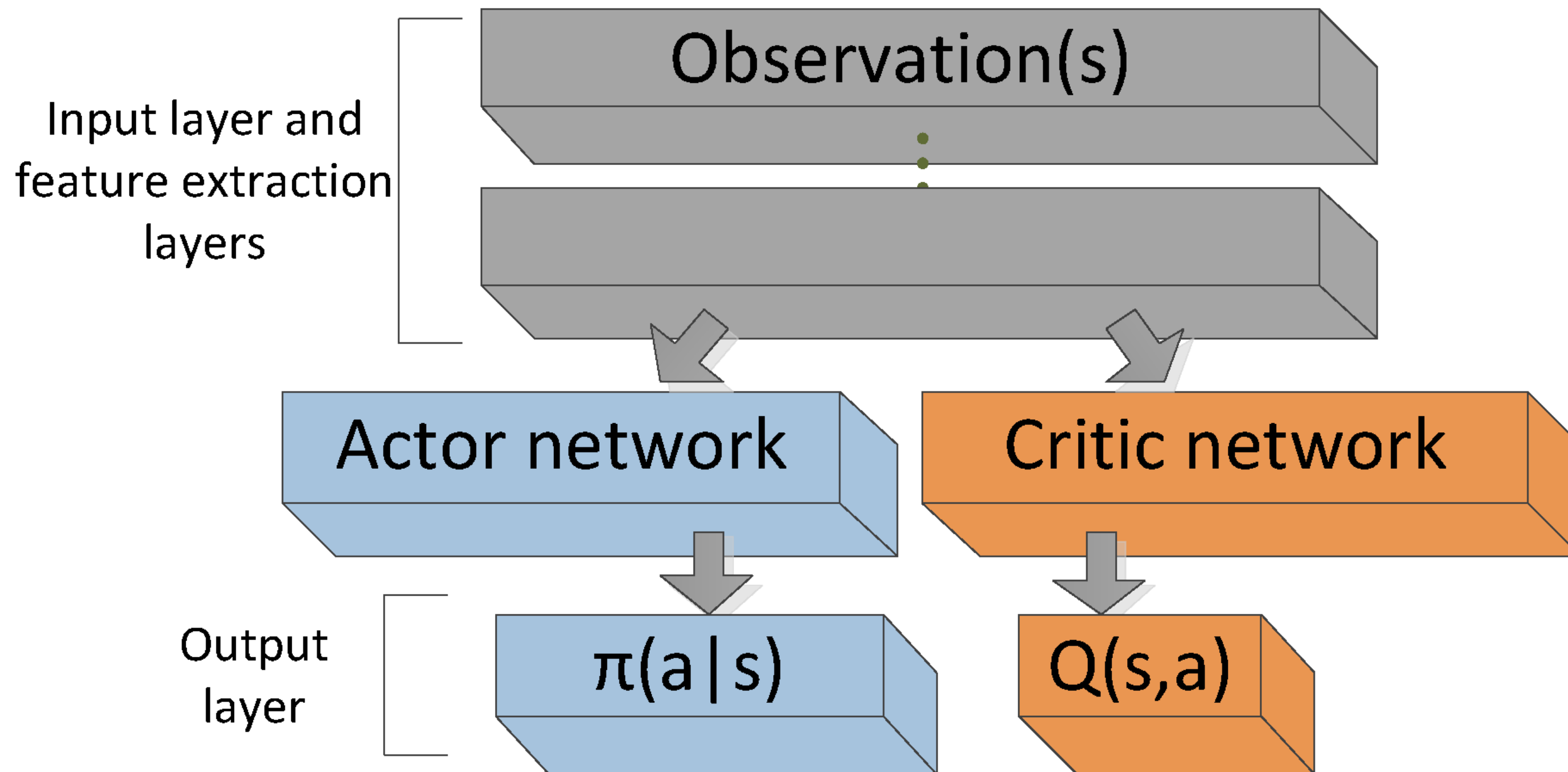
ϕ : свой набор параметров

В DQN была Target Network, тут просто не проводим градиенты.

