Г.В. Чернов
Международная лаборатория экспериментальной и поведенческой экономики, НИУ ВШЭ; ФГБУН Институт психологии РАН, Москва

И.С. Сусин
Международная лаборатория экспериментальной и поведенческой экономики, НИУ ВШЭ, Москва

Модели обучения в играх: обзор[1]

Аннотация. В данном обзоре анализируются центральные идеи и современное состояние экономической теории обучения в играх. В рамках теории игр обучение можно рассматривать и как альтернативу равновесиям, и как способ, позволяющий разобраться с их природой. Вне этих рамок теория обучения в играх показывает экономический анализ (включая классический сюжет олигополии по Курно) в новом свете, предоставляет интересные теоретические задачи, нетривиальна с эконометрической точки зрения и может изучаться экспериментальными методами. Теория обучения в играх связывает экономику с другими (подчас неожиданными) научными дисциплинами: биологией, философией рациональности и компьютерными науками. Однако существующие обзорные работы не рассчитаны на русскоязычных читателей и читателей без математического образования. Данный обзор направлен на заполнение этих пробелов и может служить введением и кратким справочником по теории обучения в играх. В статье анализируются не только модели, но и подходы к их классификации. Теоретические выкладки иллюстрируются конкретными примерами теоретико-игровых ситуаций. Помимо теоретических моделей особое внимание уделено проблемам их эмпирической и экспериментальной проверки, предлагаются основания новой исследовательской программы и обосновываются предположения о перспективах их будущего развития и роли в экономической теории.

Ключевые слова: *обучение с подкреплением, фиктивная игра, рациональное обучение, ограниченная рациональность, алгоритмы обучения.*

Классификация JEL: C70, D84.

DOI: 10.31737/2221-2264-2019-44-4-3

1. Введение

Данный обзор посвящен одному из самых значимых подходов к анализу экономического поведения в теории игр — моделям обучения. Эти модели не получили до сих пор должного освещения в русскоязычной литературе, несмотря на то что зародились почти в то же время, что и теория игр. Ранние формальные модели относятся к 1950-м годам (например, (Brown, 1951; Bush, Mosteller, 1955)), а неформальный анализ можно увидеть уже у Огюстена Курно в 1838 г. (Cournot, 1960).

Модели обучения и сегодня активно развиваются как самостоятельная область, ценная для теории игр в целом. Во-первых, в наблюдениях или экспериментах по любой динамической (в частности, повторяющейся) игре поведение участников не обязательно сразу ока-

[1] Авторы выражают глубокую признательность А.В. Белянину за консультации и помощь в подготовке текста, а также полезную научную дискуссию. Авторы также хотели бы поблагодарить Й. Оксслера, С.Г. Коковина А.Д. Суворова, а так же рецензента за ценные комментарии. Статья подготовлена в результате проведения работы в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ) и с использованием средств субсидии в рамках государственной поддержки ведущих университетов Российской Федерации «5-100», а также при поддержке Российского научного фонда (проект 17-78-30035).

Журнал НЭА,
№4 (44), 2019,
с. 77–125

зывается равновесным, что требует применения специальных моделей. Во-вторых, с теоретической точки зрения равновесий в играх может быть много, что ставит естественную задачу выбора наилучших (Van Damme, 1991). В частности, критерием сравнения равновесий может быть вероятность того, что игроки сойдутся именно к нему из случайного начального состояния. Характеристика таких равновесий требует понимания динамики взаимодействия участников и правил, по которым они выбирают свои решения. Все это требует явной модели обучения и динамики игры, позволяющей предсказывать исход социального взаимодействия.

В современной теории некооперативных игр равновесный анализ заключается в поиске неподвижной точки функций наилучшего отклика (равновесия Нэша), а в случае их множественности — неподвижных точек, которые обладают особыми желательными свойствами[2]. И хотя для экономической теории существование, единственность и свойства равновесия в конкретной экономической модели могут представлять немалые технические трудности, часто неявно предполагается, что для агентов в реальной экономической ситуации это равновесие не только очевидно, но и является общим знанием. Наивное обоснование этого неявного предположения состоит в том, что если равновесие единственно, то рано или поздно игроки придут именно к нему, а если равновесий много — то к лучшим из этих равновесий (почему и необходимо рафинирование). Моделирование обучения в играх создает возможность протестировать этот тезис и предоставляет альтернативу аксиоматическому поиску рафинированных равновесий. Именно поэтому значимая часть литературы об обучении в играх посвящена анализу сходимости различных правил обучения в играх к равновесному исходу и сравнению эффективности разных моделей (алгоритмов обучения) в части этих предсказаний.

Другая причина моделирования процесса обучения — ограниченность вычислительных возможностей реальных игроков. Этими ограничениями нельзя пренебречь даже для игр с полной информацией (например, такой как шахматы) и для игроков, способных перебрать значительное число продолжений любой игровой позиции (таких как современные суперкомпьютеры). Одним из очевидных возможных решений, по которому пошло большинство исследователей, является наблюдение за поведением игроков-людей, выделение основных закономерностей игры и отображение этого наблюдения в эвристические и субоптимальные критерии и, наконец, перенос последних в формальные правила обучения. В некотором смысле такой подход нарушает предположения о рациональности игрока, однако взамен при минимальных допущениях позволяет с некоторым приближением имитировать наблюдаемое поведение.

Не в последнюю очередь интерес к таким моделям и связан с эмпирическими (прежде всего экспериментальными) исследовани-

[2] Это называется «рафинирование» (усиление, refinement) равновесий (Van Damme, 1991).

Журнал НЭА,
№4 (44), 2019,
с. 77–125

ями. Они убедительно показывают, что в ряде случаев реальные люди ведут себя не так, как предсказывают классические решения, такие как равновесие Нэша. Так, в экспериментах с играми «Конкурс красоты» (Nagel, 1995) и «Ультиматум» (Güth et al., 1982) игроки регулярно и существенно отклоняются от единственного равновесия даже в тех случаях, когда игра повторяется достаточно большое число раз и стимулы в игре составляют значимую часть их доходов.

Модели индивидуального обучения характеризуются способностью игроков обновлять свое поведение от раунда к раунду, поскольку в каждом раунде необходимо учитывать стратегическую неопределенность и новую информацию, помогающую ее разрешить. В терминах (Brandenburger, 1996), в игре присутствует не только *структурная неопределенность* (относительно выигрышей других игроков), но и *стратегическая неопределенность* (относительно их действий и вер). Игроки могут наблюдать платежи своих соперников с некоторым шумом (Foster, Young, 2001) или не наблюдать вовсе. Даже в играх с полной информацией реальным участникам могут быть не известны и *отношение* оппонентов к материальным платежам (выигрышам в эксперименте), и их *веры* (убеждения, beliefs) относительно намерений оппонентов. Поэтому поведение в играх с обучением может быть довольно сложным — игроки не только учатся реагировать друг на друга по отдельности, они решают эту задачу одновременно и реакция одного игрока на обучение другого влияет на обучение. Микроэкономические модели индивидуального обучения — не единственный подход к таким задачам, поэтому очертим рамки данного обзора и уточним, что остается за его рамками.

Равновесная динамика. Необходимо отделять модели с оптимальной динамической оптимизацией (восходящей к работам Беллмана (Bellman, 1957)) или динамическим равновесием от моделей обучения. Модель Курно, которую мы обсуждаем как модель обучения, может быть задана и как равновесная. Тогда предполагается, что адаптация происходит только в процессе принятия решения, когда фирмы оценивают последовательность наилучших откликов друг на друга. Но в равновесной модели в первом же периоде ожидается, что будет выбран равновесный профиль стратегий, несмотря на мотив обучения. Поэтому мы разделяем мотивацию равновесных моделей неравновесными соображениями и неравновесные модели, фокусируясь на последних. Упомянем, что значительное число исследований посвящены тому, какие ограничения на правила обучения и реакцию игроков устанавливают условия, при которых наилучшие стратегии сходятся к равновесию Нэша (Kalai, Lehrer, 1993), коррелированному равновесию (Aumann, 1987; Foster, Vohra, 1997), самоподтверждающемуся равновесию (Fudenberg, Levine, 1993), множеству, устойчивому относительно наилучшего ответа (Basu, Weibull, 1991). Связи моделей обучения с равновесными концепциями теории игр, во-первых, предлагают обосно-

Журнал НЭА,
№4 (44), 2019,
с. 77–125

вание для равновесных решений, а во-вторых, позволяют сравнивать эффективность разных решений и моделей (алгоритмов) обучения, в том числе в свете экспериментальных результатов.

Игры в развернутой форме. В статье мы будем рассматривать повторяющиеся игры в нормальной форме и опустим игры в развернутой форме. На это есть две причины:

1) хороший обзор результатов для повторяющихся игр в развернутой форме сделан в (Fudenberg, Levine, 2009) (многие из этих результатов были получены Д. Фуденбергом);
2) хотя анализ таких игр более технически изощрен и предоставляет некоторые любопытные результаты (среди них возможность самоподтверждающихся равновесий, self-confirming equilibria), эта область стоит несколько особняком и слабо влияет на магистральную линию теоретических и эмпирических исследований (для более эволюционного подхода к таким играм можно порекомендовать (Cressman, 2003)).

Макроэкономические ожидания. В макроэкономике обучение и формирование ожиданий агентов анализируются на агрегатном уровне и обычно подразумевают равновесные ожидания или адаптивные модели, похожие на рассматриваемые нами (обзоры этой литературы имеются в работах (Evans, Honkapohja, 2001)).

Эволюционные модели. Приводимые в данном обзоре модели обучения отличаются не только от моделирования равновесного поведения (т.е. возможных итогов такого процесса, а не самого процесса поиска равновесия), но и от эволюционных моделей, рассматривающих дорациональные (например, биологические) механизмы обучения типа репликаторной динамики (Taylor, Jonker, 1978; Hofbauer, Sigmund, 1998) (где принятие рациональных решений достигается на уровне популяции (подробнее см. (Sandholm, 2010))).

Глубокое обучение. Модели алгоритмического обучения активно развиваются в компьютерных науках (Cesa-Bianchi, Lugosi, 2006; Sutton, Barto, 2018), где, в частности, особое внимание уделяется так называемому глубокому обучению с подкреплением (deep reinforcement learning). Примером успешной работы таких алгоритмов могут служить компьютерные алгоритмы AlphaGo и AlphaZero (Silver et al., 2018), достигшие свехчеловеческого уровня игры в нескольких классических играх с полной информацией (го и шахматы).

Приведем другие обзоры, посвященные обучению в играх. Прежде всего необходимо отметить работы (Fudenberg, Levine, 1998a; Young, 2004). Первая считается классической, хотя уже несколько устарела, вторая, выросшая из серии лекций, относительно легко читается и освещает предметную область.

Философский анализ рациональности и обучения (в том числе в играх) дан в (Hutteger, 2017), а характерный для компьютерных наук и предсказательной статистики — в (Cesa-Bianchi, Lugosi, 2006).

Журнал НЭА,
№4 (44), 2019,
с. 77–125

Следует упомянуть работы (Marimon, 1997; Опойцев, 1977; Fudenberg, Levine, 1998b; Fudenberg, Levine, 2009, 2016). В (Erev, Haruvy, 2013) изложен взгляд на данную теорию, близкий экспериментальной и поведенческой экономике. Нельзя не отметить и статью (Nachbar, 1990), которая наиболее емко описывает основные результаты данной области.

Обзор состоит из следующих разделов. Классификация последующих моделей обучения и мотивировка последующего деления даны в разд. 2; разд. 3 посвящен классу рациональных, а разд. 4 — ограниченно рациональных моделей. В разд. 5 мы суммируем роль экспериментов для моделей обучения. В разд. 6 делаются выводы по рассмотренным моделям. В приложении приводится развернутый пример одной из важнейших моделей рационального обучения.

2. Классификации моделей обучения

Поскольку в литературе не сложилось единого подхода к классификации различных алгоритмов обучения (Marimon, 1997; Nachbar, 2009; Fudenberg, Levine, 2009, 2016), мы рассматриваем классификацию как дополнительный способ глубже понять полученные в литературе результаты.

Первое, важное для понимания данной области, деление — на детерминированные и стохастические алгоритмы. В обоих случаях ответы на действия оппонента наилучшие, но детерминированный игрок всегда играет наилучший на текущий момент отклик. В стохастическом алгоритме это не обязательно; его выбор вероятностный, но наилучший отклик играется с повышенной вероятностью. Довольно поверхностной модификацией функции отклика можно преобразовать детерминированный алгоритм в стохастический или обратно, поэтому может показаться, что и данная классификация не важна, однако наличие случайности в поведении значимо для сходимости к равновесию.

Так же важны различия между алгоритмами, которые используют всю доступную информацию и отвечают на нее наилучшим в текущий момент образом (*рациональными*) и теми, которые в силу внешних или внутренних ограничений и особенностей этого не делают (*ограниченно рациональными*). Соответственно, обычно подразумевается, что рациональные алгоритмы детерминированы (хотя, как мы уже заметили, возможны модификации). При определенных допущениях примерами рационального обучения могут служить модели «многорукого бандита», байесовское обучение и «фиктивная игра» с неограниченной памятью, о которых речь пойдет далее.

Что касается ограниченной рациональности, этот класс моделей разбивается на подклассы по типу наиболее важных ограничений. Особую роль играют ограничения на вычислительные ресурсы игрока и память, а также неопределенность в целях или намерениях

Журнал НЭА,
№4 (44), 2019,
с. 77–125

оппонента. Вычислительные ограничения очевидны в шахматах: реальные игроки не способны просчитать все возможные варианты и достоверно найти оптимальный ход в каждой возможной ситуации в игре. Для модели такого поведения важно не только общее число возможных сравниваемых ходов, но и эвристики, позволяющие вычислять ветвление последствий только самых перспективных ходов и, тем самым, уменьшить объем вычислений. Мы иллюстрируем работу таких моделей на примере стандартной модели обучения с подкреплением.

Ограничения памяти ввести проще всего: для этого достаточно предположить, что игроки забывают то, что произошло достаточно давно, и выстраивают стратегию только исходя из последних нескольких раундов. Это предположение можно обосновать тем, что в нестационарной среде опыт далекого прошлого не так важен (поскольку среда могла успеть поменяться) и учитывать его не имеет особого смысла. Например, можно интерпретировать классическую для психологической науки кривую забывания Эббингауза (Ebbinghaus, 2013) как адаптивный механизм, дисконтирующий редкий случайный опыт и закрепляющий часто повторяющийся. Примером таких моделей может служить фиктивная игра с ограниченной памятью и модель наилучшего отклика Курно.

Неопределенность целей оппонента означает, например, что игрок не знает матрицу игры (ценит ли оппонент исход А выше или ниже исхода Б). Она схожа с неизвестностью реакции оппонента на действия данного игрока, когда игрок не знает, как будущие действия оппонента будут зависеть от поведения игрока в текущем раунде (например, будет ли оппонент отвечать на кооперацию кооперацией). Выделим две категории: 1) выбор действий оппонента для принятия решений случаен, т.е. не учитывает выбора самого игрока, принимающего решения; 2) игрок влияет на то, какую информацию об оппоненте и своих будущих действиях он получает. К последней категории относятся обучение с взвешенным по опыту притяжением (Experimental Weighted Attraction — EWA) и модели I-SAW, вводящие в модель EWA инерционность и экспериментирование.

После краткого описания основных обозначений мы излагаем модели в следующем порядке: начинаем с рациональных моделей против пассивного оппонента, потом рассматриваем последовательно все более сложные модели игры против активного оппонента, завершая классическим байесовским обучением и его достоинствами и недостатками. Затем переходим к ограниченно-рациональным правилам — калибровке, направленному обучению, обучению с подкреплением, обучению с взвешенным по опыту притяжением (EWA) и I-SAW. Оба раздела начинаются со статистического, не вполне теоретико-игрового взгляда на задачу, что помогает подчеркнуть специфику обучения именно в играх (табл. 1).

Журнал НЭА,
№4 (44), 2019,
с. 77–125

Таблица 1

Навигатор по классификации моделей

Правило обучения	Принцип отклика	Веры/ подкрепление	Параметры модели	Оригинальная формулировка модели
Динамика Курно, пп. 3.3.1	Детерминистический, реакция на предшествующий ход	Основанные на верах об оппоненте (Belief-Based)	–	Cournot A., 1838 г.
Фиктивная игра, пп. 3.3.2	Детерминистический, реакция на эмпирические частоты	Основанные на верах об оппоненте (Belief-Based)	Параметр ослабевания памяти	Brown G.W., 1951 г.
Байесовское обучение, п.3.5	Детерминистический, реакция на эмпирические частоты	Основанные на верах об оппоненте (Belief-Based)	Зависит от спецификации модели	Ramsey F., 1926 г.
Модель многорукого бандита, п. 3.2	Детерминистический, реакция на эмпирические частоты	Основанные на верах об оппоненте (Belief-Based)	Коэффициент дисконтирования, полезность от экспериментирования	Robbins Дж., 1952 г.
Калибровка, п. 4.1	Стохастический, реакция на эмпирические частоты	Основанные на подкрепление	–	Foster D.P., Vohra R., 1998 г.
Направленное обучение, п. 4.2	Стохастический, реакция на вознаграждение	И веры, и подкрепление	Зависит от спецификации модели	Selten R., Stoecker R., 1986 г.
Обучение с подкреплением, п. 4.3	Стохастический, реакция на вознаграждение	Основанные на подкрепление	Параметр ослабевания памяти, параметр отсечения, параметр локального экспериментирования	Roth A.E., Erev I., 1995 г.
Обучение с взвешенным по опыту притяжением (EWA), п. 4.4	Стохастический, реакция на вознаграждение	И веры, и подкрепление	Дисконтирование, сила предыдущего опыта, вес гипотетического выигрыша, чувствительность к притяжению, “форма” предыдущего притяжения	Camerer C., Ho H.T., 1999 г.
I-SAW (инерционность и случайность в EWA), п. 4.5	Стохастический, реакция на вознаграждение	И веры, и подкрепление	Склонность к инерции, длинна памяти, склонность к исследованию	Nevo I., Erev I., 2012 г.

