

# Лекция 8

## Обучение с подкреплением

Никита Юдин, iudin.ne@phystech.edu

Московский физико-технический институт  
Физтех-школа прикладной математики и информатики

3 апреля 2024



# Напоминание: Policy Gradient

$$J(\pi_\theta) := \mathbb{E}_{\mathcal{T} \sim \pi_\theta} \sum_{t \geq 0} \gamma^t r_t$$

# Напоминание: Policy Gradient

$$J(\pi_\theta) := \mathbb{E}_{\mathcal{T} \sim \pi_\theta} \sum_{t \geq 0} \gamma^t r_t$$

$$\nabla_\theta J(\pi_\theta) \approx \mathbb{E}_{\mathcal{T} \sim \pi_\theta} \underbrace{\sum_{t \geq 0} \overbrace{\nabla_\theta \log \pi_\theta(a_t | s_t)}^{\text{логарифм правдоподобия}} \left( \underbrace{Q^{\pi_\theta}(s_t, a_t) - V^{\pi_\theta}(s_t)}_{\substack{\text{бэйзлайн} \\ A^{\pi_\theta}(s_t, a_t)}} \right)}_{\substack{\text{данные порождённые } \pi_\theta \\ \text{оценка критика}}} \\ (\text{сэмплирование пар } s, a \text{ из траекторий } \mathcal{T} \sim \pi_\theta)$$

# Напоминание: Policy Gradient

$$J(\pi_\theta) := \mathbb{E}_{\mathcal{T} \sim \pi_\theta} \sum_{t \geq 0} \gamma^t r_t$$

$$\nabla_\theta J(\pi_\theta) \approx \mathbb{E}_{\mathcal{T} \sim \pi_\theta} \underbrace{\sum_{t \geq 0} \overbrace{\nabla_\theta \log \pi_\theta(a_t | s_t)}^{\text{логарифм правдоподобия}} \left( \underbrace{Q^{\pi_\theta}(s_t, a_t) - V^{\pi_\theta}(s_t)}_{\substack{\text{бэйзлайн} \\ A^{\pi_\theta}(s_t, a_t)}} \right)}_{\substack{\text{данные порождённые } \pi_\theta \\ \text{оценка критика}}} \\ (\text{сэмплирование пар } s, a \text{ из траекторий } \mathcal{T} \sim \pi_\theta)$$

Всё бы ничего, но это оценка вида *on-policy!*

# Существует более эффективный Policy Gradient?

$$\nabla_{\theta} J(\pi_{\theta}) = \frac{1}{1 - \gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi_{\theta}}(s)} \mathbb{E}_{a \sim \pi_{\theta}(a|s)}}_{\text{данные порождённые } \pi_{\theta}} \nabla_{\theta} \log \pi_{\theta}(a | s) A^{\pi_{\theta}}(s, a)$$

Предположим:

- хотим оптимизировать  $\pi_{\theta}$  (вычислить градиент для текущего  $\theta$ );

# Существует более эффективный Policy Gradient?

$$\nabla_{\theta} J(\pi_{\theta}) = \frac{1}{1 - \gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi_{\theta}}(s)} \mathbb{E}_{a \sim \pi_{\theta}(a|s)}}_{\text{данные порождённые } \pi_{\theta}} \nabla_{\theta} \log \pi_{\theta}(a | s) A^{\pi_{\theta}}(s, a)$$

Предположим:

- хотим оптимизировать  $\pi_{\theta}$  (вычислить градиент для текущего  $\theta$ );
- есть данные (траектории) порождённые  $\pi^{\text{old}}$ ;

# Существует более эффективный Policy Gradient?

$$\nabla_{\theta} J(\pi_{\theta}) = \frac{1}{1 - \gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi_{\theta}}(s)} \mathbb{E}_{a \sim \pi_{\theta}(a|s)}}_{\text{данные порождённые } \pi_{\theta}} \nabla_{\theta} \log \pi_{\theta}(a | s) A^{\pi_{\theta}}(s, a)$$

Предположим:

- хотим оптимизировать  $\pi_{\theta}$  (вычислить градиент для текущего  $\theta$ );
- есть данные (траектории) порождённые  $\pi^{\text{old}}$ ;
  - т.е. можем оценить  $\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)}$  и  $\mathbb{E}_{a \sim \pi^{\text{old}}(a|s)}$ ;

# Существует более эффективный Policy Gradient?

$$\nabla_{\theta} J(\pi_{\theta}) = \frac{1}{1 - \gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi_{\theta}}(s)} \mathbb{E}_{a \sim \pi_{\theta}(a|s)}}_{\text{данные порождённые } \pi_{\theta}} \nabla_{\theta} \log \pi_{\theta}(a | s) A^{\pi_{\theta}}(s, a)$$

Предположим:

- хотим оптимизировать  $\pi_{\theta}$  (вычислить градиент для текущего  $\theta$ );
- есть данные (траектории) порождённые  $\pi^{\text{old}}$ ;
  - т.е. можем оценить  $\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)}$  и  $\mathbb{E}_{a \sim \pi^{\text{old}}(a|s)}$ ;
  - т.е. можем обучить  $V^{\pi^{\text{old}}}(s)$  и, следовательно, оценить  $A^{\pi^{\text{old}}}(s, a)$ ;

# Существует более эффективный Policy Gradient?

$$\nabla_{\theta} J(\pi_{\theta}) = \frac{1}{1 - \gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi_{\theta}}(s)} \mathbb{E}_{a \sim \pi_{\theta}(a|s)}}_{\text{данные порождённые } \pi_{\theta}} \nabla_{\theta} \log \pi_{\theta}(a | s) A^{\pi_{\theta}}(s, a)$$

Предположим:

- хотим оптимизировать  $\pi_{\theta}$  (вычислить градиент для текущего  $\theta$ );
- есть данные (траектории) порождённые  $\pi^{\text{old}}$ ;
  - т.е. можем оценить  $\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)}$  и  $\mathbb{E}_{a \sim \pi^{\text{old}}(a|s)}$ ;
  - т.е. можем обучить  $V^{\pi^{\text{old}}}(s)$  и, следовательно, оценить  $A^{\pi^{\text{old}}}(s, a)$ ;



TRPO: применить оптимизацию эффективнее SGD!

# Соединяя две политики



Давайте **изменим награду** с помощью  
функции ценности другой политики!

# Соединяя две политики



Давайте **изменим награду** с помощью  
функции ценности другой политики!

Используя **телескопические суммы**: для произвольной последовательности  $a_t$  с условием  $\lim_{t \rightarrow \infty} a_t = 0$ :

$$\sum_{t \geq 0}^{\infty} (a_{t+1} - a_t) = -a_0$$

# Соединяя две политики



Давайте **изменим награду** с помощью  
функции ценности другой политики!

Используя **телескопические суммы**: для произвольной последовательности  $a_t$  с условием  $\lim_{t \rightarrow \infty} a_t = 0$ :

$$\sum_{t \geq 0}^{\infty} (a_{t+1} - a_t) = -a_0$$

Для произвольной траектории  $\mathcal{T} := s_0, a_0, s_1, a_1, \dots$  и произвольной политики  $\pi$ :

$$-V^\pi(s_0) = \sum_{t \geq 0} [\gamma^{t+1} V^\pi(s_{t+1}) - \gamma^t V^\pi(s_t)] \quad (1)$$

# Relative Performance Identity: доказательство

$$V^{\pi_\theta}(s) - \textcolor{violet}{V}^{\pi^{\text{old}}}(s) =$$

# Relative Performance Identity: доказательство

$$V^{\pi_\theta}(s) - \mathcal{V}^{\pi^{\text{old}}}(s) = \mathbb{E}_{T \sim \pi_\theta | s_0=s} \sum_{t \geq 0} \gamma^t r_t - \mathcal{V}^{\pi^{\text{old}}}(s) =$$

# Relative Performance Identity: доказательство

$$\begin{aligned} V^{\pi_\theta}(s) - \mathcal{V}^{\pi^{\text{old}}}(s) &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \sum_{t \geq 0} \gamma^t r_t - \mathcal{V}^{\pi^{\text{old}}}(s) = \\ &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \left[ \sum_{t \geq 0} \gamma^t r_t - \mathcal{V}^{\pi^{\text{old}}}(s_0) \right] = \end{aligned}$$

# Relative Performance Identity: доказательство

$$\begin{aligned} V^{\pi_\theta}(s) - \textcolor{violet}{V}^{\pi^{\text{old}}}(s) &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \sum_{t \geq 0} \gamma^t r_t - \textcolor{violet}{V}^{\pi^{\text{old}}}(s) = \\ &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \left[ \sum_{t \geq 0} \gamma^t r_t - \textcolor{violet}{V}^{\pi^{\text{old}}}(s_0) \right] = \\ \{\text{телескопические суммы (1)}\} &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \left[ \sum_{t \geq 0} \gamma^t r_t + \sum_{t \geq 0} \left[ \gamma^{t+1} \textcolor{violet}{V}^{\pi^{\text{old}}}(s_{t+1}) - \gamma^t \textcolor{violet}{V}^{\pi^{\text{old}}}(s_t) \right] \right] = \end{aligned}$$

# Relative Performance Identity: доказательство

$$\begin{aligned} V^{\pi_\theta}(s) - \mathcal{V}^{\pi^{\text{old}}}(s) &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \sum_{t \geq 0} \gamma^t r_t - \mathcal{V}^{\pi^{\text{old}}}(s) = \\ &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \left[ \sum_{t \geq 0} \gamma^t r_t - \mathcal{V}^{\pi^{\text{old}}}(s_0) \right] = \\ \{\text{телескопические суммы (1)}\} &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \left[ \sum_{t \geq 0} \gamma^t r_t + \sum_{t \geq 0} \left[ \gamma^{t+1} \mathcal{V}^{\pi^{\text{old}}}(s_{t+1}) - \gamma^t \mathcal{V}^{\pi^{\text{old}}}(s_t) \right] \right] = \\ \{\text{перегруппировка членов}\} &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \sum_{t \geq 0} \gamma^t \left( r_t + \gamma \mathcal{V}^{\pi^{\text{old}}}(s_{t+1}) - \mathcal{V}^{\pi^{\text{old}}}(s_t) \right) = \end{aligned}$$

# Relative Performance Identity: доказательство

$$\begin{aligned} V^{\pi_\theta}(s) - \mathcal{V}^{\pi^{\text{old}}}(s) &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \sum_{t \geq 0} \gamma^t r_t - \mathcal{V}^{\pi^{\text{old}}}(s) = \\ &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \left[ \sum_{t \geq 0} \gamma^t r_t - \mathcal{V}^{\pi^{\text{old}}}(s_0) \right] = \\ \{\text{телескопические суммы (1)}\} &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \left[ \sum_{t \geq 0} \gamma^t r_t + \sum_{t \geq 0} \left[ \gamma^{t+1} \mathcal{V}^{\pi^{\text{old}}}(s_{t+1}) - \gamma^t \mathcal{V}^{\pi^{\text{old}}}(s_t) \right] \right] = \\ \{\text{перегруппировка членов}\} &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \sum_{t \geq 0} \gamma^t \left( r_t + \gamma \mathcal{V}^{\pi^{\text{old}}}(s_{t+1}) - \mathcal{V}^{\pi^{\text{old}}}(s_t) \right) = \\ \{\text{по свойству } \mathbb{E}_x f(x) = \mathbb{E}_x \mathbb{E}_x f(x)\} &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \sum_{t \geq 0} \gamma^t \left( r_t + \gamma \mathbb{E}_{s_{t+1}} \mathcal{V}^{\pi^{\text{old}}}(s_{t+1}) - \mathcal{V}^{\pi^{\text{old}}}(s_t) \right) = \end{aligned}$$

# Relative Performance Identity: доказательство

$$\begin{aligned} V^{\pi_\theta}(s) - \mathcal{V}^{\pi^{\text{old}}}(s) &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \sum_{t \geq 0} \gamma^t r_t - \mathcal{V}^{\pi^{\text{old}}}(s) = \\ &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \left[ \sum_{t \geq 0} \gamma^t r_t - \mathcal{V}^{\pi^{\text{old}}}(s_0) \right] = \\ \{\text{телескопические суммы (1)}\} &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \left[ \sum_{t \geq 0} \gamma^t r_t + \sum_{t \geq 0} \left[ \gamma^{t+1} \mathcal{V}^{\pi^{\text{old}}}(s_{t+1}) - \gamma^t \mathcal{V}^{\pi^{\text{old}}}(s_t) \right] \right] = \\ \{\text{перегруппировка членов}\} &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \sum_{t \geq 0} \gamma^t \left( r_t + \gamma \mathcal{V}^{\pi^{\text{old}}}(s_{t+1}) - \mathcal{V}^{\pi^{\text{old}}}(s_t) \right) = \\ \{\text{по свойству } \mathbb{E}_x f(x) = \mathbb{E}_x \mathbb{E}_x f(x)\} &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \sum_{t \geq 0} \gamma^t \left( r_t + \gamma \mathbb{E}_{s_{t+1}} \mathcal{V}^{\pi^{\text{old}}}(s_{t+1}) - \mathcal{V}^{\pi^{\text{old}}}(s_t) \right) = \\ \{\text{по определению Q-функции}\} &= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0 = s} \sum_{t \geq 0} \gamma^t \left( Q^{\pi^{\text{old}}}(s_t, a_t) - \mathcal{V}^{\pi^{\text{old}}}(s_t) \right) \end{aligned}$$

# Relative Performance Identity: доказательство

$$V^{\pi_\theta}(s) - V^{\pi^{\text{old}}}(s) = \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0=s} \sum_{t \geq 0} \gamma^t r_t - V^{\pi^{\text{old}}}(s) =$$

$$= \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0=s} \left[ \sum_{t \geq 0} \gamma^t r_t - V^{\pi^{\text{old}}}(s_0) \right] =$$

$$\{\text{телескопические суммы (1)}\} = \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0=s} \left[ \sum_{t \geq 0} \gamma^t r_t + \sum_{t \geq 0} [\gamma^{t+1} V^{\pi^{\text{old}}}(s_{t+1}) - \gamma^t V^{\pi^{\text{old}}}(s_t)] \right] =$$

$$\{\text{перегруппировка членов}\} = \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0=s} \sum_{t \geq 0} \gamma^t (r_t + \gamma V^{\pi^{\text{old}}}(s_{t+1}) - V^{\pi^{\text{old}}}(s_t)) =$$

$$\{\text{по свойству } \mathbb{E}_x f(x) = \mathbb{E}_x \mathbb{E}_x f(x)\} = \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0=s} \sum_{t \geq 0} \gamma^t (r_t + \gamma \mathbb{E}_{s_{t+1}} V^{\pi^{\text{old}}}(s_{t+1}) - V^{\pi^{\text{old}}}(s_t)) =$$

$$\{\text{по определению Q-функции}\} = \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0=s} \sum_{t \geq 0} \gamma^t (Q^{\pi^{\text{old}}}(s_t, a_t) - V^{\pi^{\text{old}}}(s_t))$$

$$\{\text{по определению advantage-функции}\} = \mathbb{E}_{\mathcal{T} \sim \pi_\theta | s_0=s} \sum_{t \geq 0} \gamma^t A^{\pi^{\text{old}}}(s_t, a_t)$$

# Меняем целевую функцию

Мы можем заменить оптимизируемый функционал на:

$$J(\pi_\theta) - J(\pi^{\text{old}}) = \frac{1}{1-\gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi_\theta}(s)} \mathbb{E}_{a \sim \pi_\theta(a|s)}}_{\begin{array}{l} \text{старый критик!} \\ (\text{хорошо: можем обучить!}) \\ \text{собрать данные с помощью } \pi_\theta \\ (\text{плохо! политика неизвестна!}) \end{array}} \overbrace{A^{\pi^{\text{old}}}(s, a)}$$

# Меняем целевую функцию

Мы можем заменить оптимизируемый функционал на:

$$J(\pi_\theta) - J(\pi^{\text{old}}) = \frac{1}{1-\gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi_\theta}(s)} \mathbb{E}_{a \sim \pi_\theta(a|s)}}_{\substack{\text{собрать данные с помощью } \pi_\theta \\ (\text{плохо! политика неизвестна!})}} \overbrace{A^{\pi^{\text{old}}}(s, a)}^{\substack{\text{старый критик!} \\ (\text{хорошо: можем обучить!})}}$$

- ✓ можно решить проблему  $\mathbb{E}_{a \sim \pi_\theta(a|s)}$  с помощью **выборки по значимости**;

# Меняем целевую функцию

Мы можем заменить оптимизируемый функционал на:

$$J(\pi_\theta) - J(\pi^{\text{old}}) = \frac{1}{1-\gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi_\theta}(s)} \mathbb{E}_{a \sim \pi_\theta(a|s)}}_{\substack{\text{собрать данные с помощью } \pi_\theta \\ (\text{плохо! политика неизвестна!})}} \overbrace{A^{\pi^{\text{old}}}(s, a)}^{\substack{\text{старый критик!} \\ (\text{хорошо: можем обучить!})}}$$

- ✓ можно решить проблему  $\mathbb{E}_{a \sim \pi_\theta(a|s)}$  с помощью **выборки по значимости**;
- ✗ неизвестно, как отсэмплировать  $\mathbb{E}_{s \sim d_{\pi_\theta}(s)}$ !

# Суррогатный функционал

Введём **суррогатный функционал**:

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta) :=$$

- с этим можно работать;
- указывает на улучшение политики  $\pi^{\text{old}}$ :
  - оптимизация  $\theta$  с фиксированной  $\pi^{\text{old}}$  обучит  $\arg\max_a A^{\pi^{\text{old}}}(s, a)$
  - оптимизация  $\theta$  с фиксированными *данными* обучит  $\pi_\theta(a | s) = 1$  если  $A(s, a) > 0$ ,  $\pi_\theta(a | s) = 0$  иначе.