Policy Iteration
напоминка:



Детали реализации: Архитектура A2C



Асинхронный A2C: A3C

Не можем использовать Replay Memory, но нам нужно декоррелировать сэмплы

Асинхронный A2C: A3C

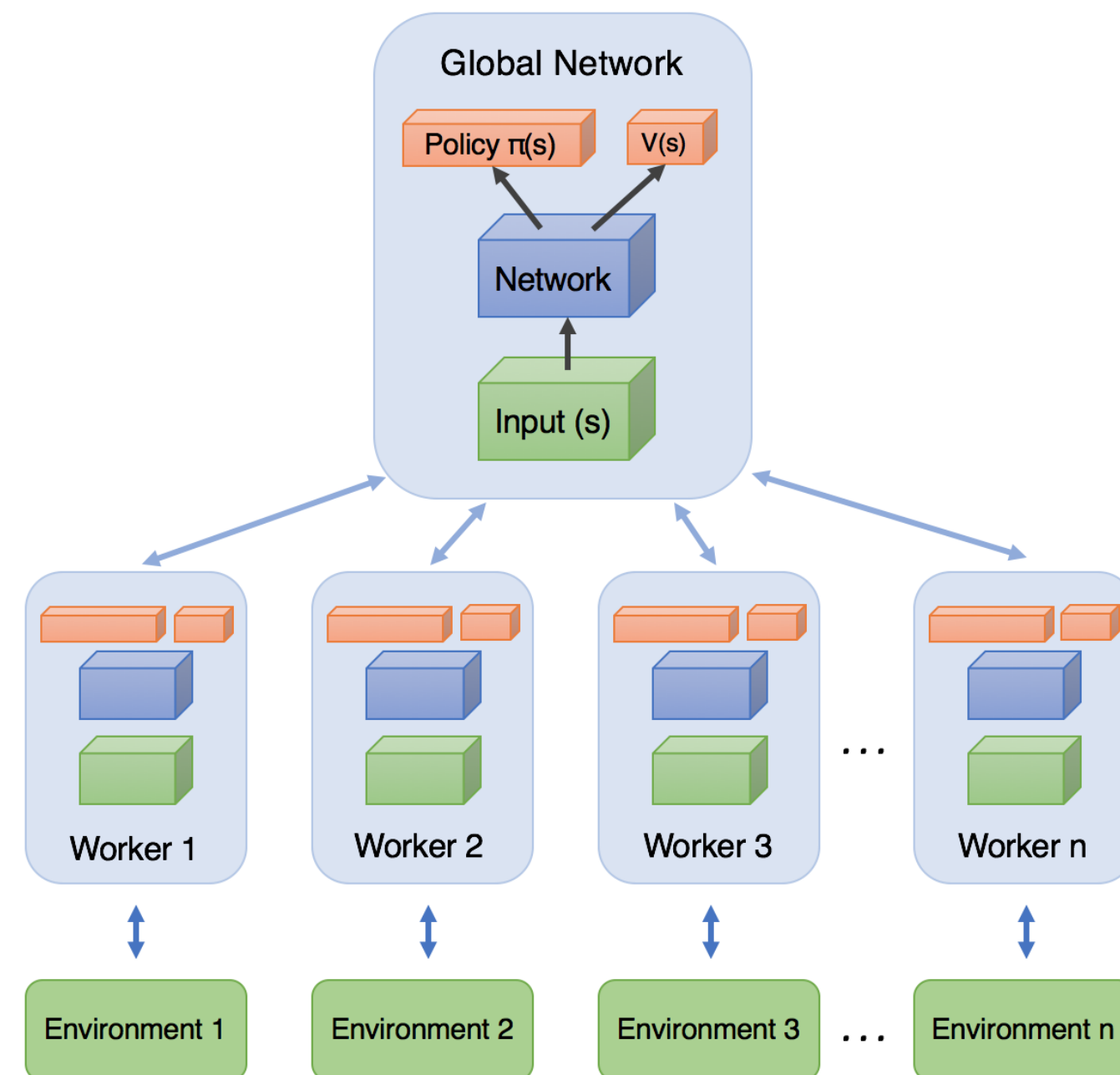
Не можем использовать Replay Memory, но нам нужно декоррелировать сэмплы

Answer: Учим на нескольких средах
одновременно!

Асинхронный A2C: A3C

Не можем использовать Replay Memory, но нам нужно декоррелировать сэмплы

Answer: Учим на нескольких средах одновременно!