Журнал НЭА,
№4 (44), 2019,
с. 77–125

3. Основные модели рационального обучения в играх

3.1. Определения и общие обозначения

Два игрока[3] играют в повторяющуюся игру, где каждое повторение (этап, раунд, период) — конечная статическая игра в нормальной форме $G = \langle \mathcal{I}, S, \{u\}, T \rangle$, $\mathcal{I} = \{1,2\}$ — множество игроков (числом $I = |\mathcal{I}|$), $S = S_1 \times S_2$ — профили стратегий игроков образуются декартовым произведением конечных множеств их чистых стратегий $S_i = \{s_{i1}, \ldots, s_{ij}\}$, $i = \{1,2\}$. Далее, $\{u\} \equiv \{u_1(S), u_2(S)\}$ — функции платежей для каждого игрока, определенных на S, а $T \leq \infty$ — число периодов (раундов) конечной или бесконечной повторяющейся игры, с типичным периодом t. Смешанная стратегия σ_i каждого игрока i — это распределение вероятностей на множестве S_i, она приписывает каждому игроку i вероятность $\sigma_i(s_{ij})$ сыграть чистую стратегию $s_{ij} \in S_i$. Множество всех смешанных стратегий игрока i образует $J-1$ мерный симплекс с типичным элементом σ_i. Множество смешанных стратегий каждого игрока включает множество чистых стратегий и определяет частоту, с которой выбирается та или иная чистая стратегия $\sigma_i^t = \{p_{i1}^t(s_{i1}), \ldots, p_{ij}^t(s_{ij})\}$, при том что $\sum_{j=1}^{J} p_{ij}^t = 1$, $p_{ij} \in \sigma_i$. Множество смешанных стратегий каждого игрока естественным образом содержит множество чистых стратегий при $\sigma_i(s_{ik}) = 1, s_{ik} \in S_i$. Для одноперiодной игры ожидаемый выигрыш игрока i определен как

$$U_i(\sigma) = \mathbb{E}_\sigma(u_i(\sigma)) = \sum_{s_i \in S_i} u_i(s_1, \ldots s_{\mathcal{I}}) \prod_{i=1}^{\mathcal{I}} \sigma_1(s_2)\sigma_1(s_2),$$

где стратегии игроков предполагаются независимыми. Аналогично выигрыш определяется для игрока 2.

Последовательность стратегий, избранных игроками в динамической игре, называется историей игры на момент t и обозначается h^t, где $h^t = \{s^1, \ldots, s^t\}$ и $s^t = \{s_1^t, \ldots, s_I^t\}$. Отображение $\xi_i : h^t \to \Delta S_i$, определяющее, какую из чистых стратегий должен выбрать игрок i в ответ на наблюдаемую историю игры, называется поведенческой стратегией игрока i. Все последующие определения естественным образом обобщаются на профили поведенческих стратегий ξ.

Стратегия $\bar{\sigma}_i$ в одноперiодной игре называется наилучшим ответом относительно профиля σ_{-i}, если $U_i(\bar{\sigma}_i, \sigma_{-i}) = \max_{\sigma_i' \in \Delta S_i} U_i(\sigma_i', \sigma_{-i})$. Множество всех наилучших ответов игрока i относительно профиля стратегий σ обозначается $B_i(\sigma)$, а множество всех наилучших ответов в профиле σ есть $B(\sigma) = \prod_{i \in \mathcal{I}} B_i(\sigma)$. Если наилучшим ответом является смешанная стратегия, каждая чистая стратегия $\bar{\sigma}_i$, которой оно придает положительную вероятность, должна давать одинаковый ожидаемый доход против σ_{-i}. В противном случае уменьшение веса на менее выгодную из чистых стратегий дало бы большую ожидаемую полезность, и $\bar{\sigma}_i$ не принадлежала бы B.

Равновесием Нэша называется такой профиль смешанных стратегий σ^*, который является взаимным наилучшим ответом, т.е. $\sigma^* \subset B(\sigma)$. В равновесии ни одному из участников не выгодно в односто-

[3] Для простоты изложения мы ограничиваемся моделями с двумя игроками и одинаковым числом стратегий J, хотя в большинстве случаев они обобщаются на любое конечное число игроков и неодинаковое число стратегий.

Журнал НЭА,
№4 (44), 2019,
с. 77–125

роннем порядке изменить свои стратегии. Теорема Нэша гласит, что в любой конечной некооперативной игре такое равновесие (в чистых или смешаных стратегиях) обязательно существует.

3.2. Оптимальная игра против Природы: модель многорукого бандита

Для иллюстрации статистической задачи обучения и ее особенностей в играх начнем с вырожденного случая обучения в играх — обучения в игре против Природы. В теории игр ключевое отличие Природы от других игроков заключается в том, что она не реагирует на действия других, стационарна по стратегиям. «Одноруким бандитом» принято называть игровой автомат с ручкой и барабанами, при случайном совпадении значений которых игрок получает приз (поскольку ожидаемый выигрыш такой игры отрицателен, автомат грабит игрока, отсюда и название).

Представим, что игрок видит перед собой несколько игровых автоматов (поэтому бандит многорукий, иногда название сокращается до «проблемы бандита», bandit problems (Bergemann, Välimäki, 2008)), про которые известно, что ожидаемый выигрыш от каждой руки может быть не эквивалентен другим. Как построить последовательность выбора рук? Каждая из машин обещает независимый от остальных и постоянный по времени (т.е. стационарный) ожидаемый выигрыш, который неизвестен игроку ни для одного игрового автомата. Информация, которая игроку доступна, появляется в результате экспериментирования — когда игрок дергает за ручку, он наблюдает результат. Однако эта информация не достается бесплатно, иначе игрок будет неограниченно экспериментировать. Существует так называемый компромисс эксплуатации-исследования (exploitation-exploration trade-off) — важнейший компромис между тем, с какой частотой стоит выбирать руку с максимальным ожидаемым выигрышем из уже опробованных, и частотой выбора еще неиспользованных рук, т.е. получения новой информации. Решение выбрать другую руку сопряжено с риском, что она будет хуже, чем наилучшая из тех, про которые информация уже накоплена. Например, если есть десять рук и про девять рук уже собрана информация, то ожидаемый выигрыш на оставшейся руке априорно равен среднему (у нас нет причин априорно полагать ее лучше или хуже других), в то время как выигрыш лучшей из девяти исследованных рук распределен, как максимум девяти таких же независимых средних. Если рук достаточно много, то, попробовав только часть из них, мы рискуем никогда не попробовать наилучшую руку. Но перебор всех рук субоптимален, поскольку каждая новая рука (как ожидается) добавляет меньше выигрыша, чем наилучшая из уже использованных.

Если руки независимы и распределение исходов для каждой из них стационарно, то такой модели достаточно, чтобы определить наилучшее решение (оно найдено в (Gittins, 1979)). На каждом ходу рассчи-

Журнал НЭА,
№4 (44), 2019,
с. 77–125

тывается так называемый индекс Гиттинса[4]. Стратегия «выбирать руку с наивысшим текущим индексом» минимизирует будущее сожаление (формально определяемое как разница между выигрышем на наилучшей и использованной рукой) от неиспользования других рук, т.е. является в байесовском смысле достоверно лучшим решением.

Формально этот индекс можно определить следующим образом (изложение дается по работе (Bergemann, Välimäki, 2008)). Рассмотрим проблему принятия решений на бесконечном горизонте с дискретным временем $t = 0, 1, \ldots$. В каждый момент необходимо выбрать между K руками, обозначим этот выбор a_t. Результатом действия $a_t = k$ будет выигрыш x_t^k. Он случаен и является реализацией случайной величины X_t^k. Последовательность выборов может менять состояние системы, обозначим это состояние через s_t. Тогда распределение величины X_t^k представляется как $F(\cdot; s_t)$, при этом независимость рук означает, что $F^k(\cdot; s_t) = F^k(\cdot; s_t^k)$. Функция перехода между состояниями $s_{t+1} = \varphi(x_t^k; s_t)$, и мы предполагаем, что переменная состояния раскладывается на K составляющих и они независимы от других рук, т.е. для всех k:

$$s_{t+1}^k = s_t^k, \text{ если } a_t \neq k;$$

$$s_{t+1}^k = \varphi(x_t^k; s_t), \text{ если } a_t = k.$$

Тогда индекс Гиттинса для момента времени τ определяется как

$$g(s_t^k) = \sup_{\tau \geq 2} \left(\mathbb{E}_x [\sum_{t=1}^{\tau-1} \beta^t X^k(s_t^k)] / \mathbb{E}_x [\sum_{t=1}^{\tau-1} \beta^t] \right),$$

где τ — текущий раунд, β — коэффициент дисконтирования, $X^k(s_t^k)$ — выигрыш для состояния s_t^k. Это нормированная на коэффициент дисконтирования ожидаемая полезность от выбора данной руки, где полезность рассчитывается с учетом изменения полезности от экспериментирования. Как мы заметим далее, вопрос оценивания полезности экспериментирования остается актуальным и для современной литературы.

Однако вычисление этого индекса может быть затруднительно, для чего постоянно разрабатываются новые приближенные методы и вариации постановки задачи, такие как более богатое описание (контекст) состояний S_t (обзор см. в (Bubeck, Cesa-Bianchi, 2012)).

Разговор об обучении нельзя завершить на этой модели, поскольку она плохо применима для игровых ситуаций с активным оппонентом, реагирующим на действия игрока. Нельзя всегда рассчитывать, что оппонент будет играть то или иное действие (дающее разные выигрыши от использования рук) независимо от действий самого игрока. То есть распределение выигрышей не только нестационарно, но и зависимо от предыдущих действий игрока, что исключает применимость успешных для «многоруких бандитов» решений в игре с оппонентом, не являющимся в теоретико-игровых терминах Природой.

[4] Можно провести аналогию с функцией Шпрага–Гранди в комбинаторных играх.

Журнал НЭА,
№4 (44), 2019,
с. 77–125

3.3. Фиктивная игра

Фиктивная игра (fictitious play) — это правило обучения, впервые описанное Дж.У. Брауном (Brown, 1951), который и ввел этот несколько неясный, но устоявшийся термин. В этом семействе моделей каждый игрок исходит из того, что противник играет стационарные (возможно, смешанные) стратегии и выбирается наилучший ответ на эмпирическую частоту стратегий противника. Каждый участник смотрит на уже сыгранную историю и действует исключительно с опорой на эти наблюдения, не принимая во внимание возможность реакции оппонента. В зависимости от точного определения "эмпирической частоты" и "наилучшего ответа" возможны разные алгоритмы с различными свойствами. Здесь мы обсудим только основные результаты.

Начнем рассмотрение с простейшей фиктивной игры и с исторически первой модели динамики обучения. В этой игре игроки реагируют на действия друг друга по принципу наилучшего ответа в модели Курно.

Две фирмы, зная издержки друг друга и рыночный спрос, одновременно определяют объемы выпуска, от которого зависит их прибыль. В каждом периоде повторяющейся игры фирмы наблюдают решения о выпуске, принятые игроками, и устанавливают свой выпуск на уровне, соответствующем наилучшему ответу на выпуск оппонента в предшествующем периоде.

3.3.1. Динамика Курно (наилучший отклик)

В знаменитой исходной работе А. Курно ((Cournot, 1960) цитируется по репринтному изданию) интересовал прежде всего не поиск равновесия, известного сейчас как равновесие Курно–Нэша, а адаптивный процесс взаимной подстройки стратегий двух фирм, конкурирующих по объемам выпуска на конкретном рынке (в книге Курно — на рынке минеральной воды). Если выпуск двух фирм обозначить через s_1, s_2, а функцию полезности через $u_i(s_1, s_2)$, $i \in \{1, 2\}$, то функцией наилучшего отклика будет $B_i(s_{-i}) = \arg\max_{\tilde{s}_i} u_i(\tilde{s}_i, s_{-i})$. В стандартном случае целевая функция $u(\cdot, s_{-i})$ строго вогнута по стратегиям оппонента, которые, тем самым, являются стратегическими субститутами. Начав с любого профиля стратегий $s^0 = \{s_1^0, s_2^0\}$, динамика Курно предполагает, что оба участника выбирают наилучший ответ на стратегию оппонента, избранную им на предыдущем ходе т.е. $s_i^t = B_i(s_{-i}^{t-1})$. В силу вогнутости целевых функций кривые наилучшего ответа убывают по стратегии оппонента и пересекаются в одной точке, что гарантирует единственность взаимного наилучшего ответа (равновесия). Такая динамика, разумеется, весьма близорука: игроки в ней реагируют только на поведение оппонента в предыдущем периоде, не пытаясь предвидеть его действий.

Расширением динамики Курно, учитывающим не только последний ход оппонента, но и всю историю игры, будет модель стационарной фиктивной игры.

Журнал НЭА,
№4 (44), 2019,
с. 77–125

3.3.2. Модель стационарной фиктивной игры

Два игрока играют повторяющуюся игру $G = \langle \mathcal{I}, S, \{u\}, T \rangle$, у каждого J стратегий $S_i = \{s_i^1, \ldots, s_i^J\}$. Кроме этого, задаются начальные веса или счетчики κ_{ik} для каждой стратегии k игрока i. В процессе формирования истории игры h^t, где $t \in 1, \ldots, T$, накапливается и статистика того, какие действия выбирал оппонент в предыдущих периодах. Игроки одновременно выбирают свои стратегии, наблюдают решения друг друга, после чего обновляют свои представления о том, как играет их оппонент, добавляя 1 к счетчику κ_{ik}^t той стратегии s_{-ik}^t, которую оппонент выбрал в этот период $t = 1,2,\ldots$:

$$\kappa_{ik}^t(s_{-ik}^t) = \kappa_{ik}^{t-1} + b, \quad b = \begin{cases} 1, \text{ если } s_{-ik} \in s_{-i}^t; \\ 0, \text{ если } s_{-ik} \notin s_{-i}^t. \end{cases} \tag{1}$$

В динамике Курно игрок i полагал, что оппонент выберет ту же стратегию, что и в прошлом периоде. В данном, более общем, случае вера i в то, что игрок $-i$ сыграет стратегию s_{-ik} в момент времени t, определяется как относительный вес этой стратегии в эмпирических частотах прошлых действий игрока $-i$: $\gamma_{ik}^t(s_{-i}^t) = \kappa_{ik}^t(s_{-i}^t) / \sum_{j=1}^{J} \kappa_{ij}^t$. В итоге игрок i выбирает наилучший отклик[5] (Best Response) на свои текущие представления о том, как играет его оппонент:

$$BR_i^t(\gamma_i^t) \in \arg\max_{\{s_{ik} \in S_i\}} \mathrm{E}\left(u_i^t(s_{ij}^t, s_{-i}^t) | \gamma_i^t\right).$$

Фиктивная игра сходится к равновесию Нэша в любой игре с нулевой суммой (Robinson, 1951), в любой невырожденной игре 2×2 (Miyasawa, 1961), в любой игре, решаемой методом итеративного исключения строго доминантных стратегий (Nachbar, 1990). Если фиктивная игра сходится к профилю чистых стратегий для всех игроков — этот профиль будет равновесием Нэша, а если для всех игроков сходятся эмпирические распределения частот γ_i^t, то профиль стратегий, к которому они сходятся, будет равновесием по Нэшу. Когда в игре есть строгое равновесие Нэша, то это равновесие является поглощающим состоянием фиктивной игры (Nachbar, 1990; Fudenberg, Levine, 2009).

3.3.3. Примеры фиктивной игры

Рассмотрим несколько примеров того, как работает фиктивная игра. В качестве первого примера возьмем простейшую игру «орлянка» (Matching pennies), представленную в табл. 2. В этой игре два игрока одновременно выбирают «орла» (Head) или «решку» (Tail). Если они выбрали одно и то же, то выигрывает первый, если же разные стороны монеты, то второй.

Таблица 2

Игра «орлянка»

	Орел (H)	Решка (T)
Орел (H)	1; –1	–1; 1
Решка (T)	–1; 1	1; –1

[5] Заметим, что допускается существование не единственного наилучшего отклика — в этом случае решение выбирается произвольным образом, например случайно.

Журнал НЭА,
№4 (44), 2019,
с. 77–125

Начнем с любого профиля начальных (нулевого периода) счетчиков, например, $\kappa_1 = (1, 2)$ и $\kappa_2 = (3, 1)$, т.е. первый игрок может считать более вероятным ход «решка» второго, а второй считает вероятным ход «орел» первого. Тогда в первый период оба участника играют *T*, и веса становятся (1, 3) и (3, 2). Вслед за этим игрок 1 должен играть *T*, а игрок 2 — *H* до тех пор, пока опыт не переубедит его, что первый чаще играет *T*, т.е. до весов (3, 4), когда он должен будет сменить стратегию на *H* в предположении, что оппонент меняет стратегию только тогда, когда новая стратегия строго лучше старой. В ответ на это игрок 1 будет накапливать свидетельства в пользу того, что его оппонент играет *H* (начиная с (1, 5) и до (6, 5)), когда он сам должен будет сменить стратегию на *H*, и так далее. В достаточно долгой перспективе эмпирические частоты такой игры сходятся к единственному равновесию Нэша $\{[1/2, 1/2],[1/2, 1/2]\}$. В этом примере концепция фиктивной игры не вызывает возражений.

Однако могут возникнуть и сложности. Не всегда сходимость эмпирических частот хорошо схватывает суть игры. Рассмотрим в качестве примера игру «камень, ножницы, бумага». В этой классической игре участники одновременно выбирают один из предметов; камень побеждает ножницы, бумага побеждает камень, ножницы побеждают бумагу (иными словами, отношение между стратегиями нетранзитивно) — и победитель забирает приз этого периода игры. Единственное равновесие Нэша в смешанных стратегиях это $\{1/3, 1/3, 1/3\}$ (табл. 3).

Таблица 3

Игра «камень, ножницы, бумага».

