NEURAL NETWORKS REINFORCEMENT LEARNING

Кетков С.

Кондратьев Н.

Макарова О.

Прибыткина Д.

Семенов С.

Part 1

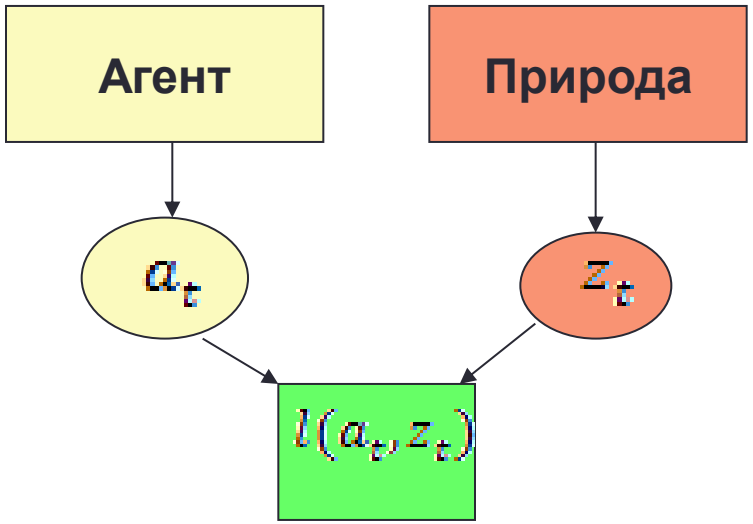
План

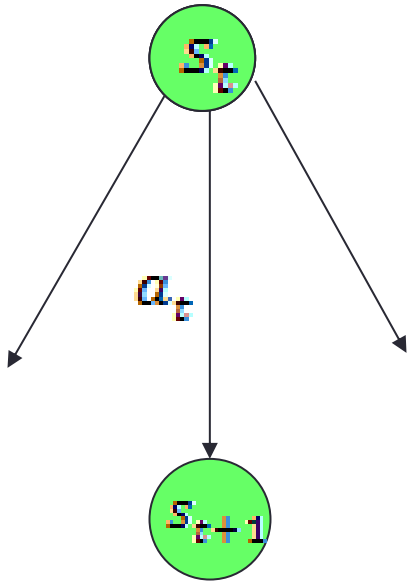
1. Постановка задачи;
2. Задача о многоруком бандите;
3. Простейшие подходы к решению;
4. Q-learning.

Задача обучения с подкреплением.

S – пространство состояний

A – пространство действий (конечное)



