J. Cvitanic, D. Prelec, S. Radas, H. Sikic, Incentive-compatible surveys via posterior probabilities, *Teor. Veroyatnost. i Primenen.*, 2020, Volume 65, Issue 2, 368–408

DOI: 10.4213/tvp5351

Том 65 ТЕОРИЯ ВЕРОЯТНОСТЕЙ И ЕЕ ПРИМЕНЕНИЯ 2020 Выпуск 2

ЦВИТАНИЧ Я.[*], *ПРЕЛЕЦ Д.*[**], *РАДАС С.*[***], *ШИКИЧ Х.*[†]

К ВОПРОСУ О СТИМУЛЬНО-СОГЛАСОВАННЫХ ИССЛЕДОВАНИЯХ С ИСПОЛЬЗОВАНИЕМ АПОСТЕРИОРНЫХ ВЕРОЯТНОСТЕЙ[1)]

Рассматривается проблема выявления правдивых ответов на вопросы некоторого исследования в случае, когда респонденты имеют общее априорное распределение, не интересующее составителя опроса. В такой постановке составителю опроса желательно иметь универсальное правило, стимулирующее респондентов отвечать правдиво при любом априорном распределении. Если дополнительно выполняются условие локальности (которое гарантирует, что платежные функции правил определяются апостериорными вероятностями фактического состояния системы) и условие достаточной гладкости, мы доказываем, что равновесная платежная функция в случае правдивых ответов респондентов является логарифмической функцией апостериорных вероятностей. Более того, респонденты должны быть упорядочены в соответствии с этими вероятностями. В заключение обсуждаются вопросы применения полученных результатов.

Ключевые слова и фразы: собственные скоринговые правила, устойчивые/универсальные правила, байесовская сыворотка правды (Bayesian Truth Serum, BTS), реализация метода, ранжирование респондентов.

DOI: https://doi.org/10.4213/tvp5351

[*]California Institute of Technology, Division of the Humanities and Social Sciences, Pasadena, CA, USA; e-mail: cvitanic@hss.caltech.edu

[**]Massachusetts Institute of Technology, Sloan School of Management, Department of Economics, Department of Brain and Cognitive Sciences, Cambridge, MA, USA; e-mail: dprelec@mit.edu

[***]The Institute of Economics, Zagreb, Republic of Croatia; Massachusetts Institute of Technology, Sloan School of Management, Cambridge, MA, USA; e-mail: sradas@eizg.hr, sradas@mit.edu

[†]University of Zagreb, Faculty of Science, Department of Mathematics, Zagreb, Republic of Croatia; e-mail: hsikic@math.hr

[1)]Работа первого автора выполнена при финансовой поддержке NSF (грант DMS-1613170). Работа второго автора выполнена при поддержке Intelligence Advanced Research Projects Activity (IARPA), Department of Interior National Business Center

1. Введение. Рассмотрим задачу выявления правдивых ответов в опросе среди байесовских агентов, имеющих общее априорное распределение[2)]. Будем предполагать, что составитель опроса "безразличен" к априорному распределению (составитель может иметь предположения относительно априорного распределения, но он их не афиширует). При этом проверить, является ли ответ конкретного респондента правдивым, не представляется возможным. Если составитель хочет учитывать достоверность ответов и то, насколько серьезно респонденты относятся к опросу, можно использовать стимульно-согласованный метод или "скоринговое правило". Будем говорить, что скоринговое правило является (строго) стимульно-согласованным, если для каждого респондента π выполнено следующее условие: в том случае, когда остальные респонденты отвечают правдиво, правдивый ответ респондента π (строго) максимизирует его скоринговый балл. Стимульно-согласованные скоринговые правила играют важную роль при изучении опросов в различных областях, преимущественно в экономике и бизнесе. Таким образом, характеризация стимульно-согласованных алгоритмов представляет ценность как для исследователей, так и для специалистов, использующих данные алгоритмы на практике. Стоит ли ожидать существование большого числа стимульно-согласованных алгоритмов, или же стимульная согласованность является довольно ограничительным требованием? В работе показано, что при выполнении естественного свойства "равновесной локальности" стимульная согласованность является скорее ограничением и приводит к широко известному алгоритму. Перейдем к подробному рассмотрению полученных в настоящей работе результатов.

В работе [24] был введен алгоритм BTS (Bayesian Truth Serum), основанный на двух параметрах, определенных для каждого респондента: декларируемый тип респондента и его предположение о распределении типов, декларируемых другими участниками опроса. Этот алгоритм обладает двумя важными свойствами [24]. Во-первых, он является строго стимульно-согласованным (incentive compatible, IC), т.е. ситуация, когда все респонденты говорят правду, является равновесием. Таким образом, все "типы", соответствующие правдивым высказываниям агентов, становятся известными[3)]. Во-вторых, равновесное значение скорингового

(контракт D11PC20058). Работа третьего автора выполнена при поддержке Marie Curie International Outgoing Fellowship в рамках 7th European Community Framework Programme (грант PIOF-GA-2013-622868-BayInno). Работа четвертого автора выполнена при поддержке Министерства науки, образования и спорта (MZOS) Республики Хорватия (грант № 037-0372790-2799) и Хорватского научного фонда (проект 3526).

[2)] Например, при маркетинговом исследовании, имеющем целью выявление спроса на определенный новый продукт.

[3)] Фактически мы показываем, что все строгие равновесия в модели BTS либо соответствуют случаю правдивых ответов агентов, либо являются перестановкой типов в случае правдивых ответов, а скоринговые баллы являются единственными с точностью до линейного преобразования.

балла респондента в алгоритме BTS, с точностью до линейного преобразования, равно логарифму его апостериорной вероятности в случае правдивого ответа. Таким образом, BTS *упорядочивает респондентов* с помощью апостериорных вероятностей. Всюду далее мы будем называть данное упорядочивание апостериорным ранжированием и будем говорить, что алгоритм BTS приводит к *логарифмическому скорингу*. Метод BTS успешно применялся во многих областях, включая исследование восприятия новых продуктов на рынке [12], экономику и психологию [15], исследование эффективности обучающих программ [20], криминологию [17].

В настоящей статье рассматриваются следующие вопросы: при каких условиях равновесные платежные функции, полученные строгим стимульно-согласованном скоринговым правилом, соответствуют логарифмическому скорингу? При каких условиях равновесие в строгом стимульно-согласованном скоринговом правиле удовлетворяет условию апостериорной локальности? Другими словами, при каких условиях равновесные платежные функции эквивалентны платежным функциям в алгоритме BTS?

Перейдем к изложению основных результатов настоящей работы. Мы определяем два условия, которым должна удовлетворять равновесная платежная функция: “апостериорная локальность” и “разделение переменных”. Основная теорема утверждает, что

$$\left.\begin{array}{l}\text{стимульная согласованность}\\ \quad + \text{апостериорная локальность}\\ \quad + \text{разделение переменных}\end{array}\right\} \rightarrow \text{логарифмический скоринг.}$$

Второй основной результат работы утверждает, что

$$\left.\begin{array}{l}\text{стимульная согласованность}\\ \quad + \text{апостериорная локальность}\end{array}\right\} \rightarrow \text{апостериорное ранжирование.}$$

Чтобы строго сформулировать данные результаты, обсудим подробнее постановку задачи. Пусть состояние системы порождено некоторым конечным множеством и бесконечной популяцией (имея в виду приложения, можно считать, что популяция конечна, но достаточно большого размера). Имеется специальный человек, который составляет вопросы и просит респондентов ответить на них. Респонденты являются игроками в байесовской игре, где каждый игрок получает от составителя опроса вознаграждение, размер которого зависит от полученного балла. Величина балла, в свою очередь, зависит от ответов респондента и ответов остальных участников опроса. Каждый игрок имеет определенный тип, принадлежащий конечному множеству. Существование различных

типов респондентов можно объяснить тем, что игроки по-разному интерпретируют поступающие сигналы. (Как показано в [26], а также в настоящей статье с использованием другого доказательства, учет только лишь типов респондентов не приводит к правдивому равновесию.)[4)] Типы респондентов являются условно независимыми и одинаково распределенными, так что существует единственное "априорное" распределение, которое описывает совместное распределение состояния системы и типов респондентов. Будем предполагать, что априорное распределение известно игрокам, но заранее не известно составителю опроса.

Напомним понятие собственных скоринговых правил в случае, когда в системе имеется только один респондент. Рассмотрим случайную величину Ω, принимающую значения в множестве $\{1,\dots,N\}$, $N>1$. Она будет представлять состояние системы. На основании деклараций респондента можно определить его мнение $\widetilde{p}=(\widetilde{p}^1,\dots,\widetilde{p}^N)$ относительно распределения величины Ω. Если реализуется исход $\Omega=i$, респонденту выплачивают $F_i(\widetilde{p})$. В случае, когда в системе имеется несколько респондентов, каждое i будет само по себе являться распределением, зависящим от мнений респондентов. В этом случае, если ответ респондента влечет $\widetilde{p}$ и реализуется исход $\Omega=i$, мы будем называть $\widetilde{p}^i$ локальной апостериорной вероятностью.

При наличии одного-единственного респондента семейство функций $\{F_i\}_{i=1,\dots,N}$ называется строго собственным скоринговым правилом, если оно является стимульно-согласованным для случая правдивых ответов. Это означает, что ожидаемый платеж респонденту достигает максимума, когда респондент говорит правду. При этом максимум достигается на его апостериорной вероятности p, т.е. для любого вектора вероятностей $\widetilde{p}\neq p$ справедливо неравенство

$$\sum_{i=1}^{N} p^i F_i(p) > \sum_{i=1}^{N} p^i F_i(\widetilde{p}). \tag{1.1}$$

Существует много собственных скоринговых правил. Общая характеризация правил и многочисленные примеры приведены в [11][5)]. Важным частным случаем является случай $F_i(p)=F_i(p^i)$, когда F_i зависит только от локальной апостериорной вероятности p^i — вероятности, назначаемой респондентом исходу $\Omega=i$, который и реализуется фактически. При этом F_i не зависит от вероятностей, соответствующих другим

[4)] Мы ожидаем, что наши результаты будут приближенно справедливы и в конечномерном случае, но для выборок больших размеров. Точная теория для конечномерного случая совершенно иная и будет рассмотрена в последующих исследованиях. Некоторые элементы конечномерной теории в теоретико-информационной постановке имеются в [7].

[5)] В работе [23] рассмотрен случай, когда у системы может быть только два состояния, а у респондентов могут быть определенные выгоды, о которых они не предполагали заранее.

гипотетическим исходам. В этом случае скоринговое правило является линейным преобразованием логарифмов p^i [28], [4]. Данное правило естественно применять в том случае, когда локальная апостериорная вероятность является мерой компетентности респондента, т.е. когда качество сигнала от респондента измеряется вероятностью, соответствующей фактическому состоянию системы.

Как отмечено выше, в случае нескольких игроков значения случайной величины Ω являются распределениями типов игроков, и эти типы могут потенциально принимать бесконечное множество значений. При этом апостериорные вероятности являются вероятностями того, что тип игрока соответствует значению величины Ω. Мы всегда будем предполагать, что число распределений конечно. На практике респондентам будет предложен дискретный набор распределений, например, "На ваш взгляд, за кандидата A проголосуют 0–10%, 10–20%, ... избирателей?". Если предположить, что наша игра имеет равновесие, строго разделяющее типы игроков, то результаты будут зависеть только от вида равновесных платежных функций и не будут зависеть от скоринговых правил, имеющих данные платежные функции. Условие апостериорной локальности постулирует, что равновесные платежные функции F_i в состоянии $\Omega = i$ являются функциями локальных апостериорных вероятностей, т.е. апостериорных вероятностей событий $\Omega = i$.