	Камень (*R*)	Ножницы (*P*)	Бумага (*S*)
Камень (*R*)	0, 0	1, –1	–1, 1
Ножницы (*S*)	–1, 1	0, 0	1, –1
Бумага (*P*)	1, –1	–1, 1	0, 0

Несложно заметить, что в этом случае участники будут отвечать на каждую стратегию сменой наилучшего ответа, как только частоты, с которыми каждый игрок играет любую из своих стратегий, побудят его оппонента сменить стратегию на выигрывающую (камень побеждает ножницы, бумага побеждает камень...). В этом случае эмпирические частоты, действительно, сойдутся к равновесию, но выигрыш игроков будет колебаться с каждым циклом смены стратегий все больше и больше, чего мы не ожидаем в равновесной игре. Заметим, что в этой игре эмпирические частоты будут сходиться к равновесным не только если играется равновесие одноперiодной игры (смешанное равновесие Нэша). Когда игроки каждый ход меняют стратегию на наилучший ответ на собственную стратегию предыдущего хода, при этом один из них будет все время выигрывать, другой — проигрывать, но эмпирические частоты будут соответствовать равновесным.

Журнал НЭА,
№4 (44), 2019,
с. 77–125

Более сильным критерием будет требование сходимости совместных эмпирических частот к профилю стратегий Нэша, в этой игре каждая позиция в матрице выигрышей встречается с одинаковой частотой $(1/9)$ в истории игры h^t. Однако и для такого критерия можно создать простое детерминированное правило, не соответствующее ожидаемому пониманию сходимости, например играть $(R,R)\to(R,S)\to(R,P)\to(S,R)\ldots,(P,P)\to(R,R)$.

Исходя из закона больших чисел, на номерах раунда $t = 1, 10, 20$ позиция (R, R) должна встречаться не чаще чем $1/9$ раз, что, очевидно, не будет выполняться.

3.4. Определение сходимости

На примере фиктивной игры видна проблема баланса определения сходимости — возможно сформулировать несколько существенно разных определений. При этом слишком сильному определению не будет соответствовать никакое правило обучения, а слишком слабому будет соответствовать почти любое, причем, на первый взгляд, разница между ними может быть невелика (Nachbar, 2009). Яркой иллюстрацией сложности нахождения такого баланса может послужить пример координационной игры (табл. 4). Если игроки начнут повторять игру равновесия Нэша, в котором они играют $(A; A)$ в нечетные периоды и $(B; B)$ в четные периоды, тогда система сходится (тривиально) к равновесию Нэша при многократной игре. При этом достигаются разные равновесия Нэша в зависимости от того, является ли изначальная дата нечетной или четной.

Таблица 4

Координационная игра “Битва полов”

	A	B
A	1, 1	0, 0
B	0, 0	1, 1

Сходимость фиктивной игры в целом не гарантирована. Классический пример приведен в (Shapley, 1964), которому соответствует, например, матрица платежей, представленная в табл. 5. Эта игра является вариантом игры «Камень, ножницы, бумага» (см. табл. 3), которая становится игрой с ненулевой суммой. Это незначительное отличие приводит к существенному сдвигу в темпах накопления весов, соответствующих эмпирическим частотам наилучшего ответа: если изначальные веса приписывают участникам играть какой-то из профилей стратегий, лежащих вне главной диагонали, то динамика фиктивной игры будет приписывать им следование по циклу $(T,M)\to(T,R)\to(M,R)\to(M,L)\to(D,L)\to(D,M)\to(T,M)\ldots$ Каждый следующий виток цикла будет требовать все больше времени, но эта динамика не сходится никогда.

Журнал НЭА,
№4 (44), 2019,
с. 77–125

Таблица 5

Игра Шепли

1 \ 2	L	M	R
T	0, 0	1, 0	0, 1
M	0, 1	0, 0	1, 0
D	1, 0	0, 1	0, 0

Представим, что один из игроков последовательно делает ходы по циклу $R \to S \to P \to R \ldots$ Его проницательный оппонент, разгадав это простое правило обновления стратегий, сможет, смещая свою стратегию в правильную сторону, всегда выигрывать, действуя по схеме $P \to R \to S \to P$. Такая стратегия также будет смешанной с частотами, соответствующими равновесию Нэша, однако средними платежами -1 для первого и 1 для второго игрока (т.е. формально сходимость к равновесию есть, но по сути первый игрок ведет себя близоруко и постоянно проигрывает).

В таком случае сходимость удобно определить не в терминах частот, а терминах стабильности средних выигрышей. Если средний выигрыш пары игроков не меняется более чем на ε, можно считать, что он сошелся. Однако сам по себе факт сходимости не накладывает ограничений ни на минимальную необходимую величину среднего выигрыша, ни на тип оппонента. Это подводит нас к понятию универсальной совместимости, или совместимости по Хэннану (Hannan, 1957). Согласно данному критерию игрок почти наверное получит по крайней мере столько же полезности, сколько мог бы получить, если бы знал заранее частоту игры стратегий оппонента σ_i (но не их порядок, не сами стратегии на каждом из раундов). Формально это можно определить как

$$\lim \sup_{T \to \infty} \left(\max_{\sigma_i} (u_i(\sigma_i, \gamma_i^t) - \frac{1}{T} \sum_t u_i(y^t(h^{t-1})) \right) \leq \varepsilon,$$

где $y^t : H \to \mathbb{R}_+$ — функция исходов. В отличие от прочих критериев универсальная сходимость позволяет судить о качестве стратегии не по теоретическим профилям, а по наблюдаемым величинам — выигрышам. Кроме того, она подразумевает способность алгоритма играть с любым типом оппонента.

Однако факт сходимости правила к равновесному исходу все еще может иметь слишком широкое пространство для интерпретаций. Например, в игре «Битва полов» на рис. 1 универсальная сходимость обеспечит игроку выигрыш в размере 2, однако при правильном чередовании стратегий игроки могли бы достигать среднего выигрыша, равного 3. Так, авторы (Mathevet, Romero, 2012) воспользовались наблюдениями (McKelvey, Palfrey, 2001) о том, что алгоритмы обучения плохо предсказывают исход экспериментов, и провели сравнение теоретических предсказаний, исходов симуляций и экспериментальных

Журнал НЭА,
№4 (44), 2019,
с. 77–125

результатов в терминах среднего выигрыша. Пример такого анализа представлен на рис. 1, где в левом столбце представлены исследуемые игры в матричном виде. Множества всех возможных платежей для каждой из игр описываются замкнутыми контурами, расположенными на графиках правее соответствующей матрицы. Во втором столбце находятся иллюстрации, включающие в себя множество платежей, которые доминируют равновесные платежи в смешанных стратегиях в соответствии с народной теоремой; в третьем столбце — результаты сходимости симулированной взвешенной фиктивной игры; в последнем — распределения платежей из экспериментальных данных.

В последних двух случаях круги обозначают платежи с координатами в центре каждого круга, а диаметры каждого круга — частоты, с которыми популяция играет соответствующий профиль стратегий. Для удобства сопоставления единичный радиус соответствует всей популяции [6].

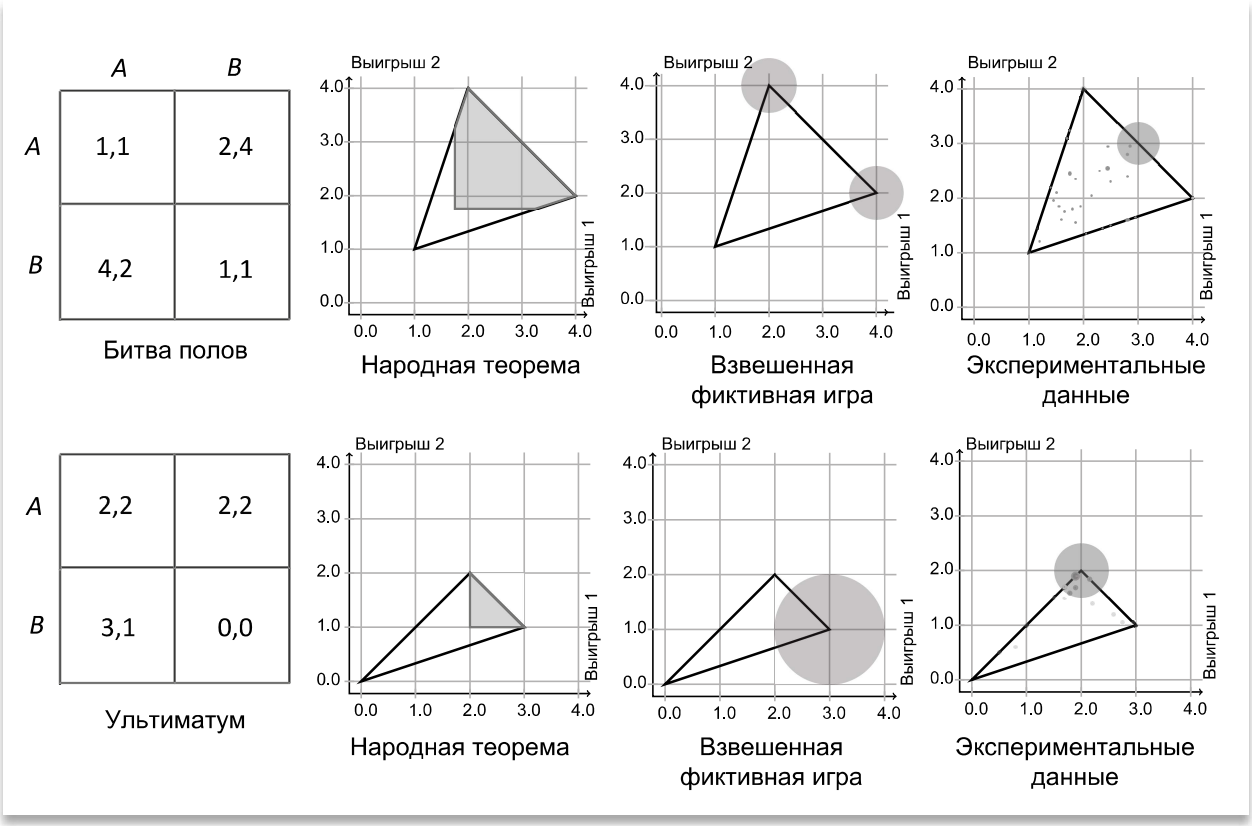

Рис. 1

Сравнительная динамика фиктивной игры по (Mathevet, Romero, 2012)

[6] Данные в третьей колонке отображают результат 1 тыс. симуляций с парами алгоритмов, запрограммированными играть взвешенную фиктивную игру с $(\phi \in (0, 1))$. Каждая симуляция продолжалась до тех пор, пока средний выигрыш пары не менялся более чем на 0,01 в 20 последовательных блоках (в частности, симуляция разбита на блоки по 100 периодов). Максимальная длина в каждом из прогонов была установлена на 100 тыс. периодов, при том что медианная длина схождения не превышала 16 800 периодов для каждой из игр. Результаты, приведенные далее на графиках, представляют средние значения, взятые за 1 тыс. смоделированных пар, за последние 1 тыс. периодов. Для экспериментальных данных размер выборки составлял 60 и 70 участников для каждой из двух игр соответственно. Данные, использованные для иллюстрации, отображают относительную частоту каждого возможного выигрыша за последние 20 периодов суперигры. Экспериментальные данные собирались в двух подгруппах с несколько разными правилами: в первой начальные 30 раундов фиксировались, а затем, начиная с 31 раунда, вероятность продолжения составляла 0,9, поэтому ожидаемая длина каждой суперигры была 40. Во второй — начиная с первого раунда, вероятность продолжения составила 0,99, поэтому ожидаемая длина составляла 100.

Журнал НЭА,
№ 4 (44), 2019,
с. 77–125

Сравнение эмпирических частот среднего выигрыша показывает, что взвешенная фиктивная игра сходится к результатам, отличным от экспериментальных данных в обеих играх, что отчетливо видно на графике. Так, например, в координационной игре "Битва полов" люди склонны достаточно быстро координироваться и переключаться между профилями (*A*, *B*), достигая тем самым идентичного среднего выигрыша в размере трех для каждого участника. Напротив, пара алгоритмов фиктивной игры сходится к менее социальному исходу (*A*, *A*), (*B*, *B*) в половине случаев (что так же демонстрирует чувствительность динамики к выбору в первом раунде или к изначальным верам). Тем самым, можно утверждать, что соответствующие модели не очень хорошо описывают поведение реальных людей, особенно в ситуациях предсказания действий противника на несколько шагов вперед (forward looking), с чем при наличии правильных стимулов, участники экспериментов отлично справляются. Если каждый игрок считает, что поведение его оппонента описывается последовательностью независимых и одинаково распределенных мультиномиальных случайных переменных, а его априорные убеждения относительно этого распределения описываются распределением Дирихле[7], то фиктивная игра соответствует более общей модели — байесовскому обучению.

3.5. Байесовское обучение

Класс алгоритмов байесовского обучения (Fudenberg, Levine, 1993; Nachbar, 2009) часто называют *рациональным обучением* (Marimon, 1995), поскольку они удовлетворяют стандартным аксиомам рациональности. Аксиоматика, задающая рациональные предпочтения (например, субъективная ожидаемая полезность Сэвиджа (Savage, 1954)), требует от игрока совместимости по верам (consistency of beliefs), в том числе это означает включение новой информации в старую систему вер согласно правилу Байеса. Однако одно соответствие некоторому набору аксиом не может служить гарантией адекватности правила обучения.

Правило Байеса широко и успешно применяется в различных областях статистики, поэтому может показаться, что следования ему рациональным игроком достаточно для успешной сходимости к равновесию (статически рациональному поведению). Однако для обучения важна структура представлений об оппоненте и некоторые детали формулировки алгоритма, что приводит к парадоксам.

Чтобы продемонстрировать особенности этого класса моделей, охарактеризуем неотъемлемые положения байесовского обучения в повторяющейся игре:

1) у каждого игрока есть априорное распределение вероятностей относительно поведенческих стратегий оппонента;
2) на основе истории игры, доступной всем игрокам после каждого периода, эти веры обновляются по правилу Байеса;

[7] Распространяется на дискретный случай игр с числом доступных стратегий больше двух, в данном случае мультиномиальное распределение и распределение Дирихле образуют семейство сопряженных распределений: для априорного распределения по Дирихле и мультиномиальной функции правдоподобия апостериорное распределение также будет распределением Дирихле.

Журнал НЭА,
№4 (44), 2019,
с. 77–125

3) в каждый момент времени каждый игрок выбирает ту стратегию поведения, которая максимизирует его ожидаемый дисконтированный выигрыш во все последующие периоды.

Даже без наложения дополнительных предположений с динамикой двух игроков в повторяющейся игре возникают сложности (в первую очередь в формализации множества априорных распределений). Однако к базовым положениям обычно добовляется ряд существенных допущений, с которых мы и начнем рассмотрение динамики байесовского обучения.

3.5.1. Модель байесовского обучения против близорукого оппонента

Базовым предположением является то, что байесовский алгоритм обучения (здесь и далее — байесовский ученик) взаимодействует со средой, никак не зависящей от его действий. Такой оппонент в экономической литературе часто называется близоруким (myopic). Данный подход можно также понимать как взаимодействие игрока с результатом действий множества других игроков, усредненным общественным действием. То есть даже если действия агента i влияют на общественный результат, то пока он не учитывает эту зависимость, он может рассматривать общественный результат как экзогенно заданный внешний мир.

В соответствии с рассматриваемым правилом обучения рациональные игроки должны обновлять свои веры по правилу Байеса по ходу истории и выбирать наилучший ответ $\sigma_i^{t+1}(h^t) \in B_i(x^t)$ во всякий период t, когда им предстоит принимать решения. Каждый платеж t игрока i зависит только от его действия $s_i^t \in S_i$ и от состояния процесса x^t, за который в играх естественно принять профиль стратегий оппонентов s_{-i}^t, так что $X = S_{-i}$ и $u \equiv u(s_i^t, x^t)$, а множество пар (s_i^t, x^t) — это история игры h^t.

В качестве примера рассмотрим игру с Природой — подбрасывание монеты, которая может быть смещенной, и где цель участника, играющего с Природой, — сделать ставку на правильную (более вероятную) сторону монеты. Каждый бросок такой монеты — это реализация случайной переменной $X = \{0,1\}$ с биномиальным распределением и неизвестным параметром θ (истинным смещением монеты). Цель игрока — определить наиболее вероятное значение $\hat{\theta}^t$ после истории h^t, что автоматически позволит ему выбрать $\sigma^t* = 0$ или 1, в зависимости от того оказывается ли $\hat{\theta}^t \gtrless 0{,}5$.

Пусть априорная оценка вероятности исхода $x \in X$ равна $\Pr(\theta)$ и имеется последовательность исходов (бросков) $\mathbf{x}$ с функцией правдоподобия $\Pr(\mathbf{x}|\theta)$. Тогда апостериорное значение θ будет определено по правилу Байеса как

$$\Pr(\theta|\mathbf{x}) = \frac{\Pr(\mathbf{x}|\theta)\Pr(\theta)}{\int \Pr(\mathbf{x}|\theta')\Pr(\theta')d\theta'}.$$

Журнал НЭА,
№4 (44), 2019,
с. 77–125

В случае с броском монеты априорная вероятность задается двухпараметрическим бета-распределением, а функция правдоподобия — биномиальным распределением, эта пара образует семейство сопряженных распределений: если априорное распределение будет бета-распределением, а функция правдоподобия биномиальной, то апостериорное распределение также будет бета-распределением и задача решается аналитически (см. разбор в Приложении). В нашем примере динамика игры будет выглядеть достаточно просто: байесовский ученик должен просто долго бросать монету, обновляя веры о параметре бета-биномиального распределения, тогда его прогноз вероятности выпадения орла в пределе сойдется к истинному распределению $\theta^t \to \theta^*$, где θ^* — истинное значение θ (Marimon, 1997). В общем случае, $\Pr(\theta|\mathbf{x}) \propto \Pr(\mathbf{x}|\theta)\Pr(\theta)$, и вычисление апостериорной вероятности будет зависеть от априорного распределения и его параметров.

