Математическая Теория Игр и её Приложения, т.18, в.2, с. 23–49

УДК 519.83
ББК 22.18

ВЛИЯНИЕ МАТРИЦЫ ВЫПЛАТ НА БОЛЬЦМАНОВСКОЕ Q-ОБУЧЕНИЕ В ПОВТОРЯЮЩИХСЯ МАТРИЧНЫХ ИГРАХ

Екатерина Е. Васильева
Физический институт им. П.Н. Лебедева РАН
119991, Москва, Ленинский проспект, 53
Андрей В. Леонидов
Алексей С. Титов
Физический институт им. П.Н. Лебедева РАН
119991, Москва, Ленинский проспект, 53
Московский физико-технический институт
14170, г. Долгопрудный, Институтский переулок, д. 9
e-mail: serebryannikovaee@lebedev.ru, leonidovav@lebedev.ru, a.titov@lebedev.ru

В работе исследуется влияние структуры матрицы выплат на динамику больцмановского Q-обучения в повторяющихся матричных играх 2×2 на примере дилеммы заключённого, охоты на оленя и интерполяции между ними. Матрицы выплат параметризуются характеристическими промежутками между элементами. Показано, что формального определения класса игры недостаточно для описания структуры стационарных состояний: фазовая диаграмма для числа асимптотических решений (являющихся равновесиями дискретного отклика) зависит от конкретных значений этих промежутков. Также в работе демонстрируется эффект «стратегического охлаждения»,

©2026 Е.Е. Васильева, А.В. Леонидов, А.С. Титов

выражающийся в смещении фазовых границ диаграммы количества решений в область более высоких температур при росте коэффициента дисконтирования из-за появления дополнительных метастабильных состояний. Продемонстрирована прямая зависимость времени пребывания агентов в таких метастабильных состояниях от величин характеристических промежутков. Полученные результаты демонстрируют, что конкретные значения параметров матрицы выплат существенно влияют на динамику и результирующие состояния больцмановского Q-обучения в повторяющихся матричных играх.

Ключевые слова: обучение с подкреплением, Q-обучение, дилемма заключённых, матрица выплат.

Поступила в редакцию: 06.03.26 *После доработки*: 03.05.26 *Принята к публикации*: 01.06.26

1. Введение

Обучение с подкреплением является одной из наиболее перспективных областей развития искусственного интеллекта. Алгоритмы, относящиеся к данному классу, находят применение в самых разных областях – робототехнике [17,11,18], телекоммуникации [21], медицине [9,6], финансах [28] и многих других. Вместе с тем как успешное практическое применение таких алгоритмов, так и, в особенности, наличие доказательств их сходимости относится преимущественно к классу задач обучения с подкреплением в стационарной среде [23]. Однако в последнее время все больший интерес исследователей направлен на анализ задач в нестационарных средах, к числу которых относится задачи с многоагентным обучением с подкреплением [22,4,1,8]. Задачи такого рода являются значительно более сложными, поскольку полученные для стационарных сред теоретические результаты относительно сходимости алгоритмов обучения с подкреплением в данном случае перестают работать [12]. Для построения содержательного описания характерных особенностей динамики мультиагентных систем необходимо в этой связи четко формализовать протоколы взаимодействия агентов. Инструментом для такого рода формализации выступает аппарат теории игр, позволяющий четко структурировать пространство состояний/решений и механизмы взаимодействия игроков. Именно поэтому особенностям применения алгоритмов обучения с подкреплением в играх в литературе

уделяется обширное внимание [1,25,3,19,27]. Ключевой целью исследований является установлении закономерностей, отражающих характерные особенности динамики повторяющихся матричных игр, в которых агентами используются различные алгоритмы обучения. В частности, важной задачей является изучение стационарных режимов такой динамики и их сравнение с соответствующими статическими теоретико-игровыми равновесиями.

Особый интерес представляет изучение алгоритмов обучения с подкреплением, относящихся к классу алгоритмов обучения, основанных на адаптивном формировании целевого уровня и его использовании на конечном количестве временных интервалов (метод временных разностей/Temporal-Difference или TD-обучение) [23]. Идея алгоритмов такого рода вдохновлена принципом максимума Беллмана и заключается в попытке предоставить агенту инструмент для учета оценки ожидаемой будущей выгоды при принятии текущего решения. Самым известным и часто применяемым на практике алгоритмом такого рода является Q-обучение [26]. Настоящая статья посвящена исследованию применения алгоритма Q-обучения в задаче о повторяющихся матричных играх.

Ниже исследуется обучение агентов в симметричных повторяющихся матричных играх 2×2 на примере *социальных дилемм* [13, 20]. Социальные дилеммы определяются как класс симметричных игр 2×2, в которых оба агента имеют две альтернативные стратегии: кооперироваться (C) или предать (D), со значениями выплат a, b, c, d, удовлетворяющими некоторому их упорядочению. Результаты взаимодействия игроков, т.е. их вознаграждение, зависят от их выбора. Все возможные исходы можно представить в виде матрицы выплат. В Таблице 1 приведен общий вид такой матрицы, где строки соответствуют стратегиям первого игрока, столбцы – второго, а числа на пересечении показывают соответствующий выигрыш агентов.

	C	D
C	a, a	b, c
D	c, b	d, d

Таблица 1. Общий вид матрицы выплат

В настоящей работе фокус делается на анализе больцмановского обучения в двух известных дилеммах – *дилемме заключённых* и *охоте на оленя*. В терминах матричных элементов a, b, c и d в Таблице 1 эти дилеммы характеризуются следующими упорядочениями:

- дилемма заключенных: $c > a > d > b$,
- охота на оленя: $a > c > d > b$.

Для изучения влияния структуры матрицы выплат на результаты обучения с подкреплением естественно рассмотреть ее параметризацию в терминах разностей ее матричных элементов – *характеристических промежутков*. Для рассматриваемой матрицы выплат в Таблице 1 ее структура в полностью характеризуется тремя характеристическими промежутками

$$g_1 = d - b,$$
$$g_2 = a - d,$$
$$g_3 = c - a.$$

Для двух рассматриваемых дилемм $g_{1,2} > 0$ и их отличие определяется знаком характеристического промежутка g_3. В дальнейшем нам будет удобно использовать обозначение $\Delta = -g_3$, так что при $\Delta < 0$ игра является дилеммой заключённых, а при $\Delta > 0$ – охотой на оленя.

Хорошо известно, что в дилемме заключённых имеется одно равновесие Нэша в чистых стратегиях, а в охоте на оленя – два равновесия Нэша в чистых стратегиях и одно в смешанных. Заметим, что структура равновесий полностью определяется соответствующим типом игры, т.е. знакоопределенностью соответствующих характеристических интервалов.

В настоящей работе изучаются результаты применения игроками алгоритма больцмановского Q-обучения в повторяющихся дилемме заключённых и охоте на оленя. Данный алгоритм описывает динамику обучения игроков в дискретном времени. Как упоминалось выше, ключевая особенность TD-методов, одним из которых является Q-обучение, является использование в процессе обучения механизма стратегической оценки будущих сценариев, называемой стратеги-

ческой эксплуатацией. Количественно роль стратегической эксплуатации параметризуется величиной коэффициента дисконтирования $\gamma \in [0, 1)$, так что при $\gamma = 0$ стратегическая эксплуатация полностью отсутствует, а большие положительные γ отвечают большему вкладу оценки будущих выплат в ценность текущего действия игрока.