Предвосхищая результаты настоящей работы, покажем, что при выполнении некоторых условий гладкости если равновесные платежные функции F_i удовлетворяют свойству, аналогичному (1.1), то разность скоринговых баллов двух респондентов с локальными апостериорными вероятностями p^i и q^i в состоянии i с точностью до слагаемых первого порядка включительно должна быть пропорциональной $\ln p^i - \ln q^i$ при $q \approx p$. Чтобы пояснить, что мы подразумеваем под этим, рассмотрим для простоты случай двух типов респондентов с локальными апостериорными вероятностями p^i и q^i соответственно. Если из высказываний игрока следует, что локальная апостериорная вероятность равна p^i, он получает выплату $F_i(p^i, q^i)$. Если из его высказываний следует, что эта вероятность есть q^i, то выплата равна $F_i(q^i, p^i)$. Стимульная согласованность семейства $\{F_i\}$ в обобщение стимульной согласованности собственных скоринговых правил (1.1) в случае одного игрока, означает, что решением задачи

$$\min_{q^i}\left\{\sum_i p^i[F_i(p^i, q^i) - F_i(q^i, p^i)] + \lambda \sum_i q^i\right\}$$

является $q^i = p^i$, где λ — множитель Лагранжа для ограничения $\sum_i q^i = 1$. Выпишем условие первого порядка для данной задачи:

$$\partial_q F_i(p^i, p^i) - \partial_p F_i(p^i, p^i) = -\frac{\lambda}{p^i},$$

где ∂_x — частная производная по переменной x. В случае одного игрока это условие принимает вид $F'(p^i) = \lambda/p^i$ и приводит к логарифмической функции, поскольку она является единственным собственным скоринговым правилом, удовлетворяющим условию локальности. В случае нескольких игроков зафиксируем p^i и разложим разность скоринговых баллов как функцию q^i до первого порядка в окрестности точки p^i. Получим

$$F_i(p^i, q^i) - F_i(q^i, p^i) \approx [\partial_q F_i(p^i, p^i) - \partial_p F_i(p^i, p^i)](q^i - p^i).$$

Наконец, объединяя полученные выше соотношения и используя разложение Тейлора для логарифма до первого порядка в окрестности точки $q^i = p^i$, $\ln p^i - \ln q^i = (p^i - q^i)/p^i$, получим

$$F_i(p^i, q^i) - F_i(q^i, p^i) \approx \lambda\left(1 - \frac{q^i}{p^i}\right) \approx \lambda(\ln p^i - \ln q^i).$$

На основе этой аппроксимации в нашей первой теореме утверждается следующее: если к условию стимульной согласованности добавить условия умеренной гладкости платежной функции и "чувствительности" разности равновесных платежных функций двух респондентов к другим игрокам, то разность между стимульно-согласованными скоринговыми баллами двух респондентов пропорциональна разности логарифмов локальных апостериорных вероятностей.

Наша вторая теорема утверждает, что любая стимульно-согласованная равновесная платежная функция игрока $F_i(p^i_k, p^i_{-k})$, является неубывающей функцией локальной апостериорной вероятности p^i_k. В теореме предполагается, что из ответов игрока следует апостериорная вероятность p^i_k, соответствующая типу k, а из ответов остальных игроков — апостериорные вероятности p^i_{-k}. Следовательно, если рассматривать p^i_k как меру компетентности участников опроса, то упорядочивание респондентов не зависит от стимульно-согласованного метода и соответствует ранжированию апостериорными вероятностями. Данный результат является весьма общим и доказывается с помощью алгебраических методов. Он обобщает также существующие результаты в случае одного респондента о монотонности, следующей из стимульной согласованности собственных скоринговых правил. См., например[6)], [19], [28]–[30].

В работе также обсуждаются вопросы применения полученных результатов. Заметим, что, хотя платежная функция вида $F_i(p^i_k, p^i_{-k})$ возникает при нахождении равновесия в теории, найти ее на практике не всегда просто. Проблема состоит в том, как найти теоретически оптимальную платежную функцию, используя только ответы респондентов

[6)] См. также [25], где определяется и тестируется более широкий класс алгоритмов, упорядочивающих респондентов в соответствии с их апостериорными вероятностями. Среди алгоритмов данного класса только про BTS известно, что он является стимульно-согласованным.

на вопросы, придуманные независимым составителем, при том, что сам опрос должен быть как можно более простым. Используя несколько более сильное условие, чем апостериорная локальность, но не используя разделение переменных, мы показываем, что платежные функции всех строго разделяющих равновесий в нашей постановке могут быть найдены с помощью некоторого опроса. Однако сам опрос может быть достаточно сложным, за исключением случая логарифмического BTS. В связи с этим напомним результат работы [24], где показано, что если скоринговые баллы респондентов подчиняются модели BTS, то равновесные скоринговые баллы имеют вид $\ln p_k^i$ (плюс слагаемое, не зависящее от респондента). Мы рассматриваем данный результат и приводим подробное доказательство. В работе также показывается, что бюджетно-сбалансированное строгое равновесие в модели BTS является разделяющим.

Связь с существующими результатами. В последние годы, начиная с работы [24], собственные скоринговые правила в контексте теории игр, также называемые "извлечение информации без верификации", стали предметом широкого изучения. В случае, когда составитель опроса знает априорное распределение типов игроков, примечательна ранняя работа [21], в которой был разработан остроумный подход к использованию собственных скоринговых правил для выявления правдивой информации. Однако, так как предположение о том, что составитель опроса знает априорное распределение респондентов, на практике невозможно, были предложены альтернативные подходы. В работе [13] авторы использовали устойчивую оптимизацию для работы с небольшими отклонениями в моделях доверия. Обобщая постановку из [24], в которой составитель опроса не знает априорного распределения, но количество участников опроса бесконечно, Витковски и Паркс [34], [33] разработали методы под общим названием "устойчивая байесовская сыворотка правды" (Robust Bayesian Truth Serum, RBTS). Данные методы работают для конечного числа респондентов, но только в случае двух типов. В работе [32] была рассмотрена общая постановка без предположений о структуре информации. В работах [26], [27] и [36] были разработаны методы, являющиеся стимульно-согласованными для любого количества респондентов, типы которых не являются бинарными. В работах [9] и [35] требования о знании составителем опроса априорных распределений респондентов были ослаблены еще больше с помощью включения в опрос множества схожих по смыслу вопросов. В [2] с помощью байесовского рынка найдено правдивое равновесие для случая бинарных типов. В [8] разработаны дополнительные стимульно-согласованные правила, которые легко объяснить респондентам. В работе [10] авторы охарактеризовали "минимальные" парно-предсказательные методы, в которых скоринговый балл зависит от типа, декларируемого респондентом, и от типа, который декларирует его определенным образом выбранный "кол-

лега". Мы же, с одной стороны, позволяем скоринговым баллам зависеть от декларируемых типов респондентов и от их предположений о том, какие типы будут декларированы другими респондентами; с другой стороны, мы используем предположение локальности скоринговых баллов. Выбор того или иного подхода определяется тем, какое из двух свойств, минимальность или локальность является наиболее важным. В работе [14] авторы разработали теоретическо-информационную концепцию для создания механизмов стимулирования, включающую в качестве частных случаев множество известных механизмов, в том числе и BTS. В [14] также показано, что свойства BTS могут быть доказаны более простым способом, если использовать связь с взаимной информацией Шеннона. Другая связь между методом BTS и теорией информации обсуждается в [7]. В работе [16] разработана "равномерно доминирующая" сыворотка правды, в которой истинный сигнал передается с шумом и имеется достаточное число респондентов и вопросов. Скоринговые правила в указанной работе зависят от того, является ответ респондента информативным или нет, и, таким образом, в некотором смысле связаны с компетентностью респондента.

Однако во всех упомянутых выше работах, за исключением [24], изучалась стимульная согласованность и не рассматривалась проблема упорядочивания респондентов. Таким образом, предложенные в этих работах методы либо не удовлетворяют нашему условию апостериорной локальности, либо требуют, чтобы составитель опроса знал априорные распределения респондентов. В центре нашего внимания будут методы, обладающие всеми тремя свойствами: они допускают IC-равновесие; они применимы и в том случае, когда составитель опроса безразличен к априорным вероятностям; они упорядочивают респондентов с помощью апостериорных вероятностей. Основной результат настоящей работы состоит в том, что в этом случае при выполнении относительно слабых условий равновесная платежная функция может иметь только логарифмический вид. Если же апостериорное ранжирование от метода не требуется, то в работах, отмеченных выше, приведено множество способов разработки IC-методов.

Проблему, изучаемую в настоящей работе, можно рассматривать как задачу разработки метода, поскольку мы стараемся описать механизмы, которые были бы стимульно-согласованными и в тоже время обладали свойствами, интересными для приложений в области выявления истинных суждений респондентов. С одной стороны, наш подход является более общим, чем просто разработка моделей[7)], поскольку мы допускаем неопределенность как в отношении имеющейся информации об игроках (типах игроков), так и в отношении фактического состояния системы,

[7)] См., например, [18], [3] и [5]. В работе [3] имеется подробный обзор существующих результатов.

причем эти две неопределенности могут быть зависимы нетривиальным образом. Именно их совместное распределение определяет все результаты. Наше основное предположение состоит в том, что респонденты имеют общее априорное представление о совместном распределении, но это представление не используется составителем опроса. Мы формулируем это скорее как методологическое, чем как содержательное требование: хотя составитель опроса может иметь предположения относительно априорного распределения, он придерживается нейтральной позиции, не используя эти предположения и не делясь ими с участниками опроса. Таким образом, составителю нужен "универсальный" метод, который бы работал для любого априорного распределения и не требовал от составителя других сведений, кроме формулировки самих вопросов с несколькими вариантами ответа[8)]. С этой точки зрения в настоящей работе исследуются устойчивые байесовские методы. С другой стороны, наша постановка является менее общей в следующем смысле: у респондентов есть только одно возможное действие — ответить на вопросы, и за это действие не предусмотрена плата. Таким образом, в работе отсутствует моделирование полезности тех или иных действий; единственная величина, связанная с полезностью действий, — это ожидаемый платеж, получаемый респондентом. Более того, при разработке некоторых устойчивых моделей не предполагают, что все респонденты знают априорное распределение. Мы же считаем, что априорное распределение может быть неизвестно только составителю опроса. В этом смысле постановка, используемая в настоящей работе, является менее общей. В заключении (раздел 5) мы обсуждаем, в каких направлениях можно попытаться обобщить наши результаты.

Оставшаяся часть работы организована следующим образом: в разделе 2 определяется модель, в разделе 3 приводятся основные теоретические результаты, в разделе 4 обсуждаются вопросы применения, в разделе 5 приводится заключение. Доказательства вынесены в приложение (раздел 6).

2. Модель, определения и предположения. В рамках нашей модели метод состоит в присвоении скоринговых баллов игрокам (респондентам) различных типов[9)]. Приложением таких моделей являются опросы: респондентов просят ответить на заранее составленные вопросы. Цель составителя опроса — выявление правдивых ответов на вопросы с несколькими вариантами ответов и упорядочивание игроков в соответствии с качеством предоставляемой ими информации. В нашей поста-

[8)] Теоретически использование "правила большинства", которое потребовало бы декларирования общего априорного распределения, может привести к равновесию, которое и выявило бы это общее априорное распределение; однако, и это обсуждается в настоящей статье, данное правило на практике не применимо.

[9)] Отрицательный балл в литературе обычно называется трансфером (см., например, [5]).

новке это будет означать упорядочивание респондентов в соответствии с их апостериорными вероятностями фактического состояния системы. Например, целью составителя опроса могла бы быть оценка стоимости определенной бутылки вина через несколько лет, он подбирает вопросы и задает их экспертам. К более общим приложениям модели относятся исследование настроений избирателей перед голосованием на выборах, предсказание политических событий, маркетинговые исследования рынка, онлайн-исследования спроса на продукты, а также любые другие приложения, где может быть проведено анкетирование, включающее вопросы с несколькими вариантами ответов[10)]. Отметим, что в рамках модели не обязательно присутствует возможность проверки правильности ответов на вопросы с несколькими вариантами ответов.