Асинхронный A2C: A3C

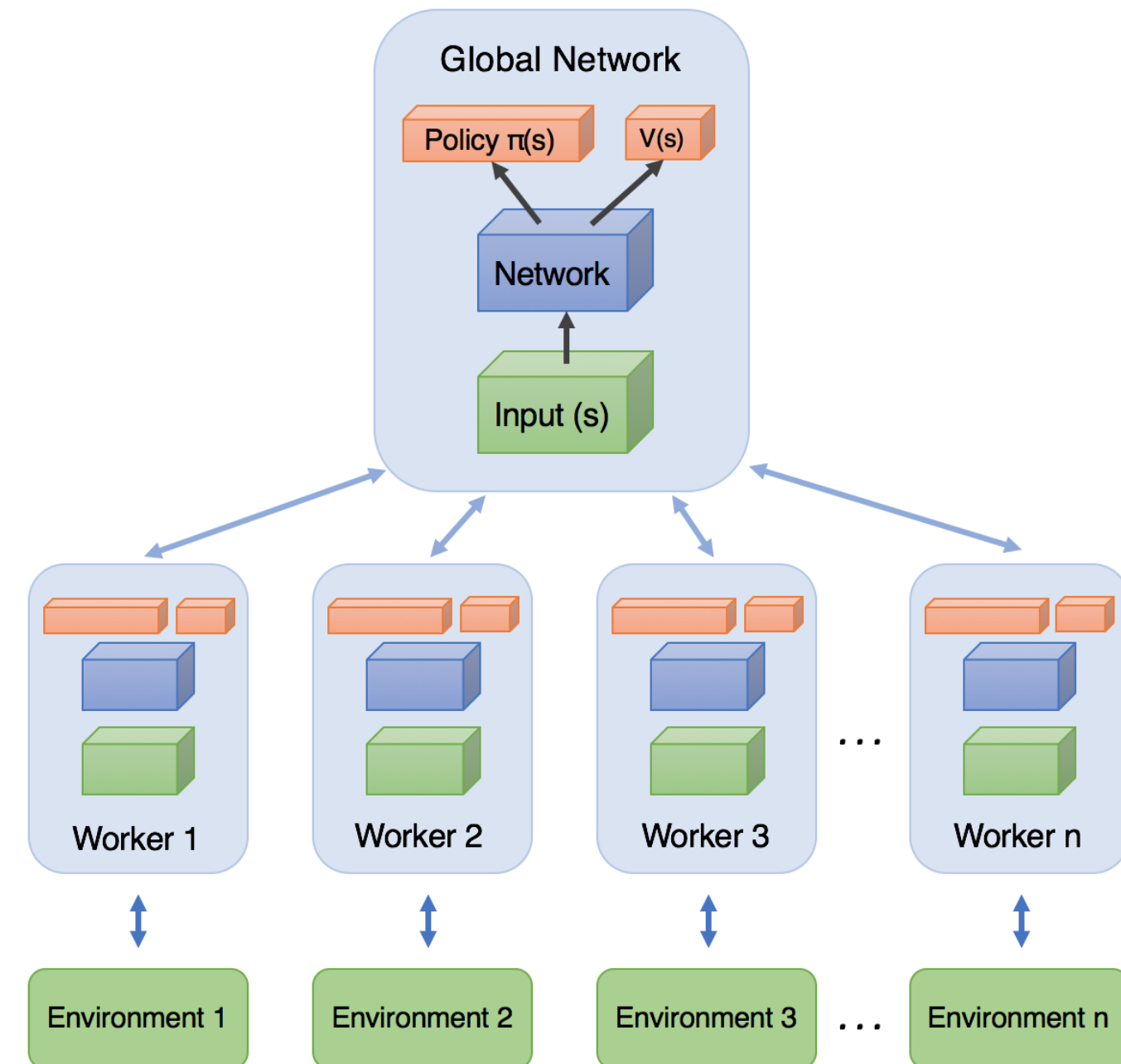
Не можем использовать Replay Memory, но нам нужно декоррелировать сэмплы

Answer: Учим на нескольких средах одновременно!