Особенности динамики больцмановского Q-обучения в повторяющихся матричных играх обсуждались в литературе на протяжении достаточно долгого времени. При отсутствии стратегической эксплуатации (при $\gamma = 0$) анализ динамики обучения можно провести для непрерывного времени [14, 10]. Можно показать, что в этом случае описание динамики обучения дается уравнениями репликаторного типа [14, 10]. В рамках такого приближения демонстрировалась сходимость к равновесию дискретного отклика (РДО), а в [10] был предсказан фазовый переход по числу стационарных состояний в зависимости от матрицы выплат. Важно отметить, что непротиворечивое описание динамики с коэффициентом дисконтирования $\gamma > 0$ возможно только в дискретном времени. Особенности дискретной динамики при $\gamma > 0$ рассматривались в работе [15], где для дилеммы заключённых в явной дискретной динамике обнаружен эффект стратегической ловушки – возникновение метастабильных состояний, отличных от асимптотического равновесия дискретного отклика. Более подробная характеристика результатов, представленных в работах [10, 15], приведена в разделе 2.2. Таким образом, вопрос о том, как конкретные параметры матрицы выплат влияют, во-первых, на структуру фазовой диаграммы, описывающей структуру пространства стационарных состояний в случае явной дискретной динамики и, во-вторых, на время пребывания игроков в ловушке, остаётся открытым. Практическую значимость такого анализа подтверждают поведенческие эксперименты, показывающие, что исходы игр у реальных людей также определяются не только формальным классом игры, но и конкретными значениями выигрышей [7, 2, 5].

В настоящейей работе подробно изучается зависимость характера дискретной динамики больцмановского Q-обучения от параметров матрицы выплат. А именно, исследовано, как характеристики явной дискретной динамики при выключенной стратегической эксплуатации ($\gamma = 0$) соотносятся, во-первых, с границами классов социальных

дилемм и, во-вторых, с асимптотическими результатами в приближении [10]. Анализируется, как эти результаты трансформируются при учете стратегической эксплуатации ($\gamma > 0$). Кроме того, демонстрируется влияние параметров матрицы выплат на промежуточное квазистационарное состояние (стратегическую ловушку), наблюдаемое на конечных временах обучения. Нетривиальность полученных эффектов является мотивацией к проведению в будущем анализа корректной непрерывной динамики, учитывающей частоту выбора действий.

В разделе 2 приведён алгоритм больцмановского Q-обучения, и представлено описание известных особенностей динамики в применении к социальным дилеммам 2×2 в пределе бесконечного времени обучения и на конечных временах. В разделе 3 мы характеризуем влияние параметров матрицы выплат на свойства явной дискретной динамики Q-обучения в отсутствие стратегической эксплуатации, т.е. при $\gamma = 0$ (раздел 3.1), а также при положительных значениях параметра γ (раздел 3.2). В разделе 4 суммируются результаты работы и обсуждаются возможные направления дальнейших исследований.

2. Больцмановское Q-Обучение

2.1. Процедура обучения

Как упоминалось выше, в настоящем исследовании изучается динамика системы взаимодействующих агентов, использующих алгоритм Q-обучения в повторяющихся социальных дилеммах 2×2. Работа продолжает исследование особенностей дискретной динамики Q-обучения, начатое в статье [15].

Алгоритм Q-обучения относится к классу алгоритмов обучения с подкреплением, в которых стратегии агентов основаны на оценке ценности Q пары состояние-действие (value-based methods, см. [23]). В рассматриваемом случае повторяющихся матричных игр, для которых матрица выплат не меняется от шага к шагу, мы, тем самым, имеем дело с системой с одним состоянием.

Как и в работе [15], рассматривается механизм «слепого» обучения с подкреплением, когда доступная агентам информация ограничена их действиями и выплатами. Рассматриваемый алгоритм обучения описывается следующим образом. В момент времени t игрок i

$(i = 1, 2)$ выбирает действие $a_i^{(t)} \in \{C, D\}$ одновременно с противником и независимо от него. Это действие выбирается в соответствии с текущей политикой $\pi_i^{(t)} = \left(p_i^{(t)}(C), p_i^{(t)}(D)\right)$, где $p_i^{(t)}(a_i)$ – вероятность выбора действия $a_i \in \{C, D\}$ на момент времени t. Игроки получают вознаграждения $r_i^{(t)} = r_i^{(t)}(a_i, a_{-i})$, взятые из матрицы выигрышей, зависящие от выбранного действия каждым из игроков. Далее Q-значения обоих игроков $(i = 1, 2)$ обновляются в соответствии с правилом обновления Q-значений игрока i для действия $a_i \in \{C, D\}$ в момент времени $t \in \{1, ..., t^{max}\}$:

$$Q_i^{(t+1)}(a_i) = \begin{cases} Q_i^{(t)}(a_i) + \alpha\Big(r_i^{(t)}(a_i, a_{-i}^{(t)}) + \gamma \max\limits_{a' \in \{C,D\}} Q_i^{(t)}(a_i') - Q_i^{(t)}(a_i)\Big), \\ \qquad \text{если } a_i = a_i^{(t)} (\text{т.е. было выбрано действие } a_i) \\ Q_i^{(t)}(a_i), \text{ в противном случае,} \end{cases} \tag{2.1}$$

где $a_{-i}^{(t)}$ – выбор оппонента в момент времени t. Таким образом, для каждого игрока в каждый момент времени происходит обновление только одного из двух значений $Q_i^{(t)}(C)$ или $Q_i^{(t)}(D)$, соответствующего текущему выбору игрока.

В момент времени $t = 0$ положим $Q_i^{(0)}(a_i) = 0 \ \forall i = 1, 2, a_i \in \{C, D\}$. В настоящей работе рассматривается больцмановское Q-обучение, предполагающее, что политика игрока i в момент времени t рассчитывается согласно распределению

$$p_i^{(t)}(a_i) = \frac{e^{\beta Q_i^{(t)}(a_i)}}{e^{\beta Q_i^{(t)}(C)} + e^{\beta Q_i^{(t)}(D)}}, \ a_i \in \{C, D\}, \tag{2.2}$$

где $\beta = \frac{1}{T}$ – параметр, отражающий интенсивность исследовательского поведения агента. Большие значения T и, соответственно, малые значения β, отвечают интенсивному исследованию игроками пространства стратегий, в том числе – неоптимальных с точки зрения текущего накопленного опыта. И наоборот, при $T \to 0$ $(\beta \to \infty)$ при выборе стратегии игроки жадно следуют своим текущим оценкам Q-значений.

Вклад $\gamma \max_{a' \in \{C,D\}} Q_i^{(t)}(a_i')$ отражает оценку игроками ценности последующего действия. Тем самым, чем больше величина параметра $\gamma \in [0, 1)$, тем более предусмотрительным (стратегическим) является их поведение. Следуя терминологии [15], механизм учета вклада

этой составляющей в оценку ценности действия и, следовательно, в политику игрока мы будем называть стратегической эксплуатацией. Эффективная процедура обучения должна быть основана на балансе между интенсивностью исследования и стратегической эксплуатации.