# Суррогатный функционал

Введём **суррогатный функционал**:

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta) := \underbrace{\frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \frac{\pi_\theta(a | s)}{\pi^{\text{old}}(a | s)}}_{\text{данные порождённые } \pi^{\text{old}}} \underbrace{A^{\pi^{\text{old}}}(s, a)}_{\substack{\text{коррекция выборкой} \\ \text{по значимости}}}$$

не требуется  
свежий критик

# Суррогатный функционал

Введём **суррогатный функционал**:

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta) := \frac{1}{1-\gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)}}_{\text{данные порождённые } \pi^{\text{old}}} \underbrace{\frac{\pi_\theta(a | s)}{\pi^{\text{old}}(a | s)}}_{\substack{\text{коррекция выборкой} \\ \text{по значимости}}} \underbrace{A^{\pi^{\text{old}}}(s, a)}_{\substack{\text{не требуется} \\ \text{свежий критик}}}$$

- с этим можно работать;

# Суррогатный функционал

Введём **суррогатный функционал**:

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta) := \frac{1}{1-\gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)}}_{\text{данные порождённые } \pi^{\text{old}}} \underbrace{\frac{\pi_\theta(a | s)}{\pi^{\text{old}}(a | s)}}_{\substack{\text{коррекция выборкой} \\ \text{по значимости}}} \underbrace{A^{\pi^{\text{old}}}(s, a)}_{\substack{\text{не требуется} \\ \text{свежий критик}}}$$

- с этим можно работать;
- указывает на улучшение политики  $\pi^{\text{old}}$ :
  - оптимизация  $\theta$  с фиксированной  $\pi^{\text{old}}$  обучит
$$\operatorname{argmax}_a A^{\pi^{\text{old}}}(s, a)$$

# Суррогатный функционал

Введём **суррогатный функционал**:

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta) := \frac{1}{1-\gamma} \underbrace{\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)}}_{\text{данные порождённые } \pi^{\text{old}}} \underbrace{\frac{\pi_\theta(a | s)}{\pi^{\text{old}}(a | s)}}_{\substack{\text{коррекция выборкой} \\ \text{по значимости}}} \underbrace{A^{\pi^{\text{old}}}(s, a)}_{\substack{\text{не требуется} \\ \text{свежий критик}}}$$

- с этим можно работать;
- указывает на улучшение политики  $\pi^{\text{old}}$ :
  - оптимизация  $\theta$  с фиксированной  $\pi^{\text{old}}$  обучит  $\underset{a}{\operatorname{argmax}} A^{\pi^{\text{old}}}(s, a)$
  - оптимизация  $\theta$  с фиксированными *данными* обучит  $\pi_\theta(a | s) = 1$  если  $A(s, a) > 0$ ,  $\pi_\theta(a | s) = 0$  иначе.

# Теоретическая граница ошибки

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta)$$

Хотим оптимизировать другой функционал. **Насколько ошибёмся?**

# Теоретическая граница ошибки

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta)$$

Хотим оптимизировать другой функционал. **Насколько ошибёмся?**

Теорема

$$|J(\pi_\theta) - J(\pi^{\text{old}}) - L_{\pi^{\text{old}}}(\theta)| \leq C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta),$$

# Теоретическая граница ошибки

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta)$$

Хотим оптимизировать другой функционал. **Насколько ошибёмся?**

Теорема

$$|J(\pi_\theta) - J(\pi^{\text{old}}) - L_{\pi^{\text{old}}}(\theta)| \leq C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta),$$

где:

$$\text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta) := \max_s \text{KL}(\pi^{\text{old}}(\cdot \mid s) \parallel \pi_\theta(\cdot \mid s))$$

# Теоретическая граница ошибки

$$J(\pi_\theta) - J(\pi^{\text{old}}) \approx L_{\pi^{\text{old}}}(\theta)$$

Хотим оптимизировать другой функционал. **Насколько ошибёмся?**

Теорема

$$|J(\pi_\theta) - J(\pi^{\text{old}}) - L_{\pi^{\text{old}}}(\theta)| \leq C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta),$$

где:

$$\text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta) := \max_s \text{KL}(\pi^{\text{old}}(\cdot \mid s) \parallel \pi_\theta(\cdot \mid s))$$

$$C := \frac{4\gamma \max_{s,a} |A^{\pi^{\text{old}}}(s, a)|}{(1 - \gamma)^2}$$

# Алгоритм миноризации-максимизации



Нашли **вариационную нижнюю оценку** для нашего функционала:

$$J(\pi_\theta) - J(\pi^{\text{old}}) \geq L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta)$$

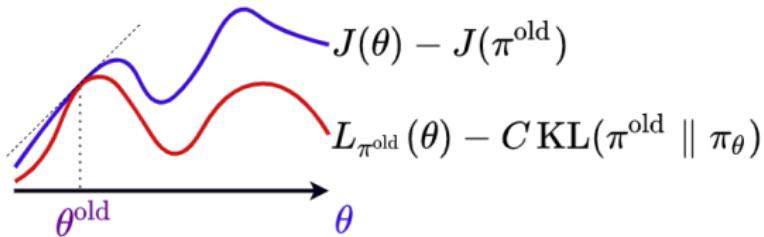
# Алгоритм миноризации-максимизации



Нашли **вариационную нижнюю оценку** для нашего функционала:

$$J(\pi_\theta) - J(\pi^{\text{old}}) \geq L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta)$$

- **Миноризация:** построить новую нижнюю оценку; в нашем случае использовать  $\pi^{\text{old}} \leftarrow \pi_\theta$ .



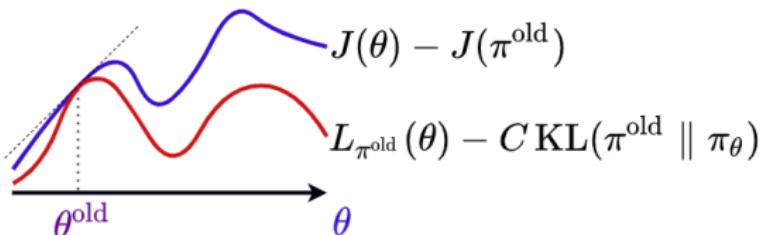
# Алгоритм миноризации-максимизации



Нашли **вариационную нижнюю оценку** для нашего функционала:

$$J(\pi_\theta) - J(\pi^{\text{old}}) \geq L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta)$$

- **Миноризация:** построить новую нижнюю оценку; в нашем случае использовать  $\pi^{\text{old}} \leftarrow \pi_\theta$ .
- **Максимизация:** оптимизация нижней оценки (произвольным способом).



# Алгоритм миноризации-максимизации



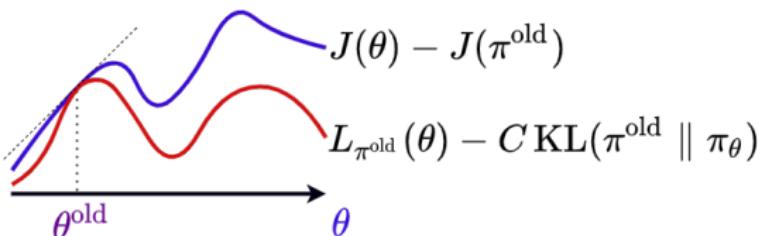
Нашли **вариационную нижнюю оценку** для нашего функционала:

$$J(\pi_\theta) - J(\pi^{\text{old}}) \geq L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta)$$

- **Миноризация:** построить новую нижнюю оценку; в нашем случае использовать  $\pi^{\text{old}} \leftarrow \pi_\theta$ .

- **Максимизация:** оптимизация нижней оценки (произвольным способом).

✓ гарантия монотонного улучшения!



# Воспользуемся принципом "проб и ошибок"

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

Проблемы:

- кроитик неидеален :(



# Воспользуемся принципом "проб и ошибок"

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

## Проблемы:

- кроитик неидеален :(
  - используем что есть...



# Воспользуемся принципом "проб и ошибок"

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

## Проблемы:

- кроитик неидеален :(
  - используем что есть...
- не умеем работать с  $\text{KL}^{\max}$  :(



# Воспользуемся принципом "проб и ошибок"

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

## Проблемы:

- кроитик неидеален :(
  - используем что есть...
- не умеем работать с  $\text{KL}^{\max}$  :(
  - меняем на  
 $\mathbb{E}_s \text{KL}(\pi^{\text{old}}(\cdot | s) \parallel \pi_\theta(\cdot | s)))$ ...



# Воспользуемся принципом "проб и ошибок"

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

## Проблемы:

- кроитик неидеален :(
  - используем что есть...
- не умеем работать с  $\text{KL}^{\max}$  :(
  - меняем на  
 $\mathbb{E}_s \text{KL}(\pi^{\text{old}}(\cdot | s) \parallel \pi_\theta(\cdot | s)))$ ...
- не знаем  $C$  :(



# Воспользуемся принципом "проб и ошибок"

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

## Проблемы:

- кроитик неидеален :(
  - используем что есть...
- не умеем работать с  $\text{KL}^{\max}$  :(
  - меняем на  $\mathbb{E}_s \text{KL}(\pi^{\text{old}}(\cdot | s) \parallel \pi_\theta(\cdot | s)))$ ...
- не знаем  $C$  :(
  - сделаем гиперпараметром...



# Воспользуемся принципом "проб и ошибок"

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

## Проблемы:

- кроитик неидеален :(
  - используем что есть...
- не умеем работать с  $\text{KL}^{\max}$  :(
  - меняем на  $\mathbb{E}_s \text{KL}(\pi^{\text{old}}(\cdot | s) \parallel \pi_\theta(\cdot | s)))$ ...
- не знаем  $C$  :(
  - сделаем гиперпараметром...
- и на деле он очень большой :(
  - Хм...

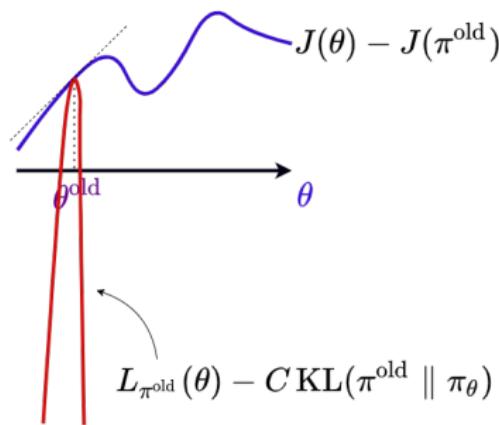


# Воспользуемся принципом "проб и ошибок"

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

## Проблемы:

- кроитик неидеален :(
  - используем что есть...
- не умеем работать с  $\text{KL}^{\max}$  :(
  - меняем на  $\mathbb{E}_s \text{KL}(\pi^{\text{old}}(\cdot | s) \parallel \pi_\theta(\cdot | s))$ ...
- не знаем  $C$  :(
  - сделаем гиперпараметром...
- и на деле он очень большой :(
  - Хм...

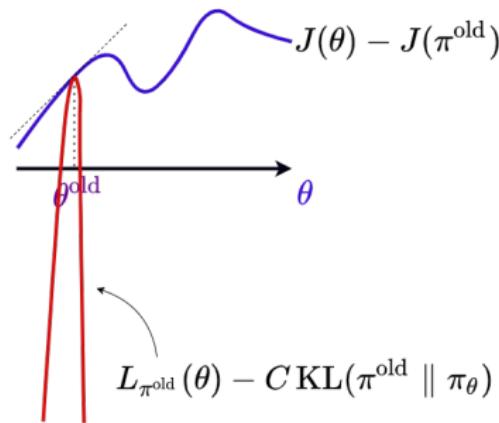


# Воспользуемся принципом "проб и ошибок"

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}^{\max}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

## Проблемы:

- критик неидеален :(
  - используем что есть...
- не умеем работать с  $\text{KL}^{\max}$  :(
  - меняем на  $\mathbb{E}_s \text{KL}(\pi^{\text{old}}(\cdot | s) \parallel \pi_\theta(\cdot | s))$ ...
- не знаем  $C$  :(
  - сделаем гиперпараметром...
- и на деле он очень большой :(
  - Хм...



Directed by  
ROBERT B. WEIDE



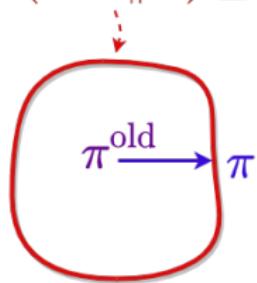
Оптимизировать  $L_{\pi^{\text{old}}}(\theta)$  с условием, что  
 $\pi_\theta$  не изменится сильно относительно  $\pi^{\text{old}}$ !



Оптимизировать  $L_{\pi^{\text{old}}}(\theta)$  с условием, что  $\pi_\theta$  не изменится сильно относительно  $\pi^{\text{old}}$ !

$$\begin{cases} L_{\pi^{\text{old}}}(\theta) \rightarrow \max_{\theta} \\ \mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \text{KL}(\pi^{\text{old}}(\cdot | s) \| \pi_\theta(\cdot | s)) \leq \delta \end{cases}$$

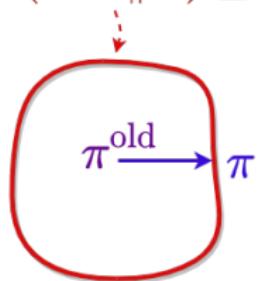
$$\text{KL}(\pi^{\text{old}} \| \pi) \leq \delta$$





Оптимизировать  $L_{\pi^{\text{old}}}(\theta)$  с условием, что  $\pi_\theta$  не изменится сильно относительно  $\pi^{\text{old}}$ !

$$\text{KL}(\pi^{\text{old}} \parallel \pi) \leq \delta$$



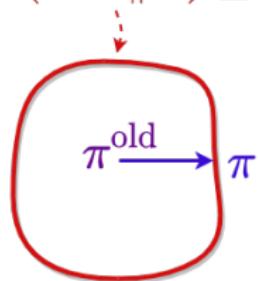
$$\begin{cases} L_{\pi^{\text{old}}}(\theta) \rightarrow \max_{\theta} \\ \mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \text{KL}(\pi^{\text{old}}(\cdot | s) \parallel \pi_\theta(\cdot | s)) \leq \delta \end{cases}$$

- $L_{\pi^{\text{old}}}(\theta)$  – модель функционала;
- условие – доверительная область этой модели;



Оптимизировать  $L_{\pi^{\text{old}}}(\theta)$  с условием, что  $\pi_\theta$  не изменится сильно относительно  $\pi^{\text{old}}$ !

$$\text{KL}(\pi^{\text{old}} \parallel \pi) \leq \delta$$



$$\begin{cases} L_{\pi^{\text{old}}}(\theta) \rightarrow \max_{\theta} \\ \mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \text{KL}(\pi^{\text{old}}(\cdot | s) \parallel \pi_\theta(\cdot | s)) \leq \delta \end{cases}$$

- $L_{\pi^{\text{old}}}(\theta)$  – модель функционала;
- условие – доверительная область этой модели;

Алгоритм TRPO решает приблизительно данную задачу. Сложная имплементация.



Использовать разложение  
Тейлора для модели и  
условия при решении задачи!



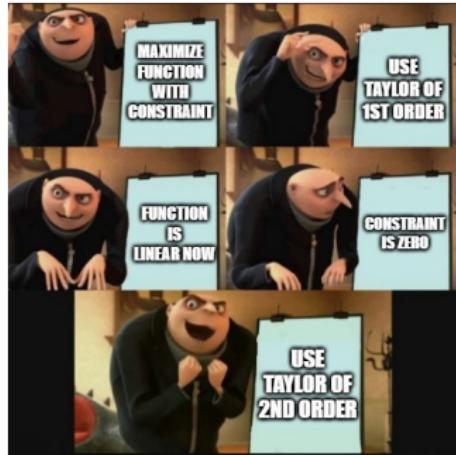
Использовать разложение  
Тейлора для модели и  
условия при решении задачи!

$$L_{\pi^{\text{old}}}(\theta) \approx g^T d \rightarrow \max_d$$
$$d := \theta - \theta^{\text{old}}, \quad g := \nabla_{\theta} L_{\pi^{\text{old}}}(\theta)|_{\theta=\theta^{\text{old}}}$$



Использовать разложение  
Тейлора для модели и  
условия при решении задачи!

$$L_{\pi^{\text{old}}}(\theta) \approx g^T d \rightarrow \max_d$$
$$d := \theta - \theta^{\text{old}}, \quad g := \nabla_{\theta} L_{\pi^{\text{old}}}(\theta)|_{\theta=\theta^{\text{old}}}$$



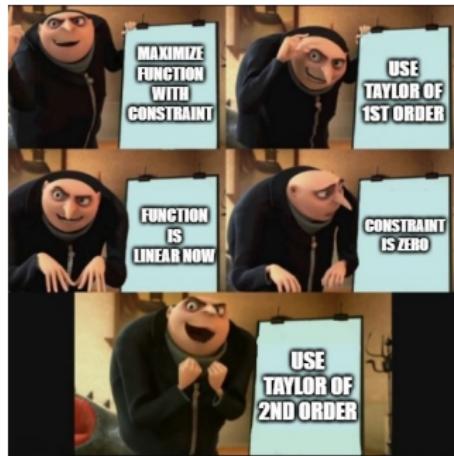


Использовать разложение  
Тейлора для модели и  
условия при решении задачи!

$$L_{\pi^{\text{old}}}(\theta) \approx g^T d \rightarrow \max_d$$
$$d := \theta - \theta^{\text{old}}, \quad g := \nabla_{\theta} L_{\pi^{\text{old}}}(\theta)|_{\theta=\theta^{\text{old}}}$$

Для  $KL$ -дивергенции, первое слагаемое 0,  
используем разложение второго порядка:

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \text{KL}(\pi^{\text{old}}(\cdot | s) \| \pi_{\theta}(\cdot | s)) \approx \frac{1}{2} d^T H d \leq \delta$$



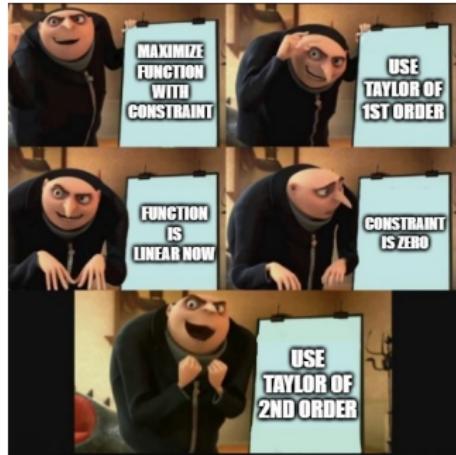


Использовать разложение  
Тейлора для модели и  
условия при решении задачи!

$$L_{\pi^{\text{old}}}(\theta) \approx g^T d \rightarrow \max_d$$
$$d := \theta - \theta^{\text{old}}, \quad g := \nabla_{\theta} L_{\pi^{\text{old}}}(\theta)|_{\theta=\theta^{\text{old}}}$$

Для  $KL$ -дивергенции, первое слагаемое 0,  
используем разложение второго порядка:

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \text{KL}(\pi^{\text{old}}(\cdot | s) \| \pi_{\theta}(\cdot | s)) \approx \frac{1}{2} d^T H d \leq \delta$$



Натуральный градиентный подъём

Решение:  $d \propto H^{-1}g$

$$\theta_{k+1} = \theta_k + \alpha H^{-1} g$$

### Внимание!

Натуральный градиентный подъём похож на методы второго порядка.

$$\theta_{k+1} = \theta_k + \alpha H^{-1} g$$

### Внимание!

Натуральный градиентный подъём похож на методы второго порядка.