3.5.2. Проблемы байесовского обучения

Даже с близоруким оппонентом оптимальные процедуры являются таковыми, только если веры агента соответствуют сложности окружающей среды. Нужно заранее иметь правдивые предположения о том, как фактическая окружающая среда устроена. Если такое представление упрощено, то и прогнозы агента могут быть далеки от правильных. Например, представим, что смещенность монетки θ не постоянна, а может изменяться от периода к периоду (мы попеременно пользуемся двумя разными монетами). Правило байесовского обучения не гарантирует нам успеха: пусть, например, последовательность исходов $X = \{0,1\}$ — простое чередование исходов $(0,1,0,1,...,0,1,...)$. Зная или угадав правило, возможно угадывать каждый следующий исход. Но байесовский ученик должен, игнорируя чередование, сойтись к оценке $\theta = 1/2$ для такой последовательности, что приведет к ошибке в половине случаев. В данном примере, в процессе формирования априорного распределение вероятности, должна быть учтена возможность зависимости смещения от времени (в том числе «по четным — одно, по нечетным — другое»). Однако в общем случае сложно составить исчерпывающий список, какие возможности следует учитывать.

Другая проблема вызвана детерминированностью и невозможностью экспериментов. Рассмотрим байесовское обучение в задаче многорукого бандита с двумя руками. В одной руке монета с вероятностью p выпадения 1 (другая сторона приносит 0), в другой — с вероятностью q, и необходимо максимизировать дисконтированную сумму выпавших единиц. Оказывается, что даже байесовские ученики со сколь угодно сложной системой априорных вер с положительной вероятностью сходятся к выбору неоптимальной руки в данной задаче (Rothschild, 1974). Это происходит, если текущая информация игрока свидетельствует в пользу одной руки и он навсегда перестает выбирать вторую, так как не получает о ней новой информации. Поскольку одна рука выбирается

Журнал НЭА,
№4 (44), 2019,
с. 77–125

только конечное число раз, оценка q может не сойтись к ее истинному значению и ученик продолжить выбирать руку с вероятностью p даже при $p < q$.

Если существуют ситуации, в которых байесовский ученик даже при достаточно строгих допущениях, ограничивающих поведение оппонента, ведет себя неоптимальным образом, остаются два вопроса.

1. Если два байесовских ученика будут играть друг с другом, сойдутся ли они к равновесию (в частности, к равновесию Нэша)?

2. Смогут ли их веры друг о друге в пределе сойтись к истинным?

3.5.3. Байесовское обучение в общем виде

Основные сложности использования байесовского обучения как универсального правила кроются в том, как может задаваться система априорных вер.

В общем виде, игрок переходит от пассивного обучения (когда он взаимодействовал с близоруким оппонентом) к активному (допускающему, что действия оппонента зависят некоторым образом от истории игры). Например, достаточно терпеливый рациональный игрок может придерживаться неограниченно сложных стратегий (Young, 2004, p. 91), таких как оптимальное экспериментирование, распознавание паттернов, преднамеренное введение в заблуждение противника, имитация близорукости и пр.

Чтобы проиллюстрировать данные сложности, перейдем к рассмотрению общей модели байесовского обучения в повторяющейся игре, такой что участники взаимодействия учатся на основании всех историй $h^t = (s^1, \ldots, s^t)$, включающих в себя все действия, выбранные участниками до периода t включительно. Множество всех возможных конечных историй обозначается H. Напоминаем, что последовательность стратегий, избранных игроками в динамической игре двух игроков, называется историей игры на момент t и обозначается h^t, где $h^t = \{s^1, \ldots, s^t\}$ и $s^t = \{s_1^t, \ldots, s_I^t\}$. Отображение $\xi_i : h^t \to \Delta S_i$, определяющее, какую из чистых стратегий должен выбрать игрок i в ответ на наблюдаемую историю игры, называется поведенческой стратегией игрока i. Предполагается, что каждый участник в каждый момент времени имеет некоторую модель поведения оппонентов, определенную как отображение из множества допустимых историй во множество стратегий его оппонентов, и обозначаемую $m_i : H_i \to \Delta(S_{-i})$. Множество возможных моделей обозначим $\mathcal{M}_i$, а множество вер, которые придают положительную вероятность множеству возможных моделей $\mathcal{M}_i$, как $\mu_i(\mathcal{M}_i)$.

Понятно, что при некоторых таких моделях ход игры сойдется к равновесию (например, если игроки изначально убеждены, что оппоненты сразу играют одно и то же равновесие Нэша). Но насколько широк класс таких моделей? Верно ли, что игроки, рационально обучающиеся по правилу Байеса, всегда сойдутся к равновесию? В общем случае этого гарантировать нельзя, однако первым существенным

Журнал НЭА,
№4 (44), 2019,
с. 77–125

результатом стал результат из (Kalai, Lehrer, 1993), где авторами была получена характеристика условий, при выполнении которых веры игроков сходятся к истинным распределениям поведенческих стратегий оппонентов, и вслед за ними смешанные стратегии игроков сходятся к равновесию Нэша.

Каждый профиль поведенческих стратегий $\xi(h^t) \equiv \xi_1(h^t) \times \ldots \times \times \xi_{\mathcal{I}}(h^t)$, реализованный после истории h^t, задает хорошо определенное распределение вероятностей на множестве возможных историй с точки зрения игрока i, которое обозначается $D_i(\xi_i)$. Веры μ_i называются *абсолютно непрерывными относительно хода игры*, если для каждого распределения вероятностей на множестве возможных историй $D_i(\xi_i)$ найдется такая модель $m_i \in \mathcal{M}_i$, что стратегия, предписанная этой моделью (обозначим ее $D_i(\mu_i, \xi_i)$), задает то же самое распределение вероятностей, т.е. $D_i(\mu_i, \xi_i) = D_i(\xi_i)$, и каждой из действительно возможных историй на момент t эти веры приписывают положительную вероятность.

Основной результат о классе сходящихся к равновесию байесовских игроков — если все стратегии $\xi_i(\mu_i)$, порожденные верами μ, абсолютно непрерывны относительно хода игры $D_i(\xi_i)$, то эти стратегии сходятся к равновесию Нэша. Сходимость понимается так, что для каждой истории распределения вероятностей заданные $\xi_i(\mu_i)$, почти наверное, совпадают с равновесными смешанными стратегиями в данной стадийной игре. Доказательство этого технически непростого результата следует из теоремы о достижимости Блэквелла–Дабинса (см. (Fudenberg, Levine, 1998a)).

Требование абсолютной непрерывности нетривиально. Предположим, в бесконечно повторяющейся дилемме заключенного оба игрока играют стратегию “Спусковой крючок” (т.е. кооперация до тех пор, пока оппонент не обманул в первый раз, а после этого навсегда отказ от кооперации) и верят в то, что оба они продолжат играть эту стратегию до тех пор, пока только один из них обманет. Если стратегии игроков таковы, что один из них достаточно скоро действительно обманет оппонента, то стратегия спускового крючка реализуется и веры оказывается абсолютно непрерывными относительно хода игры. Если игроки никогда не обманывают, у них никогда не возникнет возможности проверить данную стратегию и абсолютная непрерывность нарушается.

Другой пример для более широкого класса игр (Nachbar, 2005): если два игрока, играющих в «Орлянку», построили некоторые корректные модели поведения своих оппонентов, то как только они распознают их стратегии, им перестанет быть выгодно придерживаться равновесной стратегии, заданной этими моделями, что опять-таки нарушает абсолютную непрерывность.

В общем случае оказывается невозможно построить класс вер, позволяющих выучить параметры истинного распределения вер оппонента, и который, как следствие, обеспечивает сходимость к равно-

Журнал НЭА,
№4 (44), 2019,
с. 77–125

весию независимо от действий оппонента (Nachbar, 2009) (для игр с неопределенностью в выигрышах идейно похожий результат получен в (Foster, Young, 2001)).

Для пояснения этого результата сначала рассмотрим пример из (Marimon, 1995). Пусть $X = \{A, B\}$, матрица выигрышей задается табл. 6, (Marimon, 1995) и μ_i удовлетворяет условию абсолютной непрерывности относительно семейства распределений $\nu \in \mathcal{N}$. Процесс, лежащий в основе байесовского обучения, это результат наилучшего отклика игрока B_i согласно его поведенческой стратегии ξ_i.

Таблица 6

Игра против неблизорукого игрока

	A	B
s_1	1	0
s_2	0	1

Рассмотрим процесс генерации данных x^t, который на наилучший ответ игрока предписывает его оппоненту играть невыгодную для игрока стратегию, т.е. $Prob(x^t = A \mid x^{t-1}) > 1/2$, если $\sigma_t(x_{t-1}) = s_2$ и $Prob(x^t = B \mid x^{t-1}) > 1/2$, если $\sigma_t(x_{t-1}) = s_1$. Если такой процесс лежит в $\mathcal{N}$, то байесовский ученик сможет его выучить, однако B_i^t больше не будет наилучшим ответом. С новым $\hat{B}_i^t$ свяжем новое невыгодное правило, все так же из $\mathcal{N}$ и т.д. Как правило, нельзя закрыть этот процесс, нет оптимальной стратегии для всех способов игры, лежащей в $\mathcal{N}$, и в то же время оставаться в этом классе $\mathcal{N}$, принимая во внимание обратную связь на оптимальную стратегию. Иными словами, игрок может предположить, что оппонент неблизорукий и поменять стратегию с s_1 на s_2, но если и второй игрок неблизорукий (оба игрока рациональны) и достаточно терпелив, то попытка выучить сложное поведение противника приводит к усложнению поведения этого противника. В некоторых типах игр (Young, 2004, p. 92) этот интерактивный эффект дает поведение, которое становится сколь угодно сложным. Это делает процесс обучения с помощью обновления вер (priors) практически невозможным.

Общая теорема (Nachbar, 2005) говорит о невозможности одновременного выполнения для рациональной стратегии обучения трех, казалось бы, естественных правил:

1) обучаемости (learnability), т.е. возможности прийти в такое состояние, в котором веры относительно игры оппонента предсказывают его следующий ход, как если бы предсказание делалось при известном истинном распределении, лежащим в основе игры оппонента;
2) богатства класса (richness) возможных траекторий на множестве всех возможных будущих историй;

Журнал НЭА,
№4 (44), 2019,
с. 77–125

3) согласованности (consistency), в смысле соответствия наилучшей стратегии игрока в каждый момент игры его верам.

Интуитивно выбор наилучших стратегий и постепенное их нащупывание означает, что некоторые стратегии необходимо исключать из рассмотрения, что нарушает требование включения богатого класса траекторий. Обучаемость и богатство класса траекторий предполагает наличие независимо распределенных путей для каждого момента принятия решений, включая слабо доминантные, что противоречит согласованности выбора (как было в выше описанном примере). Этот результат устойчив и для ε-равновесных стратегий (Nachbar, 2005). Сам по себе он не означает невозможности обучения и говорит лишь о том, что сходящийся алгоритм обучения не может отвечать одновременно трем вышеупомянутым свойствам. Если же не предполагать, что игрок что-то знает о функции выигрышей оппонента (такая постановка задачи обучения получила название несвязанной, или расцепленной динамики (uncoupled dynamics)). В (Hart, MasColell, 2001) показано, что сходимость к равновесию Нэша против любого произвольно заданного алгоритма обучения невозможна для любого правила обучения вообще, а не только для детерминированных алгоритмов на основе правила Байеса.

4. Основные модели ограниченно-рационального обучения в играх

4.1. Калибровка

Начнем обсуждение ограниченно рациональных правил со статистически примечательного правила калибровки.

Калибровка прогнозов. Обучение в играх тесно смыкается с задачами предиктивной статистики: и прогноз погоды, и прогноз поведения оппонента требуют анализа предыдущего опыта и могут включать в себя как точечные прогнозы («Завтра будет дождь», «Оппонент выберет стратегию “Решка”»), так и вероятностные («Вероятность дождя завтра составляет 85%», «Наилучшим ответом на данную историю игры будет 85% вероятность выбора “Решки”»). Сложно ли для предсказания поведения оппонента получить правильный вероятностный прогноз?

Откалиброванность прогноза определяется как соответствие предсказанных вероятностей событий и частот, с которыми эти события случаются (Foster, Vohra, 1998). Например, если четверть дней в году были дождливыми, откалиброванность прогноза тем лучше, чем ближе была предсказанная вероятность дождя к 0,25. Заметим, что предсказание бинарной последовательности 01010101 может быть хорошо откалибровано и предсказанием 0,5 вероятности каждого исхода, и точным прогнозом. Однако плохо откалиброванный прогноз ‘10101010’ хотя и ошибается каждый раз, схватывает суть процесса (чередование нулей и единиц) лучше, чем прогноз, который будет

Журнал НЭА,
№4 (44), 2019,
с. 77–125

предсказывать вероятность появления единицы, равную 0,33. Эти проблемы аналогичны проблемам определения сходимости эмпирических частот игры к равновесным.

Тем не менее небесполезен вопрос, возможно ли построить правило прогнозирования, которое будет хорошо откалибровано на любых будущих путях развития предсказываемой последовательности. Ответ для детерминированных правил прост — нет, для детерминированного правила, каким бы сложным оно не было, всегда найдется последовательность, на которой оно плохо откалибровано. Ответ для правил со случайностью зависит от источника предсказываемой последовательности: это может быть Природа; оппонент с противоположными интересами (adversarial) и возможностью выбрать следующий элемент последовательности, зная прогноз для этого элемента; и промежуточный случай, когда оппонент может менять последовательность по ходу игры, но знает только распределение предсказаний, а не точные прогнозы на следующий ход. Как и можно ожидать, в игре против всезнающего оппонента возможность рандомизировать не помогает, а задача многорукого бандита показывает, что в игре против Природы ответ будет положительным. Для адаптивного оппонента в (Foster, Vohra, 1998) получен нетривиальный результат, что возможно построить такое правило предсказания произвольной последовательности, порождаемой невсезнающим оппонентом, что это правило будет хорошо откалибровано. Альтернативное доказательство существования калиброванных рандомизированных правил может быть получено с помощью теоремы о минимаксе (подробнее см. (Foster, Vohra, 1997)). Этот результат породил даже вопрос, как можно проверить, не является ли некоторый эксперт-предсказатель шарлатаном, не обладающим никакой информацией о процессе генерации данных, но использующим это правило для создания видимости хорошей откалиброванности (подробнее см. раздел про калибровку в обзоре (Nachbar, 2009)).

4.2. Направленное обучение

Направленное обучение (directed learning) — один из наиболее общих способов задания ограниченно-рационального правила обучения с вероятностной функцией выбора действия (Selten, Stoecker, 1986; Selten, Buchta, 1999; Selten et al., 2005). Классический пример направленного обучения — стрельба по мишени. Когда при стрельбе есть несколько попыток, стрелок может оценить, в каком направлении результат выстрела был бы лучше, и для следующего выстрела сместить точку прицеливания в этом же направлении (потому и направленное).

Формально можно дать качественное описание тремя условиями (Selten, 2004):

1) для обучения необходимо дискретное время $t = 1,\ldots,T$;
2) должен существовать действительный параметр υ_t который игрок выбирает в каждом периоде;

Журнал НЭА,
№ 4 (44), 2019,
с. 77–125

3) должна осуществляться обратная связь, позволяющая корректировать значение параметра относительно предыдущего выбора.

Например, рассмотрим игру «Дилемма заключенных» (табл. 7).

По обратной индукции (D, D) — единственное равновесие в повторяющейся игре с конечным числом раундов (будем называть всю совокупность раундов суперигрой). Однако эмпирически участники экспериментов с этой игрой склонны к кооперации в значительном числе раундов.

Таблица 7

Игра «Дилемма заключенных»

	C	D
C	5, 5	0, 7
D	7, 0	2, 2

С точки зрения модели направленного обучения (Selten, Stoecker, 1986) это называется молчаливой кооперацией. Участники готовы кооперироваться вплоть до последнего раунда в суперигре, пока оппонент не отклоняется. Но каждый участник, естественно, прогнозирует, что в конце игры оппонент отклонится. Поэтому он вынужден решать, в какой момент отклониться ему самому, чтобы не жалеть об упущенной выгоде. Выбирая раунд в суперигре, в котором игрок начнет сам отклоняться от кооперативного поведения, он будет руководствоваться опытом, полученным из предыдущих раундов. Намеченный период старта отклонения и будет параметром направленного обучения в данной игре.

Поскольку направленное обучение — это качественная, а не количественная теория, для каждой отдельной игры ему соответствуют некие стохастические функции, предписывающие игроку вместо случайного выбора играть более часто тот выбор, который смещает игрока в правильном направлении. (Примеры более конкретных спецификаций направленного обучения (Selten, Buchta, 1999; Cason, Friedman, 1999; Sadrieh, 1998).)

4.3. Обучение с подкреплением

В направленном обучении правило переопределения вероятности действия зависит от параметра, обуславливающего направление, в котором двигается игрок. Что если такое направление невозможно задать в явном виде? Обучение с подкреплением — важный класс ограниченно-рациональных моделей, подходящий как раз для такого случая.

Естественный для теории игр неэкономический источник вдохновения — биология, предоставляет две развитые парадигмы адаптации. Одна из них заимствована вместе с названием в эволюционной теории игр, но мы рассмотрим вторую, в которой адаптация происходит не между поколениями организмов, а для одного и того же организма — теорию обучения с подкреплением (reinforcement learning).

Журнал НЭА,
№4 (44), 2019,
с. 77–125

Она берет свое начало в работах Ивана Павлова по формированию условных рефлексов, но за век активных исследований перешагнула пределы собственно биологии и стала неотъемлемой частью компьютерных наук и психологии. В ее основании лежит простой психологический принцип обратной связи, т.е. выбор и закрепление в наблюдаемом поведении тех действий, ответом от внешней среды на которые являются положительные стимулы. При этом если в современной психологии к бихевиоризму, который берет эту модель за единственное основание всех психологических процессов, относятся со скепсисом, то в машинном обучении это активно развивающаяся область исследований (для более подробного знакомства с этой областью рекомендуем (Sutton, Barto, 2018)).