2.1. Модель. Игроки проиндексированы с помощью параметра $\pi \in R$, где R — бесконечное счетное множество[11)]. Состояние системы является случайной величиной Ω, принимающей значения в множестве $\{1, \ldots, N\}$, $N > 1$[12)]. У каждого игрока есть тип, количество различных типов равно $M > 1$. Типы игроков можно понимать как случайные сигналы, получаемые игроками о состоянии системы. Тип игрока π является случайной величиной, которая обозначается через T^π и принимает значения $t^\pi \in \{1, \ldots, M\}$. Мы будем рассматривать скоринговые методы, в которых игрок π дает ответ $a^\pi \in \mathbf{R}^K$ (a — от англ. *action*), имеющий размерность K, где K — фиксированное натуральное число.

Ответ a^π будет обычно включать декларируемый респондентом тип (выбор одного из возможных ответов), а также ответы на некоторые другие вопросы для стимулирования респондента говорить правду[13)]. Ответ респондента может включать также его суждение об априорном распределении типов и состояний системы (определения см. ниже). Иными

[10)] Большое количество примеров приведено в работах [24] и [25].

[11)] Предположение о бесконечности числа игроков нами делается по нескольким причинам. Во-первых, мы не хотим накладывать ограничений на форму платежной функции вне равновесия; для этого мы будем пользоваться тем, что при наличии бесконечного числа игроков форма равновесной платежной функции не меняется, когда игрок одного типа имитирует равновесную стратегию другого типа. Во-вторых, достигнуть истинности типов намного сложнее в случае конечного числа игроков. Также в случае конечного числа игроков сложно найти равновесные платежные функции, используя начальные данные, встречающиеся на практике. Случай конечного числа игроков будет нами исследован в будущих работах. Наконец, наличие бесконечного числа игроков позволит применить теорему де Финетти для нашей модели.

[12)] Строго говоря, данное предположение является приближением. На практике может быть бесконечное множество состояний системы, имеющее мощность континуума. В примере с определением стоимости бутылки вина состоянием системы может быть процент экспертов, считающих, что бутылка будет стоить более тысячи долларов.

[13)] В разделе 4 мы увидим, что другой вопрос может быть о проценте респондентов, выбравших определенный вариант ответа из списка предложенных.

словами, респондента могут попросить высказать свои априорные предположения. Мы постулируем следующее.

Предположение 2.1. (i) Семейство сигналов T^π, $\pi \in R$, является семейством взаимозаменяемых случайных величин. Случайные величины T^π, $\pi \in R$, являются независимыми одинаково распределенными при условии состояния Ω системы.

(ii) Если респондент π выбирает ответ a^π, а ответы остальных респондентов обозначены через $a^{-\pi}$, то его скоринговый балл определяется *скоринговой функцией* $f(a^\pi, a^{-\pi})$. Порядок ответов респондентов в $a^{-\pi}$ не влияет на значение функции f, т.е. она симметрична по аргументам в $a^{-\pi}$.

Условие (i) означает, что порядок, в котором мы рассматриваем игроков, не имеет значения (с точки зрения вероятностного распределения всей последовательности). Более того, по теореме де Финетти из предположения о взаимозаменяемости следует вторая часть предположения 2.1, (i), о том, что существует случайная величина Ω такая, что величины T^π независимы и одинаково распределены при условии Ω (см., например, [1] или [6]).

Свойство симметричности в (ii) является естественным ограничением, если учитывать тот факт, что составитель опроса не делает различия между типами, которые по условию (i) являются взаимозаменяемыми.

Всюду далее мы будем предполагать, что игроки являются риск-нейтральными, т.е. задача каждого игрока состоит в максимизации его ожидаемого платежа[14)].

Априорные и апостериорные вероятности. Совместное распределение типов и состояний системы задается $(M \times N)$-матрицей $Q = [q_k^i]$, где

$$q_k^i = \Pr(T^\pi = k,\, \Omega = i).$$

Заметим, что Q на самом деле не зависит от π. Это является следствием предположения о взаимозаменяемости T^π.

Будем считать, что матрица Q известна всем игрокам, и называть ее общим априорным распределением. Однако при составлении опроса матрица Q не используется. Фактически составитель опроса может даже не знать числа N состояний системы. Единственное, что должен знать составитель опроса, это M. Например, M нужно знать при составлении

[14)] Обычно при разработке моделей только типы респондентов считаются случайными величинами с априорными распределениями, известными и составителю опроса. В этом смысле наша модель является более общей, так как мы считаем случайными не только типы, но и состояния системы, причем корреляция между ними невырождена. С другой стороны, рассматриваемая нами модель является менее общей в том смысле, что у игроков есть только одно возможное действие — выбирать ответ на вопрос.

вопросов с несколькими вариантами ответов — количество вариантов ответов должно быть равно количеству типов[15)].

Матрица Q определяет маргинальные распределения типов, которые мы будем называть *вероятностями типов*. А условные вероятности состояний системы при фиксированных типах будем называть *апостериорными вероятностями*. Данные вероятности обозначаются через s_k и z_k^i соответственно:

$$s_k = \Pr(T^\pi = k) \quad \text{и} \quad z_k^i = \Pr(\Omega = i \mid T^\pi = k).$$

Мы предполагаем, что маргинальные вероятности типов и состояний системы строго положительны. Апостериорные вероятности образуют матрицу $Z = [z_k^i]_{k=1,i=1}^{M,N}$. Заметим, что z_k^i не зависят от π, для любого $k \in \{1, \dots, M\}$ выполнено равенство $\sum_{i=1}^N z_k^i = 1$ и любую матрицу, обладающую данным свойством, можно рассматривать как Z-матрицу апостериорных вероятностей некоторого совместного распределения Q. Вектор $(s_1, \dots, s_M)$ обозначается через S.

2.2. Равновесные платежные функции и стимульная согласованность. В литературе по скоринговым правилам обычно предполагается только один респондент, которого просят дать апостериорное распределение случайной величины Ω, т.е. z^i. Если реализуется исход $\Omega = i$, респонденту выплачивается величина $F_i(z)$. Семейство функций $\{F_i\}_{i=1,\dots,N}$ называется строго собственным скоринговым правилом, если оно является стимульно согласованным для случая правдивых ответов респондентов, т.е. ожидаемый платеж респонденту достигает максимума, когда респондент говорит правду; это означает, что для любого $\widetilde{p} \neq p$ мы имеем $\sum_{i=1}^N p^i F_i(p) > \sum_{i=1}^N p^i F_i(\widetilde{p})$.

В нашей постановке с бесконечным множеством респондентов *мы рассматриваем только платежные функции, допускающие строго разделяющее байесовское равновесие Нэша* (*strictly separating Bayesian Nash equilibrium*, SSNE, *см. определение* 2.1 *ниже*), *в котором равновесными платежами являются функции* $F_i\colon (0,1)^{2M} \to \mathbf{R}$ *вида* $F_i(z_k^i, z_{-k}^i; s_k, s_{-k})$, где $z_{-k}^i = (z_1^i, \dots, z_{k-1}^i, z_{k+1}^i, \dots, z_M^i)$. Назовем данное свойство *апостериорной локальностью*.

Сформулируем сказанное выше более точно. Назовем чистой стратегией для игрока π отображение σ^π, которое типу игрока ставит в соответствие его ответ a^π. Мы будем допускать только чистые стратегии. Обозначим через $\sigma(t)$ множество всех чистых стратегий, через $\sigma^\pi(t^\pi)$ чистые стратегии игрока π и через $\sigma^{-\pi}(t^{-\pi})$ все чистые стратегии, за

[15)] Чтобы обойти проблему незнания общего априорного распределения, составитель опроса может попросить каждого игрока указать априорное распределение и строго наказывать тех игроков, чьи ответы отличаются от остальных. Однако такой подход вряд ли будет работать на практике — скорее всего, большинство ответов будут отличаться друг от друга, и тогда составителю опроса придется наказать большинство респондентов.

исключением стратегий игрока π. Скоринговый балл игрока π есть $f(\sigma^\pi(t^\pi), \sigma^{-\pi}(t^{-\pi}))$, где f — скоринговая функция, являющаяся отображением из множества ответов респондентов в множество вещественных чисел. Функция $f(\cdot,\cdot)$ имеет один и тот же вид для всех N и Q.

Мы предполагаем, что игроки стремятся максимизировать ожидаемый скоринговый балл. В основном будут рассматриваться только "бюджетно-сбалансированные" платежные механизмы, т.е. такие, для которых сумма скоринговых баллов всех игроков равна нулю с вероятностью единица[16)].

Приведем определение равновесия.

Определение 2.1. (i) Пусть Q — матрица априорных вероятностей. Будем говорить, что скоринговая функция f допускает строгое (байесовское) равновесие Нэша (strict (Bayesian) Nash equilibrium, SNE), если существует множество $\sigma = \sigma_Q$ чистых стратегий такое, что для всех π, t^π, $t^{-\pi}$, t^γ справедливо следующее: для любого ответа $a^\pi \neq \sigma^\pi(t^\pi)$ имеем

$$\mathbf{E}[f(a^\pi, \sigma^{-\pi}(t^{-\pi})) \mid T^\pi = t^\pi] < \mathbf{E}[f(\sigma^\pi(t^\pi), \sigma^{-\pi}(t^{-\pi})) \mid T^\pi = t^\pi],$$

где математическое ожидание берется относительно (условного) распределения случайной величины Ω.

В этом случае множество стратегий σ называется SNE. Если равновесие является также разделяющим (т.е. дополнительно к свойству выше справедлива импликация $\sigma^\pi(t^\pi) = \sigma^\gamma(t^\gamma) \Rightarrow t^\pi = t^\gamma$), то σ называется строго разделяющим (байесовским) равновесием Нэша (strictly separating (Bayesian) Nash equilibrium, SSNE).

(ii) Будем говорить, что скоринговая функция f является универсальным разделяющим скоринговым правилом (universal separating scoring rule, USSR), если для любого Q существует хотя бы одно бюджетно-сбалансированное SNE σ_Q и каждое бюджетно-сбалансированное SNE является SSNE.

Далее в статье будет показано, что бюджетно-сбалансированный логарифмический скоринг может быть реализован с помощью USSR-функции (точнее, с помощью BTS), которая удовлетворяет предположениям ниже о равновесных платежных функциях.

Всюду далее в работе мы будем рассматривать только USSR-функции f или их не бюджетно-сбалансированные аналоги такие, что существует по крайней мере одно SSNE. Будем использовать обозначение F_i

[16)] Стоит отметить, что в бюджетно-сбалансированной игре участники знают, что они могут получить отрицательные "платежи", и поэтому некоторые игроки могут не захотеть участвовать. На практике платежи часто осуществляются не в денежном выражении, а в виде скоринговых баллов. Каждый участник будет получать неотрицательную выплату, состоящую из фиксированной и переменной частей. Размер переменной части выплаты зависит от скорингового балла участника или от его ранга, присвоенного на основании балла. Таким образом, на практике может использоваться не само бюджетно-сбалансированное скоринговое правило, а его модификация.

для платежной функции в SSNE, соответствующей f при фактическом состоянии системы i. Следующее дополнительное условие на равновесные платежные функции F_i, допускающие SSNE, является ключевым.

Предположение 2.2 (апостериорная локальность). При любых $k \in \{1,\dots,M\}$ и $i \in \{1,\dots,N\}$, если $T^\pi = k$ и $\Omega = i$, то для равновесного скорингового балла (но не обязательно для внеравновесного скорингового балла) игрока π имеет место представление

$$f(\sigma_Q^\pi(k), \sigma_Q^{-\pi}(t^{-\pi})) = F_i(z_k^i, z_{-k}^i; s_k, s_{-k}),$$

где $F_i\colon (0,1)^{2M} \to \mathbf{R}$.

