

Exploration in deep RL

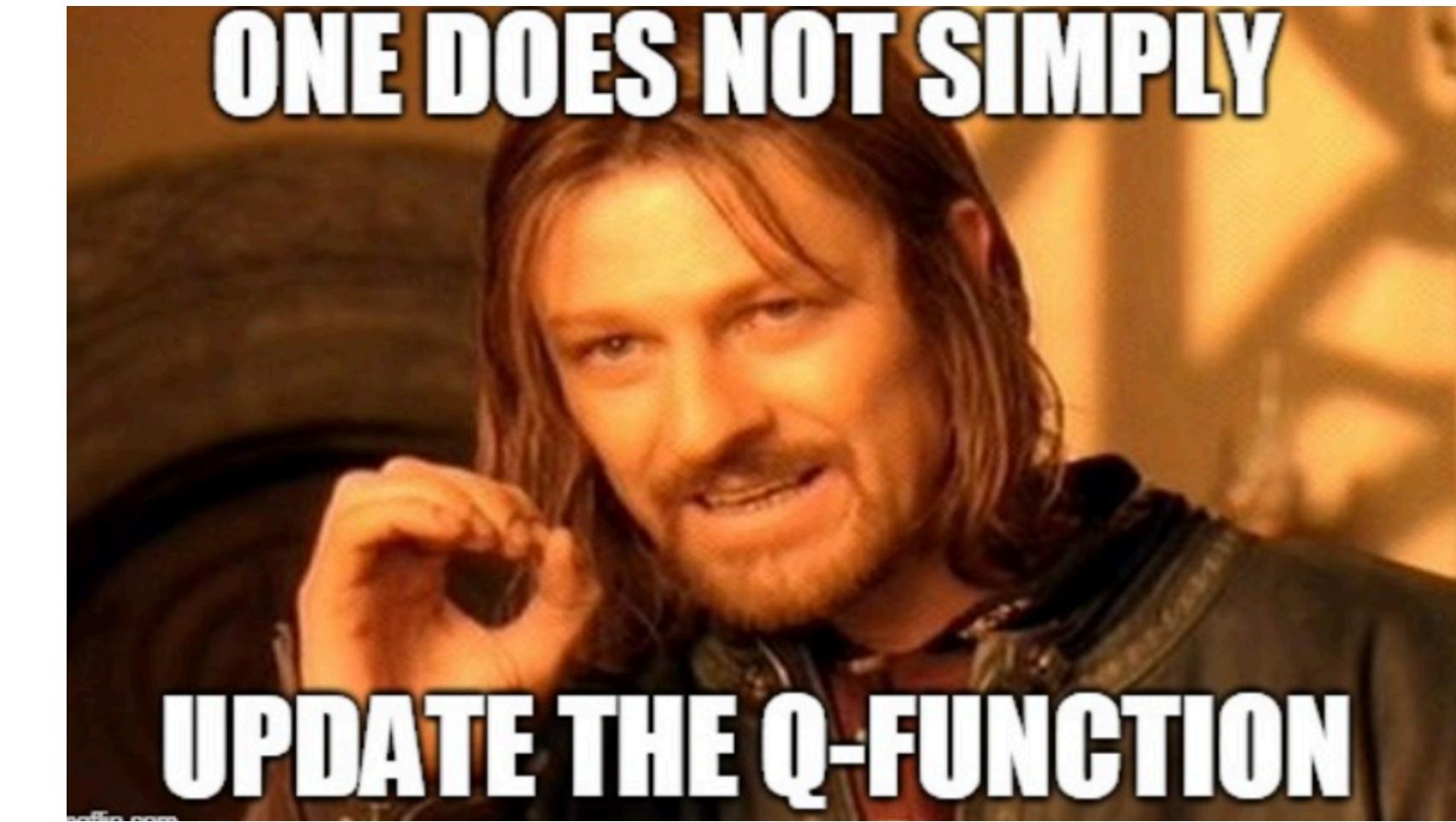