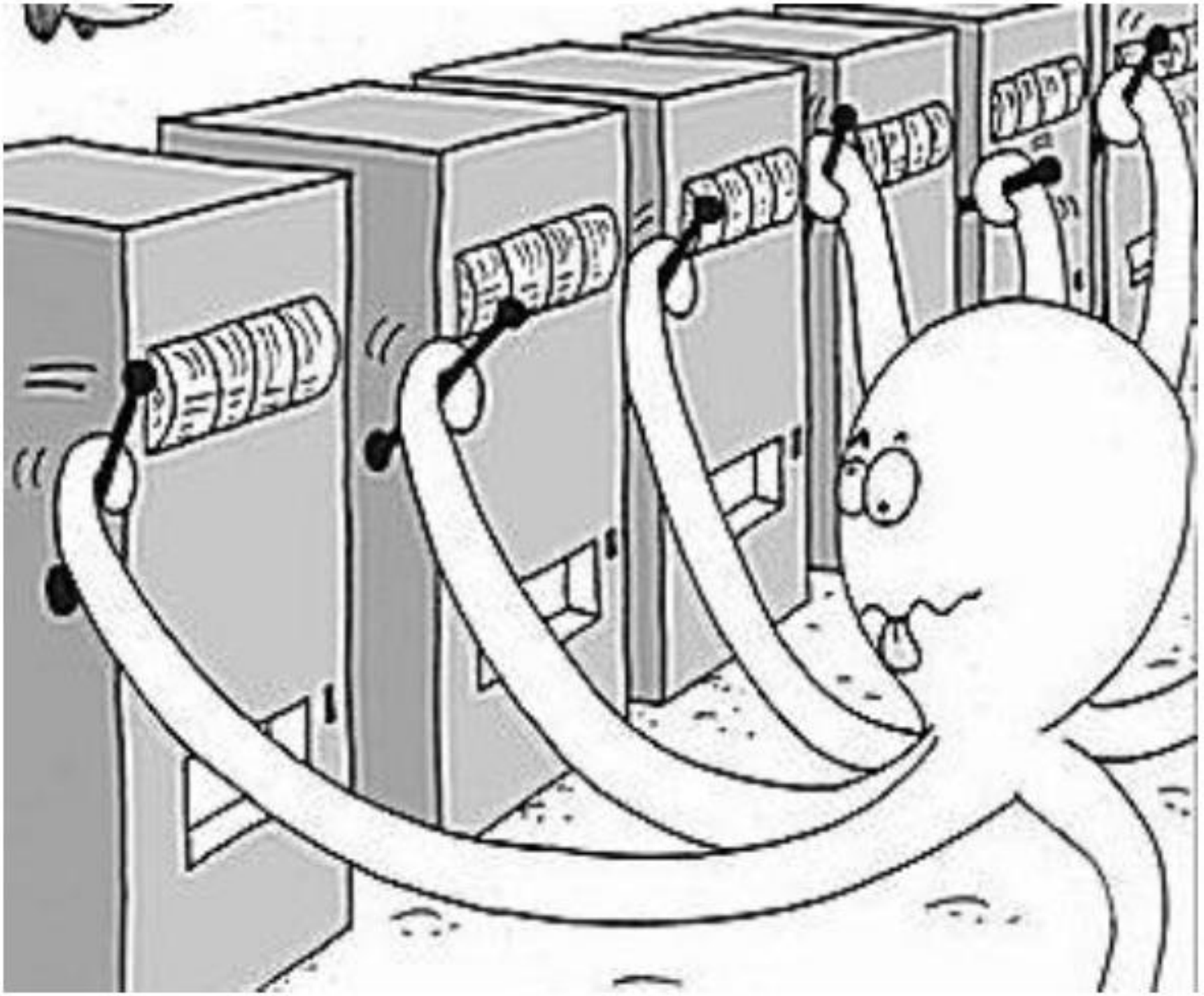


$l(a_t, z_t)$ - *функция потерь*

Предположения о виде обратной связи

1. Full information case (агент получает информацию о потерях за все возможные действия в среде)

2. Bandit feedback (агент получает информацию лишь о совершенных им действиях)



Многорукий бандит

Стохастический многорукий бандит

Известные параметры: число ручек N, и, возможно, число эпизодов $T \geq N$.

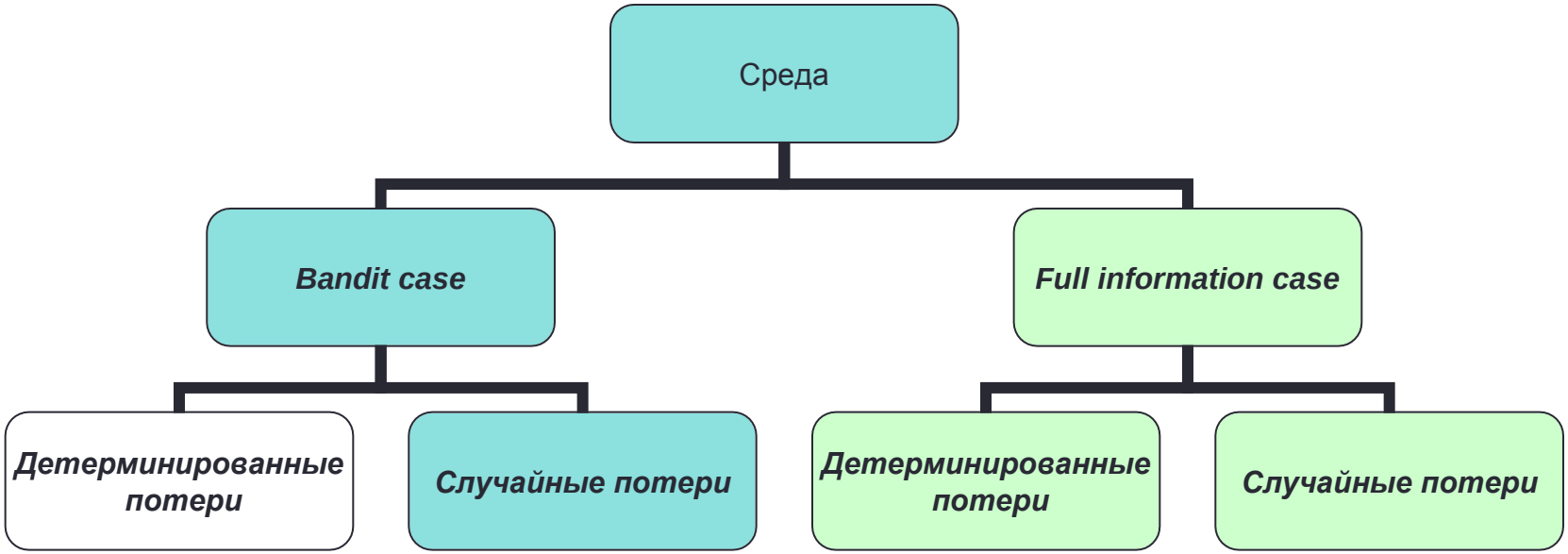
Неизвестные параметры: распределения вероятностей $\nu_1, \nu_2, \ldots, \nu_N \in [0,1]$.

Каждый раунд $t = 1, 2, \ldots$

(1) Предсказатель выбирает ручку $I_t \in \{1, \ldots, N\}$;

(2) Получив I_t, природа формирует $z_t(I_t) \in [0,1]$ согласно распределению вероятностей ν_{I_t}, независимо от предыстории, и предъявляет его предсказателю.

Предположения о среде



Предположения о природе штрафов

1. Стохастическая природа:

$$\forall i \in E \; \mathcal{M}\{z_t(i)\} \equiv \mu_i$$

z_t

2. Враждебная природа:

$$z_t = z_t(a_1, z_1; a_2, z_2; \ldots; a_{t-1}, z_{t-1})$$

Марковские процессы (MDP)

- S – множество состояний
- A – множество действий
- P_{sa} - вероятности переходов из $s \in S$ при выполнении действия a
- $\gamma \in [0, 1)$ - коэффициент дисконтирования
- $R : S \times A \mapsto \mathbb{R}$ - функция ценности

Цель обучения с подкреплением

- Общая ценность:

$$R_{total} = R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \ldots$$

- Выбираем действие так, чтобы максимизировать ожидаемое значение общей ценности:

$$E(R_{total}) \rightarrow max$$

Простейшие подходы

- ε – жадная стратегия;
- *softmax* стратегия (распределение Гиббса).

Основная проблема