Исторически обучение с подкреплением берет начало от известного в психологии эффекта Торндайка (Thorndike, 1911, 1927): «Действия, приносящие удовлетворительный эффект в конкретной ситуации, будут более вероятно выбраны повторно, а приводящие к дискомфортному эффекту, менее вероятно будут выбраны снова». Сама формулировка модели обучения с подкреплением приписывается психологам Р. Бушу и Ф. Мостеллеру (Bush, Mosteller, 1955). В экономике применение моделей обучения с подкреплением берет начало в работе (Erev, Roth, 1998).

4.3.1. Модель

Рассмотрим модель обучения с подкреплением на повторяющейся игре $G = \langle \mathcal{I}, S, \{u\}, T \rangle$[8]. У каждого игрока есть J статегий $S_i = \{s_i^1, \ldots, s_i^J\}$. По аналогии с моделью фиктивной игры для стратегии каждого игрока задается склонность (propensity) q_{ij}^t играть каждую стратегию в раунде t, начальная склонность q_{ij}^0 и обновление по правилу: $q_{ij}^{t+1} = q_{ij}^t + (u(s_j) - u_{min}(s))$, где $(u(s_j) - u_{min}(s))$ — дополнительный выигрыш от использованной стратегии $u(s)$ относительно минимального выигрыша $u_{min}(s)$. Если стратегия не была выбрана в данном раунде, то склонность играть ее остается без изменений. Вероятность выбрать стратегию в следующем раунде задается по аналогии с фиктивной игрой как относительная склонность $p_{ik}^t = q_{ik}^t \Big/ \sum_{j=1}^{J} q_{ij}^t$. Отметим, что в отличие от фиктивной игры на склонность играть ту или иную стратегию оказывает влияние не только факт успеха стратегии в прошлом, но и величина выигрыша. Таким образом, кривая обучения вначале будет иметь крутой наклон, но будет становиться все более пологой со временем (при увеличении t). Величина исходных склонностей является единственным параметром в классической формулировке модели, но в многочисленных расширениях часто появляются дополнительные.

Приведем несколько естественных примеров[9] (см., например, (Watson, 2017) и более современный пример формальной модели в (Erev et al., 1995)):

[8] Заметим, что она может быть задана и для более широкого класса игр и ситуаций.

[9] Впервые они появились в бехавиористской литературе 1930-х годов (приведены ссылки на репринтные издания книг, малодоступных в оригинальном издании).

Журнал НЭА, №4 (44), 2019, с. 77–125

1) параметр отсечения (cutoff) ϑ (Erev et al., 1995): вероятности $p_{ij} < \vartheta$ принимаются равными нулю. Если низкие вероятности неотличимы для игрока от нуля, существенно улучшается сходимость (в базовой версии модели число раундов, необходимых для достижения равновесного исхода, может превышать 10 000);

2) параметр локального экспериментирования, или генерализации. Параметр, который влияет на величину увеличения склонности играть ту или иную стратегию (из всей величины x только доля $1-\varepsilon$ прибавляется к выбранной стратегии. Оставшаяся часть ε прибавляется к наиболее близкой стратегии. При этом ε интерпретируется как локальные эксперименты или ошибки. Стратегии при этом должны интерпретироваться в одной размерности: цена, количество, сумма, предоставленная другому игроку, и т.д.);

3) параметр новизны, или ослабевания памяти (игроки склонны придавать большую значимость недавним событиям). Формально он задается как склонность с поправкой $1-\phi$ (где ϕ невелико). Данный параметр гарантирует, что новые наблюдения вносят вклад в общий процесс обучения, даже если уже накоплен обширный опыт. В классической модели с нарастанием склонностей увеличивается инерция: нужно больше новых наблюдений, чтобы заметить, что обстановка изменилась.

4.3.2 Индивидуальное эволюционное обучение

Простота базовой спецификации и очевидная практическая применимость обучения с подкреплением позволяет формулировать и применять самые различные модификации базового принципа. Мы ограничимся рассмотрением одного такого подхода, индивидуальным эволюционным обучением (individual evolutionary learning, IEL, (Arifovic, Ledyard, 2004)), в котором склонность играть ту или иную стратегию заменена на механизм создания выборки, схожий по принципу работы с биологическими репликаторами.

В модели индивидуального эволюционного обучения помимо стандартного набора стратегий S в повторяющейся игре $G = \langle \mathcal{I}, S, \{u\}, T \rangle$ рассматривается подмножество $\mathcal{O}_i^t \in S_i$ оперативных стратегий, содержащих в себе K альтернатив. Через задание оперативных стратегий алгоритм имитирует ограниченную память, поскольку могут играться только стратегии из этого подмножества. Стратегия o_i^t выбирается игроком i по принципу подкрепления, т.е. действие выбирается с вероятностью $p_{i\kappa}^t \in \sigma_i^t$ из $\mathcal{O}_i^t$ для каждого раунда t. Алгоритм делится на три этапа: экспериментирование (исследование), репликация и выбор.

Экспериментирование происходит в начале каждого периода t и добавляет к рассмотрению новые альтернативы, которые иначе никогда бы не имели возможности быть опробованными. Это гарантирует сохранение определенного разнообразия. Для каждого $\kappa = 1, \ldots, K$

Журнал НЭА,
№4 (44), 2019,
с. 77–125

с вероятностью ρ случайно отобранная стратегия из S_i заменяет собой $o^t_{i\text{к}}$. Оригинальная работа (Arifovic, Ledyard, 2004) фокусировалась на модели с непрерывным множеством стратегий S_i таких, что $J \to \infty$. В такой модели значение новой стратегии задается через нормальное распределение со средним, равным значению заменяемой стратегии и стандартным отклонением $o \sim N(o^t_{i\text{к}}, \varsigma)$. Стандартное отклонение, как и ρ, является оцениваемым параметром.

Репликация усиливает стратегии, которые были бы хорошим выбором в предыдущих раундах. Учитывая функцию платежей $\{u\}$, для каждого $\text{к} = 1,\ldots,K$, $o^t_{i\text{к}}$ выбирается следующим образом. Из $\mathcal{O}^t_i$ равновероятно достаются два элемента, например o^t_{im} и o^t_{il}. Далее

$$o^{t+1}_{i\text{к}} = \begin{Bmatrix} o^t_{im} \\ o^t_{il} o \end{Bmatrix}, \text{ если } \begin{Bmatrix} u(o^t_{im}) \geq u(o^t_{il}) \\ u(o^t_{im}) < u(o^t_{il}) \end{Bmatrix}.$$

Процесс репликации в периоде $t + 1$ придает больше веса как альтернативам, встречающимся в оперативном множестве неоднократно, так и альтернативам, которые принесли бы высокий выигрыш, будь они выбраны в данном раунде. Как и в репликаторной динамике, со временем множество $\mathcal{O}^t_i$ включает все больше одинаковых стратегий, так как большинство альтернатив копируют наиболее эффективную.

Выбор практически идентичен классической модели обучения с подкреплением

$$p^t_{i\text{к}} = (u(o^t_{i\text{к}}) - u_{min}(o)) \Big/ \left(\sum_{k=1}^{K} (u(o^t_{ik}) - u_{min}(o)) \right).$$

Единственное отличие — опыт распространяется только на предыдущий период и не накапливается в склонностях, как в модели обучения с подкреплением, а процесс обучения происходит за счет репликации. Начальные вероятности задаются с весами $1/K$ и, в результате, в модели три оцениваемых параметра: K, ς и ρ.

Модель индивидуального эволюционного поведения (IEL) хорошо зарекомендовала себя для описания поведения участников в играх с непрерывным числом стратегий (Anufriev, Arifovic, Ledyard, 2013), где даже при дискретизации пространства стратегий (например, возможности сделать ставку от 1 до 100 долларов только в целых долларах) число стратегий все еще велико для одновременного удержания их в голове агентом. Эта модель похожа на базовую модель обучения с подкреплением, но может проявить себя при оперировании более сложными типами стратегий, включающими в себя условные реакции на действия оппонента, затрагивающие более чем один период (например, стратегия “око за око”). Подобные модификации новы для литературы (например, модификация обучения с подкреплением (Ioannou, Romero, 2014) или модификация самого IEL (Arifovic, Ledyard, 2018)), и вопрос полноценного сравнения сложных моделей обучения на сопоставимых игровых пространствах (и классы эквивалентности игр и моделей) остается открытым.

Журнал НЭА,
№4 (44), 2019,
с. 77–125

4.3.3. Свойства модели обучения с подкреплением

В играх 2×2 с постоянной суммой, двумя игроками и единственным равновесием обучение с подкреплением сходится к цене игры (Beggs, 2005). Обучение с подкреплением и стохастическая фиктивная игра сходятся (но не к равновесию Нэша) в играх 2×2, играх с нулевой суммой и кооперативных играх (Hofbauer, Hopkins, 2005), при этом скорость сходимости у стохастической версии фиктивной игры выше, чем у обучения с подкреплением (Benaim, Hirsch, 1999).

Для иллюстрации сложности интерпретации динамики обучения с подкреплением снова воспользуемся данными из (Mathevet, Romero, 2012), в которых дано сравнение теоретического предсказания результатов симуляции двух игроков с правилом игры обучение с подкреплением, и результаты игры участников экспериментов в терминах среднего выигрыша. Уточненное описание процедуры проведения эксперимента и компьютерных симуляций приведено в п. 3.4.

На рис. 2, как и на рис. 1 в левом столбце приведены игры в матричном виде. Множества всех возможных платежей для каждой из игр описываются замкнутыми контурами, представленными на графиках правее соответствующей матрицы. Во втором столбце слева находится множество платежей, которые доминируют равновесные платежи в смешанных стратегиях в соответствии с народной теоремой. График в третьем столбце показывает сходимость симуляции игры пары алгоритмов обучения с подкреплением против друг друга, а в последнем —

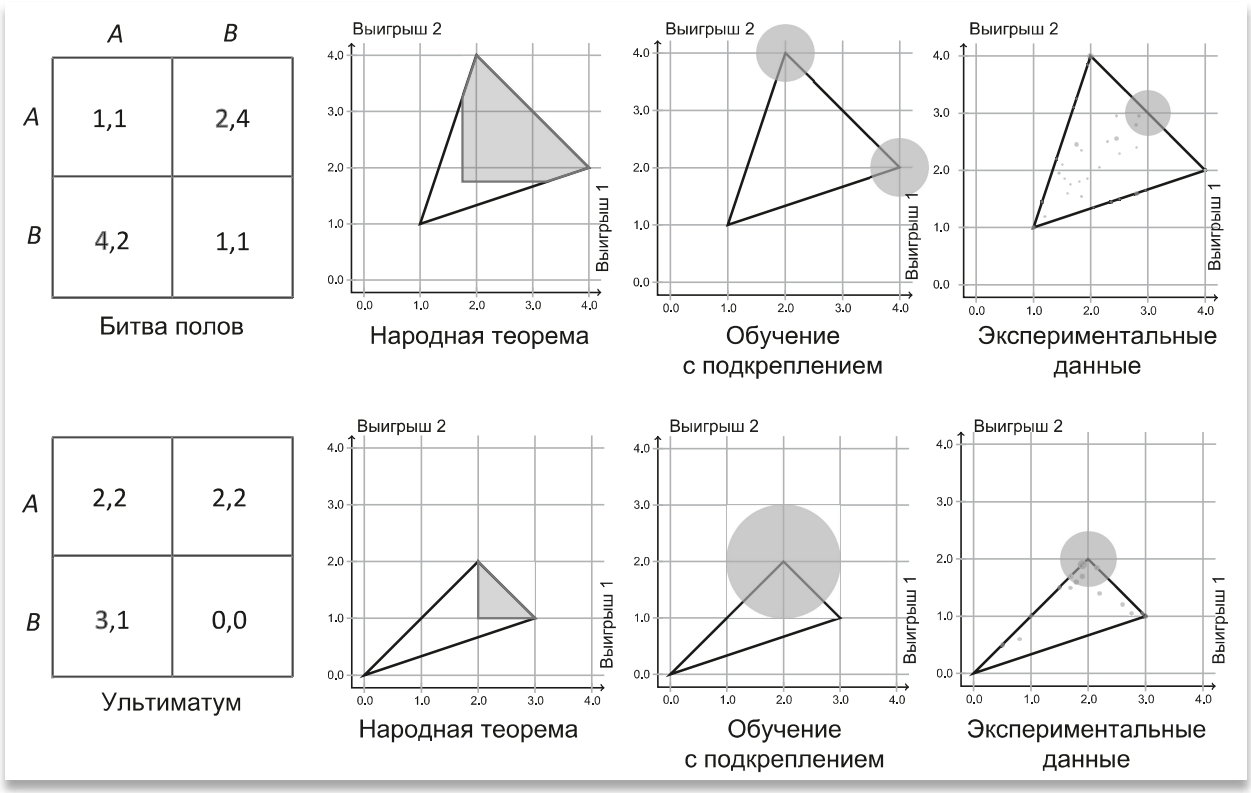

Рис. 2

Сравнительная динамика обучения с подкреплением по (Mathevet, Romero, 2012)

Журнал НЭА,
№4 (44), 2019,
с. 77–125

распределение платежей для экспериментальных данных. В последних двух случаях круги обозначают платежи с координатами в центре каждого круга, а диаметры каждого круга — частоты, с которыми популяция играет соответствующий профиль стратегий. Для удобства сопоставления единичный радиус соответствует всей популяции.

Можно заметить, что в первой игре результаты экспериментов не согласуются с прогнозом простого правила обучения с подкреплением. В игре «Ультиматум» обучение с подкреплением предсказывает половину результатов того, к чему приходят участники экспериментов. В данном случае ограничения рациональности в алгоритме улавливают тенденцию участников к играть более справедливый исход (2, 2), и усиливают ее. Это можно интерпретировать как шаг в сторону более точного описания поведения индивидов, хотя и требующий более точной настройки.

4.4. Обучение со взвешенным по опыту притяжением (EWA)

Оба подхода — и фиктивная игра, и обучение с подкреплением — отражают некоторые свойства реального обучения, но не описывают его целиком. Фиктивная игра обращает пристальное внимание на поведение оппонента, но не насколько (преположительно) наилучший отклик на это поведение действительно хорош. Обучение с подкреплением, напротив, не отслеживает поведения оппонента и концентрируется на успешности собственных действий. Поэтому естественным представляется поиск подхода, который бы объединял их в одном алгоритме. Такой алгоритм был создан и получил название опытно-взвешенного притяжения (Experience-weighted attraction, обычно в виде аббревиатуры EWA) (Camerer, Ho, 1999). Он позволил получить (как частные случаи) и фиктивную игру, и обучение с подкреплением, и их линейные комбинации. За это пришлось заплатить большим числом параметров, что вызвало критику о необходимости переобучения (overfitting) таких моделей.

Ответом на такую критику стала модель STEWA (Ho, Camerer, Chong, 2007), фиксирующая часть параметров на разумном уровне. Обучение со взвешенным по опыту притяжением — это одна из первых моделей, включающая элементы обучения с подкреплением и фиктивную игру с психологической интерпретацией. Наличие простых моделей внутри EWA удобно и по практическим соображениям, ведь они автоматически могут быть протестированы внутри модели, и если именно они точнее описывают поведение игроков, EWA должна это показать. Более того, также становится возможным пронаблюдать сходство и различия простых моделей в данных.

4.4.1. Формальная постановка

Базовые модельные предположения напоминают обучение с подкреплением. Стандартные предположения — повторяющаяся

Журнал НЭА,
№4 (44), 2019,
с. 77–125

игра $G=\langle \mathcal{I},S,\{u\},T\rangle$ с временем $t\in 1,...,T$ и J стратегиями. Вводятся два дополнительных обозначения: параметр, нормирующий опыт предыдущих периодов $N(t)$ (с начальным значением $N(0)$), и притяжение (attractions) $A_{ij}^{t}(s_{ij})$ вместо склонностей p_{ij}. Предыдущий опыт дисконтируется с коэффициентом ρ, и параметр $N(t)$ следует правилу $N(t)=\rho N(t-1)+1,\quad t>1$. $A_{ij}^{t}(s_{ij})$ — это притяжение стратегии j игрока i в момент времени t (с начальными притяжениями $A_{ij}^{0}(s_{ij})$). Подсчет и обновление A_{ij} включает три составляющих: дисконтирование старого притяжения $\phi N(t-1)A_{ij}^{t-1}(s_{ij})$ (ϕ — параметр дисконтирования), учет результата текущего раунда $u_{ij}(s_{ij}^{t},s_{-ij}^{t})$ и нормализацию опыта $N(t)$:

$$A_{ij}^{t}(s_{ij})=\frac{\phi N(t-1)A_{ij}^{t-1}+\left[\delta+(1-\delta)\,\mathbb{I}(s_{ij},s_{i}^{t})\right]u_{ij}(s_{ij}^{t},s_{-ij}^{t})}{\rho N_{i}(t-1)+1}, \qquad (2)$$

где $\mathbb{I}$ — индикатор использования стратегии, δ — определяет сравнительный вес выигрыша от выбранных и невыбранных стратегий в функции притяжения. Например, если $\delta=0$, то они будут учитываться, как в модели обучения с подкреплением (учитываем выбранную стратегию), а если $\delta=1$, то как в модели фиктивной игры (учитываем все стратегии). $A_{ij}^{0}(s_{ij})$ и $N(0)$ позволяют регулировать скорость обучения в первые раунды игры или асимметрию в привлекательности начальных стратегий, отражая начальные знания игрока.

По аналогии с обучением с подкреплением, вероятность выбора стратегии j в раунд t задается функцией выбора, зависящей от притяжений. В оригинальной статье (Camerer, Ho, 1999) предлагается несколько способов задания вероятности, но за основную берется логистическая $p_{ij}^{t+1}=e^{\lambda A_{ij}^{t}}\Big/\sum_{w=1}^{k}e^{\lambda A_{iw}^{t}}$. Всего в модели EWA насчитывается 6 параметров (см. табл. 1), поэтому модель легко подгоняет многие возможные особенности, что подтверждается симуляциями (Salmon, 2001). Несмотря на большое число параметров, EWA демонстрирует лучшие результаты оценки, чем более простые аналоги, даже с введенными штрафами на их число. В (2) ρ и ϕ отвечают за дисконтирование; $N(0)$ — сила предыдущего опыта; δ — вес гипотетического выигрыша; $A_{ij}^{0}(s_{ij})$ — форма предыдущего притяжения; λ — чувствительность к притяжению.