Допустим,  $h$  – размерность  
 $\theta$ ...

**X**  $H$  – матрица  $h \times h$   
(гессиан  
KL-дивергенции);

$$\theta_{k+1} = \theta_k + \alpha H^{-1} g$$

### Внимание!

Натуральный градиентный подъём похож на методы второго порядка.

Допустим,  $h$  – размерность  
 $\theta$ ...

- ☒  $H$  – матрица  $h \times h$   
(гессиан  
KL-дивергенции);
- ☒ Как обращать?

$$\theta_{k+1} = \theta_k + \alpha H^{-1} g$$

### Внимание!

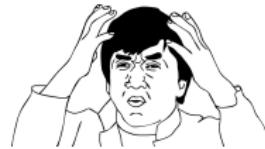
Натуральный градиентный подъём похож на методы второго порядка.

Допустим,  $h$  – размерность  
 $\theta$ ...

**X**  $H$  – матрица  $h \times h$

(гессиан  
KL-дивергенции);

**X** Как обращать?



$$\theta_{k+1} = \theta_k + \alpha H^{-1} g$$

### Внимание!

Натуральный градиентный подъём похож на методы второго порядка.

Допустим,  $h$  – размерность  
 $\theta$ ...

**X**  $H$  – матрица  $h \times h$

(гессиан  
KL-дивергенции);

**X** Как обращать?



Решать следующую СЛАУ:



$$Hd = g$$

с помощью **метода сопряжённых градиентов.**

$$\theta_{k+1} = \theta_k + \alpha H^{-1} g$$

### Внимание!

Натуральный градиентный подъём похож на методы второго порядка.

Допустим,  $h$  – размерность  
 $\theta$ ...

**X**  $H$  – матрица  $h \times h$

(гессиан  
KL-дивергенции);

**X** Как обращать?



Решать следующую СЛАУ:



$$Hd = g$$

с помощью **метода сопряжённых градиентов**.

1. вычислить  $g$   
(как в обычном градиентном спуске);

$$\theta_{k+1} = \theta_k + \alpha H^{-1} g$$

### Внимание!

Натуральный градиентный подъём похож на методы второго порядка.

Допустим,  $h$  – размерность  
θ...

✗  $H$  – матрица  $h \times h$   
(гессиан  
KL-дивергенции);

✗ Как обращать?



Решать следующую СЛАУ:



$$Hd = g$$

с помощью **метода  
сопряжённых градиентов.**

1. вычислить  $g$   
(как в обычном градиентном спуске);
2. инициализировать  $d_0$  произвольно;
3. для  $i = 0, 1, 2 \dots M$ :

$$\theta_{k+1} = \theta_k + \alpha H^{-1} g$$

### Внимание!

Натуральный градиентный подъём похож на методы второго порядка.

Допустим,  $h$  – размерность  
θ...

✗  $H$  – матрица  $h \times h$   
(гессиан  
KL-дивергенции);

✗ Как обращать?



Решать следующую СЛАУ:



$$Hd = g$$

с помощью **метода сопряжённых градиентов**.

1. вычислить  $g$   
(как в обычном градиентном спуске);
2. инициализировать  $d_0$  произвольно;
3. для  $i = 0, 1, 2 \dots M$ :
  - метод сопряжённых градиентов приближённо вычисляет  $d_{i+1}$  с помощью  $g$  и вектора  $Hd_i$ .

$$\theta_{k+1} = \theta_k + \alpha H^{-1} g$$

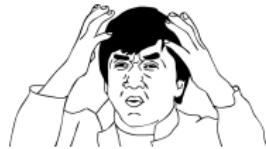
### Внимание!

Натуральный градиентный подъём похож на методы второго порядка.

Допустим,  $h$  – размерность  
 $\theta$ ...

✗  $H$  – матрица  $h \times h$   
(гессиан  
KL-дивергенции);

✗ Как обращать?



Решать следующую СЛАУ:



$$Hd = g$$

с помощью **метода сопряжённых градиентов**.

1. вычислить  $g$   
(как в обычном градиентном спуске);
2. инициализировать  $d_0$  произвольно;
3. для  $i = 0, 1, 2 \dots M$ :
  - метод сопряжённых градиентов приближённо вычисляет  $d_{i+1}$  с помощью  $g$  и вектора  $Hd_i$ .
4. с помощью линейного поиска найти  $\alpha_k$ ;

$$\theta_{k+1} = \theta_k + \alpha H^{-1} g$$

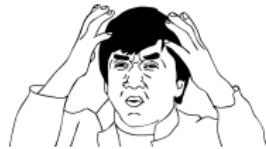
### Внимание!

Натуральный градиентный подъём похож на методы второго порядка.

Допустим,  $h$  – размерность  
 $\theta$ ...

✗  $H$  – матрица  $h \times h$   
(гессиан  
KL-дивергенции);

✗ Как обращать?



Решать следующую СЛАУ:



$$Hd = g$$

с помощью **метода сопряжённых градиентов**.

1. вычислить  $g$   
(как в обычном градиентном спуске);
2. инициализировать  $d_0$  произвольно;
3. для  $i = 0, 1, 2 \dots M$ :
  - метод сопряжённых градиентов приближённо вычисляет  $d_{i+1}$  с помощью  $g$  и вектора  $Hd_i$ .
4. с помощью линейного поиска найти  $\alpha_k$ ;
5.  $\theta_{k+1} = \theta_k + \alpha_k d_M$

# Линейный поиск

Как определить шаг  $\alpha_k$ ?  
(он связан с  $\delta$  в подходах с доверительной  
областью)

$$\theta_{k+1} = \theta_k + \alpha_k d_k$$



# Линейный поиск

Как определить шаг  $\alpha_k$ ?  
(он связан с  $\delta$  в подходах с доверительной областью)

$$\theta_{k+1} = \theta_k + \alpha_k d_k$$



Использовать **бэктрекинг** для поиска  $\alpha_k$ , так что:



$$L_{\pi^{\text{old}}}(\theta_{k+1}) > 0, \quad \mathbb{E}_s \text{KL}(\pi^{\text{old}}(\cdot | s) \| \pi_{\theta_{k+1}}(\cdot | s)) \leq \delta$$

# Линейный поиск

Как определить шаг  $\alpha_k$ ?  
(он связан с  $\delta$  в подходах с доверительной областью)

$$\theta_{k+1} = \theta_k + \alpha_k d_k$$



Использовать **бэктрекинг** для поиска  $\alpha_k$ , так что:



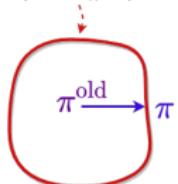
$$L_{\pi^{\text{old}}}(\theta_{k+1}) > 0, \quad \mathbb{E}_s \text{KL}(\pi^{\text{old}}(\cdot | s) \| \pi_{\theta_{k+1}}(\cdot | s)) \leq \delta$$

например, уменьшить  $\alpha_k$  вдвое  
пока не выполняются условия.

# Trust Region Policy Optimization (TRPO)

$$\begin{cases} L_{\pi^{\text{old}}}(\theta) \rightarrow \max_{\theta} \\ \text{KL}(\pi^{\text{old}} \parallel \pi_{\theta}) \leq \delta \end{cases}$$

$$\text{KL}(\pi^{\text{old}} \parallel \pi) \leq \delta$$

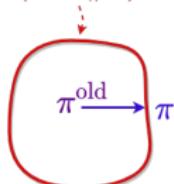


- ✓ **робастно:** предотвращает большие изменения;

# Trust Region Policy Optimization (TRPO)

$$\begin{cases} L_{\pi^{\text{old}}}(\theta) \rightarrow \max_{\theta} \\ \text{KL}(\pi^{\text{old}} \parallel \pi_{\theta}) \leq \delta \end{cases}$$

$$\text{KL}(\pi^{\text{old}} \parallel \pi) \leq \delta$$



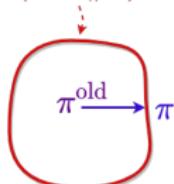
- ✓ **робастно:** предотвращает большие изменения;



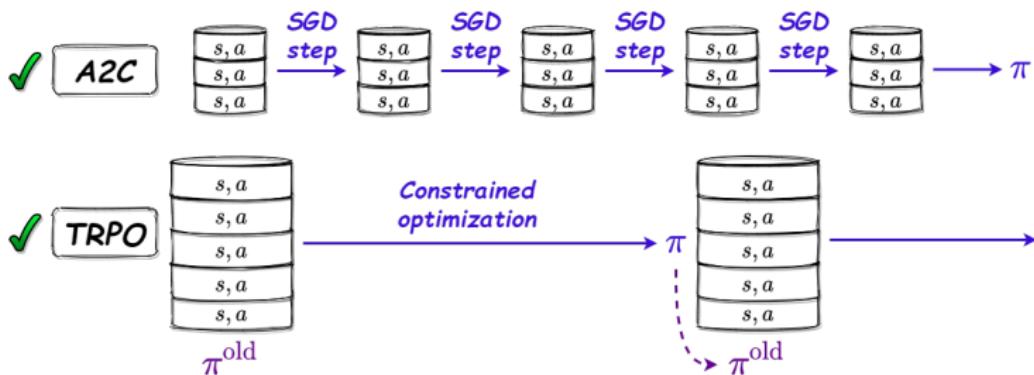
# Trust Region Policy Optimization (TRPO)

$$\begin{cases} L_{\pi^{\text{old}}}(\theta) \rightarrow \max_{\theta} \\ \text{KL}(\pi^{\text{old}} \parallel \pi_{\theta}) \leq \delta \end{cases}$$

$$\text{KL}(\pi^{\text{old}} \parallel \pi) \leq \delta$$



- ✓ **робастно:** предотвращает большие изменения;

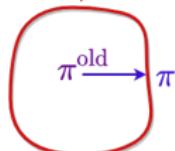


# Trust Region Policy Optimization (TRPO)

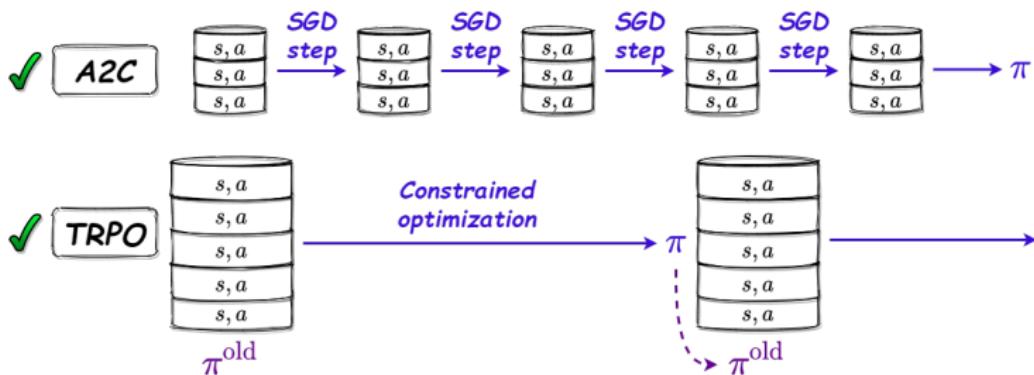
$$\begin{cases} L_{\pi^{\text{old}}}(\theta) \rightarrow \max_{\theta} \\ \text{KL}(\pi^{\text{old}} \| \pi_{\theta}) \leq \delta \end{cases}$$

- ✓ **робастно:** предотвращает большие изменения;

$$\text{KL}(\pi^{\text{old}} \| \pi) \leq \delta$$



- ✗ **актора и критика не получится архитектурно скрестить;**
- ✗ **вычислительно дорого;**
- ✗ **сложно :(|**

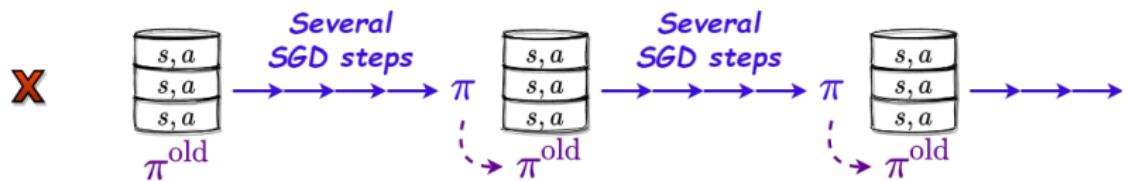


# Proximal Policy Optimization (PPO): пайплайн

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \frac{\pi_\theta(a | s)}{\pi^{\text{old}}(a | s)} A^{\pi^{\text{old}}}(s, a) - C \text{KL}(\pi^{\text{old}} \| \pi_\theta) \rightarrow \max_{\theta}$$

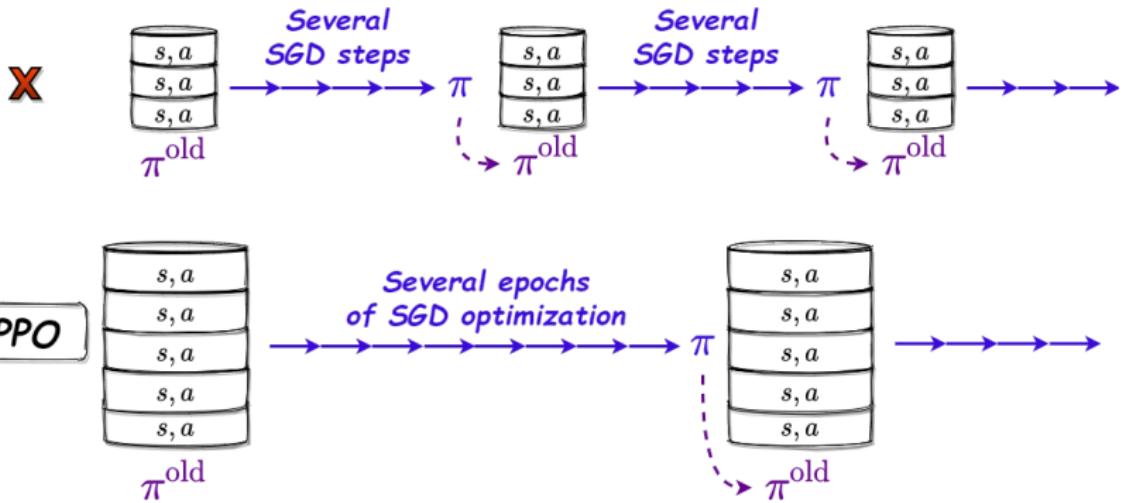
# Proximal Policy Optimization (PPO): пайплайн

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \frac{\pi_{\theta}(a | s)}{\pi^{\text{old}}(a | s)} A^{\pi^{\text{old}}}(s, a) - C \text{KL}(\pi^{\text{old}} \| \pi_{\theta}) \rightarrow \max_{\theta}$$



# Proximal Policy Optimization (PPO): пайплайн

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \frac{\pi_{\theta}(a | s)}{\pi^{\text{old}}(a | s)} A^{\pi^{\text{old}}}(s, a) - C \text{KL}(\pi^{\text{old}} \| \pi_{\theta}) \rightarrow \max_{\theta}$$



# Клиппирование (ограничение) оптимизируемого функционала

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

Суррогатный функционал:

$$\rho(\theta) := \frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}$$

$$L_{\pi^{\text{old}}}(\theta) := \mathbb{E}_{s,a} \rho(\theta) A^{\pi^{\text{old}}}(s, a)$$

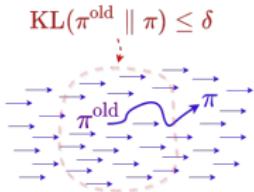
# Клиппирование (ограничение) оптимизируемого функционала

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

Суррогатный функционал:

$$\rho(\theta) := \frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}$$

$$L_{\pi^{\text{old}}}(\theta) := \mathbb{E}_{s,a} \rho(\theta) A^{\pi^{\text{old}}}(s, a)$$



# Клиппирование (ограничение) оптимизируемого функционала

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

Суррогатный функционал:

$$\rho(\theta) := \frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}$$

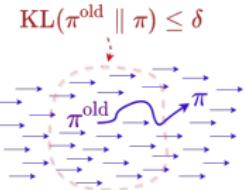
$$L_{\pi^{\text{old}}}(\theta) := \mathbb{E}_{s,a} \rho(\theta) A^{\pi^{\text{old}}}(s,a)$$

Клиппированный суррогатный  
функционал:

$$\rho^{\text{clip}}(\theta) := \text{clip}(\rho(\theta), 1 - \epsilon, 1 + \epsilon)$$



$$L_{\pi^{\text{old}}}^{\text{clip}}(\theta) := \mathbb{E}_{s,a} \rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s,a)$$



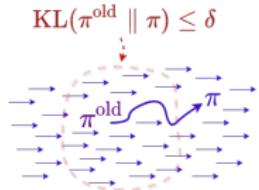
# Клиппирование (ограничение) оптимизируемого функционала

$$L_{\pi^{\text{old}}}(\theta) - C \text{KL}(\pi^{\text{old}} \parallel \pi_\theta) \rightarrow \max_{\theta}$$

Суррогатный функционал:

$$\rho(\theta) := \frac{\pi_\theta(a \mid s)}{\pi^{\text{old}}(a \mid s)}$$

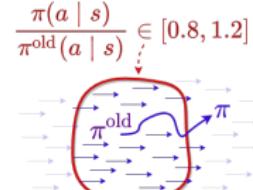
$$L_{\pi^{\text{old}}}(\theta) := \mathbb{E}_{s,a} \rho(\theta) A^{\pi^{\text{old}}}(s, a)$$



Клиппированный суррогатный функционал:

$$\rho^{\text{clip}}(\theta) := \text{clip}(\rho(\theta), 1 - \epsilon, 1 + \epsilon)$$

$$L_{\pi^{\text{old}}}^{\text{clip}}(\theta) := \mathbb{E}_{s,a} \rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s, a)$$