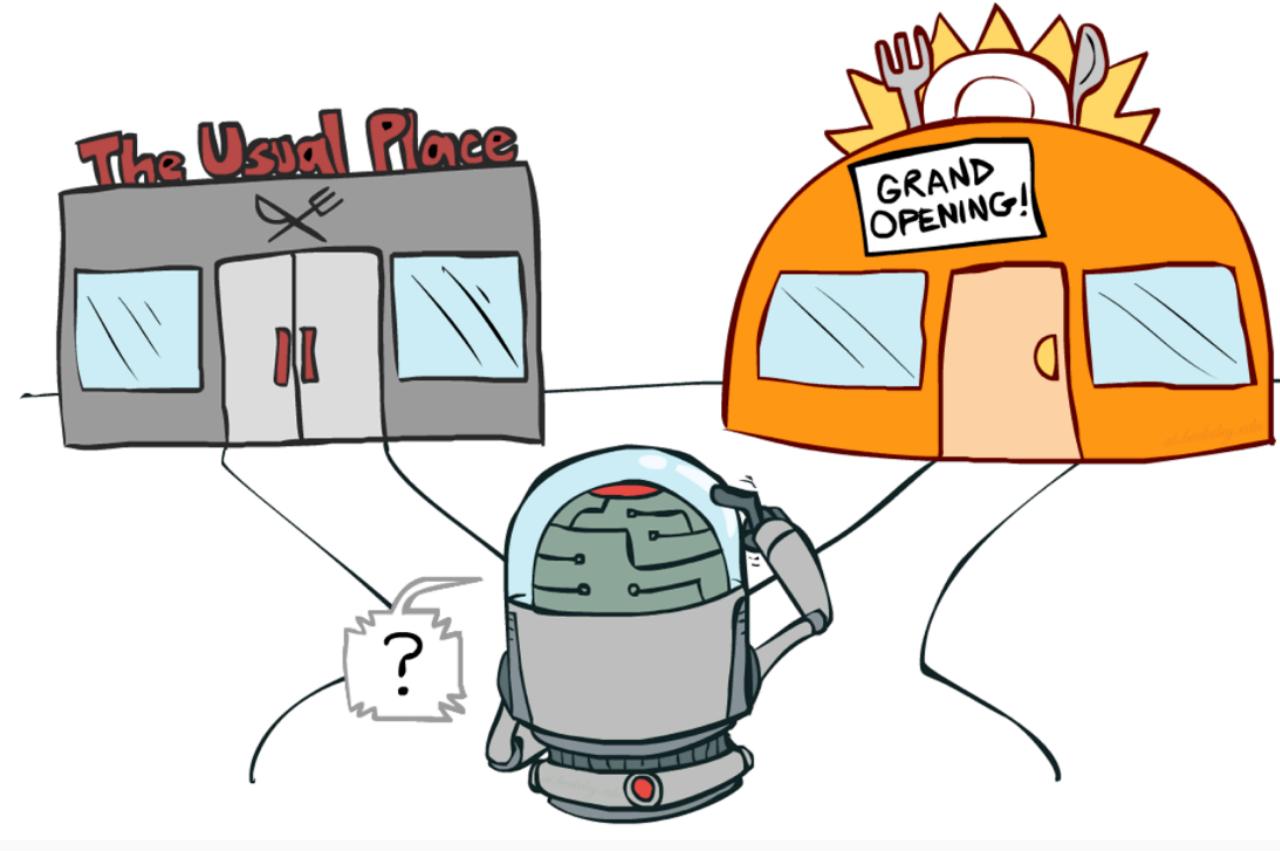
Барановская Дарья
БПМИ 181