Взаимодействие с каждой средой и обучение происходит **асинхронно**

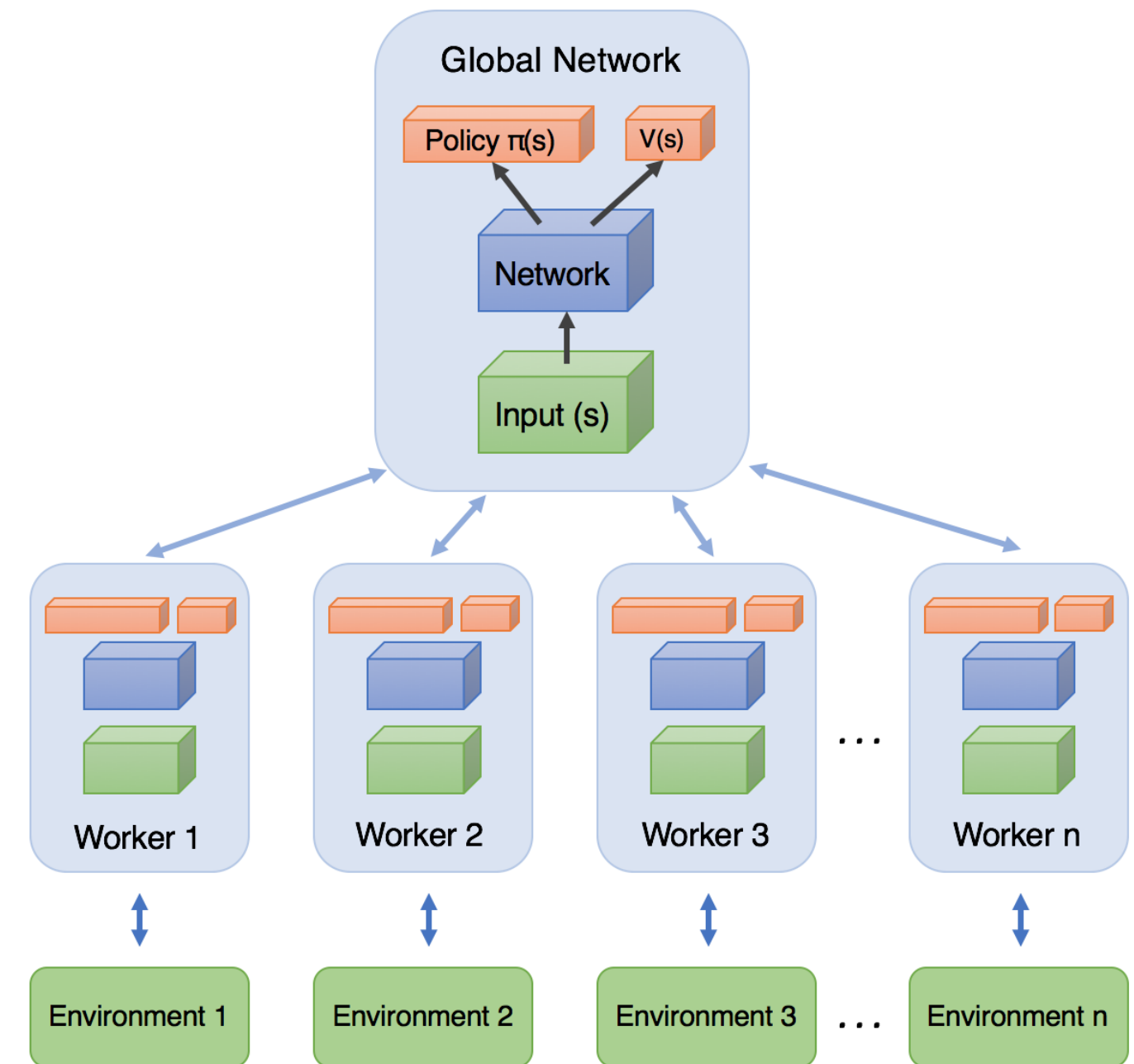
Каждый рабочий:

- Получает параметры модели из единого сервера параметров
- Генерирует траектории
- Считает Градиенты
- Отправляет градиенты обратно в сервер параметров