# Вспоминая интуицию с нижней оценкой

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min \left( \underbrace{\rho(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{исходный член}}, \underbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{с клиппированным весом значимости сэмпла}} \right) - \underbrace{C \text{KL}(\pi^{\text{old}} \| \pi_\theta)}_{\text{«регуляризатор»}} \rightarrow \max_{\theta}$$

# Вспоминая интуицию с нижней оценкой

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min \left( \underbrace{\rho(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{исходный член}}, \underbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{с клиппированным весом значимости сэмпла}} \right) - \underbrace{C \text{KL}(\pi^{\text{old}} \| \pi_\theta)}_{\text{«регуляризатор»}} \rightarrow \max_{\theta}$$

Знак Advantage	Направление	Значение веса	Градиент
$A^{\pi^{\text{old}}}(s, a) \geq 0$			

# Вспоминая интуицию с нижней оценкой

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min \left( \underbrace{\rho(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{исходный член}}, \underbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{с клиппированным весом значимости сэмпла}} \right) - \underbrace{C \text{KL}(\pi^{\text{old}} \| \pi_\theta)}_{\text{«регуляризатор»}} \rightarrow \max_{\theta}$$

Знак Advantage	Направление	Значение веса	Градиент
$A^{\pi^{\text{old}}}(s, a) \geq 0$	$\pi_\theta(a   s) \uparrow$		

# Вспоминая интуицию с нижней оценкой

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min \left( \underbrace{\rho(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{исходный член}}, \underbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{с клиппированным весом значимости сэмпла}} \right) - \underbrace{C \text{KL}(\pi^{\text{old}} \| \pi_\theta)}_{\text{«регуляризатор»}} \rightarrow \max_{\theta}$$

Знак Advantage	Направление	Значение веса	Градиент
$A^{\pi^{\text{old}}}(s, a) \geq 0$	$\pi_\theta(a   s) \uparrow$	$\rho(\theta) > 1.2$	

# Вспоминая интуицию с нижней оценкой

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min \left( \underbrace{\rho(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{исходный член}}, \underbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{с клиппированным весом значимости сэмпла}} \right) - \underbrace{C \text{KL}(\pi^{\text{old}} \| \pi_\theta)}_{\text{«регуляризатор»}} \rightarrow \max_{\theta}$$

Знак Advantage	Направление	Значение веса	Градиент
$A^{\pi^{\text{old}}}(s, a) \geq 0$	$\pi_\theta(a   s) \uparrow$	$\rho(\theta) > 1.2$	0

# Вспоминая интуицию с нижней оценкой

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min \left( \underbrace{\rho(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{исходный член}}, \underbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{с клиппированным весом значимости сэмпла}} \right) - \underbrace{C \text{KL}(\pi^{\text{old}} \| \pi_\theta)}_{\text{«регуляризатор»}} \rightarrow \max_{\theta}$$

Знак Advantage	Направление	Значение веса	Градиент
$A^{\pi^{\text{old}}}(s, a) \geq 0$	$\pi_\theta(a   s) \uparrow$	$\rho(\theta) > 1.2$ $\rho(\theta) < 0.8$	0

# Вспоминая интуицию с нижней оценкой

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min \left( \underbrace{\rho(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{исходный член}}, \underbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{с клиппированным весом значимости сэмпла}} \right) - \underbrace{C \text{KL}(\pi^{\text{old}} \| \pi_\theta)}_{\text{«регуляризатор»}} \rightarrow \max_{\theta}$$

Знак Advantage	Направление	Значение веса	Градиент
$A^{\pi^{\text{old}}}(s, a) \geq 0$	$\pi_\theta(a   s) \uparrow$	$\rho(\theta) > 1.2$ $\rho(\theta) < 0.8$	0 тот же

# Вспоминая интуицию с нижней оценкой

$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min \left( \underbrace{\rho(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{исходный член}}, \underbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{с клиппированным весом значимости сэмпла}} \right) - \underbrace{C \text{KL}(\pi^{\text{old}} \| \pi_\theta)}_{\text{«регуляризатор»}} \rightarrow \max_{\theta}$$

Знак Advantage	Направление	Значение веса	Градиент
$A^{\pi^{\text{old}}}(s, a) \geq 0$	$\pi_\theta(a   s) \uparrow$	$\rho(\theta) > 1.2$ $\rho(\theta) < 0.8$	0 тот же
$A^{\pi^{\text{old}}}(s, a) < 0$	$\pi_\theta(a   s) \downarrow$	$\rho(\theta) > 1.2$ $\rho(\theta) < 0.8$	

# Вспоминая интуицию с нижней оценкой

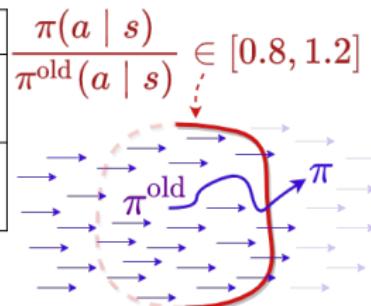
$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min \left( \underbrace{\rho(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{исходный член}}, \underbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{с клиппированным весом значимости сэмпла}} \right) - \underbrace{C \text{KL}(\pi^{\text{old}} \| \pi_\theta)}_{\text{«регуляризатор»}} \rightarrow \max_{\theta}$$

Знак Advantage	Направление	Значение веса	Градиент
$A^{\pi^{\text{old}}}(s, a) \geq 0$	$\pi_\theta(a   s) \uparrow$	$\rho(\theta) > 1.2$ $\rho(\theta) < 0.8$	0 тот же
$A^{\pi^{\text{old}}}(s, a) < 0$	$\pi_\theta(a   s) \downarrow$	$\rho(\theta) > 1.2$ $\rho(\theta) < 0.8$	тот же 0

# Вспоминая интуицию с нижней оценкой

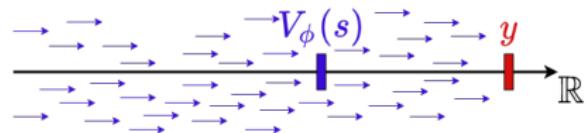
$$\mathbb{E}_{s \sim d_{\pi^{\text{old}}}(s)} \mathbb{E}_{a \sim \pi^{\text{old}}(a|s)} \min \left( \underbrace{\rho(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{исходный член}}, \underbrace{\rho^{\text{clip}}(\theta) A^{\pi^{\text{old}}}(s, a)}_{\text{с клиппированным весом значимости сэмпла}} \right) - \underbrace{C \text{KL}(\pi^{\text{old}} \parallel \pi_\theta)}_{\text{«регуляризатор»}} \rightarrow \max_{\theta}$$

Знак Advantage	Направление	Значение веса	Градиент
$A^{\pi^{\text{old}}}(s, a) \geq 0$	$\pi_\theta(a   s) \uparrow$	$\rho(\theta) > 1.2$ $\rho(\theta) < 0.8$	0 тот же
$A^{\pi^{\text{old}}}(s, a) < 0$	$\pi_\theta(a   s) \downarrow$	$\rho(\theta) > 1.2$ $\rho(\theta) < 0.8$	тот же 0



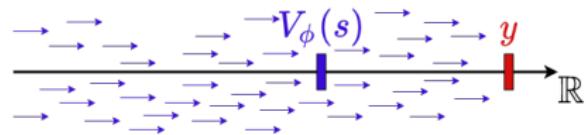
# Клиппированная функция потерь критика

$$\text{Loss}(\phi) := (y - V(\phi))^2 =$$



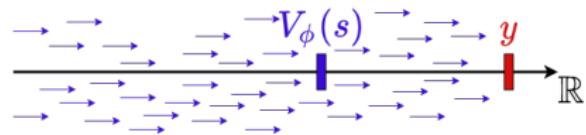
# Клиппированная функция потерь критика

$$\begin{aligned}\text{Loss}(\phi) &:= (y - V(\phi))^2 = \\ &= (y - V^{\text{old}} + V^{\text{old}} - V(\phi))^2\end{aligned}$$

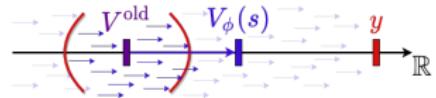


# Клиппированная функция потерь критика

$$\begin{aligned}\text{Loss}(\phi) &:= (y - V(\phi))^2 = \\ &= (y - V^{\text{old}} + V^{\text{old}} - V(\phi))^2\end{aligned}$$

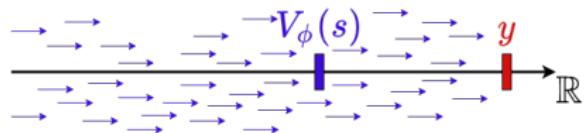


$$\text{Loss}^{\text{clip}}(\phi) := (y - V^{\text{old}} + \text{clip}(V^{\text{old}} - V(\phi), -\hat{\epsilon}, \hat{\epsilon}))^2$$

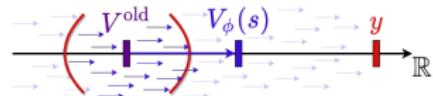


# Клиппированная функция потерь критика

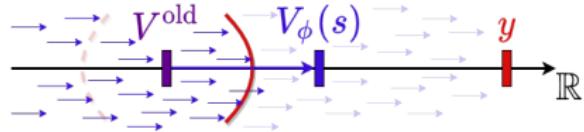
$$\begin{aligned}\text{Loss}(\phi) &:= (y - V(\phi))^2 = \\ &= (y - V^{\text{old}} + V^{\text{old}} - V(\phi))^2\end{aligned}$$



$$\text{Loss}^{\text{clip}}(\phi) := (y - V^{\text{old}} + \text{clip}(V^{\text{old}} - V(\phi), -\hat{\epsilon}, \hat{\epsilon}))^2$$



$$\max(\text{Loss}(\phi), \text{Loss}^{\text{clip}}(\phi))$$



# Компромисс между смещением и разбросом

Дана траектория  $s, r, s', r', s'', r'' \dots s^{(M)}$  по политике  $\pi$  и  
приближение  $V^\pi(s)$

# Компромисс между смещением и разбросом

Дана траектория  $s, r, s', r', s'', r'' \dots s^{(M)}$  по политике  $\pi$  и  
приближение  $V^\pi(s)$

выполним **оценку advantage (credit assignment)** для пары  $s, a$   
(хорошее ли решение принято было?)

# Компромисс между смещением и разбросом

Дана траектория  $s, r, s', r', s'', r'' \dots s^{(M)}$  по политике  $\pi$  и  
приближение  $V^\pi(s)$

выполним **оценку advantage (credit assignment)** для пары  $s, a$   
(хорошее ли решение принято было?)

Для актора:

$$\nabla := \rho(\theta) \nabla_\theta \log \pi_\theta(a | s) \underbrace{\Psi(s, a)}_{\substack{\text{оценка} \\ \text{advantage}}}$$

Для критика:

$$\underbrace{y_Q}_{\substack{\text{целевое значение} \\ \text{для регрессии}}} := \Psi(s, a) + V(s)$$

# Компромисс между смещением и разбросом

Дана траектория  $s, r, s', r', s'', r'' \dots s^{(M)}$  по политике  $\pi$  и приближение  $V^\pi(s)$

выполним **оценку advantage (credit assignment)** для пары  $s, a$   
(хорошее ли решение принято было?)

Для актора:

$$\nabla := \rho(\theta) \nabla_\theta \log \pi_\theta(a | s) \underbrace{\Psi(s, a)}_{\substack{\text{оценка} \\ \text{advantage}}}$$

Для критика:

$$y_Q := \underbrace{\Psi(s, a) + V(s)}_{\substack{\text{целевое значение} \\ \text{для регрессии}}}$$

	$\Psi(s, a)$	Смещение	Разброс
Монте-Карло	$\Psi_{(\infty)}(s, a) := r + \gamma r' + \gamma^2 r'' + \dots - V(s)$	0	высокий
1-шаг	$\Psi_{(1)}(s, a) := r + \gamma V(s') - V(s)$	высокое	низкий

# Компромисс между смещением и разбросом

Дана траектория  $s, r, s', r', s'', r'' \dots s^{(M)}$  по политике  $\pi$  и приближение  $V^\pi(s)$

выполним **оценку advantage (credit assignment)** для пары  $s, a$   
(хорошее ли решение принято было?)

Для актора:

$$\nabla := \rho(\theta) \nabla_\theta \log \pi_\theta(a | s) \underbrace{\Psi(s, a)}_{\substack{\text{оценка} \\ \text{advantage}}}$$

Для критика:

$$y_Q := \underbrace{\Psi(s, a) + V(s)}_{\substack{\text{целевое значение} \\ \text{для регрессии}}}$$

	$\Psi(s, a)$	Смещение	Разброс
Монте-Карло $N$ -шагов 1-шаг	$\Psi_{(\infty)}(s, a) := r + \gamma r' + \gamma^2 r'' + \dots - V(s)$ $\Psi_{(N)}(s, a) := r + \gamma r' + \dots + \gamma^N V(s^{(N)}) - V(s)$ $\Psi_{(1)}(s, a) := r + \gamma V(s') - V(s)$	0 промежуточное высокое	высокий промежуточный низкий

# Компромисс между смещением и разбросом

Дана траектория  $s, r, s', r', s'', r'' \dots s^{(M)}$  по политике  $\pi$  и приближение  $V^\pi(s)$

выполним **оценку advantage (credit assignment)** для пары  $s, a$   
(хорошее ли решение принято было?)

Для актора:

$$\nabla := \rho(\theta) \nabla_\theta \log \pi_\theta(a | s) \underbrace{\Psi(s, a)}_{\substack{\text{оценка} \\ \text{advantage}}}$$

Для критика:

$$y_Q := \underbrace{\Psi(s, a) + V(s)}_{\substack{\text{целевое значение} \\ \text{для регрессии}}}$$

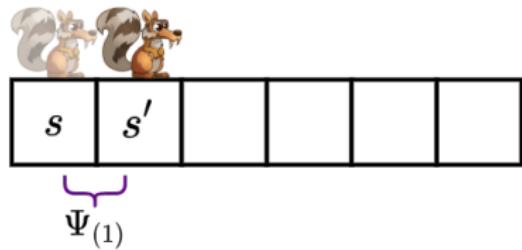
	$\Psi(s, a)$	Смещение	Разброс
Монте-Карло $N$ -шагов 1-шаг	$\Psi_{(\infty)}(s, a) := r + \gamma r' + \gamma^2 r'' + \dots - V(s)$ $\Psi_{(N)}(s, a) := r + \gamma r' + \dots + \gamma^N V(s^{(N)}) - V(s)$ $\Psi_{(1)}(s, a) := r + \gamma V(s') - V(s)$	0 промежуточное высокое	высокий промежуточный низкий

Проблема: выбор  $N$ .

# Взгляд назад: идея

$N$ -шаговое обновление:

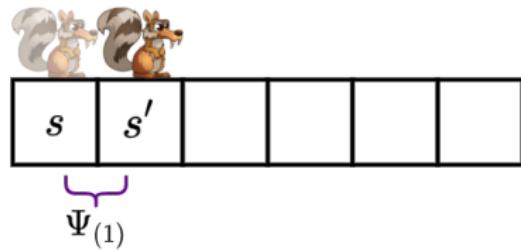
$$V(s) \leftarrow V(s) + \alpha \Psi_{(N)}(s, a)$$



# Взгляд назад: идея

$N$ -шаговое обновление:

$$V(s) \leftarrow V(s) + \alpha \Psi_{(N)}(s, a)$$



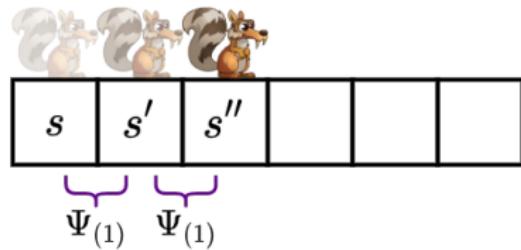
$$V(s) \leftarrow V(s) + \alpha \overbrace{(r + \gamma V(s') - V(s))}^{\Psi_{(1)}(s, a)}$$

## Взгляд назад: идея

$N$ -шаговое обновление:

$$V(s) \leftarrow V(s) + \alpha \Psi_{(N)}(s, a)$$

Как сделать из 1-шагового обновления  
2-шаговое?



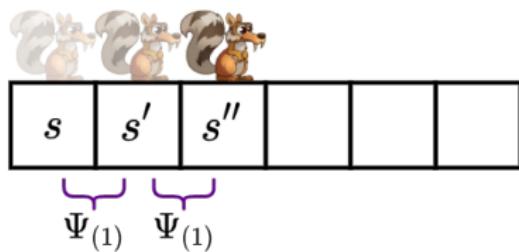
$$V(s) \leftarrow V(s) + \alpha \overbrace{(r + \gamma V(s') - V(s))}^{\Psi_{(1)}(s, a)}$$

# Взгляд назад: идея

$N$ -шаговое обновление:

$$V(s) \leftarrow V(s) + \alpha \Psi_{(N)}(s, a)$$

Как сделать из 1-шагового обновления  
2-шаговое?



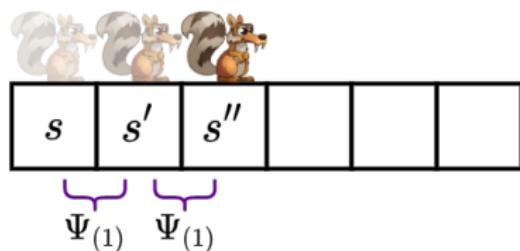
$$V(s) \leftarrow V(s) + \alpha \overbrace{(r + \gamma V(s') - V(s))}^{\Psi_{(1)}(s, a)} + \alpha \overbrace{(\gamma r' + \gamma^2 V(s'') - \gamma V(s'))}^{\gamma \Psi_{(1)}(s', a')}$$

## Взгляд назад: идея

$N$ -шаговое обновление:

$$V(s) \leftarrow V(s) + \alpha \Psi_{(N)}(s, a)$$

Как сделать из 1-шагового обновления  
2-шаговое?



