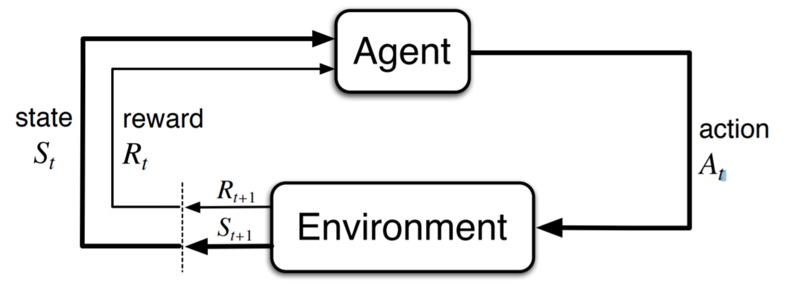
Применение обучения с подкреплением в экономическом моделировании

Всеволод Даниелян

Обучение с подкреплением и MDP

Обучение с подкреплением — разновидность машинного обучения, рассматривает агентов, обучающихся максимизировать свои совокупные вознаграждения, взаимодействуя со средой. Позволило достичь успехов при решении многих задач — от игры в старкрафт до робототехники. Задача обычно формулируется в виде марковского процесса принятия решений (MDP):



Агент наблюдает текущее состояние s_t, выбирает действие a_t, и переходит в следующее состояние s_{t+1}, которое определяется вероятностями перехода $P(s_{t+1}|a_t, s_t)$. При переходе агент также получает вознаграждение $r_t(s_t, a_t, s_{t+1})$. Также подразумевается выполнение Марковского свойства:

$$P(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \ldots) = P(s_{t+1}|s_t, a_t), \tag{1}$$

которое гарантирует, что вся информация, необходимая для принятия решения, содержится в текущем состоянии S_t. Действия выбираются в соответствии со стратегией агента: $a_t \sim \pi(\cdot|s_t)$

Траектория τ — это последовательность состояний и действий: $\tau = (s_0, a_0, s_1, a_1, \ldots)$. Дисконтированная сумма вознаграждении при следовании по траектории τ:

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t \tag{2}$$

Уравнения Беллмана и Q-learning

Полезность текущего состояния и действия при следовании стратегии π:

$$Q^{\pi}(s, a) = \mathop{\mathbb{E}}_{\tau \sim \pi} [R(\tau)|s_0 = s, a_0 = a], \tag{3}$$

Задача - выучить стратегию, максимизирующую математическое ожидание будущих вознаграждений:

$$\pi^* = \arg\max_{\pi} \mathop{\mathbb{E}}_{\tau \sim \pi} [R(\tau)] = \arg\max_{\pi} \mathop{\mathbb{E}}_{s \sim \rho^{\pi}} [Q(s, a)] \tag{4}$$

Уравнения Беллмана для оптимальной стратегии, связывает Q-функции текущего и будущего периода:

$$Q^*(s, a) = \mathbb{E}\left[r(s, a) + \gamma \max_{a'} Q^*(s', a')|s, a\right]. \tag{5}$$

Мы можем аппроксимировать Q-функцию нейросетью, и обучить ее с помощью минимизации следующей функции потерь, основанной на уравнениях Беллмана:

$$L(\theta) = \mathop{\mathbb{E}}_{s,a,r,s'} \left[(Q^*_{\theta}(s, a) - y)^2\right] \tag{6}$$

$$y = r(s, a) + \gamma \max_{a'} Q^*_{\theta}(s', a'), \tag{7}$$

где θ - параметры аппроксимирующей функции (веса нейросети). Обучаем градиентным спуском.

DDPG

Пусть у агента детерминистическая стратегия μ, тогда:

$$a = \mu(s) \tag{8}$$

Если аппроксимируем ее нейросетью с параметрами θ, то можем взять градиент относительно этих параметров:

$$\nabla_{\theta} \mathop{\mathbb{E}}_{s \sim \rho^{\mu}} [Q(s, a)] = \mathop{\mathbb{E}}_{s \sim \rho^{\mu}} [\nabla_a Q(s, a) \nabla_{\theta} \mu(s)] \tag{9}$$

Дальше можем использовать градиентное восхождение. Как учить Q-функцию уже знаем.

Общая идея алгоритма - выбираем действия с помощью актора μ, сохраняем пройденные траектории в буфер, берем оттуда случайную выборку, обновляем веса критика Q. Далее с помощью критика обновляем веса актора, используя ту же выборку.

Когда много агентов — MADDPG

Пусть теперь вместо одного агента у нас их N. Можем применить наивный подход - просто обучать их независимо друг от друга используя, например, DDPG. Плохо тем, что для каждого агента окружение перестанет быть стационарным и марковское свойство выполняться не будет. С точки зрения агента, если другие агенты меняют стратегии - меняются вероятности перехода, а история начинает играть роль влияя на стратегии. Выход - для каждого агента учить централизованного критика, учитывающего действия других агентов: $Q_i^{\mu_i}(s_1, \ldots, s_N, a_1, \ldots, a_N)$. В остальном алгоритм работает как DDPG. Для обучения нужно знать только прошлые действия и состояния других агентов, в остальном обучение независимо.