Асинхронный A2C: A3C

Взаимодействие с каждой средой и обучение происходит **асинхронно**

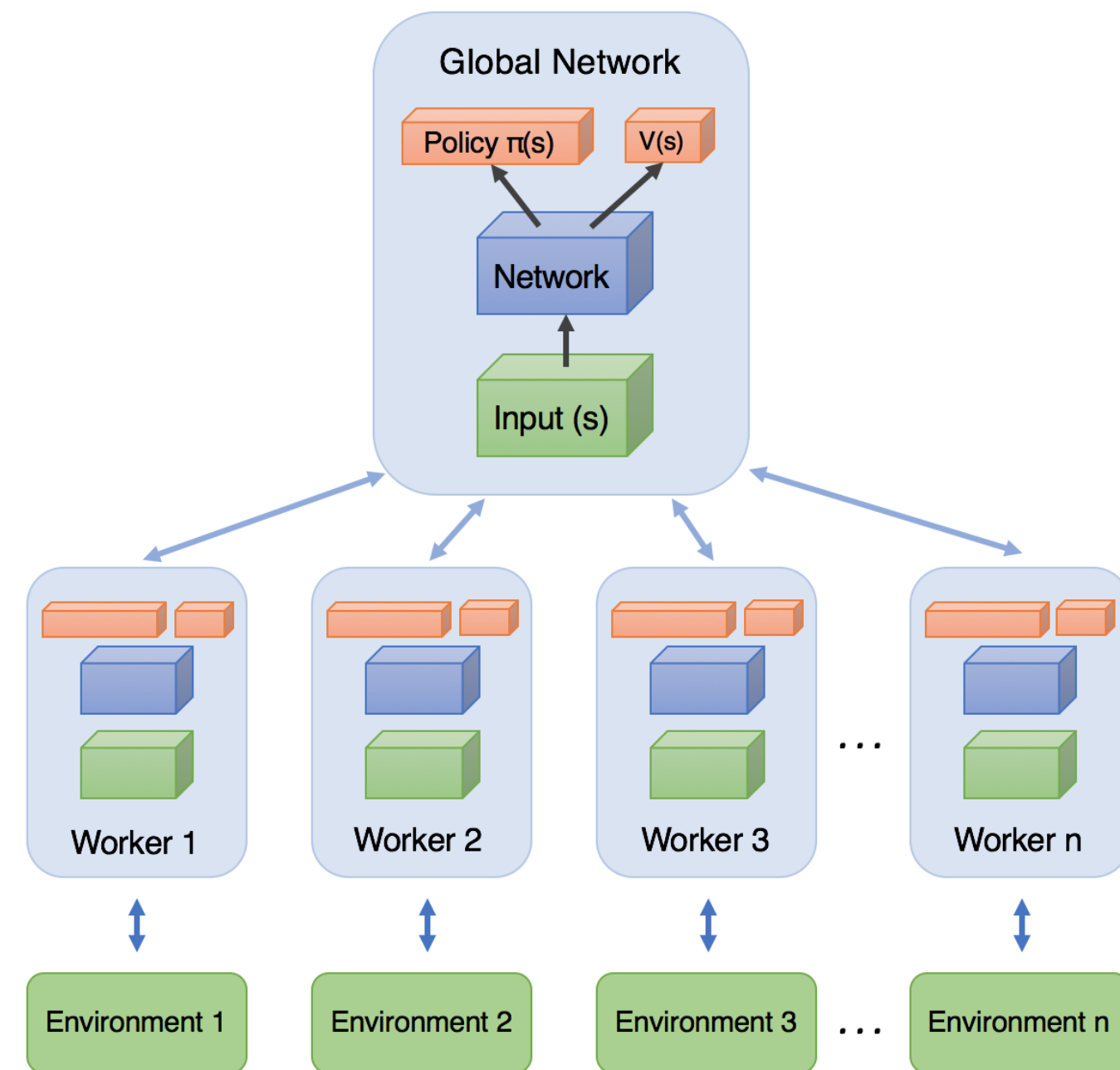


Асинхронный A2C: A3C

Взаимодействие с каждой средой и обучение происходит **асинхронно**

Преимущества:

- Работает быстрее (реальное время обучения)



Асинхронный A2C: A3C

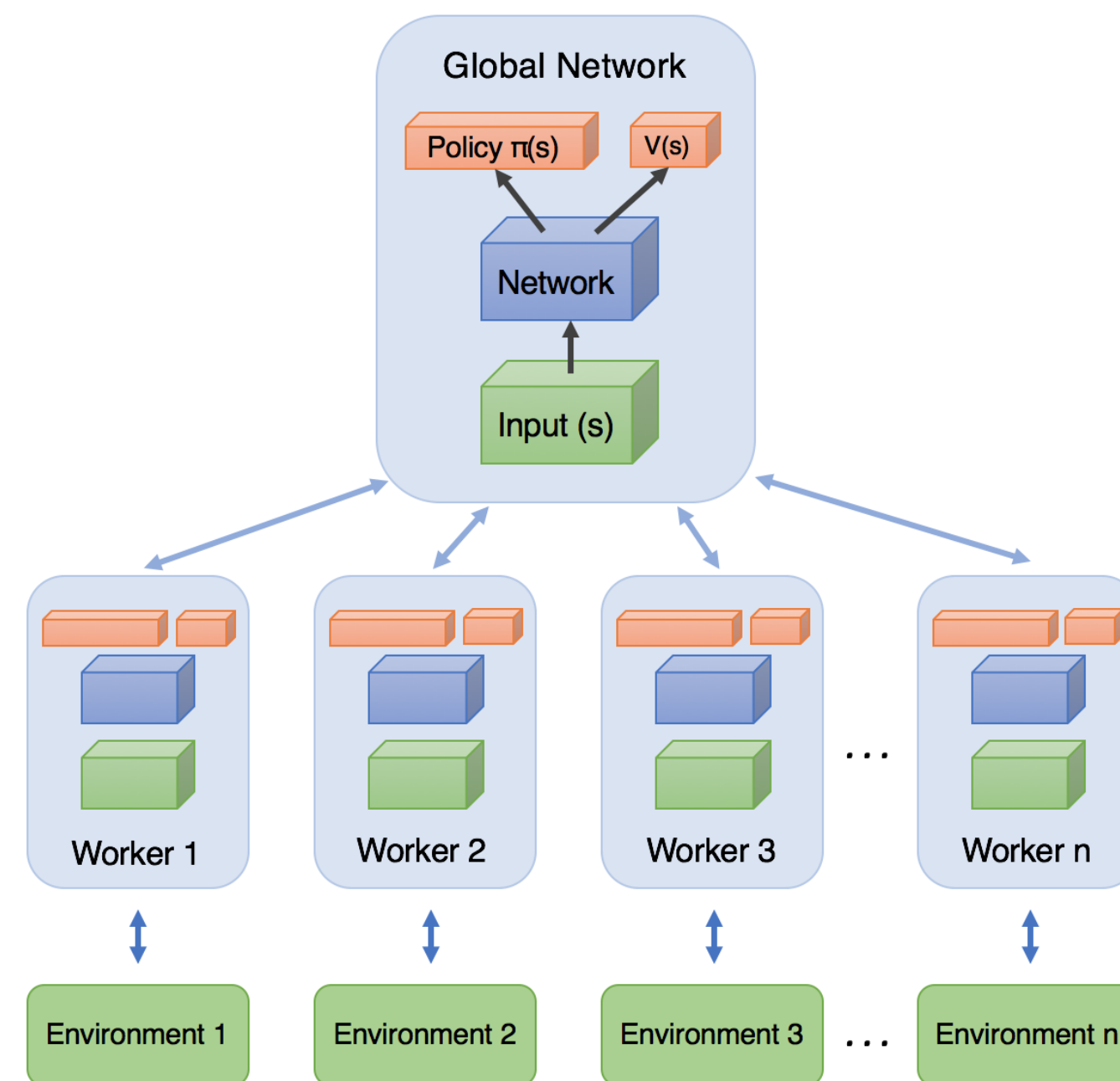
Взаимодействие с каждой средой и обучение происходит **асинхронно**