Данное предположение обсуждается в конце настоящего раздела. Мы не изучаем вопрос единственности равновесия и рассматриваем только те равновесия, в которых платежные функции (но не обязательно внеравновесные платежные функции) имеют вид, приведенный выше. Более того, мы будем предполагать, что полученные равновесные платежные функции удовлетворяют свойствам, приведенным в следующем определении.

Определение 2.2. Семейство $\{F_i\}$ функций вида $F_i(z_k^i, z_{-k}^i; s_k, s_{-k})$ называется апостериорно-локальной равновесной платежной системой (posterior-local equilibrium payoff system, PLEPS), если выполнены следующие условия.

(i) *Симметричность*: для любых $x, y \in (0,1)$, любых $z_2,\dots,z_M$, $s_2,\dots,s_M \in (0,1)$ и любой перестановки Π множества $\{2,\dots,M\}$ имеем

$$\begin{aligned} &F_i(x, z_2, z_3, \dots, z_M; y, s_2, \dots, s_M) \\ &\qquad = F_i(x, z_{\Pi(2)}, z_{\Pi(3)}, \dots, z_{\Pi(M)}; y, s_{\Pi(2)}, s_{\Pi(3)}, \dots, s_{\Pi(M)}). \end{aligned}$$

(ii) *Стимульная согласованность*, *строго разделяющее неравенство*: для любой Z-матрицы, любого S-вектора и всех $k, j \in \{1,\dots,M\}$ таких, что $(z_k^1,\dots,z_k^N) \neq (z_j^1,\dots,z_j^N)$, выполнено неравенство

$$\sum_{i=1}^N z_k^i F_i(z_k^i, z_{-k}^i; s_k, s_{-k}) > \sum_{i=1}^N z_k^i F_i(z_j^i, z_{-j}^i; s_j, s_{-j}). \tag{2.1}$$

Условие симметричности (i) означает, что равновесный скоринговый балл для типа k не зависит от порядка расположения других типов. Это согласуется с предположением 2.1, (ii), о симметричности скоринговой функции f. В условии (ii) неявно предполагается, что игроки являются риск-нейтральными и максимизируют свой ожидаемый скоринговый балл. Из предложения 2.1 (см. далее) следует, что условие (ii) автоматически выполнено, если F_i являются равновесными платежными функциями в случае правдивых ответов респондентов.

Изучим подробнее вопрос о предполагаемом виде равновесных платежных функций F_i.

Замечание 2.1. Ключевым предположением в настоящей работе является то, что скоринговый балл игрока при равновесии зависит от его апостериорной вероятности z^i при фактическом состоянии системы, равном i (z^i называется локальной апостериорной вероятностью). Это оправдано, если апостериорная вероятность является хорошей мерой компетентности игрока. Заметим, однако, что здесь речь идет о компетентности игрока при фактическом состоянии системы для одного конкретного исследования, а не о средней точности по результатам нескольких исследований. Рассматриваемая экспертиза является фактической, получающейся в результате принимаемого игроком сигнала, а не прогнозируемой, когда оценивается качество полученного сигнала.

В некоторых случаях составителю опроса требуется знать распределение типов. Это так, например, в случае выборов или в случае исследования рынка для нового продукта, когда мы пытаемся оценить процент участников, проголосующих за каждого кандидата или готовых купить предложенный продукт. В таких случаях интуитивно понятно, что респондент с более высокой апостериорной вероятностью является лучшим экспертом — у него наиболее высокая вероятность оказаться правым насчет фактического распределения ответов. Это напоминает концепцию оценок максимального правдоподобия, которые максимизируют вероятность события, которое действительно происходит. Более того, если опрос состоит из нескольких этапов (например, при исследовании рынка), то на первом этапе методом, используемым для получения платежных функций PLEPS, можно выделить экспертов и на следующих этапах опрашивать уже только их, тем самым сократив расходы на проведение исследования. Именно такие приложения мы в первую очередь и имеем в виду. В ряде других приложений, например, когда экономистов спрашивают, превысит ли инфляция в текущем году некоторый фиксированный уровень, уже не так очевидно, что более высокая локальная апостериорная вероятность на распределении типов означает более высокую компетентность эксперта. Это происходит потому, что в данном примере респонденты, лучше других оценивающие процент своих коллег, которые предскажут высокий уровень инфляции, не обязательно окажутся лучшими в оценке самой инфляции. В таких случаях более подходящими могут быть скоринговые правила, отличные от правил, использующих платежные функции PLEPS. В частности, если составитель опроса не ставит перед собой цели выявить экспертов, а его интересуют только правдивые ответы на вопросы, в предположениях модели можно исключить некоторые в высшей степени разумные скоринговые правила. Этот факт также отмечен в работах, упомянутых в обзоре литературы во введении.

Стоит отметить, что в настоящей работе мы стремимся к тому, чтобы равновесные платежные функции, которые описывают степень компе-

тентности игроков, были как можно проще. Поэтому мы считаем, что платежная функция F не зависит от локальных вероятностей, которые можно получить из априорного распределения. С другой стороны, мы позволяем платежной функции зависеть от вероятностей s_k, s_{-k} по той причине, что в приложениях данные вероятности будут совпадать с частотами типов, которые в свою очередь можно использовать для того, чтобы сделать метод бюджетно-сбалансированным. На самом деле для бюджетной сбалансированности достаточно иметь зависимость от локальных условных вероятностей $s_k^i = \Pr(T^\pi = k \mid \Omega = i)$, но для большей общности мы допускаем зависимость от s_k, s_{-k} (везде за исключением раздела 4, где мы обсуждаем применение полученных результатов).

Естественным образом возникает следующий вопрос: для каждой ли PLEPS-функции F существует скоринговое правило f, которое реализует F в равновесии? В разделе 4 мы показываем, что ответ на данный вопрос положительный в случае, когда F_i, вместо того, чтобы зависеть от, возможно, всех s_k, s^{-k}, зависит только от $s_k^i = \Pr(T^\pi = k \mid \Omega = i)$. Другим естественным вопросом является единственность равновесия, реализующего F для заданной f. Далее будет показано, что единственность имеет место для случая скорингового правила BTS.

Следующий результат является простым, но в то же время ключевым для нас. Он показывает, что собой представляет скоринговый балл для типа, который имитирует равновесную стратегию другого типа. Подчеркнем, что для данного результата требуется наличие бесконечного числа игроков.

Предложение 2.1. *Предположим, что для некоторой игры, фактические платежные функции которой заданы* PLEPS*-функциями* $\{F_i\}$*, существует строго разделяющее байесовское равновесие Нэша. Тогда, если респондент типа* k *отклоняется от равновесия, используя стратегию типа* $j \neq k$*, его платежная функция равна* $F_i(z_j^i, z_{-j}^i; s_j, s_{-j})$*. Таким образом, если игрок типа* k *имитирует равновесную стратегию типа* j*, его платежная функция будет задаваться реализацией равновесия, соответствующей типу* j*.*

Данное утверждение справедливо в силу того, что каждый тип представлен бесконечным числом игроков, равновесные платежные функции являются строго разделяющими, а скоринговая функция f симметрична относительно ответов. Доказательство предложения 2.1 см. в разделе 6.

Следующий отрицательный результат справедлив для случая конечного числа игроков (доказательство см. в разделе 6)[17].

Предложение 2.2. *Предположим* (*только в рамках данного предложения*)*, что в системе имеется конечное, но не меньшее двух число*

[17] Более тщательный анализ случая конечного числа игроков мы откладываем на будущее.

игроков. Тогда не существует бюджетно-сбалансированной системы функций PLEPS.

Отметим также, что в случае конечного числа игроков даже не бюджетно-сбалансированная версия метода BTS не является стимульно-согласованной.

Практическая реализация ожидаемых и фактических платежных функций. Даже если отождествить состояния системы с возможными эмпирическими частотами ответов, процесс нахождения апостериорных вероятностей состояний системы посредством вопросов респондентам на практике, по всей видимости, окажется чрезмерно сложным, поскольку от респондентов будет требоваться предоставить распределение для всех возможных эмпирических частот. Поэтому на практике составитель опроса, который хочет иметь метод, дающий фактические платежные функции F_i в случае, когда игроки говорят правду, будет искать способ найти фактические платежные функции, обещая заплатить игрокам в соответствии с их ожидаемым скоринговым баллом; а это требует от игроков значительно более простой информации, чем их мнение о распределении эмпирических частот. Мы обсудим это в разделе 4, а здесь лишь заметим следующее. Рассмотрим модельный пример системы PLEPS — логарифмическую платежную функцию скорингового правила:

$$F_i(z_k^i, z_{-k}^i; s_k, s_{-k}) = \ln(z_k^i).$$

В работе [24] показано, что бюджетно-сбалансированная версия данной платежной функции может быть реализована следующим образом: респондентов (в бесконечном числе) просят не только декларировать свой тип (выбрав ответ среди нескольких вариантов), но и высказать их мнение относительно процента других типов в выборке, т.е. указать эмпирические частоты каждого варианта ответа. Участникам опроса гораздо проще дать оценки эмпирических частот, чем оценить вероятностное распределение эмпирических частот. В логарифмическом случае это означает, что респондента типа k не спрашивают о вероятностях z_k^i и ему не обещают выплатить $\ln z_k^i$, однако его просят дать более простую информацию, которая определит его ожидаемый скоринговый балл с помощью специальной функции f (называемой байесовской сывороткой правды, Bayesian Truth Serum); при этом значение скорингового балла оказывается равным $\ln z_k^i$ в случае, когда игроки разыгрывают правдивое равновесие.

3. Всевозможные равновесные платежные функции.

3.1. Логарифмические равновесные платежные функции. Мы приведем в этом пункте примеры семейств функций PLEPS и рассмотрим вопрос о том, являются ли логарифмические равновесные платежные функции и их простейшие модификации единственно возможными системами PLEPS.

3.1.1. *Модельный пример — логарифмическая функция.* Каноническим примером PLEPS (без учета бюджетной сбалансированности) является логарифмическая функция:

$$F_i(z_k^i, z_{-k}^i; s_k, s_{-k}) = \ln z_k^i.$$

Точнее, равновесная платежная функция игрока есть логарифм апостериорной вероятности состояния системы при заданном типе игрока. Хорошо известно (и легко проверить), что данная функция удовлетворяет строго разделяющему неравенству (2.1). Это следует из хорошо известного *неравенства Гиббса*, согласно которому для вектора вероятностей $(p^1, \ldots, p^N)$ мы имеем

$$0 = \min_{q^i \geqslant 0, \sum_i q^i = 1} \sum_{i=1}^{N} p^i [\ln p^i - \ln q^i]. \tag{3.1}$$

Последнее равенство можно проверить, рассмотрев задачу

$$0 = \min_{q^i} \left\{ \sum_{i=1}^{N} p^i [\ln p^i - \ln q^i] + \lambda \sum_i q^i \right\}, \tag{3.2}$$

где λ является множителем Лагранжа для ограничения $\sum_i q^i = 1$. Условиями первого порядка являются $p^i/q^i = \lambda$, которые выполняются при $q^i = p^i$.

Возникает вопрос, является ли логарифмическая функция единственно возможной PLEPS (без учета бюджетной сбалансированности). Ответ в общем случае отрицательный, и далее мы приводим соответствующий контрпример. Однако затем мы показываем, что при выполнении некоторых не слишком ограничительных условий логарифмическая равновесная платежная функция уже является единственно возможной PLEPS.