Экономическая модель

Я применил MADDPG к динамической дуополии Курно. В этой модели два агента конкурируют по выпуску, выбирая его одновременно в каждом периоде. Выпуск дискретизирован и выбирается из конечного множества возможных выпусков $\{0, 10, \ldots, 1000\}$. Цена в каждом периоде определяется функцией спроса:

$$P_t(q_t^1, q_t^2) = 500 - \frac{q_t^1 + q_t^2}{2}, \tag{10}$$

где q_t^1, q_t^2 - выпуски соотвествующих агентов в период t. Агенты воспринимают цену, как непрерывную величину, и поэтому в случае необходимости функцию спроса можно сделать стохастической или подчиняющейся определенной динамике. В каждом периоде агенты получают прибыль:

$$r_t^i(q_t^i) = P_t q_t^i - 50 q_t^i \tag{11}$$

В начале каждого периода агенты наблюдают цену и прибыль за предыдущий период. Цель каждого агента - максимизировать свою совокупную дисконтированную прибыль. Промежуток времени не был ограничен сверху, но для удобства обучения в действительности использовались эпизоды длиной в 100 периодов. Агентам во время обучения были известны выпуски других агентов в прошлых периодах и эпизодах.

Результаты

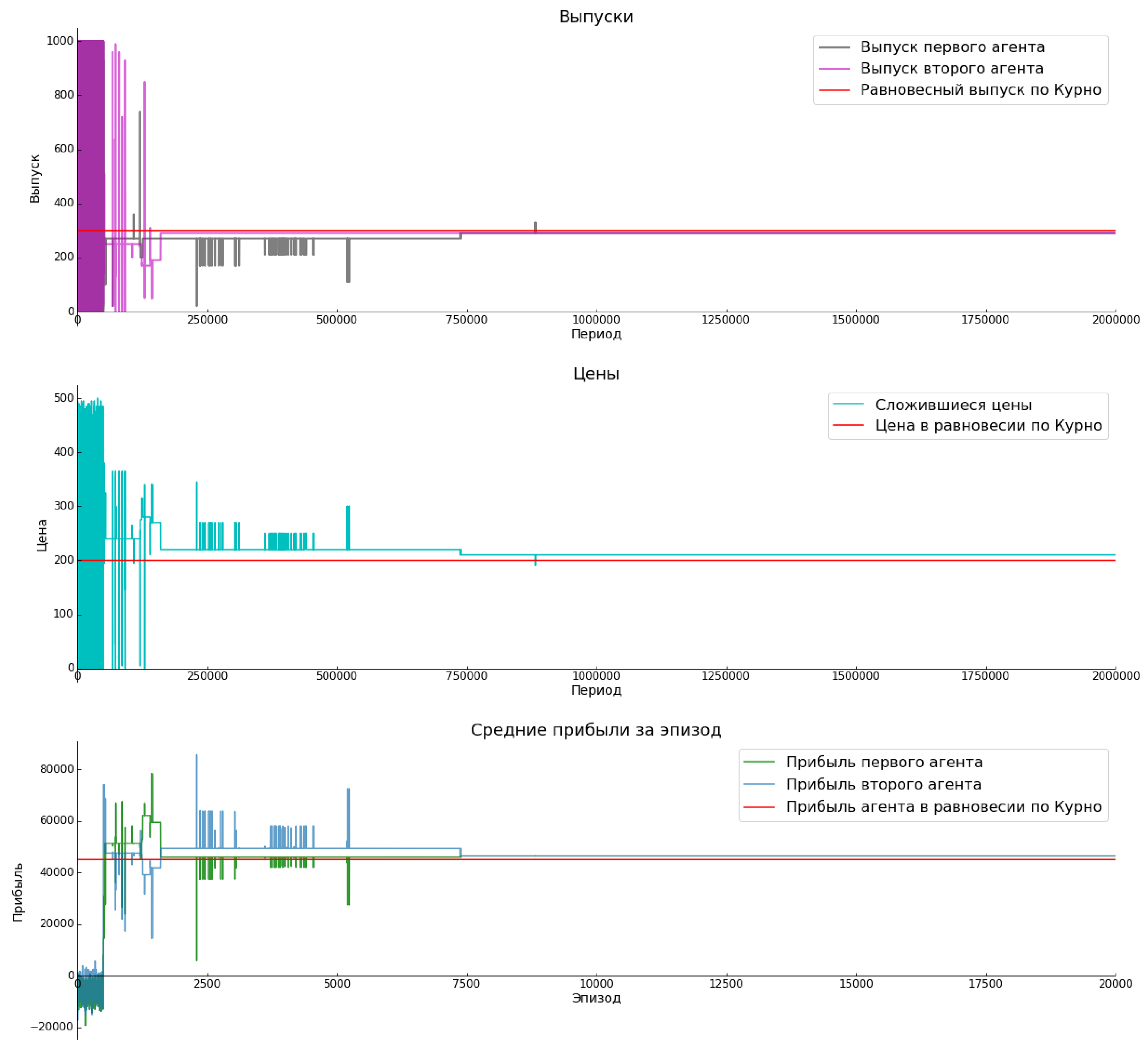


Fig. 2: $\gamma = 0.3$

γ	Цена	Прибыль 1-го агента	Прибыль 2-го агента
0	200	45000	45000

Tab. 1: Статическое равновесие Курно

γ	Цена	Прибыль 1-го агента	Прибыль 2-го агента
0	245	42900	56550
0.3	210	46400	46400
0.6	210	44800	48000
0.9	290	50400	50400

Tab. 2: Характеристики равновесия при различных γ

Агенты в алгоритме MADDPG действуют скорее рефлекторно, чем сознательно. Знание о модели и стратегиях других агентов используется только имплицитно. Само решение принимается, опираясь только на информацию о состоянии, в нашем случае цене. Такой подход больше подошел бы для моделирования действий индивидов в обстоятельствах, требующих быстрого принятия решения, чем для моделирования действий экономических агентов уровня фирм.