Преимущества:

- Работает быстрее (реальное время обучения)

Недостатки:

- Для N асинхронных рабочих нужно хранить $N+1$ копий параметров модели
- Проблема протухших градиентов



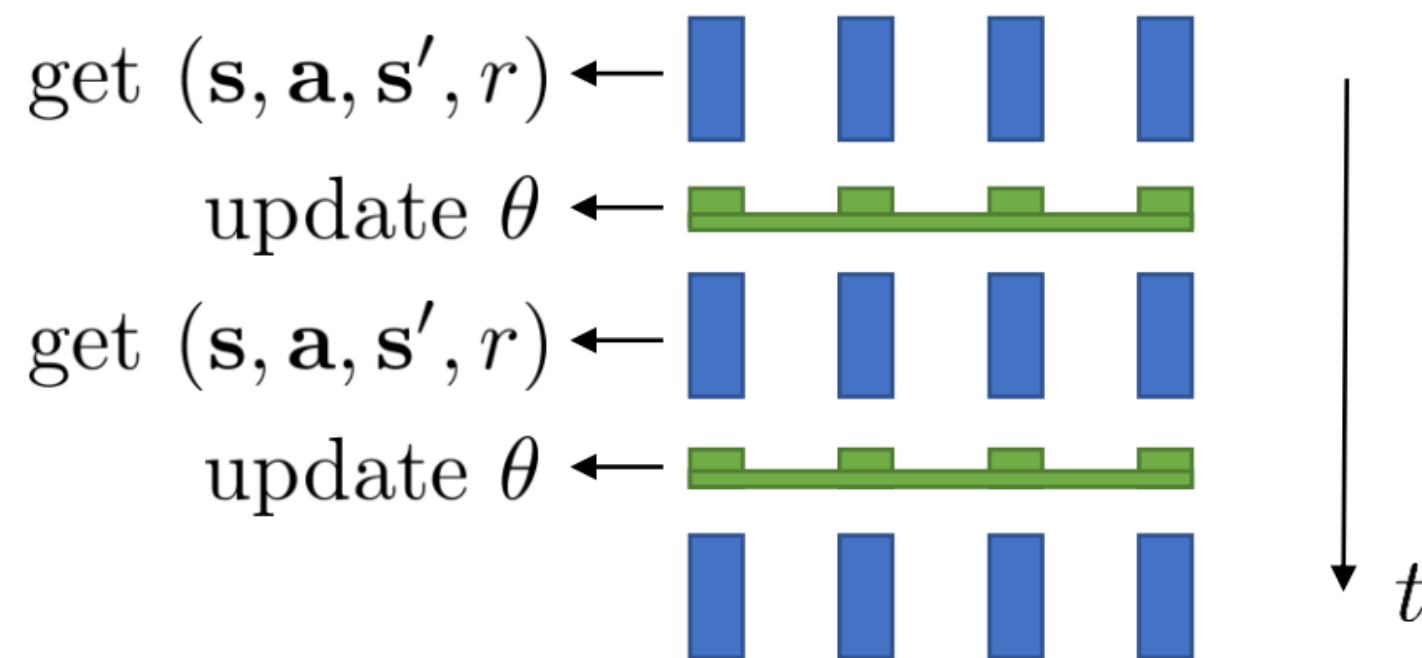
Синхронный параллельный A2C

Эту версию обычно называют A2C... снова...