Ряд важных характерных особенностей явной дискретной динамики больцмановского Q-обучения описан в работе [15]. Так, в этой работе показано, во-первых, что в пределе бесконечного времени обучения политики игроков описываются РДО для рассматриваемой игры; и, во-вторых, описан эффект стратегической ловушки обучения, заключающейся во возникновении промежуточных квазистационарных аттракторов, отличных от аттрактора в пределе бесконечного времени.

2.2. Динамика Q-обучения в социальных дилеммах 2x2: ключевые результаты работ [10, 15]

В работах [10, 15] демонстрируется, что асимптотическим равновесным состоянием больцмановского Q-обучения соответствуют равновесия дискретного отклика игры, которой обучаются агенты. Для квазистационарных Q-значений в [15] были получены следующие выражения: пусть a_i – стратегия, максимизирующая $Q_i^*(\cdot)$, т.е. $\max\limits_{a_i'} Q_i^*(a_i') = Q_i^*(a_i)$, и $a_i' \neq a_i$ – другая стратегия. Тогда выполняется:

$$\begin{aligned} Q_i^*(a_i) &= \frac{\mathbb{E}_{\pi_{-i}^*} r_i(a_i, a_{-i})}{1-\gamma}, \\ Q_i^*(a_i') &= \frac{\mathbb{E}_{\pi_{-i}^*} r_i(a_i', a_{-i}) + \gamma \mathbb{E}_{\pi_{-i}^*}\left[r_i(a_i, a_{-i}) - r_i(a_i', a_{-i})\right]}{1-\gamma}. \end{aligned} \tag{2.3}$$

Теперь, подставляя квазистационарные Q-значения в формулу для вероятности выбора (2.2), получим следующие выражения для квазистационарных вероятностей:

$$p_i^*(a_i) = \frac{1}{1+e^{\beta\left[Q_i^*(a_i') - Q_i^*(a_i))\right]}} = \frac{1}{1+e^{\beta \mathbb{E}_{\pi_{-i}^*}\left[r_i(a_i', a_{-i}) - r_i(a_i, a_{-i})\right]}}. \tag{2.4}$$

Квазистационарная вероятность для альтернативного действия a_i' равна $p_i^*(a_i') = 1 - p_i^*(a_i)$. Выражение (2.4) в точности совпадает с соответствующей формулой, определяющей равновесие дискретного

отклика для зашумленной дилеммы заключенных с гумбелевским шумом.

Важным наблюдением является то, что разность $Q_i^*(a_i') - Q_i^*(a_i)$ не зависит от коэффициента дисконтирования γ и определяется исключительно ожидаемой выгодой от выбора действия a_i относительно a_i' при условии, что противник придерживается стационарной политики π_{-i}^*. Тем самым, в случае больцмановского правила вычисления стратегий из Q-значений в асимптотике, наличие положительного коэффициента дисконтирования γ влияет лишь на абсолютные значения функций ценности $Q_i^*(a_i)$, но не на итоговое распределение вероятностей действий игроков. Асимптотическая сходимость к РДО наблюдается при любых значениях параметра дисконтирования. Данный результат показывает, что даже в условиях нестационарной среды в пределе бесконечного времени обучения слепое обучение с подкреплением приводит агентов к равновесию.

Вместе с тем, на конечных временах наблюдаются важные особенности динамики Q-обучения, ключевым образом отличающие ее от описанной ранее в литературе непрерывной репликаторной динамики. А именно, в работе [15] описан эффект *стратегической ловушки обучения*, состоящий в том, что при сравнительно больших значениях параметра γ на конечных временах наблюдается существенное изменение динамики политик игроков по сравнению со случаем, когда стратегическая эксплуатация отсутствует ($\gamma = 0$). Для дилеммы заключённых данный эффект состоит в том, что игроки на длительное время «застревают» в состоянии, статистически неотличимом от равновесия Нэша в чистых стратегиях для классической дилеммы заключённых без шума (D, D). Иными словами, несмотря на то, что предельным аттрактором на бесконечном времени по-прежнему остается РДО, на конечных временах наблюдается принципиально иное поведение игроков.

Механизм ловушки обусловлен главным образом зависимостью частоты обновления Q-функций для разных действий от вероятности их выбора (эффект, который не учитывается приближением [10]): когда один из игроков начинает преимущественно выбирать D, его $Q_i^{(t)}(D)$ быстро растет, что характерно на раннем этапе обучения. При этом вероятности выбора кооперации $p_i^{(t)}(C)$ становятся экстре-

мально малыми ($\approx 10^{-4}$ в симуляциях), что приводит к редкому обновлению $Q_i^{(t)}(C)$ (2.1). В результате стратегическая эксплуатация полностью подавляет механизмы исследования, и система на достаточно длительном временном интервале фиксируется в чистом равновесии Нэша. Для выхода из ловушки агенту необходимо многократно выбрать действие C, вероятность которого в ловушке крайне мала. Поэтому накопление подобного опыта требует значительного времени. Кроме того, как только одному из игроков удается выйти из ловушки, вероятность выбора действия C для второго игрока с большой вероятностью сокращается еще больше до значений порядка $\approx 10^{-6}$ (в симуляциях). Такое соотношение политик $p_2^{(t)}(C) \approx 10^{-6}$ и $p_2^{(t)}(D) \approx 1 - 10^{-6}$ для второго игрока делает его выход из ловушки практически невозможным в пределах разумного времени моделирования. Таким образом, стратегическая ловушка обучения представляет собой метастабильное состояние, в котором система застревает из-за наличия сильной стратегической эксплуатации и подавления стохастического исследования.

2.3. Структура асимптотических состояний динамики Q-обучения

В работе [10] рассмотрен структурный фазовый переход для решений по параметрам матрицы выплат в динамике репликатора, являющейся приближенным описанием алгоритма Q-обучения при $\gamma = 0$ и в предположении равенства частот обновления Q-функций игроков, отвечающих разным действиям. Как следует из результатов работы [15], предельное состояние такой динамики корректно характеризует предельное состояние реальной дискретной динамики Q-обучения. Поскольку ключевая задача настоящего исследования состоит в изучении зависимости характеристических особенностей дискретной динамики Q-обучения от структуры матрицы выплат, целесообразно провести сравнение фазовой диаграммы, полученной аналитически в приближении [10], с реальными симуляциями дискретной динамики Q-обучения. Для этой цели в текущем разделе мы воспроизводим результаты [10] в терминах, удобных для последующего сопоставления с наблюдаемой дискретной динамикой.

Изучим переход между классами игр дилеммы заключённых и охоты на оленя при изменении величины характеристического про-

межутка $g_3 = c - a$, переобозначенного для краткости $\Delta = -g_3$. В частности, рассмотрим количество стационарных точек динамики в зависимости от значений характеристических промежутков.