3.1.2. *Другие примеры семейств* PLEPS. Сначала отметим, что существуют вариации логарифмических равновесных платежных функций, дающие эквивалентные скоринговые баллы при выполнении условия бюджетной сбалансированности. Например, для функции G, симметричной по всем аргументам, и для некоторой константы K определим функцию

$$F(z_k, z_{-k}) = \ln z_k - K \sum_{j \neq k} \ln z_j + G(z_1, \ldots, z_M)$$

(мы опускаем зависимость от состояния системы i). Тогда функция F будет удовлетворять PLEPS. Это можно проверить тем же способом, что и для задачи (3.2). Однако на самом деле данная функция не сильно

отличается от логарифмической равновесной платежной функции, если мы настаиваем на выполнении условия бюджетной сбалансированности. Действительно, легко убедиться в том, что если мы добавим к данной функции константу, которая сделает ее бюджетно-сбалансированной, мы получим те же самые равновесные платежные функции, что и в логарифмическом случае.

Приведем пример функций PLEPS, которые имеют слагаемые более высоких порядков. Наличие данных слагаемых делает эти функции отличными от логарифмических PLEPS, даже если приводить их к бюджетно-сбалансированным функциям.

Пример 3.1. Рассмотрим случай трех типов, $M = 3$, и положим

$$p^i = z^i_k, \qquad (q^i, r^i) = z^i_{-k}.$$

Определим следующую функцию:

$$F(p, q, r) = K \ln p + p^4 - 2p^3(q + r) - 6p(qr^2 + q^2 r).$$

Нетрудно проверить, что для достаточно больших K данная функция удовлетворяет строго разделяющему неравенству (2.1). Действительно, условия первого порядка для соответствующей оптимизационной задачи Лагранжа

$$\min_{q^i} \left\{ \sum_i p^i [F(p^i, q^i, r^i) - F(q^i, p^i, r^i)] + \lambda \sum_i q^i \right\}$$

имеют вид

$$p^i [\partial_p F(q^i, p^i, r^i) - \partial_q F(p^i, q^i, r^i)] = \lambda, \tag{3.3}$$

где $\partial_x F$ — частная производная функции F по переменной x, а λ — множитель Лагранжа. Эти условия первого порядка выполняются для рассматриваемой функции при $q^i = p^i$. Для достаточно больших K условия первого порядка также являются достаточными условиями оптимальности, поскольку будут выполнены условия оптимальности второго порядка. Отсюда следует справедливость (2.1).

Замечание 3.1. Хотя существуют “странные” PLEPS-функции F_i как в примере выше, для всех них, как было отмечено во введении, разность между двумя равновесными платежными функциями пропорциональна разности логарифмических платежных функций с точностью до слагаемых первого порядка включительно. Это верно и в том случае, когда F_i зависит от вероятностей типов s_k, если выполнены условия приводимой далее леммы 3.1.

В следующем пункте мы установим условия, при которых отсутствуют слагаемые второго порядка и бюджетно-сбалансированная логарифмическая платежная функция является единственной бюджетно-сбалансированной PLEPS-функцией.

3.2. Когда равновесные платежные функции будут логарифмическими? В этом пункте мы предполагаем, что[18)] $N \geqslant 3$. Как было показано выше, разность между фактическими скоринговыми баллами двух типов равна разности логарифмов скоринговых баллов с точностью до слагаемых первого порядка включительно. Найдем условия, при которых слагаемые более высоких порядков будут отсутствовать, т.е. фактически любая PLEPS-функция будет логарифмической равновесной платежной функцией.

Мы будем действовать следующим образом:

(i) сформулируем предположение на смешанную производную второго порядка от разности равновесных скоринговых баллов двух типов;

(ii) покажем, что из сформулированного предположения следует существование аддитивного представления равновесной платежной функции заданного типа в виде суммы двух слагаемых, одно из которых не зависит от апостериорных вероятностей других типов, а второе симметрично относительно типов;

(iii) покажем, что при условии достаточной гладкости из полученного представления следует, что равновесные платежные функции будут логарифмическими.

Для удобства записи мы продолжаем считать, что $M = 3$, и используем введенные выше обозначения p^i, q^i, r^i для локальных апостериорных вероятностей трех типов. Через s_p, s_q, s_r обозначаются соответствующие вероятности, образующие распределение типа. Это не приводит к потери общности, рассуждения останутся теми же и для M, больших трех.

Сформулируем необходимое нам предположение о "разделении переменных". В свете рассмотренного выше приближения первого порядка неудивительно, что оно касается свойств второго порядка равновесной платежной функции. В частности, оно является более слабым, чем предположение о том, что разность равновесных платежных функций двух типов не зависит от остальных типов.

Предположение 3.1. Для всех i и любого набора вероятностей типов s_p, s_q, s_r вторая смешанная производная

$$\partial_{pq}[F_i(p^i, q^i, r^i; s_p, s_q, s_r) - F_i(q^i, p^i, r^i; s_q, s_p, s_r)]$$

разности скоринговых баллов двух типов с локальными апостериорными вероятностями p^i и q^i соответственно (в предположении, что она существует) не зависит от локальных апостериорных вероятностей r^i других типов.

В данном предположении говорится, что на (смешанную) "чувствительность" разности равновесных платежных функций к соответствующим типам остальные типы не влияют.

[18)] Хорошо известно, что при $N = 2$ существуют квадратичные скоринговые правила, являющиеся строго разделяющими при любом априорном распределении.

Сформулируем утверждение об аддитивном представлении (доказательство см. в разделе 6).

Предложение 3.1. *Рассмотрим систему* PLEPS-*функций* $\{F_i\}$, *для которой выполняется предположение* 3.1. *Тогда, если для некоторого* $p^0 \in (0,1)$ *и для любых фиксированных вероятностей типа* s_p, s_q, s_r *функция* $F_i(p^i, q^i, r^i; s_p, s_q, s_r)$ *допускает разложение в бесконечный ряд Тейлора в окрестности точки* $(p^i, q^i, r^i) = (p^0, p^0, p^0) \in (0,1)^3$, *то справедливо следующее аддитивное представление*:

$$F_i(p^i, q^i, r^i; s_p, s_q, s_r) = G_i(p^i; s_p, s_q, s_r) + H_i(p^i, q^i, r^i; s_p, s_q, s_r), \qquad (3.4)$$

где H_i — *функция, симметричная по всем парам* (p^i, s_p), (q^i, s_q), (r^i, s_r), $i = 1, \ldots, N$.

Основной результат этого пункта состоит в следующем.

Теорема 3.1. *Рассмотрим систему* PLEPS-*функций* $F_i(p^i, q^i, r^i; s_p, s_q, s_r)$, $i = 1, \ldots, N$, *для которых выполнены условия предложения* 3.1. *Предположим, что функция* F_i *такова, что* G_i *симметрична по всем переменным* s_k *при каждой фиксированной переменной* p^i, $i = 1, \ldots, N$. *Тогда для некоторых функций* λ *и* B_i *от вероятностей типов* $S = (s_p, s_q, s_r)$ *справедливо равенство*

$$G_i(p^i; s_p, s_q, s_r) = \lambda(S) \ln p^i + B_i(S).$$

В частности, если соответствующая система PLEPS *является бюджетно-сбалансированной, то равновесная платежная функция для респондента, чьи локальные апостериорные вероятности равны* p^i, *имеет вид*

$$F_i(p^i, q^i, r^i; s_p, s_q, s_r) = \lambda(S) \ln p^i - \lambda(S) \sum_{t=p,q,r} s_t^i \ln t^i, \qquad (3.5)$$

где s_o^i — *апостериорная вероятность типа быть равным* o, *где* $o = p, q, r$, *при условии, что состояние системы есть* i.

Замечание 3.2. Снова подчеркнем, что данный результат получен с наложением ограничений только на равновесные свойства скорингового правила. Никаких ограничений на внеравновесные свойства не накладывалось.

Доказательство теоремы 3.1. Так как F_i является системой PLEPS, оно удовлетворяет разделяющему неравенству (2.1). В силу установленной симметричности H_i функция G_i удовлетворяет неравенству аналогичного типа. Отсюда следует, что

$$0 = \min_{q^i} \left\{ \sum_i p^i G_i(p^i; s_p, s_q, s_r) - \sum_i p^i G_i(q^i; s_p, s_q, s_r) \right\}. \qquad (3.6)$$

Как показано в работе [28], из данного свойства следует, что функция G_i непрерывно дифференцируема по p^i, $i = 1, \ldots, N$. Тогда по лемме 3.1

(см. ниже), в которой устанавливается условие первого порядка для данной минимизационной задачи, существует множитель Лагранжа $\lambda(S)$, не зависящий от p и такой, что (мы опускаем зависимость от i)

$$\lambda(S)\frac{1}{p^i} = \partial_p G(p^i; s_p, s_q, s_r).$$

Отсюда следует, что функция G_i имеет логарифмический вид. Равенство (3.5) устанавливается прямой подстановкой. Теорема 3.1 доказана.

Следующая лемма об "оптимизации по Лагранжу" доказана в разделе 6. Она дает условие первого порядка для IC-проблемы минимизации в (3.7) ниже.

Лемма 3.1. *Рассмотрим, в обозначениях выше, функции* $F_i(p^i, q^i, r^i; s_p, s_q, s_r)$, $i = 1, \ldots, N$, *которые непрерывно дифференцируемы по переменным* p^i, q^i *и, для любых фиксированных* p^i, q^i, r^i *симметричны по всем переменным* s. *Рассмотренное ранее строго разделяющее неравенство* (2.1) *можно записать в виде*

$$0 = \min_{q^i}\left\{\sum_i p^i F_i(p^i, q^i, r^i; s_p, s_q, s_r) - \sum_i p^i F_i(q^i, p^i, r^i; s_p, s_q, s_r)\right\}, \quad (3.7)$$

т.е. минимум по вероятностям q^i *достигается при* $q^i = p^i$. *Тогда существует функция* $\lambda(S) = \lambda(s_p, s_q, s_r)$ *такая, что для любых* i, p^i, q^i, r^i, s_p, s_q, s_r *имеет место равенство*

$$\lambda(S) = p^i[\partial_p F_i(p^i, p^i, r^i; s_p, s_q, s_r) - \partial_q F_i(p^i, p^i, r^i; s_p, s_q, s_r)]. \quad (3.8)$$

3.3. Ранжирование с использованием (локальных) апостериорных вероятностей. В этом пункте мы покажем, что платежные функции PLEPS упорядочивают игроков в соответствии с относительным рангом соответствующих локальных апостериорных вероятностей. Таким образом, когда мы используем скоринговый метод, приводящий к равновесию с платежными функциями PLEPS, составитель опроса будет знать, какие игроки являются лучшими экспертами, полагая уровень локальных апостериорных вероятностей равным уровню экспертизы[19)]. Отметим, что одним из ключевых предположений здесь является то, что равновесные скоринговые баллы зависят только от локальных апостериорных вероятностей фактического состояния системы.

Основной результат этого пункта состоит в следующем.

Теорема 3.2. *Пусть* $\{F_i\}$ — *система платежных функций* PLEPS. *Тогда* F_i *являются строго возрастающими как функции апостериорных вероятностей фактического состояния системы, т.е.* (*при любой*

[19)] Если игроки не были упорядочены с помощью локальных апостериорных вероятностей, то перед началом игры участники могут пытаться избежать сбора информации о фактическом состоянии системы, что является нежелательным.

матрице априорного распределения Q)

$$\begin{aligned} &\textit{если } j,k \in \{1,\dots,M\} \ u\ z_k^i > z_j^i, \\ &\qquad \textit{то } F_i(z_k^i, z_{-k}^i; s_k, s_{-k}) > F_i(z_j^i, z_{-j}^i; s_j, s_{-j}). \end{aligned} \tag{3.9}$$

Другими словами, пусть составитель опроса хочет определить относительный уровень экспертизы игроков, получающих взаимозаменяемые сигналы. Тогда достаточно создать такую скоринговую систему, которая допускает только равновесия, реализованные посредством функций PLEPS. Таким образом, неравенство (2.1) не только гарантирует строгую разделенность типов, но и, как следствие, дает метод ранжирования на основе апостериорных вероятностей.