Напомним, но уже в табличном виде отношения между EWA и вложенными в нее моделями (при $\delta=1$ параметр $N(0)$ не имеет значения и может быть любым) (табл. 8).

Для иллюстрации сложности интерпретации динамики обучения со взвешенным по опыту притяжением снова воспользуемся сравнением среднего выигрыша для теоретического предсказания, результатов симуляции двух EWA-игроков и результатов игры участников экспериментов из (Mathevet, Romero, 2012). Уточненное описание процедуры проведения эксперимента и компьютерных симуляций приведено в п. 3.4.

Журнал НЭА,
№4 (44), 2019,
с. 77–125

Таблица 8

Ограничения на параметры EWA, соответствующие важным частным случаям

ϕ	δ	ρ	$N(0)$	Модель
1	1	1	–	Фиктивная игра
0	1	0	–	Наилучший отклик по Курно
$\phi \in (0,1)$	1	ϕ	–	Взвешенная фиктивная игра
$\phi \in [0,1]$	0	0	1	Кумулятивное подкрепление
$\phi \in [0,1]$	0	ϕ	$1/(1-\phi)$	Усредненное подкрепление

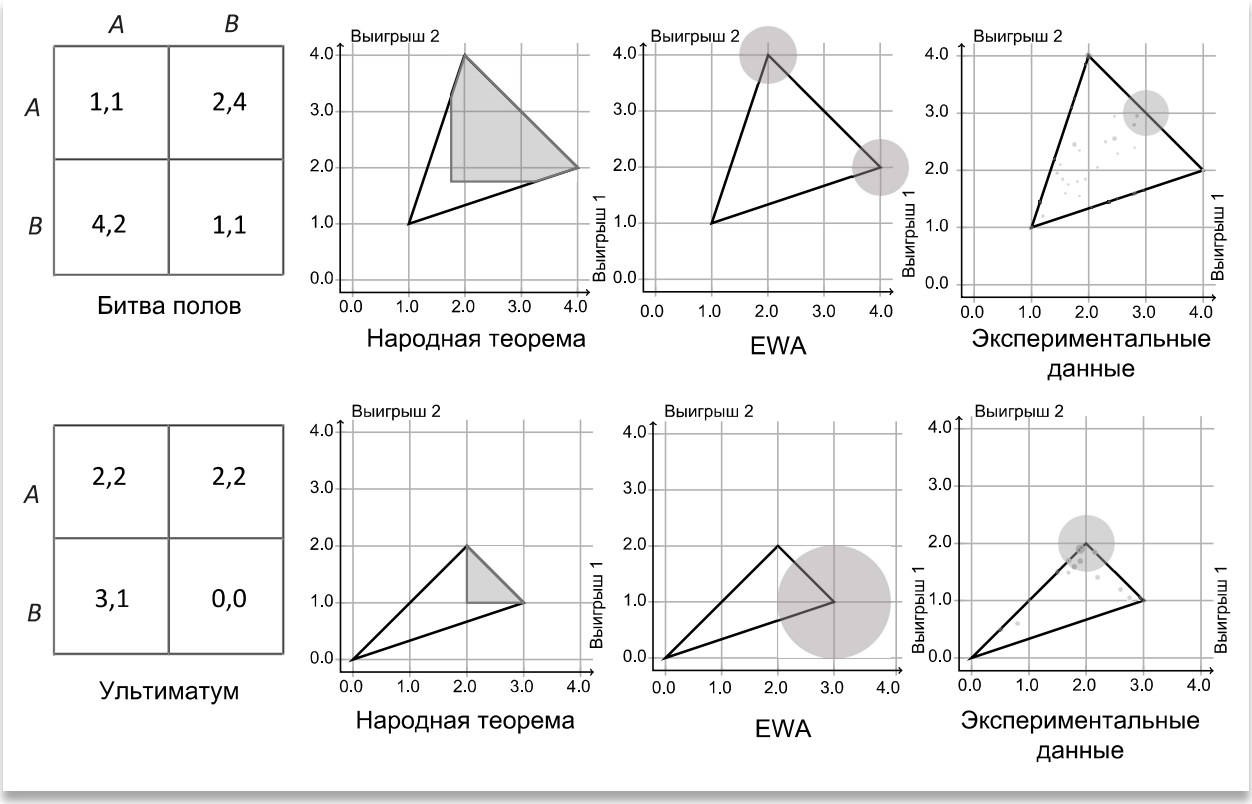

Рис. 3

Сравнительная динамика EWA по (Mathevet, Romero, 2012)

На рис. 3 в левом столбце приведены игры в матричном виде. Множества возможных платежей для каждой игры описываются замкнутыми контурами, представленными на графиках справа от матрицы. Во втором столбце – множество платежей, которые доминируют равновесные платежи в смешанных стратегиях в соответствии с народной теоремой. В третьем столбце – сходимость пар алгоритмов взвешенного по опыту притяжения (EWA), а в последнем – распределения плате-

Журнал НЭА,
№4 (44), 2019,
с. 77–125

жей для экспериментальных данных. В последних двух случаях кругами обозначены платежи профиля стратегий, соответствующего центру каждого круга. Диаметры каждого круга отражают частоту, с которой популяция играет соответствующий профиль стратегий. Для удобства сопоставления единичный радиус соответствует всей популяции.

Результаты симуляций, как и в случае с предыдущими алгоритмами, демонстрируют несовпадение с экспериментами. Теоретически обучение с взвешенным по опыту притяжением могло бы давать предсказание, близкое к обучению с подкреплением, при параметрах $\delta = 0$, однако при средних параметрах оно демонстрирует динамику, больше похожую на фиктивную игру. Такой результат можно интерпретировать и как излишнее усложнение, поскольку алгоритм более склонен использовать слишком много параметров для достаточно узкого пространства стратегий.

4.4.2. Самонастраивающаяся модель EWA (STEWA)

Модифицированная модель STEWA (self-tuning EWA) (Camerer, Ho, Chong, 2007) содержит только один параметр, а остальные самонастраиваются. Начальные параметры ($N(0)$, A_{ij}^0) приравниваются к 1, поскольку скорее отражают характеристику стратегической ситуации, чем процесс обучения. Оставшиеся два параметра дисконтирования (ϕ,ρ) были объединены в $\phi_i(t)$. Без изменений остаются δ и λ. Учет предыдущего опыта в модифицированной модели задается детектором изменений — функцией $\phi_i(t)$ на основе индекса неожиданности. Он отражает разницу между действиями игроков на протяжении всей игры и последними действиями.

Определим вектор средней истории игры игроков-оппонентов, содержащий относительную частоту игры их J стратегий, его элементом будет $\eta_{ij}(t) = \sum_{t=1}^{T} \mathbb{I}(s_{ij}, s_{-i}(t)) \,/\, t$.

Последняя фактическая история игры образует другой вектор $r_{ij}(t) = \mathbb{I}(s_{ij}, s_{-i}(t))$. Индекс неожиданности $\Omega_i(t)$ — это сумма квадратов отклонений между этими двумя векторами $\Omega_i(t) = \sum_{j=1}^{J} (\eta_{ij}(t) - r_{ij}(t))^2$. Он принимает значения между нулем (стабильная среда) и двумя (максимально неожиданный исход). Другими словами, индекс неожиданности отражает степень изменения самого последнего наблюдения относительно стабильной истории. Теперь функция детектора изменений задается как $\phi_i(t) = 1 - \Omega_i(t) / 2$.

Приведем пример из (Camerer, Ho, Chong, 2007). Представим, что ваш оппонент постоянно играет одну и ту же стратегию на протяжении нескольких раундов, а потом внезапно меняет ее. В этом случае $\phi_i(t)$ будет равен $(2t-1) / t^2$, и чем больше повторяется выбор, тем большей неожиданностью будет появление другой стратегии (для $t = 2,\ 3,\ 5$ и 10 значение составит 0,75; 0,56 и 0,19).

Журнал НЭА,
№4 (44), 2019,
с. 77–125

Функция внимания, как еще один самонастраивающийся параметр, имеет схожую, психологически обоснованную природу. Ее идея состоит в том, что внимание игрока переключается на стратегию, которая не была выбрана, но гипотетический выигрыш по ней превысил текущий. Функция внимания $\delta(t)$ равна 1, если $u_i(s_{ij}, s_{-i}(t)) \geq u_i(t)$ и 0 — в обратном случае.

Модель STEWA можно считать вершиной классической теории моделей обучения в играх, однако она не позволяет ответить на многие естественные вопросы и оставляет широкое поле для дальнейших исследований. Одним из таких вопросов будет формальное определение и изучение экспериментирования в играх — поведения, которое не является локально оптимальным, но оптимально в долгосрочном периоде, поскольку позволяет больше узнать о поведении и реакции оппонента на действия игрока. Как мы видели на протяжении обзора, позитивные результаты в теории обучения так или иначе указывают на необходимость такого поведения для достижения равновесия. Однако его сложно исследовать, поскольку, чтобы заметить экспериментирование как отклонение от оптимального поведения, нужно определить, какое поведение оптимально в рамках модели. Для разных моделей оптимальным (и, соотвественно, экспериментальным) может оказаться разное поведение даже на одинаковом наборе данных. Мы завершим обзор теоретических моделей на одном примечательном современном примере модели с явным экспериментированием.

4.5. I-SAW (инерционность и случайность в EWA)

Модель I-SAW (Erev, Haruvy, 2013) разрабатывалась как альтернатива подходу EWA, так же основываясь на принципе взвешивания стратегий и отражая ряд психологических регулярностей, замеченных в экспериментах. В каждый период игрок с разной вероятностью попадает в один из режимов. Таких режимов реагирования три: исследование, инерция и эксплуатация.

4.5.1. Модель

Снова используется игра $G = \langle \mathcal{I}, S, \{u\}, T \rangle$ двух игроков с двумя стратегиями: $I = 2$ и $S_i = \{s_{i1}, s_{i2}\}$. Вектор истории игры задается отношением $h^t = \{s^1, \ldots, s^t\}$ и h_{-1}. Каждый игрок описывается набором параметров $(\rho_A, \varepsilon, \pi_i, \mu_i, \iota, w_i)$, среди которых только $\rho_A \in [0, 1]$ одинаков, а $\varepsilon_A \sim U[0, \varepsilon]$, $\varepsilon_A \sim U[0, \varepsilon]$, $\pi_i \sim U[0, \pi]$, $\mu_i \sim U[0, \mu]$, $\iota_i \sim U[0, \iota]$, $w_i \sim U[0, w]$ меняются от игрока к игроку.

Поведение игроков делится на три фазы: исследование, эксплуатация и инерция. Общая схема работы модели представлена на рис. 4.

Исследование. Под исследованием будет пониматься случайный выбор стратегии s_{i1} с вероятностью ρ_i и стратегии s_{i2} с вероятностью $1 - \rho_i$.

Журнал НЭА,
№4 (44), 2019,
с. 77–125

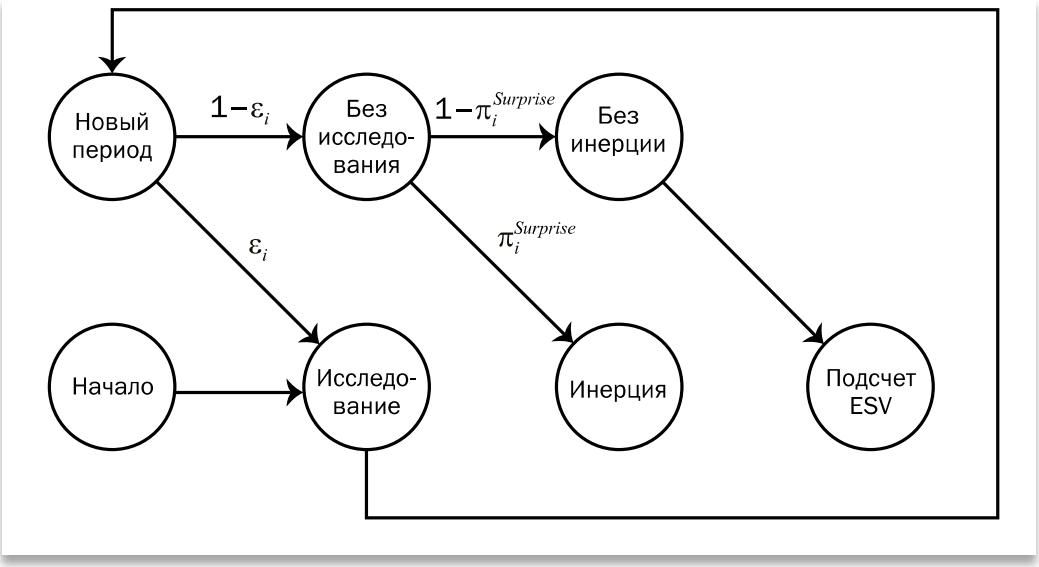

Рис. 4

Схематичное представление модели I-SAW

Эксплуатация. Зададим функцию оцененной субъективной ценности (Estimated Subjective Value (ESV)) для обеих стратегий, максимальное значение которой игрок и будет использовать для выбора текущей стратегии. ESV от альтернативной стратегии j для периода $t>1$ составляет $ESV(j,t)=(1-w_i)(\overline{S})+w_i(\overline{G})$, где $\overline{G}$ (grand mean) — средний выигрыш от альтернативы j на всем множестве $(t-1)$ предыдущих испытаний, $\overline{S}$ (sample mean) — средний выигрыш от альтернативы j в μ_i последних испыаниях (например μ_i может принимать значения 1, ..., 4, но всегда меньше, чем T); μ_i и w_i — индивидуальные параметры игроков.

Инерция. Люди склонны повторять свой выбор (инерционность игрока, что и определяет название режима). Когда игра предсказуема для игрока, он чаще повторяет предыдущий ход. В модели склонность к повторению задается как убывающая функция от неожиданности (surprise). Переход игрока в режим инерции после второго периода $t>1$ определяется вероятностью $\varphi_i^{Surprise(t)}$: $p(Inertia(t+1))=\varphi_i^{Surprise(t)}$, где $0<\varphi_i<1$ — индивидуальный параметр склонности к инерции, возводимый в степень функции неожиданности (*Surprise*). Она задается следующим образом: чем больше разрыв (*Gap*) между ожидаемым и полученным исходом, тем больше неожиданность (*Surprise*):

$$Gap(t)=\frac{1}{4}\left[\sum_{j=1}^{2}\left|s_j(t-1)-s_j(t)\right|+\sum_{j=1}^{2}\left|\overline{G}(t-1)-s_j(t)\right|\right],$$

$$\overline{Gap}(t+1)=\overline{Gap}(t)(1-1/\chi)+Gap(t)(1/\chi),$$

$$Surprise(t)=Gap(t)/[\overline{Gap}(t)+Gap(t)],$$

где χ — ожидаемое число испытаний.

Журнал НЭА,
№4 (44), 2019,
с. 77–125

Такая парадигма «неожиданность—триггер—смена», по мнению (Nevo, Erev, 2012), объясняет эффекты большего веса недавней истории игры, наблюдаемые в экспериментах с людьми.

5. Эмпирические исследования моделей обучения

После перечисления различных моделей, естественно, возникает вопрос, какая модель правильная. С одной стороны, этот вопрос можно задать просто как нормативный, как правильно действовать в игровой ситуации с неопределенностью того или иного рода. С другой стороны, ответ на нормативный вопрос требует сначала ответ на позитивный вопрос, как понять, какой моделью руководствуется оппонент. К сожалению, на текущий момент в исследованиях этого вопроса заметна стагнация.

В поведении участников экспериментов в повторяющихся играх обнаружены эмпирические регулярности, не объясняемые совместно ни одной известной моделью. В качестве примеров можно привести недооценку участниками вероятности наступления редких событий (Kahneman, Tversky, 1979), инерцию (склонность повторять предыдущий выбор) (Suppes, Atkinson, 1960), эффект новизны (Estes, 1964; Nevin, 1988; Erev, Haruvy, 2005; Cooper, Kagel, 2008).

Авторы (Erev, Roth, 1999) выделяют три когнитивных фактора принятия решений в повторяющейся игре в экспериментальных условиях. Первый фактор связан со способом задания когнитивных стратегий игрока — они могут формироваться как выбор одного из доступных действий (например, выбрать правую или левую клавишу), но могут включать более сложные построения (например ответить взаимностью, имитировать последнее действие оппонента и пр.). Второй фактор — соотношение между исследованием и эксплуатацией в демонстрируемом поведении (т.е. того как достигается компромисс между сбором и использованием имеющейся информации для достижения наилучших результатов). Третий фактор — это анализ человеком предыдущего опыта. Чтобы эмпирические закономерности лучше учитывались моделями, исследователями было создано несколько более сложных и параметризированных моделей, в число которых, например, входят Experience Weighted Attraction (EWA) и ее модификации (Camerer, Ho, 1999), γ-Weighted Beliefs (Cheung, Friedman, 1997) и I-SAW (Nevo, Erev, 2012).

Эмпирически обоснованные модели (так же, как и более простые модели обучения) склонны не учитывать первые два фактора (Erev, Haruvy, 2013). Это происходит во многом потому, что они чрезмерно зависят от контекста. Поэтому ряд предположений ограничивает применение таких моделей. Во-первых, в повторяемых стратегических ситуациях для стратегий, рассматриваемых агентами, нужна возможность аппроксимации элементарным набором действий. Во-вторых, исследование игры не должно приносить дополнительную информа-

Журнал НЭА,
№4 (44), 2019,
с. 77–125

цию (что справедливо, например, для игр на общее благо, однако уже существенно ограничивает интерпретируемость результатов, полученных в рамках изучения поведения людей в социальных дилеммах).