Матрицы выплат игроков будем считать симметричными, поэтому стационарные точки динамики также будут симметричными. Введём обозначение $x = p_i^*(C)$. Тогда, используя параметризацию через характеристические промежутки g_1 и Δ, выражение (2.4) для неподвижной точки динамики можно переписать в следующем виде:

$$x = \frac{1}{1 + exp(\beta(x(-\Delta - g_1) + g_1))}. \tag{2.5}$$

Для начала отметим, что значения РДО явно выражаются через значения g_1 и g_3 (Δ), но прямой зависимости от g_2 не наблюдается. Далее, будем искать решения следующего уравнения:

$$K(x) = \beta(x(\Delta + g_1) - g_1) - \log(\frac{x}{1-x}) = 0. \tag{2.6}$$

Сделаем замену

$$A = \beta(\Delta + g_1), \quad B = -\beta g_1. \tag{2.7}$$

Тогда

$$K(x) = Ax + B - \log(\frac{x}{1-x}) = 0. \tag{2.8}$$

Рассмотрим производную и приравняем её к 0:

$$\frac{d}{dx}\left[Ax + B - \ln\left(\frac{x}{1-x}\right)\right] = A - \frac{1}{x(1-x)} = 0. \tag{2.9}$$

Отсюда

$$A = \frac{1}{x(1-x)}. \tag{2.10}$$

На интервале $(0, 1)$ данная формула эквивалентна следующей:

$$Ax^2 - Ax + 1 = 0. \tag{2.11}$$

Тогда координаты точек касания:

$$x_{1,2} = \frac{A \pm \sqrt{A(A-4)}}{2A}. \tag{2.12}$$

Сделаем замену

$$L = \sqrt{A(A-4)}. \tag{2.13}$$

Из формы функции очевидно, что $x_1 = \frac{A-L}{2A}$ отвечает минимуму, а $x_2 = \frac{A+L}{2A}$ – максимуму. Тогда у $K(x)$ будет существовать три решения, если $K(x_1) < 0$, а $K(x_2) > 0$, два решения, если хотя бы одно из $K(x_1), K(x_2)$ равно нулю и одно решение иначе. Заметим, что из (2.10) в точках экстремумов верно

$$x(1-x) = \frac{1}{A}. \tag{2.14}$$

И тогда (2.8) можно переписать в виде

$$K(x) = Ax + B - \log(Ax^2). \tag{2.15}$$

Найдём критические точки $K(x_1) = 0$ и $K(x_2) = 0$:

$$K(x) = Ax + B - \log(Ax^2) = 0. \tag{2.16}$$

Тогда

$$B = 2\log(A \pm L) - \frac{1}{2}(A \pm L) - \log(A) - 2\log(2), \tag{2.17}$$

и

$$\beta g_1 = \frac{1}{2}(A \pm L) - \log(\frac{4A}{(A \pm L)^2}). \tag{2.18}$$

Значения β и Δ, удовлетворяющие (2.18), будут являться критическими точками фазового перехода между состояниями, когда система имеет одно решение и три решения. Поскольку явный вид данной функции неизвестен, воспользуемся численным приближением. На Рис. 1 представлена фазовая диаграмма в плоскости Δ–β, где красным цветом выделена область, в которой у динамики есть три решения, серым – одно решение, а точки фазового перехода (отвечающие соответствующим решениям уравнения), в которых имеется два решения, обозначены синей пунктирной и голубой кривыми. При этом параметр $g_1 = 1$.

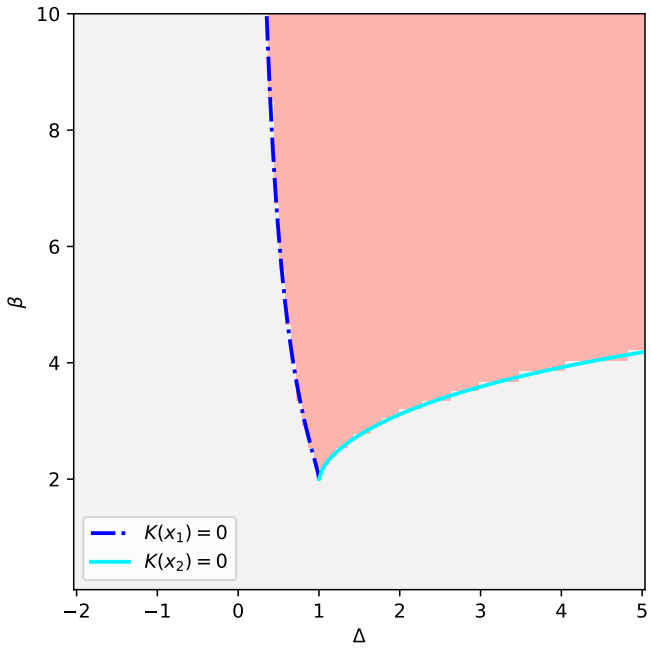


Рисунок 1. Фазовая диаграмма количества стационарных точек Q-Обучения в плоскости Δ-β. Серым выделена область параметров, в которой существует одно решение, красным – в которой существует три решения. Синей пунктирной линией и голубой линией показаны решения уравнения (2.18) для x_1 и x_2.

Таким образом, фазовая диаграмма показывает, что на количество решений в динамике репликатора влияет не класс игры, а конкретные значения матрицы выплат. Только при полной заморозке $\beta = \infty$ фазовый переход будет происходить в точке $\Delta = 0$. При конечных β существует область $[0, \Delta_{crit}(\beta)]$ в которой несмотря на то, что формально игра поменяла свой класс, поведение не будет меняться. Также из фазовой диаграммы видно, что при $\beta < 2$ фазового перехода вообще не будет, и динамика будет иметь одно стационарное состояние независимо от того, является игра дилеммой заключённых или охотой на оленя.

3. Анализ влияния структуры матрицы выплат на динамику обучения/игры

3.1. Динамика в отсутствие стратегической эксплуатации

В данном разделе исследуется зависимость динамики больцмановского Q-обучения от параметров матрицы выплат социальных дилемм в отсутствие механизма стратегической эксплуатации, т.е. в случае $\gamma = 0$. Как говорилось выше, попытка такого рода анализа предпринималась в работе [10], где производился анализ зависимости динамики репликатора от параметров матрицы выплат. Однако, как отмечалось, динамика, описанная в [10], является приближенным описанием реальной динамики Q-обучения при $\gamma = 0$, поскольку в ней упускается один из важных аспектов обучения – различие частот выбора действий C и D игроками и, как следствие, различие в скорости обновления соответствующих Q-значений.

В данном разделе рассмотрим результаты численных симуляций при варьирующихся значениях параметров β и Δ и фиксированном значении $\gamma = 0$ и сравним эти результаты с описанными в разделе 2.3 теоретическими результатами.

В ходе симуляций значения β варьировались в интервале $[1, 6]$ с шагом 0.1, а значения Δ в интервале $[-2, 3]$ с тем же шагом. Для каждой комбинации β и Δ на сетке мы проводили 30 симуляций обучения. В каждой симуляции инициализация значений Q-функций проводилась нулевыми значениями $Q_i^{(0)}(a_i) = 0,\ i = 1, 2,\ a_i = C, D$. Соответственно, политики игроков инициализируются как $p_i^{(0)}(C) = p_i^{(0)}(D) = 0.5,\ i = 1, 2$. Затем запускается описанный в разделе 2.1 алгоритм Q-обучения на временном горизонте $t^{max} = 10000$. Результат каждой из симуляций будем относить к одному из трех классов по описанным далее правилам:

- сходимость к *нижнему* аттрактору: к этому классу отнесем симуляции, в которых в результате обучения вероятности $p_i^{(t)}(C)$, $i = 1, 2$, сходятся к некоторому аттрактору, который не превосходит фиксированной величины $p^{low}(\beta)$[1];

[1]При проведении экспериментов была принята простая пороговая функция: $p^{low} = 0.35$ при $\beta < 2$ и $p^{low} = 0.2$ при $\beta \geq 2$, однако конкретный выбор порогов в разумных интервалах качественно не меняет приведенные выводы.