02.03.2021

План рассказа

- 1. Exploration vs Exploitation**
- 2. Классические стратегии**
- 3. Основные проблемы exploration:** проблема трудного исследования, проблема шумного телевизора
- 4. Внутренние награды за исследования:**
 - A. Исследование на основе подсчета
 - B. Исследование на основе предсказательной модели(prediction model)
 - C. Случайные нейросети
 - D. Физические свойства
- 5. На основе памяти:**
 - A. Эпизодическая память
 - B. Прямое исследование

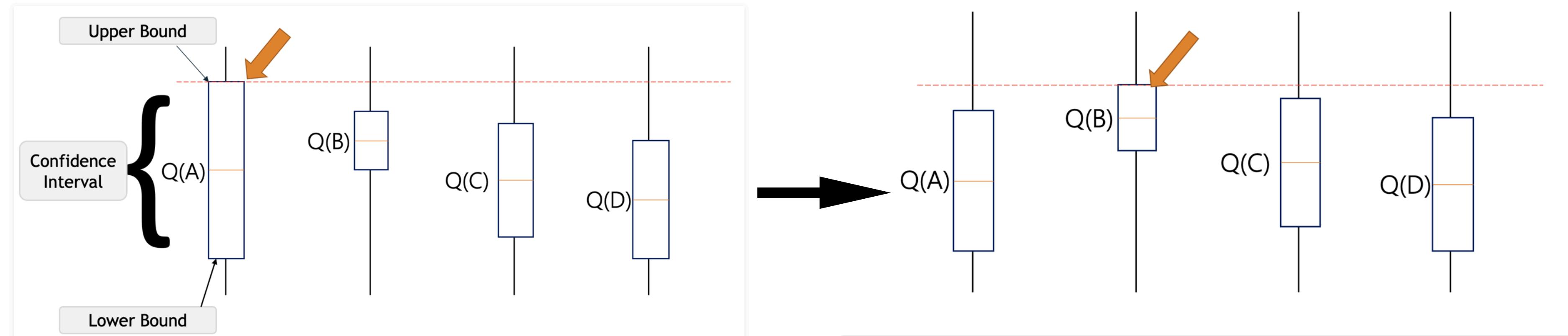
Exploration vs Exploitation



- Эксплуатации - использование лучшего из известных вариантов.
Исследование - риск чтобы собрать информацию о неизвестных вариантах.
- Лучшая долгосрочная стратегия может включать в себя краткосрочные жертвы

Классические стратегии

- **Epsilon-жадная стратегия:** (С вероятностью epsilon агент проводит случайное исследование, а с вероятностью 1-epsilon эксплуатирует известную лучшую стратегию)
- **Верхний доверительный интервал:** (Агент оптимистично выбирает действие с наивысшей верхней границей доверительного интервала. Делая это агент либо получает наивысшее вознаграждение либо исследует действие, о котором знает меньше всего)

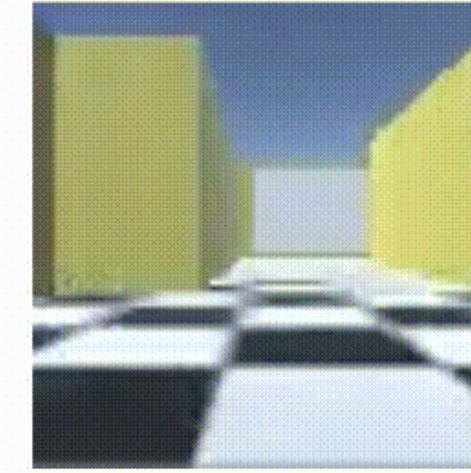


Основные проблемы exploration

- Проблема сложного исследования (в основном в среде с очень редким или даже обманчивым вознаграждением)
- Noisy-TV problem (агент находит источник случайности и продолжает наблюдать за ним, получая небольшие награждения за исследование)



Agent in a maze with a noisy TV



Agent in a maze without a noisy TV

Как решать эти проблемы?

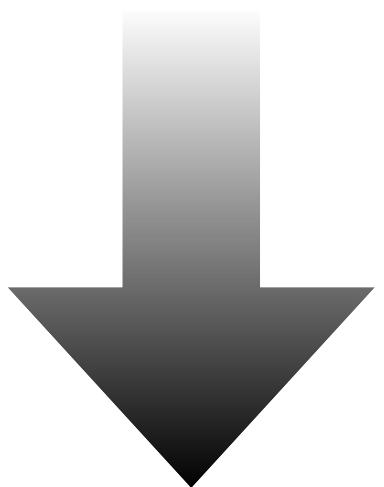
Метод внутреннего вознаграждения

$$r_t = r_t^e + \beta r_t^i$$

r_t^e - внешнее вознаграждение среды во время t

r_t^i - внутренний бонус за исследование во время t

β - гиперпараметр, регулирующий баланс между исследованием и эксплуатацией



$$Q(s_t, a_t) \leftarrow (1 - \alpha) Q(s_t, a_t) + \alpha \left(r + \gamma \max_a Q(s_{t+1}, a) \right)$$

Исследование на основе предсказательной модели

Forward dynamics

Intelligent Adaptive Curiosity

$s_i(t)$ - данные с i -го сенсора в момент времени t

$m_i(t)$ - i -й параметр действий в момент времени t

$SM(t)$ - матрица всех параметров в момент времени t

$(SM(t), S(t + 1))$ - экземпляр, собранный роботом

R_n - регион, с которым ассоциируется эксперт E_n . Экспертом может быть нейронная сеть, SVM, Байесовская машина

$\tilde{S}(t + 1)$ - сенсорное состояние, предсказанное экспертом E_n

$S(t + 1)$ - сенсорное состояние, которое наблюдалось на самом деле

θ - параметр сглаживания

τ - параметр временного окна

$e_n(t + 1) = ||S(t + 1) - \tilde{S}(t + 1)||^2$ - squared error

$\langle e_n(t) \rangle = \frac{1}{\theta + 1} \sum_{i=0}^{\theta} e_n(t - i)$ - mean error rate