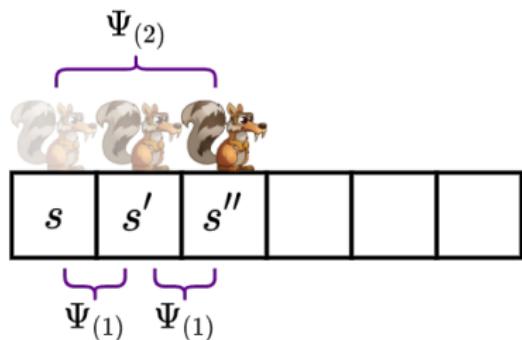
$$V(s) \leftarrow V(s) + \alpha \overbrace{\left( r + \underbrace{\gamma V(s')} - V(s) \right)}^{\Psi_{(1)}(s, a)} + \alpha \overbrace{\left( \gamma r' + \underbrace{\gamma^2 V(s'')} - \gamma V(s') \right)}^{\gamma \Psi_{(1)}(s', a')} =$$

# Взгляд назад: идея

$N$ -шаговое обновление:

$$V(s) \leftarrow V(s) + \alpha \Psi_{(N)}(s, a)$$

Как сделать из 1-шагового обновления  
2-шаговое?



$$\begin{aligned} V(s) &\leftarrow V(s) + \alpha \overbrace{\left( r + \underbrace{\gamma V(s')}_{\Psi_{(1)}(s, a)} - V(s) \right)}^{\Psi_{(1)}(s, a)} + \alpha \overbrace{\left( \gamma r' + \underbrace{\gamma^2 V(s'')}_{\gamma \Psi_{(1)}(s', a')} - \gamma V(s') \right)}^{\gamma \Psi_{(1)}(s', a')} = \\ &= V(s) + \alpha \Psi_{(2)}(s, a) \end{aligned}$$

# Взгляд назад: идея

$N$ -шаговое обновление:

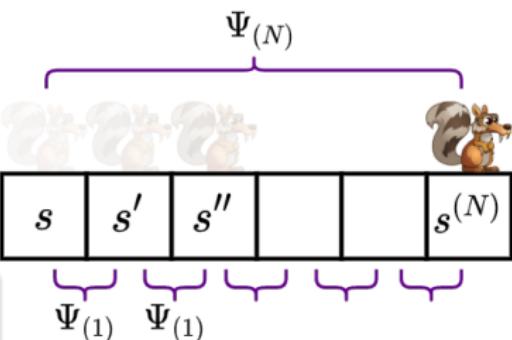
$$V(s) \leftarrow V(s) + \alpha \Psi_{(N)}(s, a)$$

Как сделать из 1-шагового обновления  
2-шаговое?

$N$ -шаговая невязка – сумма 1-шаговых

$$\Psi_{(N)}(s, a) = \sum_{t=0}^N \gamma^t \Psi_{(1)}(s^{(t)}, a^{(t)})$$

$$V(s) \leftarrow V(s) + \alpha \overbrace{(r + \gamma V(s') - V(s))}^{\Psi_{(1)}(s, a)} + \alpha \overbrace{(\gamma r' + \gamma^2 V(s'') - \gamma V(s'))}^{\gamma \Psi_{(1)}(s', a')} = V(s) + \alpha \Psi_{(2)}(s, a)$$



# Eligibility Traces (посещаемость)

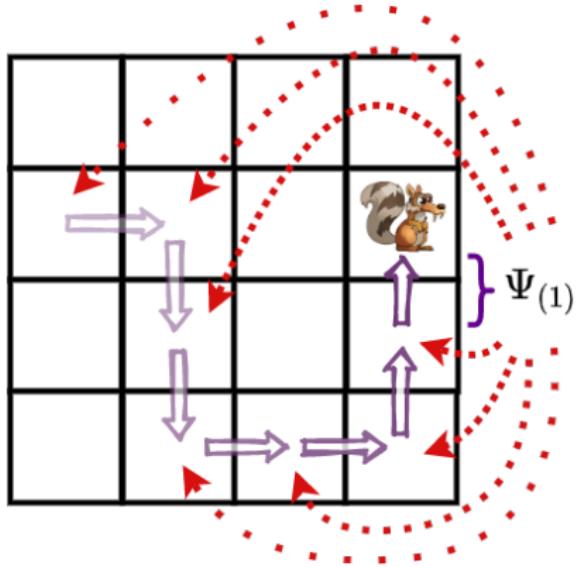


Использовать 1-шаговую TD-ошибку при обновлении  $V(s)$  для **всех** состояний

# Eligibility Traces (посещаемость)



Использовать 1-шаговую TD-ошибку при обновлении  $V(s)$  для всех состояний



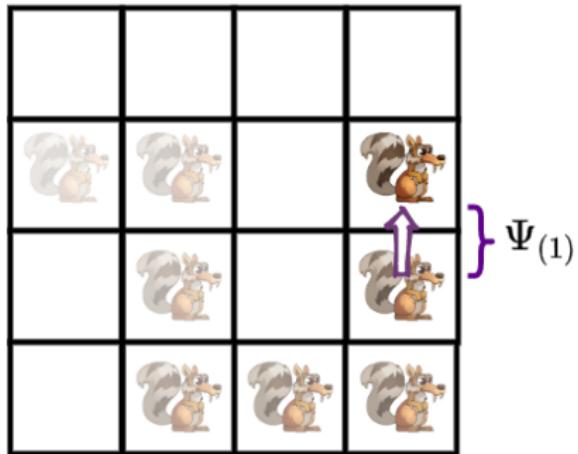
# Eligibility Traces (посещаемость)



Использовать 1-шаговую TD-ошибку при обновлении  $V(s)$  для **всех** состояний

Определить **eligibility trace** (посещаемость)  $e(s)$  как скаляр в обновлении:

$$\forall s: V(s) \leftarrow V(s) + \alpha e(s) \Psi_{(1)}$$



# Eligibility Traces (посещаемость)



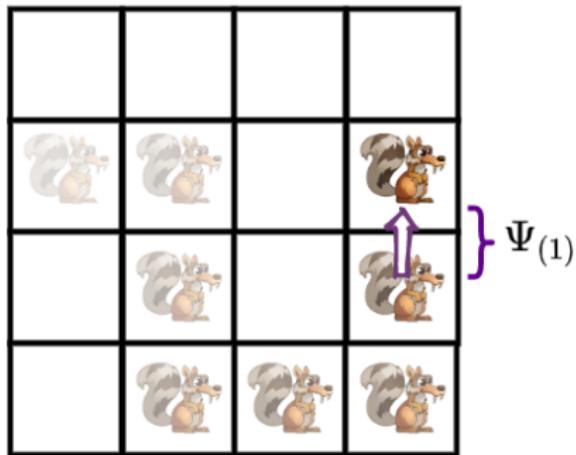
Использовать 1-шаговую TD-ошибку при обновлении  $V(s)$  для **всех** состояний

Определить **eligibility trace** (посещаемость)  $e(s)$  как скаляр в обновлении:

$$\forall s: V(s) \leftarrow V(s) + \alpha e(s) \Psi_{(1)}$$

Онлайн «Монте-Карло» обновления:

- $\forall s: e(s) := 0$  в начале каждого эпизода



# Eligibility Traces (посещаемость)



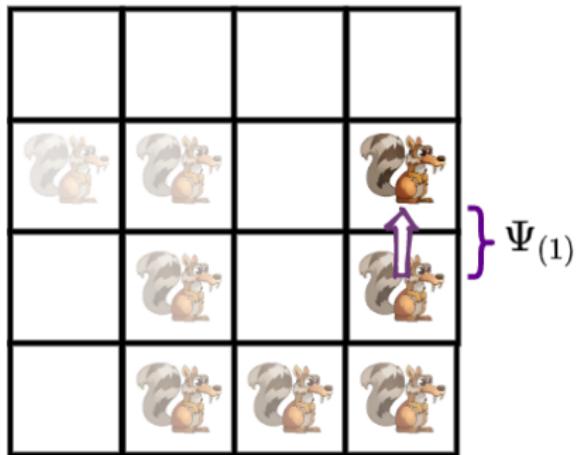
Использовать 1-шаговую TD-ошибку при обновлении  $V(s)$  для **всех** состояний

Определить **eligibility trace** (посещаемость)  $e(s)$  как скаляр в обновлении:

$$\forall s: V(s) \leftarrow V(s) + \alpha e(s) \Psi_{(1)}$$

Онлайн «Монте-Карло» обновления:

- $\forall s: e(s) := 0$  в начале каждого эпизода
- $e(s) \leftarrow e(s) + 1$  после посещения  $s$



# Eligibility Traces (посещаемость)



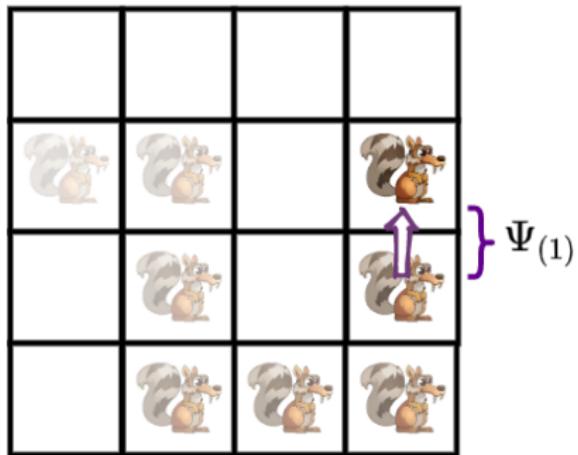
Использовать 1-шаговую TD-ошибку при обновлении  $V(s)$  для **всех** состояний

Определить **eligibility trace** (посещаемость)  $e(s)$  как скаляр в обновлении:

$$\forall s: V(s) \leftarrow V(s) + \alpha e(s) \Psi_{(1)}$$

Онлайн «Монте-Карло» обновления:

- $\forall s: e(s) := 0$  в начале каждого эпизода
- $e(s) \leftarrow e(s) + 1$  после посещения  $s$
- $\forall s: e(s) \leftarrow \gamma e(s)$  после каждого шага



# TD(1) и TD(0)

## TD(1)

Ввод: политика  $\pi$

Инициализировать  $V(s)$

произвольно

Инициализировать  $e(s) = 0$

наблюдаем  $s_0$

для  $k = 0, 1, 2 \dots$

- выбираем действие  $a_k \sim \pi$ ,  
наблюдаем  $r_k, s_{k+1}$

# TD(1) и TD(0)

## TD(1)

Ввод: политика  $\pi$

Инициализировать  $V(s)$

произвольно

Инициализировать  $e(s) = 0$

наблюдаем  $s_0$

для  $k = 0, 1, 2 \dots$

- выбираем действие  $a_k \sim \pi$ ,  
наблюдаем  $r_k, s_{k+1}$
- $\Psi_{(1)} := r_k + \gamma V(s_{k+1}) - V(s_k)$

# TD(1) и TD(0)

## TD(1)

Ввод: политика  $\pi$

Инициализировать  $V(s)$

произвольно

Инициализировать  $e(s) = 0$

наблюдаем  $s_0$

для  $k = 0, 1, 2 \dots$

- выбираем действие  $a_k \sim \pi$ ,  
наблюдаем  $r_k, s_{k+1}$
- $\Psi_{(1)} := r_k + \gamma V(s_{k+1}) - V(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$

# TD(1) и TD(0)

## TD(1)

Ввод: политика  $\pi$

Инициализировать  $V(s)$

произвольно

Инициализировать  $e(s) = 0$

наблюдаем  $s_0$

для  $k = 0, 1, 2 \dots$

- выбираем действие  $a_k \sim \pi$ ,  
наблюдаем  $r_k, s_{k+1}$
- $\Psi_{(1)} := r_k + \gamma V(s_{k+1}) - V(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s: V(s) \leftarrow V(s) + \alpha e(s) \Psi_{(1)}$

# TD(1) и TD(0)

## TD(1)

Ввод: политика  $\pi$

Инициализировать  $V(s)$

произвольно

Инициализировать  $e(s) = 0$

наблюдаем  $s_0$

для  $k = 0, 1, 2 \dots$

- выбираем действие  $a_k \sim \pi$ ,  
наблюдаем  $r_k, s_{k+1}$
- $\Psi_{(1)} := r_k + \gamma V(s_{k+1}) - V(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s: V(s) \leftarrow V(s) + \alpha e(s) \Psi_{(1)}$
- $\forall s: e(s) \leftarrow \gamma e(s)$

# TD(1) и TD(0)

TD(1)

**Ввод:** политика  $\pi$

**Инициализировать**  $V(s)$   
произвольно

**Инициализировать**  $e(s) = 0$   
наблюдаем  $s_0$   
**для**  $k = 0, 1, 2 \dots$

- выбираем действие  $a_k \sim \pi$ ,  
наблюдаем  $r_k, s_{k+1}$
- $\Psi_{(1)} := r_k + \gamma V(s_{k+1}) - V(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s: V(s) \leftarrow V(s) + \alpha e(s) \Psi_{(1)}$
- $\forall s: e(s) \leftarrow \gamma e(s)$

TD(0)

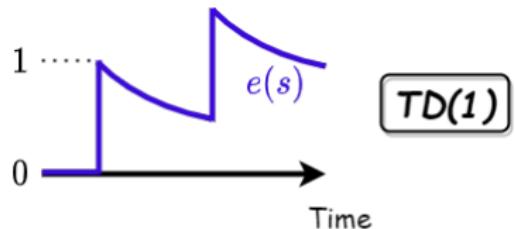
**Ввод:** политика  $\pi$

**Инициализировать**  $V(s)$   
произвольно

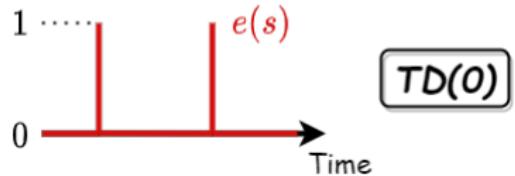
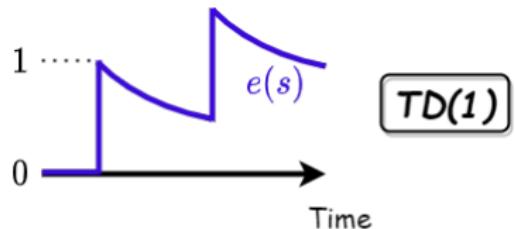
**Инициализировать**  $e(s) = 0$   
наблюдаем  $s_0$   
**для**  $k = 0, 1, 2 \dots$

- выбираем действие  $a_k \sim \pi$ ,  
наблюдаем  $r_k, s_{k+1}$
- $\Psi_{(1)} := r_k + \gamma V(s_{k+1}) - V(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s: V(s) \leftarrow V(s) + \alpha e(s) \Psi_{(1)}$
- $\forall s: e(s) \leftarrow 0 \cdot \gamma e(s)$

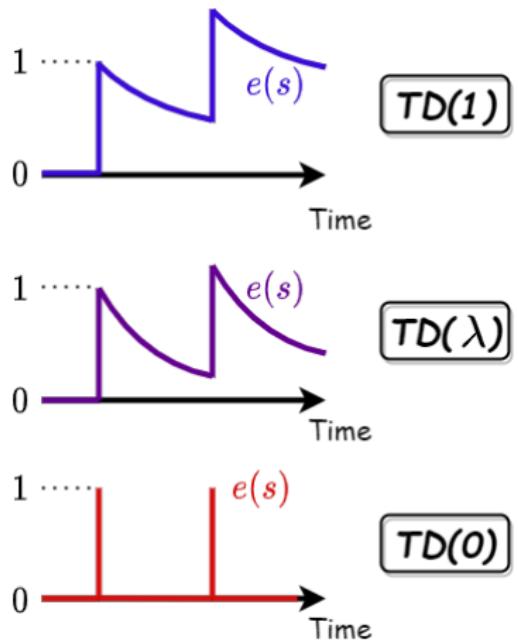
# $\text{TD}(\lambda)$



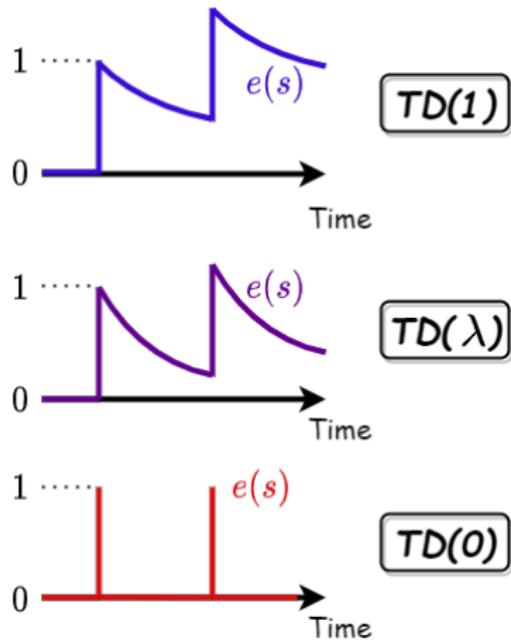
# $\text{TD}(\lambda)$



# $TD(\lambda)$



# $TD(\lambda)$



$TD(\lambda)$

Ввод: политика  $\pi$

Инициализировать  $V(s)$  произвольно

Инициализировать  $e(s) = 0$

наблюдаем  $s_0$

для  $k = 0, 1, 2 \dots$

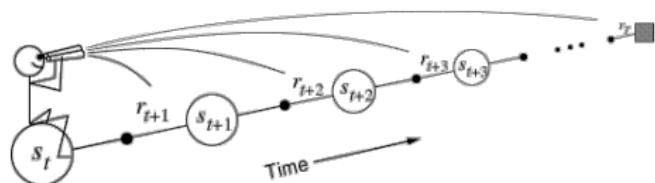
- выбираем действие  $a_k \sim \pi$ ,  
наблюдаем  $r_k, s_{k+1}$
- $\Psi_{(1)} := r_k + \gamma V(s_{k+1}) - V(s_k)$
- $e(s_k) \leftarrow e(s_k) + 1$
- $\forall s: V(s) \leftarrow V(s) + \alpha e(s) \Psi_{(1)}$
- $\forall s: e(s) \leftarrow \lambda \gamma e(s)$

# Взгляд вперёд или взгляд назад

## Взгляд вперёд

Дать оценку настоящему по известному будущему

«Хорошо ли решение или плохо на основе результата?»

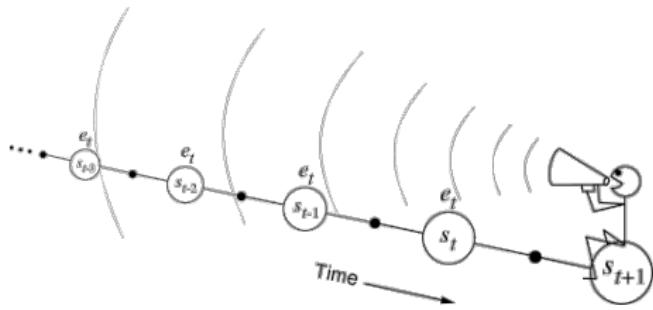
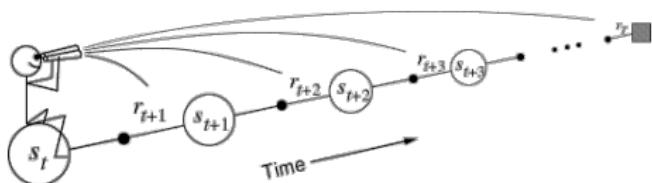


# Взгляд вперёд или взгляд назад

## Взгляд вперёд

Дать оценку настоящему по известному будущему

«Хорошо ли решение или плохо на основе результата?»