- сходимость к *верхнему* аттрактору: к этому классу отнесем симуляции, в которых в результате обучения вероятности $p_i^{(t)}(C)$, $i = 1, 2$, сходятся к некоторому аттрактору, который не менее фиксированной величины p^{high}[2];

- к третьему классу относятся прочие симуляции.

На Рис. 2 приведена диаграмма, где частота первого вида выделена синим цветом, второго – красным и третьего – зелёным. Кроме того, для простоты сравнения теоретических и экспериментальных результатов, на диаграмме отмечена граница фаз с Рис. 1.

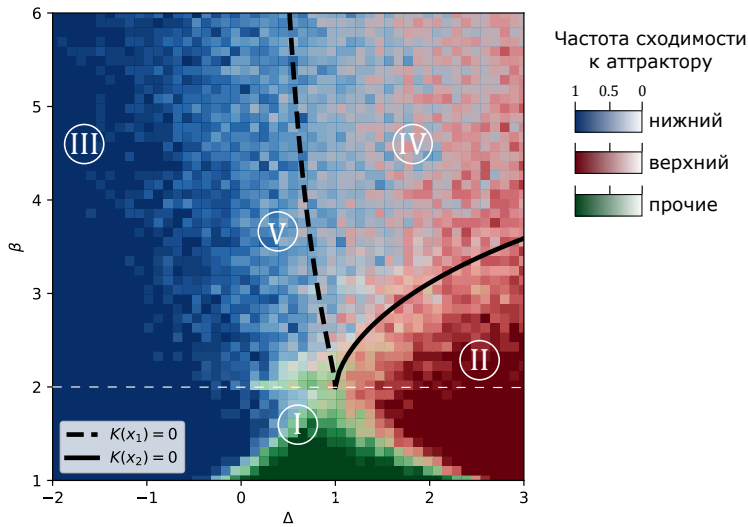


Рисунок 2. Диаграмма частот исходов симуляций в плоскости Δ-β. Каждой точке сопоставляются частоты трех видов исходов симуляций: синим – сходимость $p_i^{(t)}(C)$ к нижнему аттрактору ($p_i^{(t)}(C) < p^{low}(\beta)$), красным – к верхнему аттрактору $p_i^{(t)}(C) > p^{high}$, зеленым выделены точки с остальными вариантами. В каждой точке усреднение проводилось по 30 симуляциям. $t^{max} = 10000$, $\alpha = 0,01, \gamma = 0$. Пунктирной и сплошной кривыми показаны решения уравнения (2.18) для x_1 и x_2 соответственно, характеризующие границы фаз с Рис. 1

[2]При проведении экспериментов мы фиксировали $p^{high} = 0.9$, однако конкретный выбор данного порога в разумном интервале качественно не меняет приведенные выводы.

Между диаграммами на Рис. 1 и 2 наблюдается явное сходство, но также видны и некоторые значительные расхождения. Выделим качественно следующие пять областей на диаграмме Рис. 2 (границы этих областей нечеткие):

I: часть фазового пространства, где $\beta < 2,\ \forall\Delta$, в которой в каждой ячейке сетки наблюдается сходимость с близкой к единице частотой к аттрактору одного из трех типов: нижнему, верхнему или прочим. На диаграмме есть небольшие переходные области между сходимостью к нижнему аттрактору и прочим аттракторам, а также между прочими аттракторами и верхним, однако эти области отражают технический артефакт классификации, подробно описанный ниже;

II: часть фазового пространства в правом нижнем углу, где $2 \leq \beta \lesssim 3,\ \Delta \gtrsim 1$, в которой наблюдается сходимость к верхнему аттрактору;

III: левая область фазового пространства при $\beta \geq 2$, где наблюдается сходимость к одному нижнему аттрактору;

IV: правый верхний угол, где наблюдается сходимость либо к нижнему, либо к верхнему аттракторам;

V: область перехода между III и IV.

Несмотря на то, что у РДО есть три решения в синей зоне Рис. 2, случаев сходимости к аттрактору, отвечающему смешанному равновесию Нэша, зафиксировано не было.

Охарактеризуем подробно динамику обучения в области I. В соответствии с теоретическими результатами, отраженными на диаграмме Рис. 1, в данной области есть одно решение РДО. На панели А Рис. 3 представлен график зависимости значения РДО от Δ при $\beta = 1$. Мы видим, что при фиксированном значении β постепенное увеличение значения параметра Δ приводит к сдвигу уровня РДО от нуля (отвечающего нижнему аттрактору в приведенной выше классификации результатов симуляций) к единице (отвечающей верхнему аттрактору в симуляциях). Это изменение постепенное, что дополнительно иллюстрируется графиками динамики обучения в этой области, приведенными на панели Б Рис. 3. Именно поэтому в области

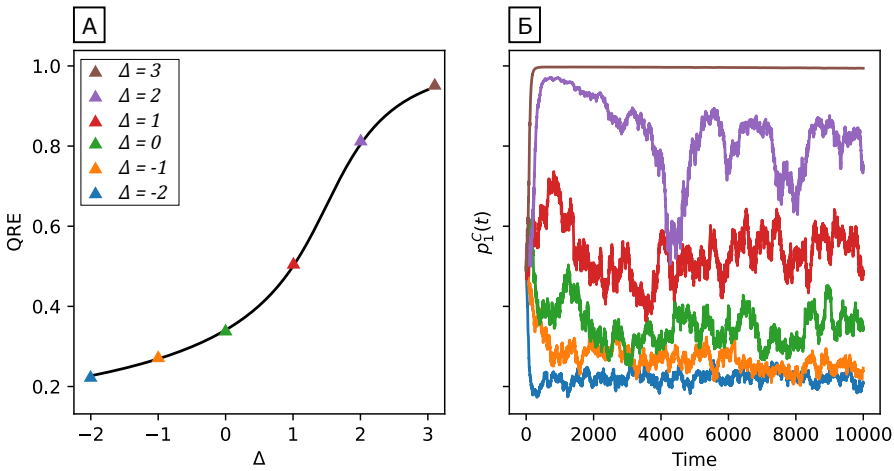


Рисунок 3. Панель А: зависимость значения РДО от Δ при $\beta = 1$. На графике выделены точки, соответствующие целым значениям Δ в диапазоне от -2 до 3. Панель Б: примеры траекторий $p_1^{(t)}(C)$ при $\beta = 1$ и целых значениях Δ в диапазоне от -2 до 3. $t^{max} = 10000$, $\alpha = 0.01$, $\gamma = 0$.

I на фазовой диаграмме Рис. 2 мы наблюдаем переход из синей области (отвечающей сходимости к нижнему аттрактору) в красную область (сходимость к верхнему аттрактору). Таким образом, динамика в этой области хорошо описывается теоретическим приближением динамики репликатора.

Области II и III также соответствуют теоретическим оценкам, приведенным на Рис. 1, однако вид траекторий обучения в этой области фазовой диаграммы отличается от траекторий из области I, представленных на Рис. 3. На Рис. 4 представлены примеры траекторий из этих областей: на панели А – из области II, на панели Б – из области III. На панели А видна явная сходимость к верхнему аттрактору, а на панели Б – к нижнему.