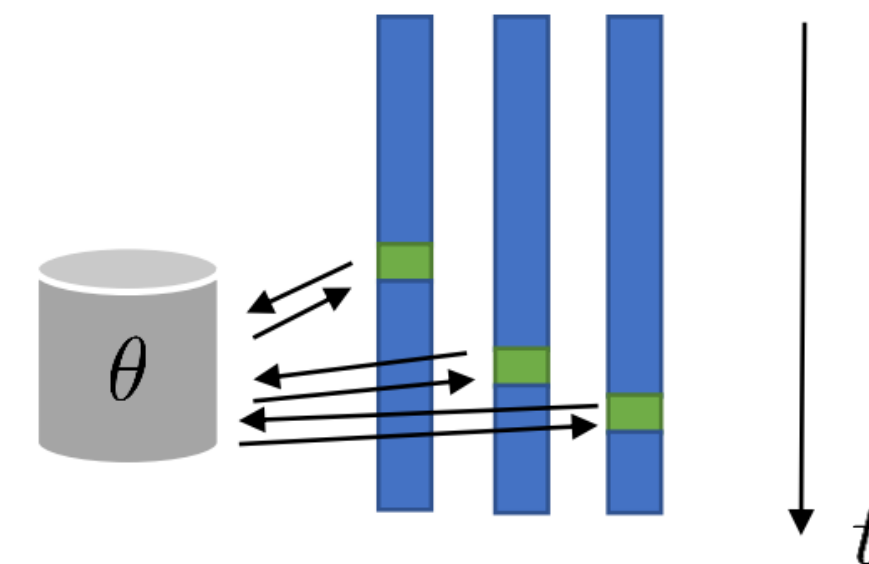
Решение проблем A3C:

- Пусть все среды работают параллельно
- **Среды синхронизируются** после каждого шага
- Можно выбирать действия **используя только одну копию параметров**
- Обновляем параметры каждые t шагов в среде/средах

synchronized parallel actor-critic



asynchronous parallel actor-critic



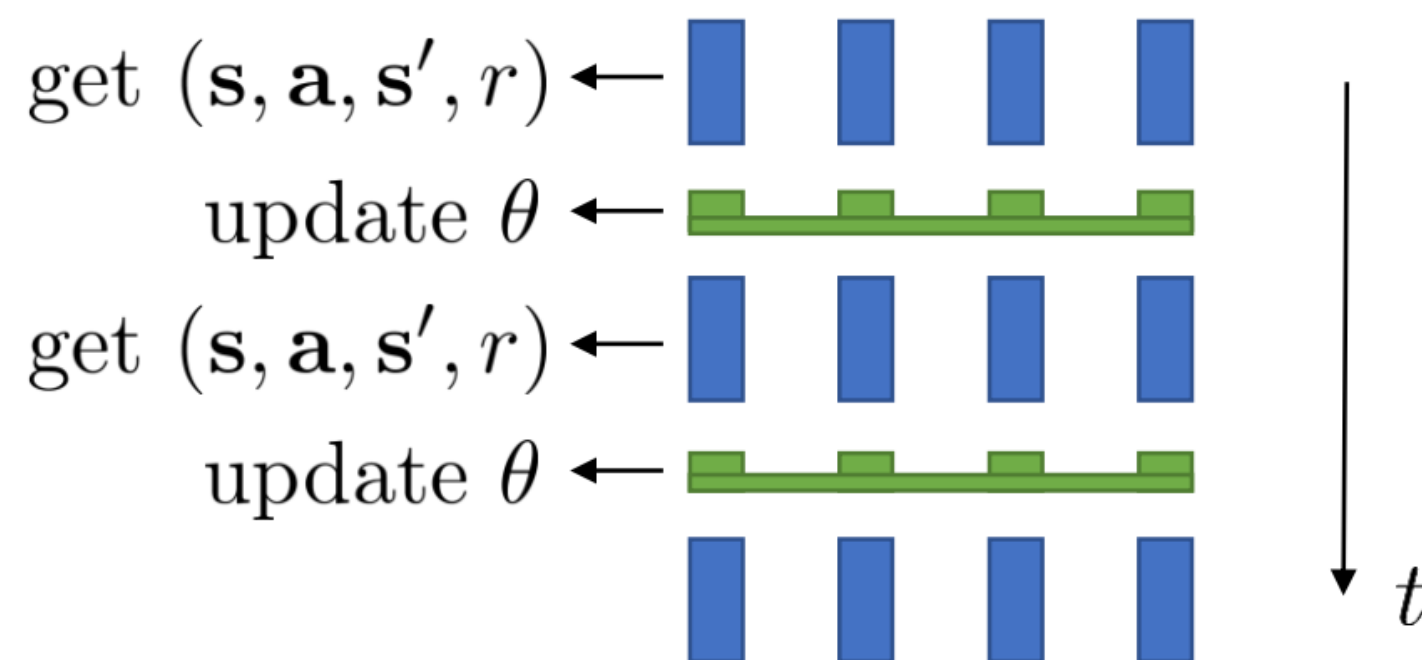
Синхронный параллельный A2C

Эту версию обычно называют A2C... снова...