## Взгляд назад

Обновить прошлые оценки с помощью настоящей информации

«Какие решения в прошлом оказали значительное влияние на настоящие?»

# Взгляд вперёд для $\text{TD}(\lambda)$

Шаг	Обновление	$\Psi_{(1)}(s, a)$	$\Psi_{(2)}(s, a)$	$\Psi_{(3)}(s, a)$	$\dots$	$\Psi_{(N)}(s, a)$
0	$\Psi_{(1)}(s, a)$	1	0	0	$\dots$	0

# Взгляд вперёд для $\text{TD}(\lambda)$

Шаг	Обновление	$\Psi_{(1)}(s, a)$	$\Psi_{(2)}(s, a)$	$\Psi_{(3)}(s, a)$	$\dots$	$\Psi_{(N)}(s, a)$
0	$\Psi_{(1)}(s, a)$	1	0	0		0
1	$\Psi_{(1)}(s, a) + \gamma \lambda \Psi_{(1)}(s', a')$	$1 - \lambda$	$\lambda$	0		0

# Взгляд вперёд для $\text{TD}(\lambda)$

Шаг	Обновление	$\Psi_{(1)}(s, a)$	$\Psi_{(2)}(s, a)$	$\Psi_{(3)}(s, a)$	$\dots$	$\Psi_{(N)}(s, a)$
0	$\Psi_{(1)}(s, a)$	1	0	0		0
1	$\Psi_{(1)}(s, a) + \gamma\lambda\Psi_{(1)}(s', a')$	$1 - \lambda$	$\lambda$	0		0
2	$\Psi_{(1)}(s, a) + \gamma\lambda\Psi_{(1)}(s', a') + (\gamma\lambda)^2\Psi_{(1)}(s'', a'')$	$1 - \lambda$	$(1 - \lambda)\lambda$	$\lambda^2$		0

# Взгляд вперёд для $\text{TD}(\lambda)$

Шаг	Обновление	$\Psi_{(1)}(s, a)$	$\Psi_{(2)}(s, a)$	$\Psi_{(3)}(s, a)$	$\dots$	$\Psi_{(N)}(s, a)$
0	$\Psi_{(1)}(s, a)$	1	0	0		0
1	$\Psi_{(1)}(s, a) + \gamma\lambda\Psi_{(1)}(s', a')$	$1 - \lambda$	$\lambda$	0		0
2	$\Psi_{(1)}(s, a) + \gamma\lambda\Psi_{(1)}(s', a') + (\gamma\lambda)^2\Psi_{(1)}(s'', a'')$	$1 - \lambda$	$(1 - \lambda)\lambda$	$\lambda^2$		0
$\vdots$						
$N$	$\sum_{t=0}^N (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)})$	$1 - \lambda$	$(1 - \lambda)\lambda$	$(1 - \lambda)\lambda^2$		$\lambda^N$

# Взгляд вперёд для TD( $\lambda$ )

Шаг	Обновление	$\Psi_{(1)}(s, a)$	$\Psi_{(2)}(s, a)$	$\Psi_{(3)}(s, a)$	$\dots$	$\Psi_{(N)}(s, a)$
0	$\Psi_{(1)}(s, a)$	1	0	0		0
1	$\Psi_{(1)}(s, a) + \gamma\lambda\Psi_{(1)}(s', a')$	$1 - \lambda$	$\lambda$	0		0
2	$\Psi_{(1)}(s, a) + \gamma\lambda\Psi_{(1)}(s', a') + (\gamma\lambda)^2\Psi_{(1)}(s'', a'')$	$1 - \lambda$	$(1 - \lambda)\lambda$	$\lambda^2$		0
$\vdots$						
$N$	$\sum_{t=0}^N (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)})$	$1 - \lambda$	$(1 - \lambda)\lambda$	$(1 - \lambda)\lambda^2$		$\lambda^N$

Эквивалентные формы обновлений TD( $\lambda$ )

$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) =$$

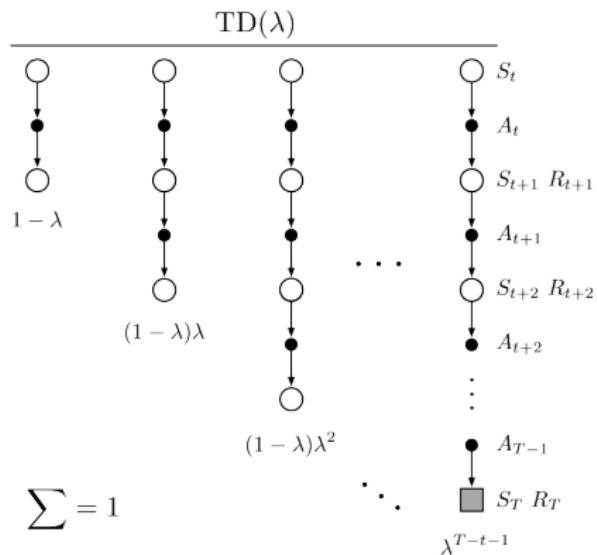
# Взгляд вперёд для $\text{TD}(\lambda)$

Шаг	Обновление	$\Psi_{(1)}(s, a)$	$\Psi_{(2)}(s, a)$	$\Psi_{(3)}(s, a)$	$\dots$	$\Psi_{(N)}(s, a)$
0	$\Psi_{(1)}(s, a)$	1	0	0		0
1	$\Psi_{(1)}(s, a) + \gamma\lambda\Psi_{(1)}(s', a')$	$1 - \lambda$	$\lambda$	0		0
2	$\Psi_{(1)}(s, a) + \gamma\lambda\Psi_{(1)}(s', a') + (\gamma\lambda)^2\Psi_{(1)}(s'', a'')$	$1 - \lambda$	$(1 - \lambda)\lambda$	$\lambda^2$		0
$\vdots$						
$N$	$\sum_{t=0}^N (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)})$	$1 - \lambda$	$(1 - \lambda)\lambda$	$(1 - \lambda)\lambda^2$		$\lambda^N$

Эквивалентные формы обновлений  $\text{TD}(\lambda)$

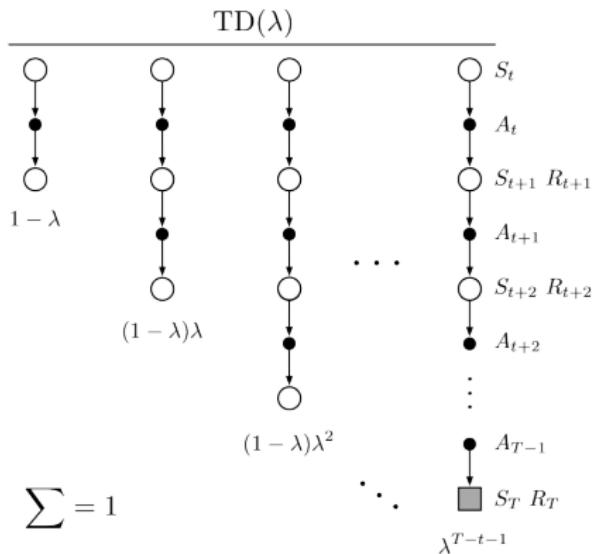
$$\sum_{t=0}^{\infty} (\gamma\lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)}) = (1 - \lambda) \sum_{N=1}^{\infty} \lambda^{N-1} \Psi_{(N)}(s, a)$$

# Generalized Advantage Estimation (GAE)



Что если для некоторой пары  $s, a$   
нам известно будущее только на  $T$   
шагов вперёд?

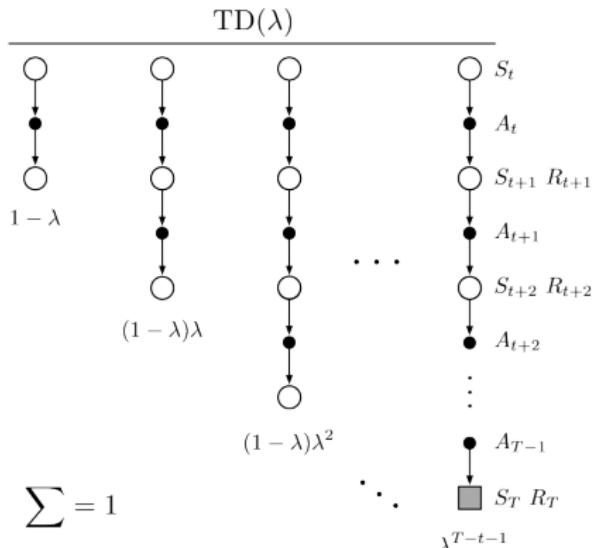
# Generalized Advantage Estimation (GAE)



Что если для некоторой пары  $s, a$  нам известно будущее только на  $T$  шагов вперёд?

$$\Psi^{\text{GAE}}(s, a) := \sum_{t=0}^T (\gamma \lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)})$$

# Generalized Advantage Estimation (GAE)



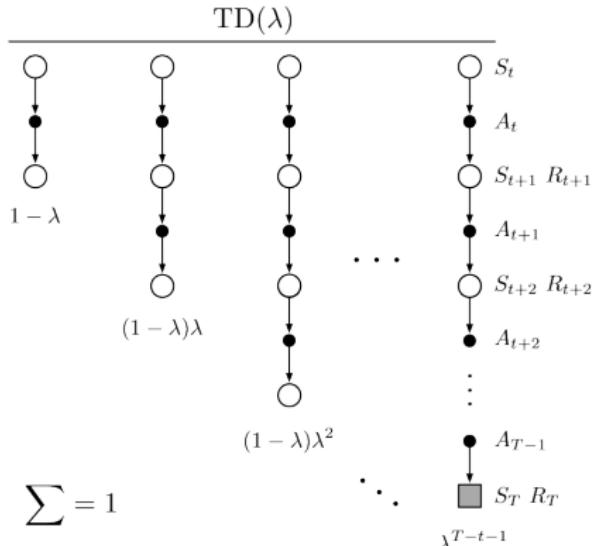
Что если для некоторой пары  $s, a$  нам известно будущее только на  $T$  шагов вперёд?

$$\Psi^{\text{GAE}}(s, a) := \sum_{t=0}^T (\gamma \lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)})$$

Используемое на практике уравнение:

$$\begin{aligned} \Psi^{\text{GAE}}(s_t, a_t) &= \Psi_{(1)}(s_t, a_t) + \\ &+ \lambda \gamma (1 - \text{done}_{t+1}) \Psi^{\text{GAE}}(s_{t+1}, a_{t+1}) \end{aligned}$$

# Generalized Advantage Estimation (GAE)



Что если для некоторой пары  $s, a$  нам известно будущее только на  $T$  шагов вперёд?

$$\Psi^{\text{GAE}}(s, a) := \sum_{t=0}^T (\gamma \lambda)^t \Psi_{(1)}(s^{(t)}, a^{(t)})$$

Используемое на практике уравнение:

$$\begin{aligned} \Psi^{\text{GAE}}(s_t, a_t) &= \Psi_{(1)}(s_t, a_t) + \\ &+ \lambda \gamma (1 - \text{done}_{t+1}) \Psi^{\text{GAE}}(s_{t+1}, a_{t+1}) \end{aligned}$$

# GAE в Advantage Actor-Critic



Длинные развертки порождают богатые GAE ансамбли.

# GAE в Advantage Actor-Critic



Длинные развертки порождают богатые GAE ансамбли.



# GAE в Advantage Actor-Critic



Длинные развёртки порождают богатые GAE ансамбли.



В A2C развёртки обычно короткие, поэтому часто выбирают  $\lambda = 1$ .  
(иногда это называют **max-trace** оценкой)

# РРО: особенности реализации

## Ключевые элементы:

- ✓ Клиппинг функции потерь политики
- ✓ Клиппинг функции потерь критика
- ✓ GAE

## Другие приёмы:

- ! Нормализация и клиппинг награды<sup>1</sup>
- Нормализация и клиппинг наблюдений (состояний)<sup>2</sup>
- Ортогональная инициализация слоёв
- Отжиг по  $\epsilon$  (параметр клиппинга)

---

<sup>1</sup> делением на скользящее стандартное отклонение собранных кумулятивных наград

<sup>2</sup> может быть критично в непрерывном управлении

# РРО: особенности реализации

## Детали пайплайна:

- ! Нормализация Advantage в мини-батчах
- Нет KL-регуляризации
- Энтропийная функция потерь

## Стандартные приёмы:

- Оптимизатор Adam, отжиг шага
- Функция активации Tanh
- ! Клиппинг градиента

## Proximal Policy Optimization (PPO)

Инициализировать  $\pi(a | s, \theta), V_\phi(s)$ ;

## Proximal Policy Optimization (PPO)

Инициализировать  $\pi(a | s, \theta)$ ,  $V_\phi(s)$ ;

for  $k = 0, 1, 2 \dots$

- собрать  $s_0, a_0, r_0, s_1, \text{done}_1, a_1 \dots s_N, \text{done}_N$  по  $\pi(a | s, \theta)$ ;  
запомнить вероятности действий как  $\pi^{\text{old}}(a_t | s_t) := \pi(a_t | s_t, \theta)$   
запомнить значения критика как  $V^{\text{old}}(s_t) := V_\phi(s_t)$

## Proximal Policy Optimization (PPO)

Инициализировать  $\pi(a | s, \theta)$ ,  $V_\phi(s)$ ;

for  $k = 0, 1, 2 \dots$

- собрать  $s_0, a_0, r_0, s_1, \text{done}_1, a_1 \dots s_N, \text{done}_N$  по  $\pi(a | s, \theta)$ ;  
запомнить вероятности действий как  $\pi^{\text{old}}(a_t | s_t) := \pi(a_t | s_t, \theta)$   
запомнить значения критика как  $V^{\text{old}}(s_t) := V_\phi(s_t)$
- вычислить 1-шаговые невязки:  
 $\Psi_{(1)}(s_t, a_t) := r_t + \gamma(1 - \text{done}_{t+1})V_\phi(s_{t+1}) - V_\phi(s_t)$

## Proximal Policy Optimization (PPO)

Инициализировать  $\pi(a | s, \theta)$ ,  $V_\phi(s)$ ;

for  $k = 0, 1, 2 \dots$

- собрать  $s_0, a_0, r_0, s_1, \text{done}_1, a_1 \dots s_N, \text{done}_N$  по  $\pi(a | s, \theta)$ ;  
запомнить вероятности действий как  $\pi^{\text{old}}(a_t | s_t) := \pi(a_t | s_t, \theta)$   
запомнить значения критика как  $V^{\text{old}}(s_t) := V_\phi(s_t)$
- вычислить 1-шаговые невязки:  
 $\Psi_{(1)}(s_t, a_t) := r_t + \gamma(1 - \text{done}_{t+1})V_\phi(s_{t+1}) - V_\phi(s_t)$
- вычислить GAE advantage-оценки:  
 $\Psi^{\text{GAE}}(s_{N-1}, a_{N-1}) := \Psi_{(1)}(s_{N-1}, a_{N-1})$
- для  $t$  от  $N - 2$  до 0:
  - $\Psi^{\text{GAE}}(s_t, a_t) :=$

## Proximal Policy Optimization (PPO)

Инициализировать  $\pi(a | s, \theta)$ ,  $V_\phi(s)$ ;

for  $k = 0, 1, 2 \dots$

- собрать  $s_0, a_0, r_0, s_1, \text{done}_1, a_1 \dots s_N, \text{done}_N$  по  $\pi(a | s, \theta)$ ;  
запомнить вероятности действий как  $\pi^{\text{old}}(a_t | s_t) := \pi(a_t | s_t, \theta)$   
запомнить значения критика как  $V^{\text{old}}(s_t) := V_\phi(s_t)$
- вычислить 1-шаговые невязки:  
 $\Psi_{(1)}(s_t, a_t) := r_t + \gamma(1 - \text{done}_{t+1})V_\phi(s_{t+1}) - V_\phi(s_t)$
- вычислить GAE advantage-оценки:  
 $\Psi^{\text{GAE}}(s_{N-1}, a_{N-1}) := \Psi_{(1)}(s_{N-1}, a_{N-1})$
- для  $t$  от  $N - 2$  до 0:
  - $\Psi^{\text{GAE}}(s_t, a_t) := \Psi_{(1)}(s_t, a_t) + \lambda\gamma(1 - \text{done}_{t+1})\Psi^{\text{GAE}}(s_{t+1}, a_{t+1})$

## Proximal Policy Optimization (PPO)

Инициализировать  $\pi(a | s, \theta), V_\phi(s)$ ;

for  $k = 0, 1, 2 \dots$

- собрать  $s_0, a_0, r_0, s_1, \text{done}_1, a_1 \dots s_N, \text{done}_N$  по  $\pi(a | s, \theta)$ ;  
запомнить вероятности действий как  $\pi^{\text{old}}(a_t | s_t) := \pi(a_t | s_t, \theta)$   
запомнить значения критика как  $V^{\text{old}}(s_t) := V_\phi(s_t)$
- вычислить 1-шаговые невязки:  
 $\Psi_{(1)}(s_t, a_t) := r_t + \gamma(1 - \text{done}_{t+1})V_\phi(s_{t+1}) - V_\phi(s_t)$
- вычислить GAE advantage-оценки:  
 $\Psi^{\text{GAE}}(s_{N-1}, a_{N-1}) := \Psi_{(1)}(s_{N-1}, a_{N-1})$
- для  $t$  от  $N - 2$  до 0:
  - $\Psi^{\text{GAE}}(s_t, a_t) := \Psi_{(1)}(s_t, a_t) + \lambda\gamma(1 - \text{done}_{t+1})\Psi^{\text{GAE}}(s_{t+1}, a_{t+1})$
- вычислить целевые значения критика:  
 $y(s_t) := \Psi^{\text{GAE}}(s_t, a_t) + V_\phi(s_t)$