Теорема 3.2 является обобщением известных результатов о том, что из стимульной согласованности собственных скоринговых правил следует монотонность (см., например, [19], [28]–[30]). В данных работах рассматривалась только неигровая версия задачи, и только с одним респондентом. Более того, при доказательстве использовались в основном аналитические методы, тогда как в настоящей работе доказательство полностью алгебраическое[20)].

Интуитивно данный результат означает, что если апостериорная вероятность состояния системы для типа A была выше, чем для B, но скоринговый балл A в данном состоянии ниже, то для A выгоднее выдать себя за тип B. Более строго, рассмотрим случай двух типов (A и B) и двух состояний системы (1 и 2). Обозначим через p_A и p_B апостериорные вероятности состояния 1 и предположим без ограничения общности, что $p_A > p_B$. В каждом состоянии i существует только два возможных скоринговых балла PLEPS, обозначим их через $F_i(p_A, p_B)$ и $F_i(p_B, p_A)$ (мы опускаем зависимость от вектора S). Пусть D_i — разность скоринговых баллов: $D_i^A = F_i(p_A, p_B) - F_i(p_B, p_A)$. Тогда в случае равновесия если апостериорная вероятность типа A в состоянии i является более высокой, то в данном состоянии скоринговый балл A будет также более высоким, а разность D_i^A будет положительной. Чтобы это показать, сначала заметим, что в силу строго разделяющего неравенства ожидаемая разность скоринговых баллов игрока A, равная взвешенному среднему D_1^A и D_2^A с весами p_A и $1 - p_A$, является положительной. Те же самые рассуждения приводят к тому, что взвешенное среднее D_1^A и D_2^A с весами p_B и $1 - p_B$ является отрицательным. В случае, когда $p_A > p_B$

[20)] Теорема 3.2 сформулирована в духе теорем, которые связывают стимульную согласованность с монотонностью относительно типов, если мы приравняем типы и апостериорные вероятности (см. [22], где приведены первые теоремы данного типа, и [31], где проведено детальное исследование вопроса). Наша постановка задачи отличается от стандартной, так как мы рассматриваем случайные состояния системы. При этом стимульная согласованность является свойством взвешенной суммы (ожидаемым значением при условии типа), а не самого значения. Как следствие, методы упомянутых работ в нашем случае не применимы.

(и, значит, $1 - p_A < 1 - p_B$), это возможно, только если $D_1^A > 0$ и $D_2^A < 0$. Таким образом, тип с более высокой апостериорной вероятностью в некотором состоянии получает более высокий скоринговый балл в этом состоянии. Иными словами, если тип с более высокой апостериорной вероятностью состояния не получает более высокий скоринговый балл в данном состоянии, он станет придерживаться стратегии другого типа. В приложении несложные соображения, приведенные выше, формулируются в виде леммы 6.1, которая затем применяется для случая произвольного числа типов и состояний системы.

4. Применение полученных результатов. В данном разделе мы сначала рассмотрим вопрос реализации любой системы функций PLEPS, с точностью до некоторых слабых ограничений. Затем мы дополним результат работы [24] о том, что алгоритм BTS дает возможность реализовать бюджетно-сбалансированные логарифмические платежные функции в (3.5) (при выполнении стандартных байесовских предположений и предположения о рациональности); см. также [25]. Кроме того, мы коснемся вопросов единственности равновесия в скоринговом правиле BTS.

4.1. Практическая реализация PLEPS. Под реализацией равновесной системы платежных функций F будем понимать составление опроса и скоринговых правил так, чтобы ассоциированная игра допускала равновесие с платежными функциями системы F.

Будем предполагать, как и выше, что имеется бесконечно много респондентов. Рассмотрим случай, когда респондентам предлагается ответить на вопрос, выбрав правильный вариант из нескольких предложенных (а именно, декларировать свой тип). Предположим, что состояния системы принимают значения в множестве вероятностных распределений ответов на вопросы.

Следующий результат показывает, что любая PLEPS-система $\{F_i\}$, зависящая только от локальных условных вероятностей $s_k^i = \Pr(T^\pi = k \mid \Omega = i)$, может быть реализована таким образом, что F_i можно использовать для вычисления значения скоринговой функции f в состоянии i.

Предложение 4.1. *Рассмотрим систему* PLEPS *с платежными функциями* $F_i(z_k^i, z_{-k}^i; s_k^i, s_{-k}^i)$ *типа* k *в равновесном состоянии* i; *т.е. платежная функция зависит, в дополнение к зависимости от* z_k^j, *только от локальных условных вероятностей* $s_k^i = \Pr(T^\pi = k \mid \Omega = i)$ *вместо возможной зависимости от всех априорных вероятностей, содержащихся в векторе* s_k. *Тогда данная система* PLEPS *может быть реализована нейтральным составителем опроса. А именно, существует набор вопросов, получив ответы на которые составитель опроса может построить оценки* $\widehat{i}$ *и* $\widehat{z_k^i}$, $\widehat{s_k^i}$ *фактического состояния системы* i *и вероятностей* z_k^i, s_k^i, *обладающие следующим свойством: если составитель опроса объявляет, что игрок, декларирующий тип* k, *по-*

лучает выплату $F_{\widehat{i}}(\widehat{z}_k^{\widehat{i}}, \widehat{z}_{-k}^{\widehat{i}}; \widehat{s}_k^{\widehat{i}}, \widehat{s}_{-k}^{\widehat{i}})$, *то ситуация, при которой игроки говорят правду, образует равновесие.*

Доказательство. Предположим, что составитель просит респондентов

(a) ответить на вопрос, выбрав правильный ответ из нескольких вариантов;

(b) определить возможные состояния системы, т.е. указать множество возможных распределений ответов на вопрос из (a) и определить ожидаемую вероятность (z) для каждого из этих распределений.

Чтобы гарантировать, что случай правдивых ответов образует равновесие, составитель опроса будет рассчитывать скоринговый балл F_i следующим образом. Оценка $\widehat{i}$ фактического состояния системы i будет задаваться частотами, с которыми участники будут выбирать каждый из возможных вариантов ответа. Полученные частоты также будут оценками $\widehat{s}_k^{\widehat{i}}$ вероятностей типов. Оценки $\widehat{z}_k^{\widehat{i}}$ будут выбраны среди всех вероятностей z_k^j, $j = 1, \ldots, N$, которые респондент приводит в (b). Получив все необходимые оценки, составитель опроса рассчитает соответствующие значения функций $F_{\widehat{i}}$, подставив $\widehat{z}_k^{\widehat{i}}$ и $\widehat{s}_k^{\widehat{i}}$ в качестве аргументов.

Теперь предположим, что все игроки, за исключением игрока π типа k, дают правдивые ответы. Если игрок π тоже отвечает правдиво, его платежная функция в состоянии i есть $F_i(z_k^i, z_{-k}^i; s_k^i, s_{-k}^i)$, поскольку величины i, z и s корректно оценены составителем опроса. Если игрок π типа k декларирует тип $j \neq k$, его платежная функция в состоянии i есть $F_i(z_j^i, z_{-j}^i; s_j^i, s_{-j}^i)$. Последнее справедливо в силу того, что количество игроков бесконечно и все игроки, за исключением π, говорят правду, откуда следует, что i, z и s снова корректно оценены составителем опроса. Из IC-неравенства (2.1) следует, что ожидаемое значение платежной функции игрока π в случае, когда он нечестен, меньше ожидаемого значения его платежной функции в случае, когда он говорит правду. Предложение 4.1 доказано.

Замечание 4.1. Полученный метод реализации не является устойчивым — на практике количество различных исходов при ответе на вопрос (b) будет больше количества типов, и различные респонденты будут рассматривать различные распределения ответов на вопрос (a) как возможные состояния системы. Таким образом, нужно будет проводить определенную группировку ответов. Более того, необходимость ответа на вопрос (b) является достаточно затратной операцией, так как респонденты должны предоставить возможные частоты ответов на (a) и распределения по этим частотам. Для бюджетно-сбалансированных логарифмических равновесных платежных функций ситуация, как показывается в следующем пункте, иная: скоринговое правило BTS, введенное в работе [24], реализует такие функции. При этом используются входные данные, которые являются более простыми, чем ответы на (b), а также

используется устойчивый метод (что избавляет от необходимости деления ответов на группы).

4.2. Реализация логарифмических равновесных платежных функций с помощью BTS. Вначале напомним определение байесовской сыворотки правды (Bayesian Truth Serum, BTS). Уточним модель, используя обозначения из раздела 2. Предположим, что имеется бесконечное (счетное) количество респондентов, проиндексированное параметром $\pi \in R$. Правдивый ответ респондента π будем представлять парой M-кортежей $(X^\pi; Y^\pi) = ((X_1^\pi, \ldots, X_M^\pi); (Y_1^\pi, \ldots, Y_M^\pi))$ случайных величин. Здесь величины X_i^π принимают значения 0 или 1, и только одна из величин в кортеже равна единице. Это можно интерпретировать как выбор одного ответа из M возможных. Случайные величины Y_i^π принимают значения в $[0,1]$ и $\sum_{i=1}^M Y_i^\pi = 1$. Величина Y_i^π представляет декларированное респондентом мнение π о доле респондентов, которые выберут i в качестве правильного ответа.

Как и в разделе 2, будем предполагать, что бесконечная последовательность $(X^\pi, \pi \in R)$ является взаимозаменяемой. Тогда по теореме де Финетти существует M-мерный (потенциально случайный) вектор

$$\overline{X} = \lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} X^\pi,$$

принимающий значения в $[0,1]^M$ и такой, что величины X^π являются независимыми при условии $\overline{X}$. Под $\overline{X}$ будем понимать состояние системы, которое ранее обозначалось через Ω.

Пусть $\overline{x}_j$ — выборочное среднее декларированных значений x_j^π величин X_j^π по всем респондентам π, а $\ln \overline{y}_j$ — выборочное среднее декларированных значений $\ln y_j^\pi$ величин $\ln Y_j^\pi$ (тогда $\overline{y}_j$ будет геометрическим средним):

$$\ln \overline{y}_j := \lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} \ln y_j^\pi.$$

Определение 4.1. Скоринговая функция BTS для респондента π имеет вид

$$\mathrm{BTS}^\pi = \sum_{j=1}^{M} x_j^\pi \ln \frac{\overline{x}_j}{\overline{y}_j} + \sum_{j=1}^{M} \overline{x}_j \ln \frac{y_j^\pi}{\overline{x}_j}.$$

В работе [24] доказано, что BTS является стимульно-согласованным методом в том смысле, что платежная функция респондента достигает максимума, когда он и все остальные респонденты говорят правду. Более того, мы можем сформулировать новый результат о "единственности": в рамках метода BTS любое бюджетно-сбалансированное строгое (байесовское) равновесие Нэша является разделяющим.

Замечание 4.2. Естественно считать, что $\ln(\overline{x}_j/\overline{y}_j) = 0$ при $\overline{x}_j = \overline{y}_j = 0$, а также что $\overline{x}_j \ln(y_j^\pi/\overline{x}_j) = 0$ при $\overline{x}_j = 0$. Заметим, что если $x_j^\pi = 0$ для всех, за исключением конечного числа, индексов π (в частности, $\overline{x}_j = 0$), то для каждого игрока π будет оптимальным предсказать $y_j^\pi = 0$. Поэтому в таком случае естественно положить $\overline{y}_j$ равным нулю[21)]. При этих соглашениях единственно возможными бюджетно-сбалансированными SNE являются разделяющие SNE, в которых игроки одного типа придерживаются одинаковых стратегий. Обоснуем это.