В области IV в симуляциях наблюдается сходимость к одному из дух аттракторов: у нижнему или к верхнему. Динамика в каждом из этих двух случаев качественно не отличается от примеров, представленных на Рис. 4. Отметим, что в симуляциях сходимости к аттрактору РДО, соответствующему сдвинутому смешанному равновесию Нэша, не наблюдается. То есть в реальной дискретной динамике обучения стратегии игроков сходятся только к двум из трёх возможных

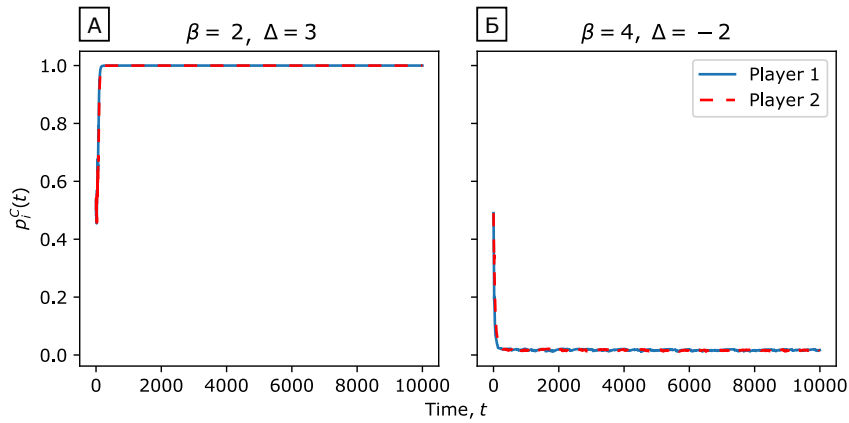


Рисунок 4. Панель А: пример траектории из области II $p_i^{(t)}(C)$ при $\beta = 2$ и $\Delta = 3$. Панель Б: пример траектории из области III $p_i^{(t)}(C)$ при $\beta = 4$ и $\Delta = -2$. Синей линией показана траектория для игрока $i = 1$, красной пунктирной – для игрока $i = 2$. $t^{max} = 10000$, $\alpha = 0.01$, $\gamma = 0$.

РДО, которые соответствуют двум сдвинутым чистым равновесиям Нэша. Особый интерес представляет область V, которая является переходной между областями III и IV.

Заметим, что траектории, сходящиеся к верхнему аттрактору появляются сильно левее теоретической границы, разделяющей области с одним и тремя решениями РДО. Данное явление может быть связано с описанным в разделе 2.2 эффектом стратегической ловушки обучения, когда игроки на конечных временах на продолжительном временном интервале оказываются в состоянии статистически неотличимом от равновесия Нэша в чистых стратегиях. Важно отметить, из формы функции $K(x)$ из уравнения (2.6) следует, что что при росте β, расстояние от РДО, отвечающее верхнему аттрактору, до чистого равновесия Нэша (C, C) сокращается быстрее, чем расстояние от РДО, отвечающего нижнему аттрактору, до чистого равновесия (D, D). То есть, при больших β, РДО, отвечающее верхнему аттрактору, численно лишь незначительно отличается от чистого равновесия, в то время как РДО, отвечающее нижнему аттрактору, явно выделяется.

Область V соответствует области теоретического фазового пере-

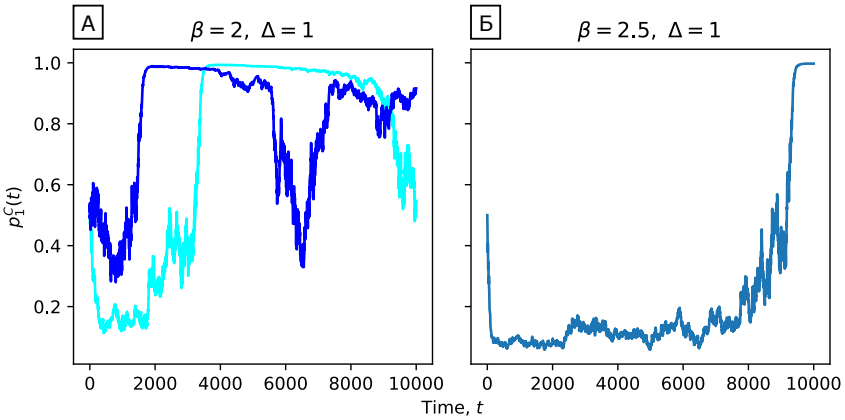


Рисунок 5. Панель А: пример траектории $p_1^{(t)}(C)$ при $\beta = 2$ и $\Delta = 1$ при двух различных запусков симуляции. Панель Б: пример траектории $p_1^{(t)}(C)$ при $\beta = 2.5$ и $\Delta = 1$. $t^{max} = 10000$, $\alpha = 0.01$, $\gamma = 0$.

хода. При больших β наблюдаемые траектории всегда сходятся к одному из двух аттракторов и остаются в нём. Однако при приближении к $\beta = 2$ появляются симуляции, в которых до нескольких раз может происходить перескок между двумя аттракторами (эти симуляции отмечены зеленым на диаграмме Рис. 2). Примеры таких симуляций приводятся на Рис. 5.

3.2. Влияние стратегической эксплуатации

Выше рассматривался случай $\gamma = 0$, т.е. динамика обучения в отсутствии стратегической эксплуатации. В текущем разделе представлен анализ влияния эффекта стратегической эксплуатации на обучение игроков. Отметим еще раз, что при $\gamma > 0$ описание в непрерывном времени некорректно и нужно анализировать свойства явной дискретной динамики.

При нескольких значениях параметра γ: $\gamma = 0.5$, 0.7 и 0.9, был проведен ряд численных экспериментов в условиях, аналогичных описанным в разделе 3.1. На Рис. 2 для указанных положительных значений параметра γ приведены диаграммы, аналогичные приведенным на Рис. 2 для $\gamma = 0$. Соответствующие диаграммы приведены на Рис. 6.

На Рис. 6 наблюдается эффект, который по аналогии с резуль-

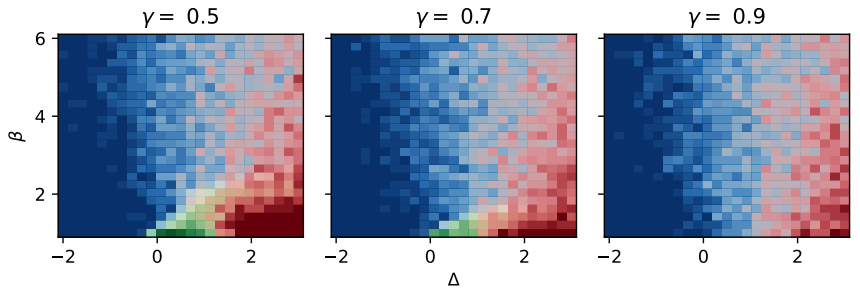


Рисунок 6. Диаграммы частот исходов симуляций в плоскости Δ-β для трех значений параметра γ: $\gamma = 0.5$ (левая панель), $\gamma = 0.7$ (средняя панель) и $\gamma = 0.9$ (правая панель). На каждой из диаграмм каждой точке сопоставляются частоты трех видов исходов симуляций: синим – сходимость $p_i^{(t)}(C)$ к нижнему аттрактору ($p_i^{(t)}(C) < p^{low}(\beta)$), красным – к верхнему аттрактору $p_i^{(t)}(C) > p^{high}$, зеленым выделены точки с остальными вариантами. В каждой точке усреднение проводилось по 30 симуляциям. $t^{max} = 10000$, $\alpha = 0,01$.