Преимущества:

- Достаточно хранить только один набор параметров
- Стабильнее A3C
 - нет **протухших градиентов**

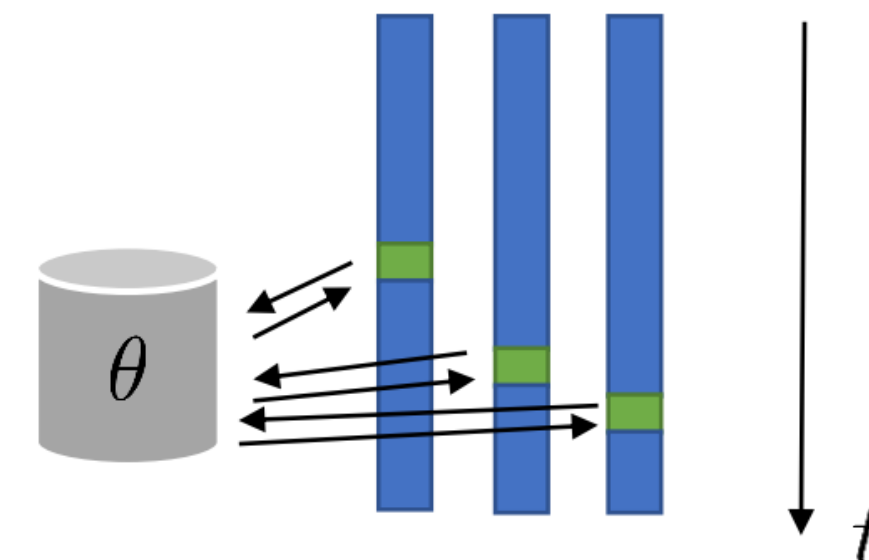
synchronized parallel actor-critic



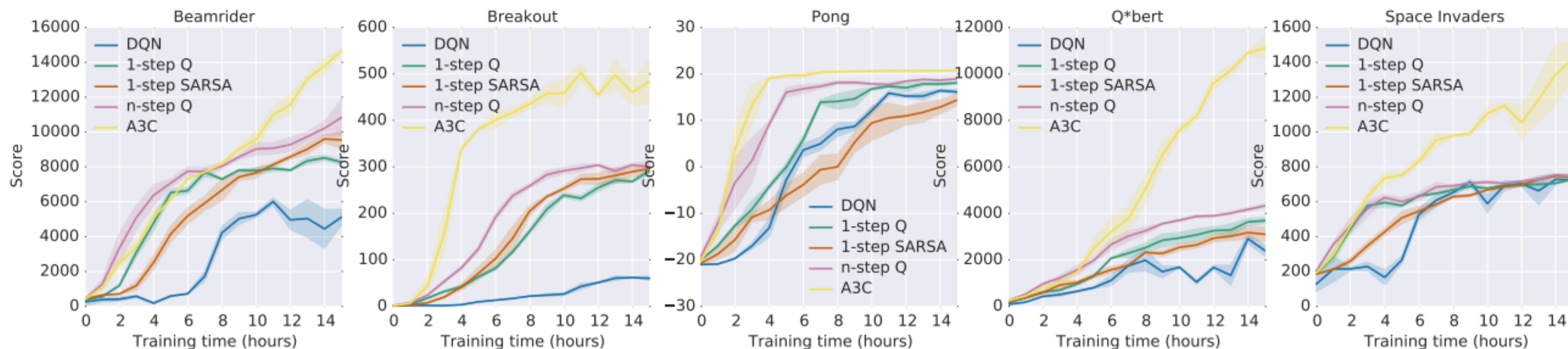
Недостатки:

- Немного медленнее, чем A3C
 - Число взаимодействий со средой в единицу времени

asynchronous parallel actor-critic



A3C/A2C Результаты:



Method	Training Time	Mean	Median
DQN	8 days on GPU	121.9%	47.5%
Gorila	4 days, 100 machines	215.2%	71.3%
D-DQN	8 days on GPU	332.9%	110.9%
Dueling D-DQN	8 days on GPU	343.8%	117.1%
Prioritized DQN	8 days on GPU	463.6%	127.6%
A3C, FF	1 day on CPU	344.1%	68.2%
A3C, FF	4 days on CPU	496.8%	116.6%
A3C, LSTM	4 days on CPU	623.0%	112.6%

Table 1. Mean and median human-normalized scores on 57 Atari games using the human starts evaluation metric. Supplementary

Спасибо за Внимание!