(i) Невозможно иметь SNE при чистых стратегиях, в которых два респондента одного типа выбирают различные стратегии и, как следствие, имеют различные ожидаемые скоринговые баллы. Действительно, предположим, что респонденты имеют различные стратегии в данном SNE. Если бы игрок 1 переключился на стратегию 2, то он по определению имел бы строго меньшее значение скорингового балла. И данное значение было бы таким же, как у игрока 2, поскольку в случае бесконечного числа игроков скоринговый балл одного игрока не зависит от действий другого игрока. Аналогично, если игрок 2 переключится на стратегию 1, то его скоринговый балл будет таким же, как и скоринговый балл 1, который, как мы показали выше, имеет строго большее значение. Отсюда следует, что игрок 2 не придерживается равновесной стратегии, с которой он начинал. Противоречие.

(ii) Два респондента различных типов не могут иметь одинаковых стратегий в SNE: если бы их стратегии были одинаковыми, то по (i) все другие игроки этих типов тоже выбрали бы данную стратегию. Это бы означало, что существует тип k, который никто не выберет, т.е. такой, что $x_k^\pi = 0$ для всех π. Так как мы предполагаем наличие бюджетной сбалансированности, то отсюда следовало бы существование игрока с неположительным скоринговым баллом. Если бы этот игрок "уклонился" в тип k, то в соответствии с принятыми выше естественными соглашениями его скоринговый балл BTS был бы равен нулю, что в слабом смысле лучше, чем не "уклоняться" в тип k. Следовательно, равновесие не было бы строгим.

В силу сделанного замечания, а также потому, что правдивое равновесие является самым важным среди строго разделяющих равновесий, всюду далее будем считать x_i и y_i правдивыми ответами.

Для удобства читателя напомним результат работы [24] о том, что при правдивом равновесии скоринговый балл BTS совпадает с бюджетно-сбалансированной логарифмической платежной функцией. Подробное доказательство см. в разделе 6.

Теорема 4.1 [24, теорема 2]**.** *В предположениях выше, если разыгрывается правдивое равновесие, то в результате скоринга методом* BTS

[21)] Это верно, поскольку при возрастании y_j скоринговый балл не меняется, тогда как при убывании y_k для $k \neq j$ скоринговый балл уменьшается, если $\overline{x}_k > 0$.

получатся бюджетно-сбалансированные логарифмические платежные функции. Точнее, в случае равновесия имеем

$$\mathrm{BTS}^{\pi} = \ln \Pr(\overline{X} = \overline{x} \mid X^{\pi} = x^{\pi}) - \lim_{n\to\infty} \frac{1}{n} \sum_{s=1}^{n} \ln \Pr(\overline{X} = \overline{x} \mid X^{\gamma} = x^{\gamma}) \tag{4.1}$$

или (*если обозначить* $x^{\pi} = k$, $x^{\gamma} = j$, $\overline{x} = i$)

$$\mathrm{BTS}^{\pi} = \ln \Pr(\Omega = i \mid T^{\pi} = k) - \sum_{j=1}^{M} \Pr(T^{\pi} = j \mid \Omega = i) \ln \Pr(\Omega = i \mid T^{\pi} = j). \tag{4.2}$$

Таким образом, скоринговый балл BTS соответствует логарифмическим PLEPS-функциям F_i. Иными словами, BTS реализует бюджетно-сбалансированные логарифмические платежные функции с помощью вопросов только двух типов: выбрать ответ из нескольких возможных вариантов и оценить процент игроков, которые выберут определенный ответ.

В заключение отметим, что основной целью данного раздела было показать привлекательность метода BTS благодаря следующим трем свойствам: BTS всегда приводит к строго разделяющему равновесию, в котором игроки одного типа имеют одинаковый скоринговый балл; результатом BTS является логарифмический скоринговый балл; метод BTS легко реализуется. Ни один другой метод, насколько нам известно, не обладает всеми тремя свойствами.

5. Заключение. В настоящей работе мы рассматриваем задачи выявления правдивых ответов для большой группы респондентов и упорядочивания респондентов в соответствии с их апостериорными вероятностями фактического состояния системы (локальными апостериорными вероятностями). При этом составитель опроса нейтрален по отношению к распределению состояний системы и к типам респондентов. Таким образом, составителю требуется разработать универсальный метод, который будет работать при любом распределении состояний системы. Один из таких методов основан на логарифмических платежных функциях. Мы доказываем следующие результаты для равновесных платежных функций, которые определяются только локальными апостериорными вероятностями и вероятностями типов: (i) в предположениях о "чувствительности" разности скоринговых баллов результатом стимульно-согласованного бюджетно-сбалансированного равновесия является логарифмическая платежная функция; (ii) для произвольного метода любое стимульно-согласованное равновесие упорядочивает респондентов в соответствии со значениями их апостериорных вероятностей фактического состояния системы. Мы также подробно рассматриваем результат работы [24] о том, что равновесные логарифмические платежные функции

могут быть реализованы с помощью алгоритма BTS. При этом другие равновесные платежные функции также могут быть реализованы, но для этого, возможно, потребуются ответы на более сложные вопросы.

Наша постановка не позволяет игрокам производить никакие действия, кроме как отвечать на вопросы исследования. Таким образом, дальнейшим развитием нашего подхода может быть более общий анализ устойчивых методов, в которых респонденты могут оценивать полезность тех или иных своих действий. В нашей модели у экспертов нет причин лгать, но есть стимул говорить правду. Можно представить себе постановку, в которой респондентам имеет смысл лгать, например когда им не важна их собственная выплата, но они хотят изменить результаты так, чтобы некоторый другой тип респондентов получил наивысший скоринговый балл. Или же постановку, в которой известны полезности, но неизвестны корреляции типов, а задача составителя опроса состоит в выявлении информации о корреляциях без того, чтобы это как-либо повлияло на полезности; например, когда составитель хочет, чтобы респонденты предсказали поведение других респондентов, и при этом не хочет, чтобы выплаты, получаемые за такие предсказания, как-либо изменили другие стимулы в игре. Наконец, наша игра является статической, тогда как многие приложения по своей природе динамические.

6. Приложение.

Доказательство предложения 2.1. Мы хотим доказать, что если игрок типа k использует равновесную стратегию типа j, то его платежная функция задается равновесием для типа j.

Прежде всего напомним, что чистая стратегия для игрока π — это отображение $\sigma^\pi(t^\pi)$, которое типу игрока ставит в соответствие его ответ a^π. Множество всех чистых стратегий респондентов обозначается через $\sigma(t)$, его элементы — $\sigma^\pi(t^\pi)$, а множество чистых стратегий всех респондентов, кроме π, обозначается через $\sigma^{-\pi}(t^{-\pi})$. Скоринговый балл игрока π равен $f(\sigma^\pi(t^\pi), \sigma^{-\pi}(t^{-\pi}))$. Обозначим через σ множество равновесных стратегий всех респондентов и определим ρ как множество стратегий, совпадающих со стратегиями из σ за тем исключением, что некоторый выделенный игрок π типа $k \neq j$ использует стратегию $\sigma^\pi(j)$, соответствующую типу j. Пусть γ обозначает игрока типа j. Тогда платежная функция имитационной стратегии, когда игрок π использует стратегию типа j, равна

$$\begin{aligned} f(\rho^\pi(k), \rho^{-\pi}(T^{-\pi})) &= f(\rho^\gamma(j), \rho^{-s}(T^{-s})) = f(\sigma^\gamma(j), \rho^{-s}(T^{-s})) \\ &= f(\sigma^\gamma(j), \sigma^{-s}(T^{-s})) \end{aligned}$$

в силу того, что $\sigma^{-s}(T^{-s})$ и $\rho^{-s}(T^{-s})$ различаются только ответом игрока π, а он не оказывает никакого влияния в случае бесконечного числа игроков. Последнее верно, так как каждый тип будет представлен бесконечным множеством игроков, а функция f симметрична относительно

их ответов. Докажем равенство выше более строгими рассуждениями. Нужно доказать, что если ровно один из $-s$ респондентов дает ответ, отличающийся от равновесного, то скоринговый балл f респондента γ не меняется. Так как σ соответствует строго разделяющему равновесию, то последовательность $\{\sigma^{-s}(T^{-s})\}$ состоит из M различных K-кортежей, каждый из которых повторяется бесконечно много раз. Если респондент типа k "уклоняется" в тип j, это означает, что одно повторение (среди бесконечного числа) K-кортежей, соответствующих типу k, становится дополнительным повторением (среди бесконечного числа) K-кортежей, соответствующих типу j. Можно определить перестановку последовательности $\{\sigma^{-s}(T^{-s})\}$, которая совпадает с "уклонившейся" последовательностью $\{\rho^{-s}(T^{-s})\}$, и из свойства симметричности функции f будет следовать справедливость доказываемого равенства. Предложение 2.1 доказано.

Доказательство предложения 2.2. Требуется доказать, что в случае конечного числа игроков не существует бюджетно-сбалансированной системы функций PLEPS.

Для простоты обозначений будем рассматривать случай $M = 2$ двух типов 1 и 2, двух или более игроков и $N = 3$ состояний системы. Состояние 1 системы имеет вид $(2, 0)$ (два игрока типа 1, ноль игроков типа 2), состояние 2 имеет вид $(1, 1)$, а состояние 3 имеет вид $(0, 2)$. Доказательство легко обобщается на случай произвольного $M > 2$, так как мы можем рассматривать только такие матрицы Q, в которых два типа находятся в отдельном блоке.

Рассмотрим матрицу Z вида

$$\begin{pmatrix} p & 1-p & 0 \\ 0 & q & 1-q \end{pmatrix},$$

где $0 < p, q < 1$ (заметим, что строки соответствуют типам, а столбцы — состояниям системы).

В случае бесконечного множества игроков любая PLEPS-система функций F_i зависит от апостериорных вероятностей, которые в свою очередь зависят от состояния i системы, соответствующего декларируемым типам. Например, если фактическое состояние системы есть $(2, 0)$, но один респондент декларирует тип 2, то платежные функции соответствуют состоянию $(1, 1)$.

Ожидаемый скоринговый балл в случае правдивого ответа для типа 1 равен

$$pF_1(p, p) + (1 - p)F_2(1 - p, q).$$

Если один респондент говорит неправду и декларирует вместо типа 1 тип 2, то его ожидаемый скоринговый балл будет равен

$$pF_2(q, 1 - p) + (1 - p)F_3(1 - q, 1 - q).$$

Следовательно, разделяющее неравенство будет иметь вид

$$pF_1(p,p) + (1-p)F_2(1-p,q) > pF_2(q,1-p) + (1-p)F_3(1-q,1-q)$$

или

$$p[F_1(p,p) - F_2(q,1-p)] + (1-p)[F_2(1-p,q) - F_3(1-q,1-q)] > 0.$$

Аналогично, когда один респондент типа 2 говорит неправду, разделяющее неравенство будет иметь вид

$$qF_2(q,1-p) + (1-q)F_3(1-q,1-q) > qF_1(p,p) + (1-q)F_2(1-p,q)$$

или

$$q[F_2(q,1-p) - F_1(p,p)] + (1-q)[F_3(1-q,1-q) - F_2(1-p,q)] > 0.$$

Предположим, что $p \neq q$. Рассмотрим без ограничения общности случай $p > q$ и применим лемму 6.1 (см. далее) к двум указанным выше разделяющим неравенствам. Мы получаем

$$F_1(p,p) - F_2(q,1-p) > 0, \qquad F_2(1-p,q) - F_3(1-q,1-q) < 0.$$

В предположении бюджетной сбалансированности должны быть выполнены равенства $F_1(p,p) = 0 = F_3(1-q,1-q)$, откуда следует, что

$$(1-p)F_2(1-p,q) + qF_2(q,1-p) = 0.$$

Заметим, что из $F_1(p,p) = 0$ следует $F_2(q,1-p) < 0$, тогда как из $F_3(1-q,1-q) = 0$ следует $F_2(1-p,q) < 0$. Получаем противоречие с последним равенством. Предложение 2.2 доказано.