татом [16] можно назвать *стратегическим охлаждением*. А именно, мы видим, что вся диаграмма Рис. 2 при росте γ смещается вниз, т.е. в область меньших значений β (больших температур). Иными словами, добавление стратегической эксплуатации качественно может быть описано эффективным уменьшением температуры при $\gamma = 0$. Данный эффект является прямым следствием эффекта стратегической ловушки обучения. При $t^{max} \to \infty$ диаграммы при любых значениях γ должны совпадать с диаграммой при $\gamma = 0$, так как финальный аттрактор вероятностей – РДО, – не зависит от значения γ. Иными словами, наблюдаемый эффект стратегического охлаждения в этой задаче – это исключительно свойство дискретной динамики на конечных временах. Однако, как отмечалось в разделе 2.2, в симуляциях агенты выходят из ловушки не одновременно, и чаще всего наблюдается односторонний выход из ловушки [15]. В результате, на практике в симуляциях на любом разумном временном горизонте будет наблюдаться зависимость, показанная на Рис. 6.

Поскольку при добавлении стратегической эксплуатации ключевые особенности динамики описываются свойствами промежуточно-

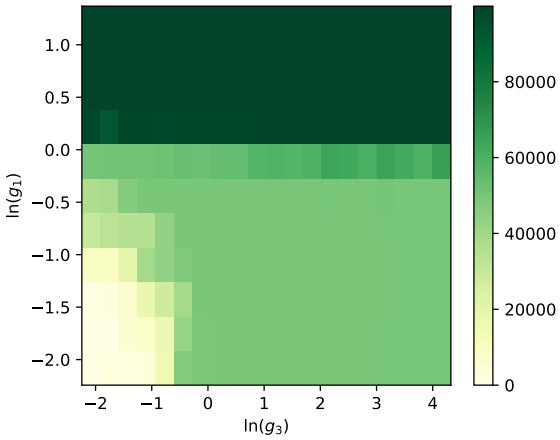


Рисунок 7. Время пребывания в стратегической ловушке обучения в плоскости $\ln(g_3)$–$\ln(g_1)$ при $g_2 = 2$. $t^{max} = 100000, \alpha = 0,1, \beta = 1, \gamma = 0,9$. Цветовая шкала демонстрирует продолжительность временного интервала, в течение которого оба игрока оказываются в стратегической ловушке обучения.

го квазистационарного состояния (стратегической ловушки обучения), а не предельного состояния на бесконечном временном горизонте, крайне важно изучить как эти свойства зависят от структуры матрицы выплат. Главная характеристика данного промежуточного состояния – это длительность временного интервала, в течение которого система оказывается в данном состоянии.

Для характеристики этого свойства был проведен ряд численных экспериментов в следующих условиях: для каждой из комбинаций значений $g_1 \in [0.125, 3.33]$ и $g_3 \in [0.125, 64]$ на сетке проводилась серия из запусков алгоритмов обучения с параметрами $\gamma = 0.9$, $g_2 = 2$, $t^{max} = 100000$, $\beta = 1$, $\alpha = 0.1$. Усредненное по запускам значение времени пребывания обоих игроков в ловушке в данном эксперименте приведено на Рис. 7. Под временем пребывания в ловушке формально мы понимаем количество итераций обучения, для которых для обоих агентов $i = 1, 2$ верно $p_i^{(t)}(C) < 0.01$.

На графике мы наблюдаем явный рост времени пребывания в ловушке при увеличении g_1 и g_3. Такое поведение можно объяснить тем, что с ростом характеристических промежутков увеличивается

средняя награда игрока. Из работы [15] известно, что игрок застревает в ловушке в результате того, что Q-значение для одного действия растет очень быстро, а для второго очень медленно, после чего оба Q-значения на определенном промежутке времени остаются константными. Формально, в этом случае для двух действий a_i и a'_i игрока i (квази-)стационарные значения удовлетворяют $Q^*_i(a_i) > Q^*_i(a'_i)$, верно выражение (2.3). Для действия a_i его (квази-)стационарное значение достигается быстро, и в это время рост $Q^{(t)}_i(a'_i)$ значительно замедляется. Тогда, из-за роста средней награды $\mathbb{E}_{\pi^{(t)}_{-i}} r_i(a_i, a_{-i})$ увеличивается расстояние между Q-значениями двух действий, что приводит к уменьшению вероятности для невыгодного действия. Время, проведенное игроком в ловушке, зависит от того, насколько близка к нулю вероятность невыгодного действия: время в ловушке увеличивается пропорционально уменьшению этой вероятности. Уменьшение вероятности – следствие роста средней награды за выгодное действие. Рост промежутков g_1 и g_3 прямо влияет на рост этой награды, что и сказывается на увеличении времени в ловушке, наблюдаемом на диаграмме Рис. 7

Таким образом, конкретные значения матрицы выплат оказывают влияние не только на квазистационарное состояние обучения в пределе бесконечного времени, но и на особенности динамики на конечных временах. При этом эта зависимость становится крайне нетривиальной при включении механизма стратегической эксплуатации.

4. Заключение

В настоящей работе проведено исследование влияния структуры матрицы выплат на динамику больцмановского Q-обучения в повторяющихся матричных играх (2×2) на примере двух классов социальных дилемм – дилеммы заключённых и охоты на оленя. Основным инструментом анализа выступила параметризация матрицы выплат с помощью характеристических промежутков $g_1 = d - b$, $g_2 = a - d$ и $g_3 = c - a$, позволяющая единообразно описывать оба класса игр, и, в частности, в терминах промежутка g_3 существует возможность рассмотреть непрерывную интерполяцию между указанными классами.

В работе проведён сравнительный анализ теоретических асимпто-

тических квазистационарных значений больцмановского Q-обучения и экспериментальных результатов симуляций в отсутствие стратегической эксплуатации ($\gamma = 0$). Теоретическая фазовая диаграмма количества аттракторов демонстрирует, что структура решений определяется не формальной принадлежностью игры к тому или иному классу, но конкретными значениями параметров матрицы, что было продемонстрировано с помощью варьирования параметра $\Delta = -g_3$, отвечающего за переход между дилеммой заключённых и охотой на оленя. Также был продемонстрирован ряд различий между теоретической диаграммой и результатами симуляций.

Другим важным результатом работы является анализ влияния положительных значений коэффициента дисконтирования $\gamma > 0$, отвечающего за учёт агентами будущих выигрышей (стратегическую эксплуатацию). Хотя асимптотическое состояние системы определяется РДО и не зависит от γ, на конечных временах динамика претерпевает существенные изменения. На фазовых диаграммах, построенных для $\gamma = 0.5,\ 0.7,\ 0.9$, зафиксирован эффект «стратегического охлаждения»: увеличение γ приводит к смещению границ фазового перехода в область более высоких температур (меньших значений β). Этот эффект является прямым следствием возникновения метастабильных состояний – стратегических ловушек обучения, в которых система на продолжительное время задерживается в окрестности чистых равновесий Нэша. Также в работе демонстрируется зависимость времени пребывания системы в стратегической ловушке от величин характеристических промежутков. При варьировании g_1 и g_3 при фиксированном g_2 видно, что увеличение обоих промежутков приводит к значительному росту продолжительности фиксации системы в метастабильном состоянии. Данный эффект объясняется ростом средней ожидаемой выгоды от доминирующего действия в ловушке, что увеличивает разрыв между Q-значениями и, как следствие, снижает вероятность выбора альтернативного действия, необходимого для выхода из ловушки.