## Proximal Policy Optimization (PPO)

Инициализировать  $\pi(a | s, \theta), V_\phi(s)$ ;

for  $k = 0, 1, 2 \dots$

- собрать  $s_0, a_0, r_0, s_1, \text{done}_1, a_1 \dots s_N, \text{done}_N$  по  $\pi(a | s, \theta)$ ;  
запомнить вероятности действий как  $\pi^{\text{old}}(a_t | s_t) := \pi(a_t | s_t, \theta)$   
запомнить значения критика как  $V^{\text{old}}(s_t) := V_\phi(s_t)$

- вычислить 1-шаговые невязки:

$$\Psi_{(1)}(s_t, a_t) := r_t + \gamma(1 - \text{done}_{t+1})V_\phi(s_{t+1}) - V_\phi(s_t)$$

- вычислить GAE advantage-оценки:

$$\Psi^{\text{GAE}}(s_{N-1}, a_{N-1}) := \Psi_{(1)}(s_{N-1}, a_{N-1})$$

- для  $t$  от  $N - 2$  до 0:

- $\Psi^{\text{GAE}}(s_t, a_t) := \Psi_{(1)}(s_t, a_t) + \lambda\gamma(1 - \text{done}_{t+1})\Psi^{\text{GAE}}(s_{t+1}, a_{t+1})$

- вычислить целевые значения критика:

$$y(s_t) := \Psi^{\text{GAE}}(s_t, a_t) + V_\phi(s_t)$$

- составить выборку  $(s_t, a_t, \Psi^{\text{GAE}}(s_t, a_t), y(s_t), \pi^{\text{old}}(a_t | s_t), V^{\text{old}}(s_t))$

## Proximal Policy Optimization (PPO) — продолжение

- проход по выборке  $n_{\text{epochs}}$  раз, сэмплируя мини-батчи размера  $B$ ; для каждого мини-батча:

## Proximal Policy Optimization (PPO) — продолжение

- проход по выборке  $n\_epochs$  раз, сэмплируя мини-батчи размера  $B$ ; для каждого мини-батча:
  - нормализация  $\Psi^{\text{GAE}}(s, a)$  в батче вычитанием среднего и делением на стандартное отклонение

## Proximal Policy Optimization (PPO) — продолжение

- проход по выборке  $n\_epochs$  раз, сэмплируя мини-батчи размера  $B$ ; для каждого мини-батча:
  - нормализация  $\Psi^{\text{GAE}}(s, a)$  в батче вычитанием среднего и делением на стандартное отклонение
  - вычислить веса выборки по значимости:
$$\rho(s, a, \theta) := \frac{\pi(a|s, \theta)}{\pi^{\text{old}}(a|s)}, \quad \rho^{\text{clip}}(s, a, \theta) = \text{clip}(\rho(s, a, \theta), 1 - \epsilon, 1 + \epsilon)$$

## Proximal Policy Optimization (PPO) — продолжение

- проход по выборке  $n\_epochs$  раз, сэмплируя мини-батчи размера  $B$ ; для каждого мини-батча:
  - нормализация  $\Psi^{\text{GAE}}(s, a)$  в батче вычитанием среднего и делением на стандартное отклонение
  - вычислить веса выборки по значимости:  
$$\rho(s, a, \theta) := \frac{\pi(a|s, \theta)}{\pi^{\text{old}}(a|s)}, \quad \rho^{\text{clip}}(s, a, \theta) = \text{clip}(\rho(s, a, \theta), 1 - \epsilon, 1 + \epsilon)$$
  - обновление параметров актора:  
$$L_1(s, a, \theta) := \rho(s, a, \theta)\Psi^{\text{GAE}}(s, a), \quad L_2(s, a, \theta) := \rho^{\text{clip}}(s, a, \theta)\Psi^{\text{GAE}}(s, a)$$
  
$$\theta \leftarrow \theta + \alpha \nabla_{\theta} \frac{1}{B} \sum_{s,a} \min(L_1(s, a, \theta), L_2(s, a, \theta))$$

## Proximal Policy Optimization (PPO) — продолжение

- проход по выборке  $n\_epochs$  раз, сэмплируя мини-батчи размера  $B$ ; для каждого мини-батча:
  - нормализация  $\Psi^{\text{GAE}}(s, a)$  в батче вычитанием среднего и делением на стандартное отклонение
  - вычислить веса выборки по значимости:  
$$\rho(s, a, \theta) := \frac{\pi(a|s, \theta)}{\pi^{\text{old}}(a|s)}, \quad \rho^{\text{clip}}(s, a, \theta) = \text{clip}(\rho(s, a, \theta), 1 - \epsilon, 1 + \epsilon)$$
  - обновление параметров актора:  
$$L_1(s, a, \theta) := \rho(s, a, \theta)\Psi^{\text{GAE}}(s, a), \quad L_2(s, a, \theta) := \rho^{\text{clip}}(s, a, \theta)\Psi^{\text{GAE}}(s, a)$$
  
$$\theta \leftarrow \theta + \alpha \nabla_{\theta} \frac{1}{B} \sum_{s,a} \min(L_1(s, a, \theta), L_2(s, a, \theta))$$
  - обновление параметров критика:  
$$\text{Loss}_1(s, \phi) := (y(s) - V_{\phi}(s))^2$$
  
$$\text{Loss}_2(s, \phi) := \left( y(s) - V^{\text{old}}(s) - \text{clip}(V_{\phi}(s) - V^{\text{old}}(s), \hat{\epsilon}, -\hat{\epsilon}) \right)^2$$
  
$$\phi \leftarrow \phi - \alpha \nabla_{\phi} \frac{1}{B} \sum_s \max(\text{Loss}_1(s, \phi), \text{Loss}_2(s, \phi))$$

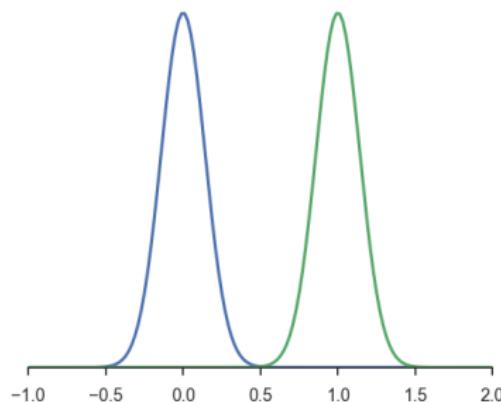


## Источники

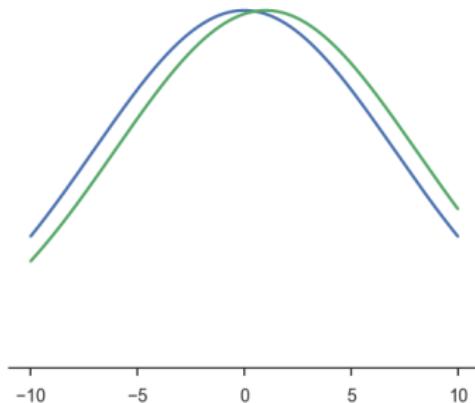
- Proximal Policy Optimization Algorithms;
- Implementation Matters in Deep Policy Gradients: A Case Study on PPO and TRPO;
- High-Dimensional Continuous Control Using Generalized Advantage Estimation;
- Sutton, Barto — Reinforcement Learning, an Introduction, ch. 12;

# Пространство параметров или пространство распределений

Две гауссианы:  $\mathcal{N}(0, 0.2), \mathcal{N}(1, 0.2)$

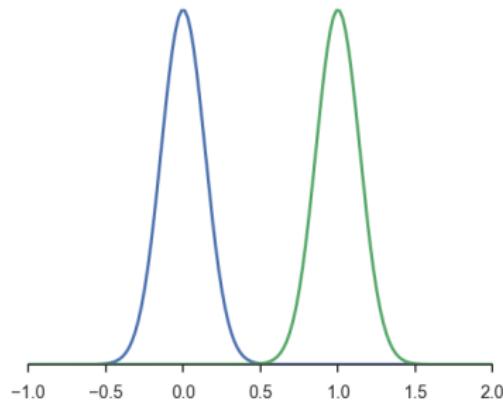


Две гауссианы:  $\mathcal{N}(0, 10), \mathcal{N}(1, 10)$

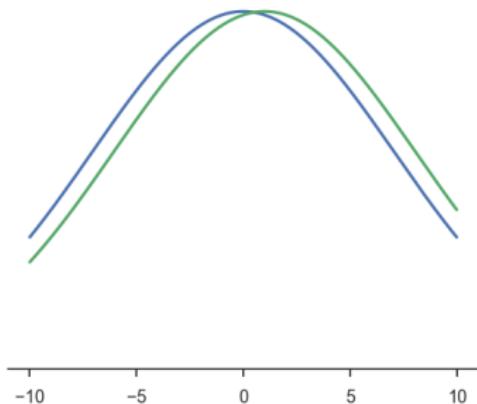


# Пространство параметров или пространство распределений

Две гауссианы:  $\mathcal{N}(0, 0.2), \mathcal{N}(1, 0.2)$



Две гауссианы:  $\mathcal{N}(0, 10), \mathcal{N}(1, 10)$



Евклидово расстояние:  $\sqrt{(\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2}$

Статистическое расстояние: например,  $\text{KL}(\mathcal{N}(\mu_1, \sigma_1) \parallel \mathcal{N}(\mu_2, \sigma_2))$

# Проблема шага: скатывание с обрыва



$$\theta_{k+1} = \theta_k + \alpha \hat{g}_k$$

Если шаг слишком велик:

- широкие шаги ведут к плохой политике;
- начиная *собирать данные* с помощью плохой политики;
- $\implies$  **качество коллапсирует**

# Проблема шага: скатывание с обрыва



$$\theta_{k+1} = \theta_k + \alpha \hat{g}_k$$

Если шаг слишком велик:

- широкие шаги ведут к плохой политике;
- начинаяем *собирать данные* с помощью плохой политики;
- $\implies$  **качество коллапсирует**



Требуется метод, измеряющий расстояние  
в пространстве распределений,  
а не в пространстве параметров!

Рассмотрим следующую задачу оптимизации:

$$G(\theta) := G(q_\theta(x)) \rightarrow \min_{\theta}$$

Рассмотрим следующую задачу оптимизации:

$$G(\theta) := G(q_\theta(x)) \rightarrow \min_{\theta}$$

Общая схема методов оптимизации с доверительной областью

- начало с произвольного  $\theta_0$ ;
- для  $k = 0, 1, 2, \dots$ :
  - построение **модели**, некоторой локальной аппроксимации  $G(\theta)$  около  $\theta_k$ :
$$m_k(\theta) \approx G(\theta)$$

Рассмотрим следующую задачу оптимизации:

$$G(\theta) := G(q_\theta(x)) \rightarrow \min_{\theta}$$

Общая схема методов оптимизации с доверительной областью

- начало с произвольного  $\theta_0$ ;
- для  $k = 0, 1, 2, \dots$ :
  - построение **модели**, некоторой локальной аппроксимации  $G(\theta)$  около  $\theta_k$ :
$$m_k(\theta) \approx G(\theta)$$
  - выбор некоторой **области доверия**, где предполагается модель достаточно точной:
$$r(\theta, \theta_k) \leq \delta$$

Рассмотрим следующую задачу оптимизации:

$$G(\theta) := G(q_\theta(x)) \rightarrow \min_{\theta}$$

Общая схема методов оптимизации с доверительной областью

- начало с произвольного  $\theta_0$ ;
- для  $k = 0, 1, 2, \dots$ :
  - построение **модели**, некоторой локальной аппроксимации  $G(\theta)$  около  $\theta_k$ :
$$m_k(\theta) \approx G(\theta)$$
  - выбор некоторой **области доверия**, где предполагается модель достаточно точной:
$$r(\theta, \theta_k) \leq \delta$$
  - поиск  $\theta_{k+1}$  в качестве решения

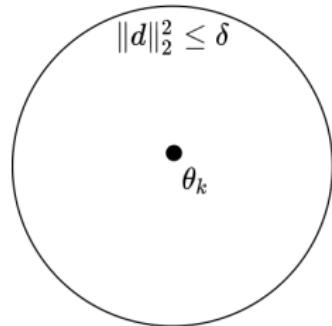
$$\begin{cases} m_k(\theta) \rightarrow \min_{\theta} \\ r(\theta, \theta_k) \leq \delta \end{cases}$$

# Натуральный градиентный спуск

$$G(\theta) := G(q_\theta(x)) \rightarrow \min_{\theta}$$

## Градиентный спуск

$$\begin{cases} G(\theta_k + d) \approx G(\theta_k) + \nabla_{\theta} G(\theta_k)^T d \rightarrow \min_d \\ \|d\|_2 \leq \delta \end{cases}$$

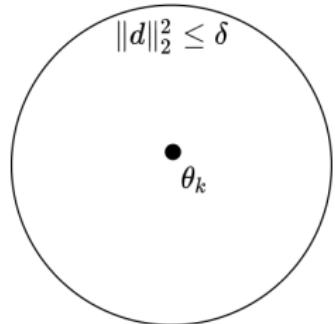


# Натуральный градиентный спуск

$$G(\theta) := G(q_\theta(x)) \rightarrow \min_{\theta}$$

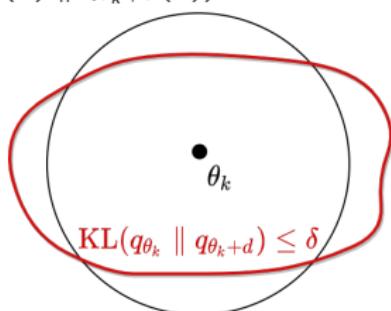
## Градиентный спуск

$$\begin{cases} G(\theta_k + d) \approx G(\theta_k) + \nabla_{\theta} G(\theta_k)^T d \rightarrow \min_d \\ \|d\|_2 \leq \delta \end{cases}$$



## Натуральный градиентный спуск

$$\begin{cases} G(\theta_k + d) \approx G(\theta_k) + \nabla_{\theta} G(\theta_k)^T d \rightarrow \min_d \\ \text{KL}(q_{\theta_k}(x) \parallel q_{\theta_k+d}(x)) \leq \delta \end{cases}$$

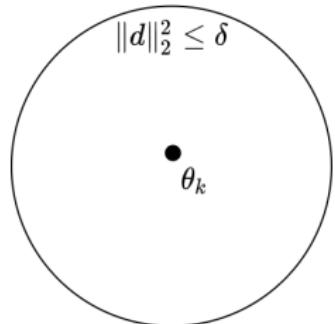


# Натуральный градиентный спуск

$$G(\theta) := G(q_\theta(x)) \rightarrow \min_{\theta}$$

## Градиентный спуск

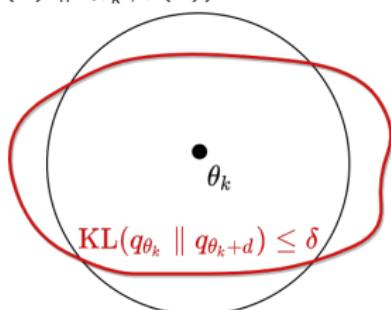
$$\begin{cases} G(\theta_k + d) \approx G(\theta_k) + \nabla_{\theta} G(\theta_k)^T d \rightarrow \min_d \\ \|d\|_2 \leq \delta \end{cases}$$



Решение:  $d \propto -\nabla_{\theta} G(\theta_k)$

## Натуральный градиентный спуск

$$\begin{cases} G(\theta_k + d) \approx G(\theta_k) + \nabla_{\theta} G(\theta_k)^T d \rightarrow \min_d \\ \text{KL}(q_{\theta_k}(x) \parallel q_{\theta_k+d}(x)) \leq \delta \end{cases}$$



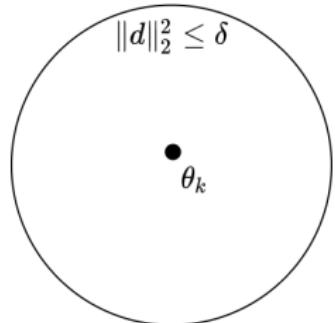
$$\text{KL}(q_{\theta_k} \parallel q_{\theta_k+d}) \leq \delta$$

# Натуральный градиентный спуск

$$G(\theta) := G(q_\theta(x)) \rightarrow \min_{\theta}$$

## Градиентный спуск

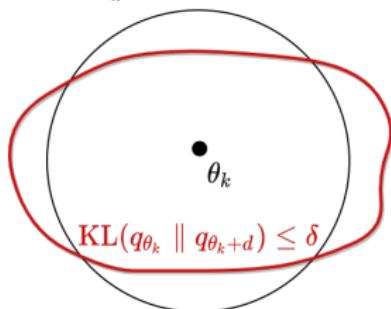
$$\begin{cases} G(\theta_k + d) \approx G(\theta_k) + \nabla_{\theta} G(\theta_k)^T d \rightarrow \min_d \\ \|d\|_2 \leq \delta \end{cases}$$



Решение:  $d \propto -\nabla_{\theta} G(\theta_k)$

## Натуральный градиентный спуск

$$\begin{cases} G(\theta_k + d) \approx G(\theta_k) + \nabla_{\theta} G(\theta_k)^T d \rightarrow \min_d \\ \text{KL}(q_{\theta_k}(x) \parallel q_{\theta_k+d}(x)) \leq \delta \end{cases}$$



Решение: ?!?



