Методы обучения с подкреплением: Возможности для применения в адаптивных стратегиях активного контроля

О. Ю. Богоявленская

Дополнительные материалы

- Tiapkin, D., Belomestny, D., Moulines, E., Naumov, A., Samsonov, S., Tang, Y., Valko, M. & Menard, P.. (2022). **From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses.** Proceedings of the 39th International Conference on Machine Learning, in Proceedings of Machine Learning Research 162:21380-21431 Available from https://proceedings.mlr.press/v162/tiapkin22a.html. https://proceedings.mlr.press/v162/tiapkin22a/tiapkin22a.pdf
- **RL: Q-обучение https://qudata.com/ml/ru/RL_Q_Learning.html**

Обучение с подкреплением

- Эксплуатация знаний vs. приобретение знаний
- MDP : агент, среда, состояния, переходы, решения (активный контроль), награды.
- Оптимальная стратегия, сожаление
- Оптимизм: бонусы vs. случайный шум и апостериорные оценки.

Обучение с подкреплением

Learning problem The agent, to which the transitions are *unknown* (the rewards are assumed to be known for simplicity), interacts with the environment during T episodes of length H, with a *fixed* initial state s_1.[5] Before each episode t the agent select a policy π^t based only on the past observed transitions up to episode $t-1$. At each step $h \in [H]$ in episode t, the agent observes a state $s_h^t \in \mathcal{S}$, takes an action $\pi_h^t(s_h^t) = a_h^t \in \mathcal{A}$ and makes a transition to a new state s_{h+1}^t according to the probability distribution $p_h(s_h^t, a_h^t)$ and receives a deterministic reward $r_h(s_h^t, a_h^t)$.

Обучение с подкреплением

value functions, denoted by $V_h^\star$ are given by the Bellman respectively optimal Bellman equations

$$Q_h^\pi(s,a) = r_h(s,a) + p_h V_{h+1}^\pi(s,a) \quad V_h^\pi(s) = \pi_h Q_h^\pi(s)$$
$$Q_h^\star(s,a) = r_h(s,a) + p_h V_{h+1}^\star(s,a) \quad V_h^\star(s) = \max_a Q_h^\star(s,a)$$

where by definition, $V_{H+1}^\star \triangleq V_{H+1}^\pi \triangleq 0$. Furthermore, $p_h f(s,a) \triangleq \mathbb{E}_{s' \sim p_h(\cdot|s,a)}[f(s')]$ denotes the expectation operator with respect to the transition probabilities p_h and $\pi_h g(s) \triangleq g(s, \pi_h(s))$ denotes the composition with the policy π at step h.

Обучение с подкреплением

Размер

Regret The quality of an agent is measured through its regret, that is the difference between what it could obtain (in expectation) by acting optimally and what it really gets,

$$\mathfrak{R}^T \triangleq \sum_{t=1}^{T} V_1^{\star}(s_1) - V_1^{\pi^t}(s_1)\,.$$

Counts $n_h^t(s,a) \triangleq \sum_{i=1}^{t} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\}$ are the number of times the state action-pair (s,a) was visited in step h in the first t episodes. Next, we define $n_h^t(s'|s,a) \triangleq \sum_{i=1}^{t} \mathbb{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}$ the number of transitions from s to s' at step h.

Пример

Пример: Frozen Lake

- Рассмотрим в качестве простого примера среду Frozen Lake из набора сред **OpenAI Gym**. Есть квадратная неизменная карта замёршего озера, часть ячеек которого содержит проруби. Необходимо провести гнома из левого верхнего угла карты (стартовое состояние) в правый нижний (целевое состояние) не провалившись в прорубь. Проблема в том, что гном не вполне трезвый и его шатает вправо-влево. Поэтому на скользком льду (в режиме **is_slippery=False** по умолчанию) он с вероятностью **1/3** сместиться туда, куда планирует, а с вероятностями **1/3** сдвинется в одно из перпендикулярных направлениях. При попытке выйти за карту гном останется на прежней ячеке (или его шатнёт в одно из перпендикулярных направлений).

Пример: Frozen Lake

- Рассмотрим в качестве простого примера среду Frozen Lake из набора сред **OpenAI Gym**. Есть квадратная неизменная карта замёршего озера, часть ячеек которого содержит проруби. Необходимо провести гнома из левого верхнего угла карты (стартовое состояние) в правый нижний (целевое состояние) не провалившись в прорубь. Проблема в том, что гном не вполне трезвый и его шатает вправо-влево. Поэтому на скользком льду (в режиме **is_slippery=False** по умолчанию) он с вероятностью **1/3** сместиться туда, куда планирует, а с вероятностями **1/3** сдвинется в одно из перпендикулярных направлениях. При попытке выйти за карту гном останется на прежней ячеке (или его шатнёт в одно из перпендикулярных направлений).

Пример

Environment For the tabular experiments we consider a simple grid-world with 5 connected rooms of size 5×5, totalling $S = 129$ states. The agent starts in the middle room. There is one small deterministic reward in the left-most room, one large deterministic reward in the rightmost room and zero reward elsewhere. The agent can take $A = 4$ actions: moving up, down, left, right. When taking an action, the agent moves in the corresponding direction with probability 0.9 and moves to a neighboring state at random with probability 0.1. The horizon is fixed to $H = 30$; see Appendix G for details. In this environment the agent must explore efficiently all the room avoiding being lured by the small reward in the leftmost room.