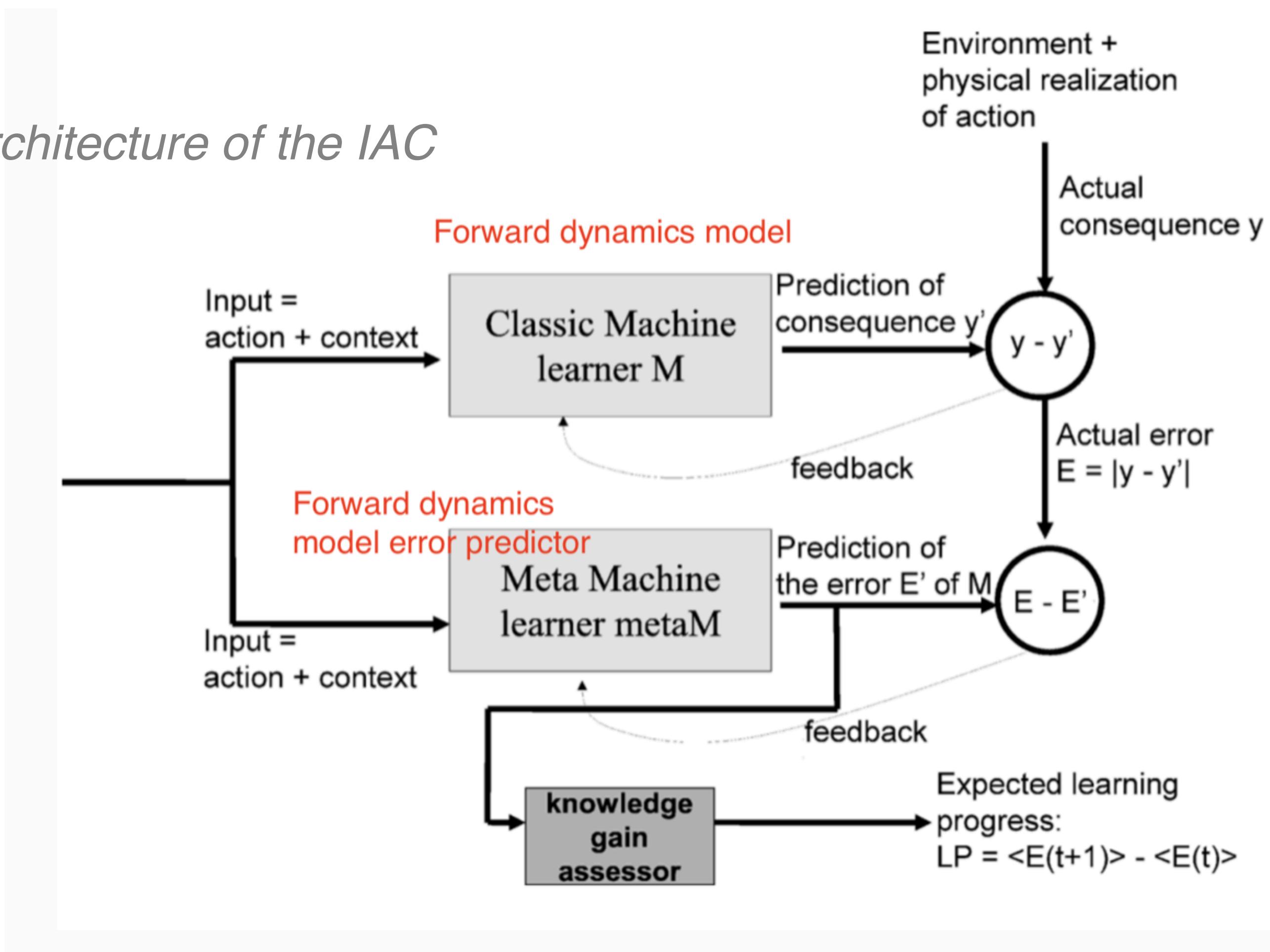
$r(t) = \langle e_n(t - \tau) \rangle - \langle e_n(t) \rangle$ - internal reward

$r(t) = \sum_i \alpha_i * r_i(t)$

Intelligent Adaptive Curiosity

Исследование на основе
предсказательной модели

Architecture of the IAC



Epsilon-greedy action selecting:

$$a = M(t)$$

$$Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma * \max_{a'}(Q(s', a')) - Q(s, a)]$$

$$r_t^i = \frac{1}{k} \sum_{i=0}^{k-1} (e_{t-i-\tau} - e_{t-i})$$

Случайные нейросети

Посмотрим, что будет, если задача прогнозирования вовсе не связана с динамикой окружающей среды. Оказывается, когда предсказание относится к случайной задаче, оно все же может помочь в исследовании