Используем разложение  
Тейлора для

$$\text{KL}(q_{\theta_k}(x) \parallel q_{\theta_k+d}(x)) \leq \delta$$



Используем разложение  
Тейлора для

$$\text{KL}(q_{\theta_k}(x) \parallel q_{\theta_k+d}(x)) \leq \delta$$

(!) Первый член в действительности равен нулю:

$$\nabla_d \text{KL}(q_{\theta_k}(x) \parallel q_{\theta_k+d}(x))|_{d=0} = 0$$



Используем разложение  
Тейлора для

$$\text{KL}(q_{\theta_k}(x) \parallel q_{\theta_k+d}(x)) \leq \delta$$

(!) Первый член в действительности равен нулю:

$$\nabla_d \text{KL}(q_{\theta_k}(x) \parallel q_{\theta_k+d}(x))|_{d=0} = 0$$

Информационная матрица Фишера

Пусть  $F(\theta_k)$  будет **матрицей Фишера** распределения  $q_\theta(x)$  в точке  $\theta_k$ :

$$F(\theta_k) := -\mathbb{E}_{x \sim q_\theta(x)} \nabla_\theta^2 \log q_\theta(x)$$



Используем разложение  
Тейлора для

$$\text{KL}(q_{\theta_k}(x) \parallel q_{\theta_k+d}(x)) \leq \delta$$

(!) Первый член в действительности равен нулю:

$$\nabla_d \text{KL}(q_{\theta_k}(x) \parallel q_{\theta_k+d}(x))|_{d=0} = 0$$

Информационная матрица Фишера

Пусть  $F(\theta_k)$  будет **матрицей Фишера** распределения  $q_\theta(x)$  в точке  $\theta_k$ :

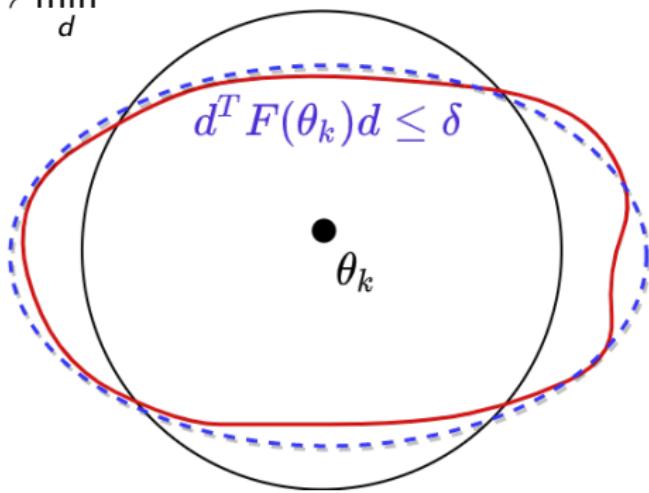
$$F(\theta_k) := -\mathbb{E}_{x \sim q_\theta(x)} \nabla_\theta^2 \log q_\theta(x)$$

Тогда приближение по Тейлору второго порядка:

$$\text{KL}(q_{\theta_k}(x) \parallel q_{\theta_k+d}(x)) \approx \frac{1}{2} d^T F(\theta_k) d$$

# Направление натурального градиента

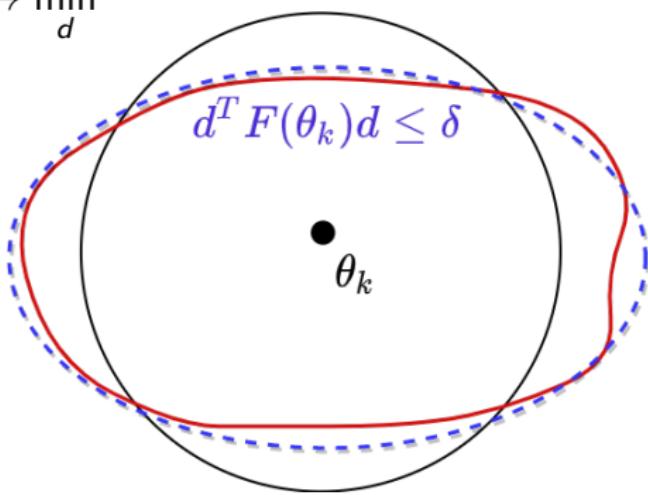
$$\begin{cases} G(\theta_k + d) \approx G(\theta_k) + \nabla_{\theta} G(\theta_k)^T d \rightarrow \min_d \\ d^T F(\theta_k) d \leq \delta \end{cases}$$



# Направление натурального градиента

$$\begin{cases} G(\theta_k + d) \approx G(\theta_k) + \nabla_{\theta} G(\theta_k)^T d \rightarrow \min_d \\ d^T F(\theta_k) d \leq \delta \end{cases}$$

Решение:  $d \propto -F^{-1}(\theta_k) \nabla_{\theta} G(\theta_k)$



# Направление натурального градиента

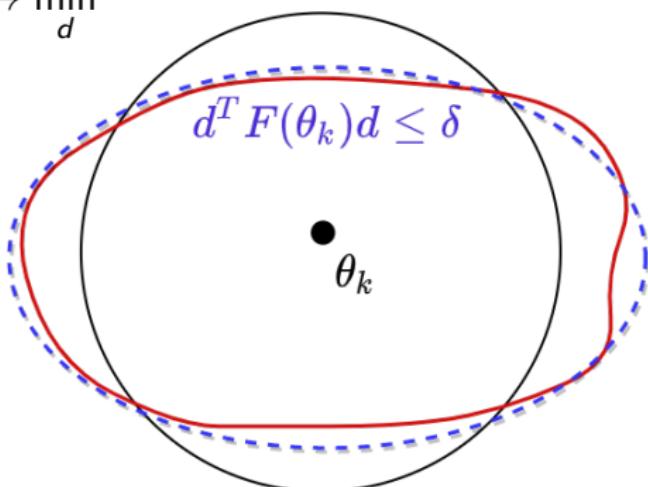
$$\begin{cases} G(\theta_k + d) \approx G(\theta_k) + \nabla_{\theta} G(\theta_k)^T d \rightarrow \min_d \\ d^T F(\theta_k) d \leq \delta \end{cases}$$

Решение:  $d \propto -F^{-1}(\theta_k) \nabla_{\theta} G(\theta_k)$

Метод натурального градиента обновляет параметры следующим образом:

$$\theta_{k+1} = \theta_k - \alpha F^{-1}(\theta_k) \nabla_{\theta} G(\theta_k),$$

где  $\alpha$  определяется выбором  $\delta$ .



## Value-based (DQN+)

✓ off-policy;

(можно использовать  
реплей-буфер)

✗ обучаем  $Q^*(s, a)$ ;

(сложный промежуточный шаг)

✗ проблемы с исследованием  
среды;

(так как Value Iteration работает с  
детерминистическими политиками)

✗ 1-шаговые целевые значения;  
(можем ли модифицировать?)

## Policy Gradient

✗ on-policy;

(данные батча бесполезны после  
обновления параметров)

✓ обучаем политику напрямую;  
(требует только  $V^\pi(s)$ , что гораздо  
проще)

✓ «естественное» исследование;  
(сэмплирование из  $\pi(a | s)$ )

✓  $\infty$ -шаговые целевые значения;  
(можно использовать GAE и для  
критика, и для актора)

# Off-policy оценка advantage

Данной траектории  $s_0, r_0, s_1, r_1, s_2, r_2 \dots s_M$  по политики  $\mu$  и приближению  $V^\pi(s)$

## Off-policy оценка advantage

Данной траектории  $s_0, r_0, s_1, r_1, s_2, r_2 \dots s_M$  по политики  $\mu$  и приближению  $V^\pi(s)$  требуется произвести **оценку advantage (credit assignment)** для пары состояние-действие  $s_0, a_0$  в **off-policy** режиме:  $\mu \neq \pi$ .

## Off-policy оценка advantage

Данной траектории  $s_0, r_0, s_1, r_1, s_2, r_2 \dots s_M$  по политики  $\mu$  и приближению  $V^\pi(s)$  требуется произвести **оценку advantage (credit assignment)** для пары состояние-действие  $s_0, a_0$  в **off-policy** режиме:  $\mu \neq \pi$ .

Было бы замечательно воспользоваться GAE:

$$\sum_{t \geq 0} (\gamma \lambda)^t \Psi_{(1)}(s_t, a_t),$$

но  $\Psi_{(1)}(s_t, a_t)$  зависит от случайных величин:  $a_0, s_0, a_1, s_2, \dots s_{t+1}$ .

## Off-policy оценка advantage

Данной траектории  $s_0, r_0, s_1, r_1, s_2, r_2 \dots s_M$  по политике  $\mu$  и приближению  $V^\pi(s)$  требуется произвести **оценку advantage (credit assignment)** для пары состояние-действие  $s_0, a_0$  в **off-policy** режиме:  $\mu \neq \pi$ .

Было бы замечательно воспользоваться GAE:

$$\sum_{t \geq 0} (\gamma \lambda)^t \Psi_{(1)}(s_t, a_t),$$

но  $\Psi_{(1)}(s_t, a_t)$  зависит от случайных величин:  $a_0, s_0, a_1, s_2, \dots s_{t+1}$ .

Внимание!

Если  $\pi(a_0 | s_0) = 0$ , то ничего не получится.



Воспользуемся коррекцией с  
помощью выборки по значимости!



Воспользуемся коррекцией с  
помощью выборки по значимости!

$$\Psi = \sum_{t \geq 0} (\gamma \lambda)^t \left( \prod_{\hat{t}=1}^t \frac{\pi(a_{\hat{t}} | s_{\hat{t}}) p(s_{\hat{t}+1} | s_{\hat{t}}, a_{\hat{t}})}{\mu(a_{\hat{t}} | s_{\hat{t}}) p(s_{\hat{t}+1} | s_{\hat{t}}, a_{\hat{t}})} \right) \Psi_{(1)}(s_t, a_t) =$$

=



Воспользуемся коррекцией с  
помощью выборки по значимости!

$$\begin{aligned}\Psi &= \sum_{t \geq 0} (\gamma \lambda)^t \left( \prod_{\hat{t}=1}^t \frac{\pi(a_{\hat{t}} | s_{\hat{t}}) \cancel{p(s_{\hat{t}+1} | s_{\hat{t}}, a_{\hat{t}})}}{\mu(a_{\hat{t}} | s_{\hat{t}}) \cancel{p(s_{\hat{t}+1} | s_{\hat{t}}, a_{\hat{t}})}} \right) \Psi_{(1)}(s_t, a_t) = \\ &= \sum_{t \geq 0} (\gamma \lambda)^t \left( \prod_{\hat{t}=1}^t \frac{\pi(a_{\hat{t}} | s_{\hat{t}})}{\mu(a_{\hat{t}} | s_{\hat{t}})} \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{\hat{t}=1}^0 \equiv 1.\end{aligned}$$



Воспользуемся коррекцией с  
помощью выборки по значимости!

$$\begin{aligned}\Psi &= \sum_{t \geq 0} (\gamma \lambda)^t \left( \prod_{\hat{t}=1}^t \frac{\pi(a_{\hat{t}} | s_{\hat{t}}) \cancel{p(s_{\hat{t}+1} | s_{\hat{t}}, a_{\hat{t}})}}{\mu(a_{\hat{t}} | s_{\hat{t}}) \cancel{p(s_{\hat{t}+1} | s_{\hat{t}}, a_{\hat{t}})}} \right) \Psi_{(1)}(s_t, a_t) = \\ &= \sum_{t \geq 0} (\gamma \lambda)^t \left( \prod_{\hat{t}=1}^t \frac{\pi(a_{\hat{t}} | s_{\hat{t}})}{\mu(a_{\hat{t}} | s_{\hat{t}})} \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{\hat{t}=1}^0 \equiv 1.\end{aligned}$$

Непрактично: очень высокая дисперсия!



Воспользуемся коррекцией с  
помощью выборки по значимости!

$$\begin{aligned}\Psi &= \sum_{t \geq 0} (\gamma \lambda)^t \left( \prod_{\hat{t}=1}^t \frac{\pi(a_{\hat{t}} | s_{\hat{t}}) \cancel{p(s_{\hat{t}+1} | s_{\hat{t}}, a_{\hat{t}})}}{\mu(a_{\hat{t}} | s_{\hat{t}}) \cancel{p(s_{\hat{t}+1} | s_{\hat{t}}, a_{\hat{t}})}} \right) \Psi_{(1)}(s_t, a_t) = \\ &= \sum_{t \geq 0} (\gamma \lambda)^t \left( \prod_{\hat{t}=1}^t \frac{\pi(a_{\hat{t}} | s_{\hat{t}})}{\mu(a_{\hat{t}} | s_{\hat{t}})} \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{\hat{t}=1}^0 \equiv 1.\end{aligned}$$

Непрактично: очень высокая дисперсия!

- «Затухающий» след:  $\mu(a|s) \gg \pi(a|s)$ :
  - типичная ситуация  $\mu$  делает примитивные случайные действия, которые  $\pi$  редко совершает. Не лечится.



Воспользуемся коррекцией с помощью выборки по значимости!

$$\begin{aligned}\Psi &= \sum_{t \geq 0} (\gamma \lambda)^t \left( \prod_{\hat{t}=1}^t \frac{\pi(a_{\hat{t}} | s_{\hat{t}}) p(s_{\hat{t}+1} | s_{\hat{t}}, a_{\hat{t}})}{\mu(a_{\hat{t}} | s_{\hat{t}}) p(\cancel{s_{\hat{t}+1} | s_{\hat{t}}, a_{\hat{t}}})} \right) \Psi_{(1)}(s_t, a_t) = \\ &= \sum_{t \geq 0} (\gamma \lambda)^t \left( \prod_{\hat{t}=1}^t \frac{\pi(a_{\hat{t}} | s_{\hat{t}})}{\mu(a_{\hat{t}} | s_{\hat{t}})} \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{\hat{t}=1}^0 \equiv 1.\end{aligned}$$

Непрактично: очень высокая дисперсия!

- «**Затухающий**» след:  $\mu(a|s) \gg \pi(a|s)$ :
  - типичная ситуация  $\mu$  делает примитивные случайные действия, которые  $\pi$  редко совершает. Не лечится.
- «**Взрывающийся**» след:  $\mu(a|s) \ll \pi(a|s)$ :
  - $\mu$  выбранное действие с малой  $\mu(a|s)$ , но вероятное для  $\pi$ .  
Причина большой дисперсии.

## Присвоение ценности: общий вид

Давайте перепишем ценность следующим образом:

$$\Psi = \sum_{t \geq 0} \gamma^t \left( \prod_{i=1}^t c_i \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{i=1}^0 \equiv 1,$$

где  $c_i$  коэффициенты «отжига следа»:

Название оценки	Коэффициенты $c_i$	Возникающая проблема
GAE	$\lambda$	только on-policy

## Присвоение ценности: общий вид

Давайте перепишем ценность следующим образом:

$$\Psi = \sum_{t \geq 0} \gamma^t \left( \prod_{i=1}^t c_i \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{i=1}^0 \equiv 1,$$

где  $c_i$  коэффициенты «отжига следа»:

Название оценки	Коэффициенты $c_i$	Возникающая проблема
GAE	$\lambda$	только on-policy
Одношаговая	0	большое смещение

## Присвоение ценности: общий вид

Давайте перепишем ценность следующим образом:

$$\Psi = \sum_{t \geq 0} \gamma^t \left( \prod_{i=1}^t c_i \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{i=1}^0 \equiv 1,$$

где  $c_i$  коэффициенты «отжига следа»:

Название оценки	Коэффициенты $c_i$	Возникающая проблема
GAE	$\lambda$	только on-policy
Одношаговая	0	большое смещение
Выборка по значимости	$\lambda \frac{\pi(a_i s_i)}{\mu(a_i s_i)}$	легко «взрывается»

## Retrace: основная теорема

$$\Psi = \sum_{t \geq 0} \gamma^t \left( \prod_{i=1}^t c_i \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{i=1}^0 \equiv 1.$$

### Теорема о Retrace

В режиме on-policy возможен выбор **произвольного** коэффициента  $c_i \in [0, 1]$ , в off-policy режиме можно выбрать **произвольный** коэффициент в следующем интервале

$$c_i \in \left[ 0, \frac{\pi(a_i | s_i)}{\mu(a_i | s_i)} \right].$$

## Retrace: основная теорема

$$\Psi = \sum_{t \geq 0} \gamma^t \left( \prod_{i=1}^t c_i \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{i=1}^0 \equiv 1.$$

### Теорема о Retrace

В режиме on-policy возможен выбор **произвольного** коэффициента  $c_i \in [0, 1]$ , в off-policy режиме можно выбрать **произвольный** коэффициент в следующем интервале

$$c_i \in \left[ 0, \frac{\pi(a_i | s_i)}{\mu(a_i | s_i)} \right].$$

- «затухающий» след: ничего не поделаешь;

## Retrace: основная теорема

$$\Psi = \sum_{t \geq 0} \gamma^t \left( \prod_{i=1}^t c_i \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{i=1}^0 \equiv 1.$$

### Теорема о Retrace

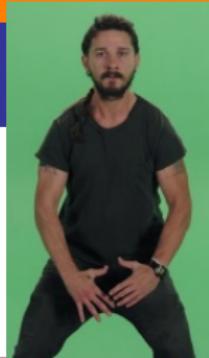
В режиме on-policy возможен выбор **произвольного** коэффициента  $c_i \in [0, 1]$ , в off-policy режиме можно выбрать **произвольный** коэффициент в следующем интервале

$$c_i \in \left[ 0, \frac{\pi(a_i | s_i)}{\mu(a_i | s_i)} \right].$$

- **«затухающий» след:** ничего не поделаешь;
- **«взрывающийся» след:** если вес из выборки по значимости больше 1, ПРОСТО КЛИППИРУЙ ЕГО!

## Retrace: основная теорема

$$\Psi = \sum_{t \geq 0} \gamma^t \left( \prod_{i=1}^t c_i \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{i=1}^0 \equiv 1.$$



### Теорема о Retrace

В режиме on-policy возможен выбор **произвольного** коэффициента  $c_i \in [0, 1]$ , в off-policy режиме можно выбрать **произвольный** коэффициент в следующем интервале

$$c_i \in \left[ 0, \frac{\pi(a_i | s_i)}{\mu(a_i | s_i)} \right].$$

- **«затухающий» след:** ничего не поделаешь;
- **«взрывающийся» след:** если вес из выборки по значимости больше 1, ПРОСТО КЛИППИРУЙ ЕГО!

## Retrace: финальный результат

$$\Psi = \sum_{t \geq 0} \gamma^t \left( \prod_{i=1}^t c_i \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{i=1}^0 \equiv 1,$$

где

$$c_i := \lambda \min \left( 1, \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \right).$$

## Retrace: финальный результат

$$\Psi = \sum_{t \geq 0} \gamma^t \left( \prod_{i=1}^t c_i \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{i=1}^0 \equiv 1,$$

где

$$c_i := \lambda \min \left( 1, \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \right).$$

Используется в:

- off-policy RL алгоритмах для теоретически корректных многошаговых целевых значений;
  - ( $\lambda = 1$ , потому что оно быстро затухает).

## Retrace: финальный результат

$$\Psi = \sum_{t \geq 0} \gamma^t \left( \prod_{i=1}^t c_i \right) \Psi_{(1)}(s_t, a_t), \quad \prod_{i=1}^0 \equiv 1,$$

где

$$c_i := \lambda \min \left( 1, \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \right).$$

Используется в:

- off-policy RL алгоритмах для теоретически корректных многошаговых целевых значений;
  - ( $\lambda = 1$ , потому что оно быстро затухает).
- дистрибутивных on-policy RL системах, где данные о градиенте от некоторых серверов могут задерживаться на несколько итераций обновления.