Группа авторов (Arifovic, McKelvey, Pevnitskaya, 2006) изучала работу известных алгоритмов (включая фиктивную игру, обучение с подкреплением и взвешенное по опыту притяжение) на экспериментальных данных (McKelvey, Palfrey, 2001). В них 48 участников сыграли 24 сессии по 50 раундов следующих игр: дилемма заключенного, игра в труса (chicken), битва полов (2 × 2 и 3 × 3), охота на оленя (stag hunt), стратегическая версия игры в ультиматум и сороконожка (centipede). Адаптивные модели обучения демонстрировали относительный успех в описании поведения субъектов в играх с единственным равновесием в смешанных стратегиях и некоторых координационных играх, но в то же время совершенно не согласовывались с поведением человека в других играх. Например, люди в повторяющейся дилемме заключенного с полной информацией часто играют «сотрудничество», даже с учетом различий в конфигурациях выигрыша. При этом фиктивная игра сходится к «предательству» всегда, независимо от значений параметров. Схожие результаты наблюдались для игры битва полов — люди быстро учатся кооперироваться. Различия в наблюдаемом поведении алгоритмов и людей были настолько большими, что авторы призвали других исследователей разработать новые модели обучения, которые больше бы опирались на экспериментальные данные.

Особые эконометрические проблемы обучения в играх отмечены в работах (Salmon, 2001; Hopkins, 2002). В (Salmon, 2001) утверждается, что при попытке сравнения различных моделей обучения в условиях, приближенных к экспериментальным, возникают значительные трудности статистического характера. Для оценки этих трудностей стоит остановиться на методе, изложенном в статье (Salmon, 2001) — сначала при помощи симуляций авторами были созданы выборки экспериментальных наблюдений, соответствующих различным моделям, которые затем оценивались на этих же данных. Основная идея состояла в большем сходстве реальной модели (при учете поправки на число параметров) с данными, действительно порожденными этой моделью. Выяснилось, что различить модели таким образом оказалось невозможно.

Вопрос о сравнении моделей можно поставить и иначе. Чем теоретически должны отличаться модели, что потом позволило бы различить их на практике? Работа (Hopkins, 2002) показывает, что стохастическая версия фиктивной игры и обучение с подкреплением содержательно (идентифицируемо) отличаются только большей скоростью обучения в случае стохастической фиктивной игры, а порождаемое ими асимптотическое поведение во многих случаях неразличимо.

Обычной интерпретацией этих результатов является идея бесполезности изучения обучения в играх с помощью экспериментов (про-

Журнал НЭА,
№4 (44), 2019,
с. 77–125

тивопоставленных теории и симуляционным методам). Неудивительно, что на горизонте в сотни наблюдений и дискретном выборе из двух-трех альтернатив тяжело различить сложные спецификации моделей.

Однако на наш взгляд, вывод о полной невозможности экспериментальной оценки моделей обучения слишком поспешен. Например, в обеих ключевых работах не рассматривается важный вопрос предиктивной силы моделей. Не вполне ясно, почему нужно отличать модели, дающие одинаковые предсказания, а для моделей, имеющих разную предсказательную силу, именно она должна служить обоснованием превосходства конкретной модели.

Заметим, что условия экспериментов, которыми ограничился (Salmon, 2001), не исчерпывают всех возможностей этого метода исследований, например использованных в (Duersch et al., 2010; Сусин, Чернов, 2018) программируемых оппонентов. Наличие роботов в игре против экспериментальных субъектов потенциально позволяет значительно улучшить идентификацию моделей, хотя степень этого возможного улучшения остается открытым вопросом.

6. Заключение

Несмотря на ряд полученных теоретических результатов о динамике поведения рациональных правил обучения в конкретных играх (Fudenberg, Levine, 2009), универсальность этих результатов все еще остается проблемой для экономической теории. Если рациональные игроки способны успешно обучиться в конкретной игре, неизвестно, насколько можно упростить модель их поведения, сохраняя этот успешный результат. Если же игроки не сходятся к равновесию, остается неотвеченным вопрос, могут ли более простые правила оказаться и более стабильными.

Отметим одно, на наш взгляд, ценное обобщающее наблюдение — теория сходимости моделей обучения раз за разом показывает, что «лучшее — враг хорошего». Игрок, который пытается играть единственный наилучший отклик на каждом ходу (что, казалось бы, соответствует критерию оптимальности), не обязательно достигнет долговременного равновесного поведения, в то время как игрок, допускающий изредка случайные, экспериментальные ходы, в пределе гарантированно его достигнет. Это справедливо для широких классов алгоритмов — и для фиктивной игры, и для байесовского обучения в целом (особенно в сравнении с обучением с подкреплением), и для калибрации.

На наш взгляд, по мере развития машинного обучения и автоматизации экономических процессов можно ожидать, что выбор экономической координации, эффективности и создания новых экономических механизмов, институтов и регуляций все больше будет полагаться на явно сформулированные модели обучения вовлеченных в экономические процессы агентов. Существующие и активно развивающиеся подходы обучению, прежде всего в компьютерных науках, должны будут

Журнал НЭА,
№4 (44), 2019,
с. 77–125

учитывать взаимодействие и взаимное построение некоторой модели оппонента экономическими агентами.

Равновесный подход уже успешно применяется в эмпирических исследованиях (Bajari, Hong, Nekipelov, 2013), но и модели обучения мало-помалу находят применение, например в эмпирических исследованиях теории отраслевых рынков по мере все большего внимания к динамическим моделям. Так, например, в статье (Doraszelski, Lewis, Pakes, 2018) модели обучения с подкреплением и фиктивной игры позволяют лучше оценить динамику рынка электроэнергии Великобритании, чем равновесные модели.

Оптимистичную оценку можно дать и попыткам ряда исследователей построить сложные модели обучения с привязкой к экспериментальным данным (Ioannou, Romero, 2014), которые открывают возможность поиска и аппроксимации поведенческих стратегий конечными автоматами, потенциал которых был отмечен еще в (Aumann, 1987).

ПРИЛОЖЕНИЕ

П.1. Байесовский алгоритм в игре против природы

Рассмотрим численно классический пример байесовского обучения — определение смещения монеты по наблюдениям. Сформулируем игру.

Природа создает монету с вероятностью выпадения орла r и затем подбрасывает эту монету N раз, генерируя ряд действий. Игрок перед каждым действием загадывает одну из сторон; он забирает единичный выигрыш при успехе, в противном случае его выигрыш равен 0. Порождаемый Природой ряд значений описывается распределением Бернулли, а функция правдоподобия — биномиальным распределением с параметрами $\theta: r$ — отражает вероятность выпадения орла (смещенность монеты), h — количество выпадения орла за N периодов и количество выпадения решки, обозначаемым за t (соответственно, $N = t + h$).

То есть вероятность, что r принимает какое-то известное значение, при известных данных $x(h = H)$ выглядит как

$$Pr\left(h = H \middle| \; r, h+t\right) = \binom{h+t}{h} r^h (1-r)^t,$$

где H — значение выпадения орлов в генеральной совокупности.

Априорное распределение плотности вероятности r обозначим $g(r) \in [0,1]$. Апостериорное распределение получается произведением функций правдоподобия $r \to p(t,h \mid r)$ и априорного распределения $g(r)$, нормализованного на вероятность выпадения орлов $p(t,h)$ в n испытаниях:

$$p(r \mid t = T, h = H) = \frac{p(t,h \mid r) g(r)}{\int p(t,h \mid r') g(r') dr'} = \frac{Pr(h = H \mid r, h+t) g(r)}{\int_0^1 Pr(h = H \mid r', h+t) g(r') r'}.$$

Поставляя формулу биномиального распределения в формулу апостериорного распределения, имеем

Журнал НЭА,
№4 (44), 2019,
с. 77–125

$$f(r \mid t=T, h=H)) = \binom{h+t}{h} r^h (1-r)^t \Big/ \left[\int_0^1 \binom{h+t}{h} r'h(1-r')dr' \right] =$$

$$= r^h (1-r)^t \Big/ \left[\int_0^1 r'h(1-r')^t dr' \right].$$

Это распределение, сопряженное априорное для биномиального распределения, называется бета-распределением, в общем случае его знаменатель выражается через бета-функцию:

$$f\left(r \middle| \ t=T, h=H\right) = \frac{1}{B(h+1, t+1)} r^h (1-r)^t.$$

Когда игрок максимизирует выигрыш, ему следует загадывать сторону с большей (согласно априорному распределению) вероятностью. Предположим, что априорное распределение r — равномерное на $[0,1]$, т.е. $g(r)=1$. Возьмем так же в качестве примера следующие значения состояния игры, пусть $n=10, h=7$, т.е. монета подброшена 10 раз и выпало 7 орлов. Что выбрать на 11 ходу? Поскольку h и t являются целыми числами, а априорное распределение — равномерно, формулу апостериорного бета-распределения также можно записать факториалами:

$$f\left(r \middle| \ t=T, h=H\right) = \frac{(t+h+1)!}{h!\,t!} r^h (1-r)^t = \frac{(10+1)!}{7!3!} r^7 (1-r)^3 = 1320 r^7 (1-r)^3,$$

$f(r \mid H=7, T=3)$ достигает своего пика при $r = h/(h+t) = 0{,}7$. Ожидаемое значение r при заданном распределении составит

$$\mathbb{E}(r) = \int_0^1 r\, f(r \mid H=7, T=3) dr = \frac{h+1}{h+t+2} = \frac{2}{3}.$$

Это значит, что байесовский игрок, выбирая наиболее вероятное событие, должен поставить в 11 ходу на орла.

ЛИТЕРАТУРА

Опойцев В.И. (1977). Равновесие и устойчивость в моделях коллективного поведения. М.: Наука.

Сусин И.С., Чернов Г.В. (2018). Распознавание эвристик и обучение в игре «Камень, ножницы, бумага»: экспериментальный подход // *Журнал экономической теории*. № 3. С. 408–420.

Anufriev M., Arifovic J., Ledyard J., Panchenko V. (2013). Efficiency of continuous double auctions under individual evolutionary learning with full or limited information // *Journal of Evolutionary Economics*. Vol. 23(3). P. 539–573.

Arifovic J., Ledyard J. (2004). Scaling up learning models in public good games // *Journal of Public Economic Theory*. Vol. 6(2). P. 203–238.

Arifovic J., Ledyard J. (2018). Learning to alternate // *Experimental Economics*. Vol. 21(3). P. 692–721.

Arifovic J., McKelvey R.D., Pevnitskaya S. (2006). An initial implementation of the Turing tournament to learning in repeated two-person games // *Games and Economic Behavior*. Vol. 57(1). P. 93–122.

Aumann R.J. (1987). Correlated equilibrium as an expression of Bayesian rationality // *Econometrica: Journal of the Econometric Society*. Vol. 55. P. 1–18.

Журнал НЭА, №4 (44), 2019, с. 77–125

Bajari P., Hong H., Nekipelov D. (2013). Game theory and econometrics: A survey of some recent research. Advances in Economics and Econometrics. 10th world congress. Vol. 3. P. 3–52.

Basu K., Weibull J.W. (1991). Strategy subsets closed under rational behavior // *Economics Letters*. Vol. 36(2). P. 141–146.

Beggs A.W. (2005). On the convergence of reinforcement learning // *Journal of Economic Theory*. Vol. 122(1). P. 1–36.

Bellman R. (1957). A Markovian decision process // *Journal of Mathematics and Mechanics*. P. 679–684.

Benaim M., Hirsch M. (1999). Mixed equilibria and dynamical systems arising from fictitious play in perturbed games // *Games Econ. Behav*. Vol. 29. P. 36–72.

Bergemann D., Välimäki J. (2008). Bandit problems. In: *The New Palgrave Dictionary of Economics*. Vol. 1–8. P. 336–340.

Brandenburger A. (1996). Strategic and structural uncertainty in games. In: Zeckhauser R.J., Keeney R.L., Sebenius J.K. (eds). "*Wise Choices: Games, Decisions, and Negotiations*". Brighton: Harvard Business School Press. P. 221–232.

Brown G.W. (1951). Iterative solution of games by fictitious play // *Activity Analysis of Production and Allocation*. Vol. 13(1). P. 374–376.

Bubeck S., Cesa-Bianchi N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems // *Foundations and Trends in Machine Learning*. Vol. 5(1). P. 1–122.

Bush R.R., Mosteller F. (1955). Stochastic models for learning. Oxford, England: John Wiley & Sons, Inc.

Camerer C., Ho H.T. (1999). Experience-weighted attraction learning in normal form games // *Econometrica*. Vol. 67(4). P. 827–874.

Cason T.N., Friedman D. (1999). Learning in a laboratory market with random supply and demand // *Experimental Economics*. Vol. 2(1). P. 77–98.

Cesa-Bianchi N., Lugosi G. (2006). Prediction, learning, and games. Cambridge: Cambridge University Press.

Cheung Y.-W., Friedman D. (1997). Individual learning in normal form games: Some laboratory results // *Games and Economic Behavior*. Vol. 19. P. 46–76.

Cooper D.J., Kagel J.H. (2008). Learning and transfer in signaling games // *Economic Theory*. Vol. 34(3). P. 415–439.

Cournot A. (1960). Researches into the mathematical principles of the theory of wealth. Trans. by N. Bacon. New York: Kelly. (Originally published by 1838).

Cressman R., Ansell C., Binmore K. (2003). Evolutionary dynamics and extensive form games. Vol. 5. Cambridge, MA: MIT Press.

Doraszelski U., Lewis G., Pakes A. (2018). Just starting out: Learning and equilibrium in a new market // *American Economic Review*. Vol. 108(3). P. 565–615.

Duersch P., Kolb A., Oechssler J., Schipper B.C. (2010). Rage against the machines: How subjects play against learning algorithms // *Economic Theory*. Vol. 43(3). P.407–430.

Ebbinghaus H. (2013). Memory: A contribution to experimental psychology // *Annals of Neurosciences*. Vol. 20(4). P. 155.

Erev I., Gopher D., Itkin R., Greenshpan Y. (1995). Toward a generalization of signal

Журнал НЭА, №4 (44), 2019, с. 77–125

detection theory to N-person games: The example of two-person safety problem // *Journal of Mathematical Psychology*. Vol. 39(4). P. 360–375.

Erev I., Haruvy E. (2005). Generality and the role of descriptive learning models // *Journal of Mathematical Psychology*. Vol. 49(5). P. 357–71.

Erev I., Haruvy E. (2013). Learning and the economics of small decisions. In: The Handbook of Experimental Economics. Vol. 2. Roth A.E., Kagel J. (eds). Princeton: Princeton University Press.

Erev I., Roth A. (1998). Predicting how people play games: Reinforcement learning in games with unique strategy mixed-strategy equilibrium // *American Economic Review*. Vol. 88. P. 848–881.

Erev I., Roth A.E. (1999). On the role of reinforcement learning in experimental games: The cognitive game-theoretic approach. In: *Games and Human Behavior: Essays in Honor of Amnon Rapoport*, edited by D. Budescu, I. Erev, and R. Zwick, 79–104. Lawrence Erlbaum Associates.

Estes W.K. (1964). Probability learning. In: "*Categories of human learning*". Cambridge: Academic Press. P. 89–128.

Foster D.P., Vohra R. (1998). Asymptotic calibration // *Biometrika*. Vol. 85. P. 379–390.

Foster D.P., Vohra R.V. (1997). Calibrated learning and correlated equilibrium // *Games and Economic Behavior*. Vol. 21(1–2). P. 40.

Foster D.P., Young H.P. (2001). On the impossibility of predicting the behavior of rational agents // *Proceedings of the National Academy of Sciences of the USA*. Vol. 98. No. 222. P. 12848–12853.

Fudenberg D., Levine D.K. (1993). Steady state learning and Nash equilibrium // *Econometrica: Journal of the Econometric Society*. Vol. 61. No. 3. P. 547–573.

Fudenberg D., Levine D.K. (1998). The theory of learning in games. Vol. 2. Cambridge, MA: MIT Press.

Fudenberg D., Levine D.K. (1998b). Learning in games // *European economic review*. Vol. 42(3–5). P. 631–639.

Fudenberg D., Levine D.K. (2009). Learning and equilibrium // *Annual Review of Economics*. Vol. 1(1). P. 385–420.

Gittins J.C. (1979). Bandit processes and dynamic allocation indices // *Journal of the Royal Statistical Society: Series B (Methodological)*. Vol. 41(2). P. 148–164.

Güth W., Schmittberger R., Schwarze B. (1982). An experimental analysis of ultimatum bargaining // *Journal of Economic Behavior and Organization*. Vol. 3. P. 367–388.

Hannan J. (1957). Approximation to Bayes' risk in repeated play. In: Dresher M., Tucker A.W., Wolfe P. (eds). *Contributions to the theory of games*. Vol. 3. Princeton: Princeton Univ. Press. P. 97–139.

Hart S., Mas-Colell A. (2001). A general class of adaptive strategies // *Journal of Economic Theory*. Vol. 98. P. 26–54.

Ho T.H., Camerer C.F., Chong J.K. (2007). Self-tuning experience weighted attraction learning in games // *Journal of Economic Theory*. Vol. 133(1). P. 177–198.

Hofbauer J., Hopkins Ed. (2005). Learning in perturbed asymmetric games // *Games and Economic Behavior*. Vol. 52(1). P. 133–152.

Журнал НЭА, №4 (44), 2019, с. 77–125

Hofbauer J., Sigmund K. (1998). Evolutionary games and population dynamics. Cambridge: Cambridge University Press.

Hopkins Ed. (2002). Two competing models of how people learn in games // *Econometrica*. Vol. 70(6). P. 2141–2166.

Hutteger S.M. (2017). The probabilistic foundations of rational learning. Cambridge: Cambridge University Press.

Ioannou C.A., Romero J. (2014). A generalized approach to belief learning in repeated games // *Games and Economic Behavior*. Vol. 87. P. 178–203.

Kahneman D., Tversky A. (1979). On the interpretation of intuitive probability: A reply to Jonathan Cohen // *Cognition*. Vol. 7(4). P. 409–411.

Kalai E., Lehrer E. (1993). Rational learning leads to Nash equilibrium // *Econometrica*. Vol. 61. P. 1019–1045.

Learning as a rational foundation for macroeconomics and finance (2013). In: Frydman R., Phelps E.S. (eds). *Rethinking expectations: The way forward for macroeconomics*. Princeton: Princeton University Press. P. 68–112.