DORA The Explorer (Directed Outreaching Reinforcement Action-Selection)

$M(S, A, P, R, \gamma)$ - Марковская модель

S - множество состояний, A - множество действий

$P(s'|s, a)$ - вероятность перехода из состояния s в состояние s' посредством действия a . $R(r|s, a)$ - распределение , из которого сэмплируется награда за эти переходы

$\pi : S \rightarrow A$ - политика

Цель агента: найти оптимальную π^* максимизирующую $Q^\pi(s, \pi(s))$

$$Q(s_t, a_t) \leftarrow (1 - \alpha) Q(s_t, a_t) + \alpha \left(r + \gamma \max_a Q(s_{t+1}, a) \right)$$

DORA The Explorer (Directed Outreaching Reinforcement Action-Selection)

Обучаем две MDP. Одна - обычная MDP, которая используется для оценки value-функции. Вторая (M') - идентичная первой, но мы предполагаем, что награда за все состояния-действия равна 0. Будем использовать алгоритм RL, чтобы «изучить» «значения действий» в этом новом MDP, которые мы обозначим как E-значения.

$$M' = (S, A, P, 0, \gamma_E)$$

Инициализируем $\forall s, a \ E(s, a) = 1$

$$E^* = 0$$

Важное отличие E-value от Q-value в том, что оно on-policy

$$E(s_t, a_t) \leftarrow (1 - \alpha_E) E(s_t, a_t) + \alpha_E (r + \gamma_E E(s_{t+1}, a_{t+1}))$$

DORA The Explorer (Directed Outreaching Reinforcement Action-Selection)

$$r(t) = \frac{1}{\sqrt{-\log(E(s_t, a_t))}} - \text{exploration bonus} \quad f - \text{epsilon-greedy or softmax}$$

Input: Stochastic action-selection rule f , learning rate α , Exploration discount factor γ_E
 initialize $Q(s, a) = 0$, $E(s, a) = 1$;

foreach *episode* **do**

- | init s ;
- | **while** *not terminated* **do**
- | | Choose $a = \arg \max_x \log f_Q(x|s) - \log \log_{1-\alpha} E(s, x)$;
- | | Observe transitions (s, a, r, s', a') ;
- | | $Q(s, a) \leftarrow (1 - \alpha) Q(s, a) + \alpha(r + \gamma \max_x Q(s', x))$;
- | | $E(s, a) \leftarrow (1 - \alpha) E(s, a) + \alpha \gamma_E E(s', a')$;
- | **end**

end

Algorithm 1: DORA algorithm using *LLL* determinization for stochastic policy f

Физические свойства

В отличие от игровых симуляторов, роботам необходимо взаимодействовать с физической средой и наблюдать последствия этих взаимодействий.

Пример: ответы на вопросы через взаимодействия

- Какой из предметов тяжелее?
- Сколько кубиков в башенке?



LEARNING TO PERFORM PHYSICS EXPERIMENTS VIA DEEP REINFORCEMENT LEARNING

**Агент - LSTM со 100
скрытыми слоями**

Трехэтапная структура среды:

- Взаимодействие:** фаза исследования, на которой агент может свободно взаимодействовать с окружающей средой и собирать информацию.
- Лейблинг:** агент отвечает на вопрос среды
- Награда:** среда положительно награждает агента за правильный ответ и отрицательно за неправильный

- Вопрос не зависит от конкретной стратегии сбора информации.
- Точность ответа – частота, с которой агент отвечает на вопрос правильно.
- В рамках одного эпизода (вопроса) агент может делать несколько действий - проводить исследование. Среда меняется на разных эпизодах.
- В рамках исследования в каждый момент времени агент может выбрать одно из 8 действий (4 поднятие одного кубика и 4 лейблинга)
- **Пример:** Задача «Какой из объектов тяжелее?» сводится к задаче многорукого бандита, где самый тяжёлый объект - лучшая рука

Список литературы

- The Noisy-TV Problem <https://arxiv.org/abs/1810.12894>
- Intelligent Adaptive Curiosity [http://citeseerx.ist.psu.edu/viewdoc/
download?doi=10.1.1.177.7661&rep=rep1&type=pdf](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.177.7661&rep=rep1&type=pdf)
- DORA the Explorer <https://arxiv.org/abs/1804.04012>
- Learning to Perform Physics Experiments via Deep Reinforcement Learning
<https://arxiv.org/abs/1611.01843>