Полученные результаты позволяют сделать вывод о том, что формальная классификация игры (дилемма заключённых или охота на оленя) не является достаточным условием для описания динамики больцмановского Q-обучения и её стационарных режимов. Конкрет-

ные численные значения матрицы выплат, описываемые характеристическими промежутками, оказывают определяющее влияние как на структуру асимптотических состояний, так и на поведение системы на конечных временах обучения при наличии стратегической эксплуатации.

Дальнейшие направления исследований могут быть связаны с построением корректной непрерывной динамики, учитывающей неравномерность частот обновления Q-функций, а также исследование поведения агентов в играх с $N > 2$ игроками.

СПИСОК ЛИТЕРАТУРЫ

1. Barfuss W., Donges J. F., Kurths J. *Deterministic limit of temporal difference reinforcement learning for stochastic games* // Physical Review E. 2019. V. 99. №. 4. P. 043305.

2. Battalio R., Samuelson L., Van Huyck J. *Optimization incentives and coordination failure in laboratory stag hunt games* // Econometrica. 2001. V. 69. №. 3. P. 749–764.

3. Bloembergen D. et al. *Evolutionary dynamics of multi-agent learning: A survey* // Journal of Artificial Intelligence Research. 2015. V. 53. P. 659–697.

4. Busoniu L., Babuska R., De Schutter B. *A comprehensive survey of multiagent reinforcement learning* // IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2008. V. 38. №. 2. P. 156–172.

5. Dubois D., Willinger M., Van Nguyen P. *Optimization incentive and relative riskiness in experimental stag-hunt games* // International Journal of Game Theory. 2012. V. 41. №. 2. P. 369–380.

6. Frommeyer T. C. et al. *Reinforcement learning and its clinical applications within healthcare: A systematic review of precision medicine and dynamic treatment regimes* // Healthcare. MDPI. 2025. V. 13. №. 14. P. 1752.

7. Harsanyi J. C., Selten R. *A general theory of equilibrium selection in games* // MIT Press Books. 1988. V. 1.

8. Huh D., Mohapatra P. *Multi-agent reinforcement learning: A comprehensive survey* // arXiv preprint arXiv:2312.10256. 2023.

9. Jayaraman P. et al. *A primer on reinforcement learning in medicine for clinicians* // NPJ digital medicine. 2024. V. 7. №. 1. P. 337.

10. Kianercy A., Galstyan A. *Dynamics of Boltzmann Q learning in two-player two-action games* // Physical Review E—Statistical, Nonlinear, and Soft Matter Physics. 2012. V. 85. №. 4. P. 041145.

11. Konar A. et al. *A deterministic improved Q-learning for path planning of a mobile robot* // IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2013. V. 43. №. 5. P. 1141–1153.

12. Laurent G. J., Matignon L., Le Fort-Piat N. *The world of independent learners is not Markovian* // International Journal of Knowledge-based and Intelligent Engineering Systems. 2011. V. 15. №. 1. P. 55–64.

13. Leibo J. Z. et al. *Multi-agent reinforcement learning in sequential social dilemmas* // arXiv preprint arXiv:1702.03037. 2017.

14. Leonardos S., Piliouras G. *Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory* // Artificial Intelligence. 2022. V. 304. P. 103653.

15. Leonidov A., Titov A., Vasilyeva E. *Strategic exploitation in Boltzmann Q-learning in the prisoners dilemma* // Chaos, Solitons & Fractals. 2025. V. 201. P. 117243.

16. Leonidov A., Vasilyeva E. *Strategic stiffening/cooling in the Ising game* // Chaos, Solitons & Fractals. 2022. V. 160. P. 112279.

17. Levine S. et al. *End-to-end training of deep visuomotor policies* // Journal of Machine Learning Research. 2016. V. 17. N. 39. P. 1–40.

18. Lin J. L. et al. *Gait balance and acceleration of a biped robot based on Q-learning* // IEEE access. 2016. V. 4. P. 2439–2449.

19. Lu Y., Yan K. *Algorithms in multi-agent systems: a holistic perspective from reinforcement learning and game theory* // arXiv preprint arXiv:2001.06487. 2020.

20. Macy M. W., Flache A. *Learning dynamics in social dilemmas* // Proceedings of the National Academy of Sciences. 2002. V. 99. №. 3. P. 7229–7236.

21. Matta M. et al. *A reinforcement learning-based QAM/PSK symbol synchronizer* // Ieee Access. 2019. V. 7. P. 124147–124157.

22. Shoham Y., Powers R., Grenager T. *If multi-agent learning is the answer, what is the question?* // Artificial intelligence. 2007. V. 171. №. 7. P. 365–377.

23. Sutton R. S. et al. *Reinforcement learning: An introduction*. Cambridge : MIT press, 1998.

24. Tallec C., Blier L., Ollivier Y. *Making deep q-learning methods robust to time discretization* // International Conference on Machine Learning. PMLR, 2019. P. 6096–6104.

25. Tuyls K., Nowe A. E*volutionary game theory and multi-agent reinforcement learning* // The Knowledge Engineering Review. 2005. V. 20. №. 1. P. 63–90.

26. Watkins C. J. C. H., Dayan P. *Q-learning* // Machine learning. 1992. V. 8. №. 3. P. 279–292.

27. Yang Y. et al. Game-theoretic multiagent reinforcement learning // arXiv preprint arXiv:2011.00583. 2020.

28. Yang H. et al. *Deep reinforcement learning for automated stock trading: An ensemble strategy* // Proceedings of the first ACM international conference on AI in finance. 2020. P. 1–8.

THE IMPACT OF PAYOFF MATRIX ON BOLTZMANN Q-LEARNING

Ekaterina E. Vasilyeva, P.N. Lebedev Physical Institute of the Russian Academy of Sciences, Cand.Sc. (serebryannikovaee@lebedev.ru),
Andrey V. Leonidov, P.N. Lebedev Physical Institute of RAS, Moscow Institute of Physics and Technology, Dr.Sc., Professor (leonidovav@lebedev.ru),
Alexey S. Titov, P.N. Lebedev Physical Institute of the RAS, Moscow Institute of Physics and Technology (a.titov@lebedev.ru).

Abstract: This study investigates the influence of the payoff matrix structure on the dynamics of Boltzmann Q-learning in repeated 2×2 matrix games, using the prisoner's dilemma, the stag hunt, and an interpolation between them as examples. The payoff matrices are parameterized by characteristic gaps between elements. It is shown that a formal definition of the game class is insufficient to describe the structure of stationary states: the phase diagram of the number of asymptotic solutions (which are quantal response equilibria) depends on the specific values of the gaps. The paper also demonstrates the effect of "strategic cooling", manifested as a shift of the phase boundaries of the solution-count diagram into the region of higher temperatures as the discounting factor increases, due to the emergence of additional metastable states. A direct dependence of the agents' residence time in these metastable states on the sizes of the characteristic gaps is demonstrated. The obtained results show that the specific values of parameters of the payoff matrix significantly affect the dynamics and the resulting states of Boltzmann Q-learning.

Keywords: reinforcement learning, Q-learning, prisoners' dilemma, payoff matrix.