Marimon R. (1997). Learning from learning in economics. In: Kreps D., Wallis K. (eds). "*Advances in economics and econometrics: Theory and applications*". Vol. 1. P. 278–315.

Mathevet L., Romero J. (2012). Predictive repeated game theory: Measures and experiments. Mimeo.

McKelvey R., Palfrey T.R. (2001). Playing in the dark: Information, learning, and coordination in repeated games. Mimeo.

Miyasawa K. (1961). On the convergence of the learning process in a 2 x 2 non-zero-sum game. Economic research program. Princeton University. Research Memorandum No. 33.

Nachbar J. (2009). Learning in games. In: "*Encyclopedia of complexity and systems science*". New York: Springer. P. 5177–5188.

Nachbar J.H. (1990). "Evolutionary" selection dynamics in games: Convergence and limit properties // *International Journal of Game Theory*. Vol. 19(1). P. 59–89.

Nachbar J.H. (2005). Beliefs in repeated games // *Econometrica* . Vol. 73(2). P. 459–480.

Nagel R. (1995). Unraveling in guessing games: an experimental study // *American Economic Review*. Vol. 85 (5). P. 1313–1326.

Nevin J.A. (1988). Behavioral momentum and the partial reinforcement effect // *Psychological Bulletin*. Vol. 103. P. 44–56.

Nevo I., Erev I. (2012). On surprise, change, and the effect of recent outcomes // *Frontiers in psychology*. Vol. 3. P. 24.

Ramsey F.P. (1926). Truth and probability. In: Braithwaite R. (ed.) "*The Foundations of Mathematics and Other Logical Essays*". London: Kegan Paul. P. 156–198.

Robbins H. (1952). Some aspects of the sequential design of experiments // *Bulletin of the American Mathematical Society*. Vol. 58(5). P. 527–535.

Robinson J. (1951). An iterative method of solving a game // *Annals of Mathematics*. Vol. 54. P. 296–301.

Roth A.E., Erev I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate terym // *Games Econ. Behav.* Vol. 8. P. 164–212.

Журнал НЭА, №4 (44), 2019, с. 77–125

Rothschild M. (1974). A two-armed bandit theory of market pricing // *Journal of Economic Theory*. Vol. 9. P. 185–202.

Sadrieh A. (1998). The alternating double auction market: A game theoretic and experimental investigation. Vol. 466. Berlin: Springer Science & Business Media.

Salmon T.C. (2001). An evaluation of econometric models of adaptive learning // *Econometrica*. Vol. 69(6). P. 1597–1628.

Sandholm W. (2010). Population games and evolutionary dynamics. Cambridge, MA: MIT Press.

Savage L.J. (1954). The foundations of statistics. New York: Wiley.

Selten R. (2004). Learning direction theory and impulse balance equilibrium. In: Cassar A., Friedman D. (eds). *Economics lab: An intensive course in experimental economics*. P. 133.

Selten R., Abbink T., Cox R. (2005). Learning direction theory and the winner's curse // *Experimental Economics*. Vol. 8(1). P. 5–20.

Selten R., Buchta J. (1999). Experimental sealed bid first price auctions with directly observed bid functions. In: Budescu D., Erev I., Zwick R. (eds). *Games and human behavior: Essays in the honor of Amnon Rapoport*. Mahwah, NJ. Lawrenz Associates. P. 79–104.

Selten R., Stoecker R. (1986). End behavior in sequences of finite Prisoner's dilemma supergames a learning theory approach // *Journal of Economic Behavior & Organization*. Vol. 7(1). P. 47–70.

Shapley L.S. (1964). Some topics in two-person games. Advances in: Dresher M., Shapley L.S., Tucker A.W. (eds). *Game Theory*. Princeton: Princeton University Press. P. 1–28.

Silver D., Hubert T., Schrittwieser J., Antonoglou I., Lai M., Guez A., Lanctot M., Sifre L., Kumaran D., Graepel T., Lillicrap T.A. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play // *Science*. Vol. 362(6419). P. 1140–1144.

Suppes P., Atkinson R.C. (1960). Markov learning models for multiperson interactions. Stanford, CA: Stanford University Press.

Sutton R.S., Barto A.G. (2018). Reinforcement learning: An introduction. Cambridge, MA: MIT Press.

Taylor P.D., Jonker L.D. (1978). Evolutionary stable strategies and game dynamics // *Mathematical Biosciences*. Vol. 40(1–2). July. P. 145–156.

Thorndike E.L. (1911). Animal intelligence. New York: Macmillan.

Thorndike E.L. (1927). The law of effect // *American Journal of Psychology*. Vol. 39. P. 212–222.

Van Damme E. (1991). Stability and perfection of Nash equilibria. Vol. 339. Berlin: Springer-Verlag.

Watson J.B. (2017). Behaviorism. Abingdon: Routledge.

Young H.P. (2004). Strategic learning and its limits. Oxford: Oxford University Press.

Поступила в редакцию 19.04.2019

Журнал НЭА,
№4 (44), 2019,
с. 77–125

REFERENCES (with English translation or transliteration)

Anufriev M., Arifovic J., Ledyard J., Panchenko V. (2013). Efficiency of continuous double auctions under individual evolutionary learning with full or limited information. *Journal of Evolutionary Economics*, 23(3), 539–573.

Arifovic J., Ledyard J. (2004). Scaling up learning models in public good games. *Journal of Public Economic Theory,* 6(2), 203–238.

Arifovic J., Ledyard J. (2018). Learning to alternate. *Experimental Economics,* 21(3), 692–721.

Arifovic J., McKelvey R.D., Pevnitskaya S. (2006). An initial implementation of the Turing tournament to learning in repeated two-person games. *Games and Economic Behavior,* 57(1), 93–122.

Aumann R.J. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica: Journal of the Econometric Society*, 55, 1–18.

Bajari P., Hong H., Nekipelov D. (2013). Game theory and econometrics: A survey of some recent research. Advances in Economics and Econometrics. 10th world congress, 3, 3–52.

Basu K., Weibull J.W. (1991). Strategy subsets closed under rational behavior. *Economics Letters,* 36(2), 141–146.

Beggs A.W. (2005). On the convergence of reinforcement learning. *Journal of Economic Theory,* 122(1), 1–36.

Bellman R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics,* 679–684.

Benaim M., Hirsch M. (1999). Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games Econ. Behav.*, 29, 36–72.

Bergemann D., Välimäki J. (2008). Bandit problems. In: *The New Palgrave Dictionary of Economics*, 1–8, 336–340.

Brandenburger A. (1996). Strategic and structural uncertainty in games. In: Zeckhauser R.J., Keeney R.L., Sebenius J.K. (eds). *Wise Choices: Games, Decisions, and Negotiations*. Brighton: Harvard Business School Press, 221–232.

Brown G.W. (1951). Iterative solution of games by fictitious play. *Activity analysis of production and allocation,* 13(1), 374–376.

Bubeck S., Cesa-Bianchi N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning,* 5(1), 1–122.

Bush R.R., Mosteller F. (1955). Stochastic models for learning. Oxford, England: John Wiley & Sons, Inc.

Camerer C., Ho H.T. (1999). Experience-weighted attraction learning in normal form games. *Econometrica,* 67(4), 827–874.

Cason T.N., Friedman D. (1999). Learning in a laboratory market with random supply and demand. *Experimental Economics,* 2(1), 77–98.

Cesa-Bianchi N., Lugosi G. (2006). Prediction, learning, and games. Cambridge: Cambridge University Press.

Cheung Y.-W., Friedman D. (1997). Individual learning in normal form games: Some laboratory results. *Games and Economic Behavior,* 19, 46–76.

Журнал НЭА, №4 (44), 2019, с. 77–125

Cooper D.J., Kagel J.H. (2008). Learning and transfer in signaling games. *Economic Theory*, 34(3), 415–439.

Cournot A. (1960). Researches into the mathematical principles of the theory of wealth. Trans. by N. Bacon. New York: Kelly. (Originally published by 1838.)

Cressman R., Ansell C., Binmore K. (2003). Evolutionary dynamics and extensive form games. Vol. 5. Cambridge, MA: MIT Press.

Doraszelski U., Lewis G., Pakes A. (2018). Just starting out: Learning and equilibrium in a new market. *American Economic Review*, 108(3), 565–615.

Duersch P., Kolb A., Oechssler J., Schipper B.C. (2010). Rage against the machines: How subjects play against learning algorithms. *Economic Theory*, 43(3), 407–430.

Ebbinghaus H. (2013). Memory: A contribution to experimental psychology. *Annals of Neurosciences*, 20(4), 155.

Erev I., Gopher D., Itkin R., Greenshpan Y. (1995). Toward a generalization of signal detection theory to N-person games: The example of two-person safety problem. *Journal of Mathematical Psychology*, 39(4), 360–375.

Erev I., Haruvy E. (2005). Generality and the role of descriptive learning models. *Journal of Mathematical Psychology*, 49(5), 357–71.

Erev I., Haruvy E. (2013). Learning and the economics of small decisions. In: Roth A.E., Kagel J. (eds). *The Handbook of Experimental Economics*. Vol. 2. Princeton University Press

Erev I., Roth A. (1998). Predicting how people play games: Reinforcement learning in games with unique strategy mixed-strategy equilibrium. *American Economic Review*, 88, 848–881.

Erev I., Roth A.E. (1999). On the role of reinforcement learning in experimental games: The cognitive game-theoretic approach. In: D. Budescu, I. Erev, R. Zwick (eds). *Games and Human Behavior: Essays in Honor of Amnon Rapoport*, 79–104. Lawrence Erlbaum Associates.

Estes W.K. (1964). Probability learning. In: *Categories of human learning*, 89–128. Academic Press.

Evans G.W., Honkapohja S. (2001). Learning as a rational foundation for macroeconomics and finance. In: Frydman R., Phelps E.S. (eds). *Rethinking expectations: The way forward for macroeconomics* 68 (2013). Princeton: Princeton University Press, 68–112.

Foster D.P., Vohra R. (1998). Asymptotic calibration. *Biometrika*, 85, 379–390.

Foster D.P., Vohra R.V. (1997). Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1–2), 40.

Foster D.P., Young H.P. (2001). On the impossibility of predicting the behavior of rational agents. *Proceedings of the National Academy of Sciences of the USA*, 98, 222, 12848–12853.

Fudenberg D., Levine D.K. (1993). Steady state learning and Nash equilibrium. *Econometrica: Journal of the Econometric Society*, 61, 3, 547–573.

Fudenberg D., Levine D.K. (1998). The theory of learning in games. Vol. 2. Cambridge, MA: MIT Press.

Fudenberg D., Levine D.K. (1998b). Learning in games. *European Economic Review*, 42(3–5), 631–639.

Журнал НЭА, №4 (44), 2019, с. 77–125

Fudenberg D., Levine D.K. (2009). Learning and equilibrium. *Annual Review of Economics*, 1(1), 385–420.

Gittins J.C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2), 148–164.

Güth W., Schmittberger R., Schwarze B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3, 367–388.

Hannan J. (1957). Approximation to Bayes' risk in repeated play. In: Dresher M., Tucker A.W., Wolfe P. (eds). *Contributions to the theory of games*. Vol. 3. Princeton: Princeton Univ. Press, 97–139.

Hart S., Mas-Colell A. (2001). A general class of adaptive strategies. *Journal of Economic Theory*, 98, 26–54.

Ho T.H., Camerer C.F., Chong J.K. (2007). Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory*, 133(1), 177–198.

Hofbauer J., Hopkins Ed. (2005). Learning in perturbed asymmetric games. *Games and Economic Behavior*, 52(1), 133–152.

Hofbauer J., Sigmund K. (1998). Evolutionary games and population dynamics. Cambridge: Cambridge University Press.

Hopkins Ed. (2002). Two competing models of how people learn in games. *Econometrica*, 70(6), 2141–2166.

Hutteger S.M. (2017). The probabilistic foundations of rational learning. Cambridge: Cambridge University Press.

Ioannou C.A., Romero J. (2014). A generalized approach to belief learning in repeated games. *Games and Economic Behavior*, 87, 178–203.

Kahneman D., Tversky A. (1979). On the interpretation of intuitive probability: A reply to Jonathan Cohen. *Cognition*, 7(4), 409–411.

Kalai E.. Lehrer E. (1993). Rational learning leads to Nash equilibrium. *Econometrica*, 61, 1019–1045.

Marimon R. (1997). Learning from learning in economics. In: Kreps D., Wallis K. (eds). *Advances in economics and econometrics: Theory and applications*, 1, 278–315.

Mathevet L., Romero J. (2012). Predictive repeated game theory: Measures and experiments. Mimeo.

McKelvey R., Palfrey T.R. (2001). Playing in the dark: Information, learning, and coordination in repeated games. Mimeo.

Miyasawa K. (1961). On the convergence of the learning process in a 2 x 2 non-zero-sum game. Economic research program. Princeton University, Research Memorandum No. 33.

Nachbar J. (2009). Learning in games. In: *Encyclopedia of complexity and systems science*. New York: Springer, 5177–5188.

Nachbar J.H. (1990). "Evolutionary" selection dynamics in games: Convergence and limit properties. *International Journal of Game Theory*, 19(1), 59–89.

Nachbar J.H. (2005). Beliefs in repeated games. *Econometrica* , 73(2), 459–480.

Nagel R. (1995). Unraveling in guessing games: an experimental study. *American Economic Review*, 85(5), 1313–1326.

Nevin J.A. (1988). Behavioral momentum and the partial reinforcement effect. *Psychological Bulletin*, 103, 44–56.

Nevo I., Erev I. (2012). On surprise, change, and the effect of recent outcomes. *Frontiers in Psychology*, 3, 24.

Opoitsev V.I. (1977). Equilibrium and stability in the collective behavior. Moscow: Nauka (in Russian).

Ramsey F.P. (1926). Truth and Probability. In: Braithwaite R. (ed.). *The foundations of mathematics and other logical essays*. London: Kegan Paul, 156–198.

Robbins H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5), 527–535.

Robinson J. (1951). An iterative method of solving a game. *Annals of Mathematics*, 54, 296–301.

Roth A.E., Erev I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games Econ. Behav.*, 8, 164–212.

Rothschild M. (1974). A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9, 185–202.

Sadrieh A. (1998). The alternating double auction market: A game theoretic and experimental investigation. Vol. 466. Berlin: Springer Science & Business Media.

Salmon T.C. (2001). An evaluation of econometric models of adaptive learning. *Econometrica*, 69(6), 1597–1628.

Sandholm W. (2010). Population games and evolutionary dynamics. Cambridge, MA: MIT Press.

Savage L.J. (1954). The foundations of statistics. New York: Wiley.

Selten R. (2004). Learning direction theory and impulse balance equilibrium. In: Cassar A., Friedman D. (eds). *Economics lab: An intensive course in experimental economics*, 133.

Selten R., Abbink T., Cox R. (2005). Learning direction theory and the winner's curse. *Experimental Economics*, 8(1), 5–20.

Selten R., Buchta J. (1999). Experimental sealed bid first price auctions with directly observed bid functions. In: D. Budescu, I. Erev, and R. Zwick (eds). *Games and Human Behavior: Essays in Honor of Amnon Rapoport*, 79–104. Lawrence Erlbaum Associates.

Selten R., Stoecker R. (1986). End behavior in sequences of finite Prisoner's dilemma supergames a learning theory approach. *Journal of Economic Behavior & Organization*, 7(1), 47–70.

Shapley L.S. (1964). Some topics in two-person games. Advances in: Dresher M., Shapley L.S., Tucker A.W. (eds). *Game Theory*. Princeton: Princeton University Press, 1–28.

Silver D., Hubert T., Schrittwieser J., Antonoglou I., Lai M., Guez A., Lanctot M., Sifre L., Kumaran D., Graepel T., Lillicrap T.A. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419), 1140–1144.

Suppes P., Atkinson R.C. (1960). Markov learning models for multiperson interactions. Stanford, CA: Stanford University Press.

Susin I.S., Chernov G.V. (2018). *Heuristics* recognition and learning in rock-paper-

Журнал НЭА,
№4 (44), 2019,
с. 77–125

scissors game: experimental study. *Journal of Economic Theory*, 3, 408–420 (in Russian).

Sutton R.S., Barto A.G. (2018). Reinforcement learning: An introduction. Cambridge, MA: MIT Press.

Taylor P.D., Jonker L.D. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1–2). July, 145–156.

Thorndike E.L. (1911). Animal intelligence. New York: Macmillan.

Thorndike E.L. (1927). The law of effect. *American Journal of Psychology*, 39, 212–222.

Van Damme E. (1991). Stability and perfection of Nash equilibria, 339. Berlin: Springer-Verlag.

Watson J.B. (2017). Behaviorism. Abingdon: Routledge.

Young H.P. (2004). Strategic learning and its limits. Oxford: Oxford University Press.

Received 19.04.2019

G.V. Chernov
HSE laboratory for experimental and behavioral economics;
Institute of Psychology of Russian Academy of Sciences, Moscow, Russia

I.S. Susin
HSE laboratory for experimental and behavioral economics, Moscow, Russia

Models of learning in games: An overview

Abstract. This survey analyzes central ideas and the current state of the economic theory of learning in games. In game theory learning can be thought of as both an alternative to equilibria and as a way to better understand the nature of equilibria. Outside of game theory, theory of learning shows economic theory (for example, the classic Cournot oligopoly) in a new light, provides interesting theoretical problems, is nontrivial from econometric perspective. It can be studied with experimental methods. It also links economics to unexpected scientific disciplines: biology, philosophy of rationality and computer science. However, existing surveys are not particularly accessible to beginners and are not accessible at all in Russian. This survey intends to fill these gaps. It can serve both as an introduction and as a short reference. We analyze issues of classification as well as the models themselves. Theoretical descriptions are illustrated with concrete examples. Special attention is devoted to the empirical and experimental work. We also draw conclusions and hypothesize on perspectives of the field and its future role in economic theory.

Keywords: *reinforcement learning, fictitious play, rational learning, bounded rationality, models of learning.*

JEL Classification: C70, D84.

DOI: 10.31737/2221-2264-2019-44-4-3