Доказательство леммы 3.1. Докажем (3.8). Согласно классическому результату об оптимизации при наличии ограничений (в нашем случае ограничением служит условие $\sum_i q^i = 1$), существует множитель Лагранжа $\lambda(\vec{p},\vec{r},s_p,s_q,s_r)$, где $\vec{p} = (p^1,\dots,p^N)$, такой, что

$$p^i[\partial_p F_i(p^i,p^i,r^i;s_p,s_q,s_r) - \partial_q F_i(p^i,p^i,r^i;s_p,s_q,s_r)] = \lambda(\vec{p},\vec{r},s_p,s_q,s_r). \tag{6.1}$$

Рассмотрим фиксированные значения i и p^i, r^i. Так как $N > 2$, мы можем положить $p^j = x$, $r^j = y$ для фиксированного, но произвольного $j \neq i$ и для любых $0 < x < 1-p^i$, $0 < y < 1-r^i$. Из соотношения выше следует, что $\lambda(\vec{p},\vec{r},S)$ фактически является функцией $\lambda(p^i,r^i,S)$, зависящей только от переменных p^i, r^i, S, и мы имеем

$$x[\partial_p F_j(x,x,y;S) - \partial_q F_j(x,x,y;S)] = \lambda(p^i,r^i;S),$$

для всех $0 < x < 1-p^i$, $0 < y < 1-r^i$. Так как мы можем выбрать p^i, r^i произвольно малыми, то при фиксированном S левая часть является постоянной для всех значений x, y в $(0,1)$. А поскольку i выбрано произвольно, получаем, что $\lambda(S)$ не зависит от значений p^i, r^i, $i = 1,\dots,N$. Лемма 3.1 доказана.

Доказательство предложения 3.1. В этом доказательстве мы опускаем зависимость от i и от s_p, s_q, s_r. Мы хотим показать, что

$$F(p,q,r) = G(p) + H(p,q,r),$$

где функция H симметрична по всем парам (p, s_p), (q, s_q), (r^j, s_{r^j}).

Для заданного $p^0 \in (0,1)$ обозначим

$$\overline{p} = p - p^0, \quad \overline{q} = q - p^0, \quad \overline{r} = r - p^0.$$

В силу гладкости и симметричности F мы можем воспользоваться разложением Тейлора и для некоторых функций a, b, c, d, e вероятностей типов получить

$$F(p,q,r) = \sum_{n=0}^{\infty} a_n \overline{p}^n + \sum_{n=1}^{\infty} (b_n^q \overline{q}^n + b_n^\pi \overline{r}^n) + \sum_{m,n=1}^{\infty} \overline{p}^m (c_{m,n}^q \overline{q}^n + c_{m,n}^\pi \overline{r}^n) + \sum_{m,n=1}^{\infty} d_{m,n} \overline{q}^m \overline{r}^n + \sum_{l,m,n=1}^{\infty} e_{l,m,n} \overline{p}^l \overline{q}^m \overline{r}^n,$$

где, ввиду симметричности,

$$b_n^q(s_p, s_q, s_r) = b_n^\pi(s_p, s_r, s_q), \qquad c_{m,n}^q(s_p, s_q, s_r) = c_{m,n}^\pi(s_p, s_r, s_q),$$
$$d_{m,n}(s_p, s_q, s_r) = d_{n,m}(s_p, s_r, s_q), \qquad e_{l,m,n}(s_p, s_q, s_r) = e_{l,n,m}(s_p, s_r, s_q).$$

Заметим, что достаточно показать, что

$$c_{m,n}^\pi = d_{m,n}, \qquad e_{l,m,n} = e_{m,l,n},$$

поскольку тогда мы можем написать

$$F(p,q,r) = \sum_{n=0}^{\infty} [a_n - b_n^q] \overline{p}^n + H(p,q,r),$$

где H — функция, симметричная по всем парам (p^i, s_p), (q^i, s_q), (r_j^i, s_{r_j}).

Рассмотрим следствия строго разделяющего неравенства (3.7), используя лемму 3.1. Имеем

$$\partial_q F(p,p,r) - \partial_p F(p,p,r) = \sum_{n=1}^{\infty} n b_n^q \overline{p}^{n-1} + \sum_{m,n=1}^{\infty} c_{m,n}^q n \overline{p}^{m+n-1} + \sum_{m,n=1}^{\infty} d_{m,n} m \overline{p}^{m-1} \overline{r}^n + \sum_{l,m,n=1}^{\infty} e_{l,m,n} m \overline{p}^{l+m-1} \overline{r}^n - \sum_{n=0}^{\infty} n a_n \overline{p}^{n-1} - \sum_{m,n=1}^{\infty} m \overline{p}^{m-1} (c_{m,n}^q \overline{p}^n + c_{m,n}^\pi \overline{r}^n) - \sum_{l,m,n=1}^{\infty} e_{l,m,n} l \overline{p}^{l+m-1} \overline{r}^n.$$

Далее,

$$
\begin{aligned}
p\,\partial_q F(p,p,r) - p\,\partial_p F(p,p,r) &= \sum_{n=1}^{\infty} n b_n^q \overline{p}^n + \sum_{m,n=1}^{\infty} c_{m,n}^q n \overline{p}^{m+n} \\
&+ \sum_{m,n=1}^{\infty} d_{m,n} m \overline{p}^m \overline{r}^n + \sum_{l,m,n=1}^{\infty} e_{l,m,n} m \overline{p}^{l+m} \overline{r}^n - \sum_{n=0}^{\infty} n a_n \overline{p}^n \\
&- \sum_{m,n=1}^{\infty} m \overline{p}^m (c_{m,n}^q \overline{p}^n + c_{m,n}^{\pi} \overline{r}^n) - \sum_{l,m,n=1}^{\infty} e_{l,m,n} l \overline{p}^{l+m} \overline{r}^n + \sum_{n=1}^{\infty} n b_n^q p^0 \overline{p}^{n-1} \\
&+ \sum_{m,n=1}^{\infty} n c_{m,n}^q p^0 \overline{p}^{m+n-1} + \sum_{m,n=1}^{\infty} d_{m,n} m p^0 \overline{p}^{m-1} \overline{r}^n \\
&+ \sum_{l,m,n=1}^{\infty} e_{l,m,n} m p^0 \overline{p}^{l+m-1} \overline{r}^n - \sum_{n=0}^{\infty} n a_n p^0 \overline{p}^{n-1} \\
&- \sum_{m,n=1}^{\infty} m p^0 \overline{p}^{m-1} (c_{m,n}^q \overline{p}^n + c_{m,n}^{\pi} \overline{r}^n) - \sum_{l,m,n=1}^{\infty} e_{l,m,n} l p^0 \overline{p}^{l+m-1} \overline{r}^n.
\end{aligned}
$$

По лемме 3.1, чтобы мы имели PLEPS, полученное выше выражение должно быть равно $(-\lambda)$ для всех p, r. Это возможно, только если выполнены следующие равенства:

– для коэффициентов при $\overline{r}^n$:

$$c_{1,n}^{\pi} = d_{1,n}; \tag{6.2}$$

– для коэффициентов при $\overline{p}\overline{r}^n$:

$$0 = c_{1,n}^{\pi} - d_{1,n} + c_{2,n}^{\pi} - d_{2,n}; \tag{6.3}$$

– для коэффициентов при $\overline{p}^2\overline{r}^n$:

$$0 = 2(d_{2,n} - c_{2,n}^{\pi}) + 3p^0(d_{3,n} - c_{3,n}^{\pi}) + p^0(e_{1,2,n} - e_{2,1,n}); \tag{6.4}$$

– для коэффициентов при $\overline{p}^3\overline{r}^n$:

$$0 = 3(d_{3,n} - c_{3,n}^{\pi}) + (e_{1,2,n} - e_{2,1,n}) + 4p^0(d_{4,n} - c_{4,n}^{\pi}) + 2p^0(e_{1,3,n} - e_{3,1,n}) \tag{6.5}$$

и т.д.

Таким образом, осталось показать, что $e_{l,m,n} = e_{m,l,n}$. А это напрямую следует из предположения 3.1, поскольку третья смешанная производная разности $F(p,q,r) - F(q,p,r)$ скоринговых баллов будет равна нулю для всех p, q, r, т.е.

$$0 = \sum_{l,m,n=1}^{\infty} lmn\; e_{l,m,n} \overline{p}^{l-1} \overline{q}^{m-1} \overline{r}^{n-1} - \sum_{l,m,n=1}^{\infty} lmn\; e_{l,m,n} \overline{q}^{l-1} \overline{p}^{m-1} \overline{r}^{n-1}.$$

Полученное соотношение завершает доказательство предложения 3.1.

Следующая лемма является ключевой при доказательстве теоремы 3.2. Она представляет собой небольшое обобщение леммы A.1 в работе [29].

Лемма 6.1 [29]. *Пусть* $0 < a \leqslant 1$, $p, q \in (0, a)$ *и* $p > q$. *Если* A, B — *вещественные числа такие, что*

$$pA + (a-p)B > 0, \qquad q(-A) + (a-q)(-B) > 0,$$

то $A > 0$ *и* $B < 0$.

Доказательство. Заметим, что $A \neq 0$. Если это не так, то два неравенства выше имеют вид $(a-p)B > 0$ и $(a-q)(-B) > 0$ — противоречие. Чтобы доказать утверждение леммы, достаточно показать, что $A > 0$. Предположим противное, т.е. что $A < 0$. Тогда $B > 0$. Из неравенства $(a-p)B > -pA$ вытекает, что $B > -pA/(a-p) > 0$. Следовательно, $0 < q(-A)+(a-q)(-B) < q(-A)+(a-q)pA/(a-p) = Aa(p-q)/(a-p) < 0$, что невозможно. Полученное противоречие показывает, что $A > 0$. Лемма 6.1 доказана.

Доказательство теоремы 3.2. Требуется доказать, что платежные функции PLEPS являются строго возрастающими функциями апостериорных вероятностей фактического состояния системы.

Мы опускаем зависимость от s_k в наших обозначениях, так как фиксирование величин s_k не накладывает ограничений на выбор двух строк матрицы Z — мы всегда можем определить матрицу Q посредством равенств $q_k^i = z_k^i s_k$.

Рассмотрим три случая в зависимости от значений M и N.

Случай 1: $M = 2, N = 2$. Матрицу Z можно записать в виде $\begin{bmatrix} z_1^1 & z_1^2 \\ z_2^1 & z_2^2 \end{bmatrix}$. Если обозначить $p := z_1^1$, $q := z_2^1$, то $Z = \begin{bmatrix} p & 1-p \\ q & 1-q \end{bmatrix}$. Пусть $p > q$ (или, эквивалентно, $1-q > 1-p$). Из IC-свойства (2.1) функций F_i следует, что

$$pF_1(p,q) + (1-p)F_2(1-p,1-q) > pF_1(q,p) + (1-p)F_2(1-q,1-p),$$
$$qF_1(q,p) + (1-q)F_2(1-q,1-p) > qF_1(p,q) + (1-q)F_2(1-p,1-q).$$

Отсюда следует, что

$$p[F_1(p,q) - F_1(q,p)] + (1-p)[F_2(1-p,1-q) - F_2(1-q,1-p)] > 0,$$
$$q[F_1(q,p) - F_1(p,q)] + (1-q)[F_2(1-q,1-p) - F_2(1-p,1-q)] > 0.$$

Возьмем $a = 1$, $A = F_1(p,q) - F_1(q,p)$, $B = F_2(1-p,1-q) - F_2(1-q,1-p)$ и применим лемму 6.1 к неравенствам выше. Получим, что $F_1(p,q) > F_1(q,p)$ и $F_2(1-p,1-q) > F_2(1-q,1-p)$, что доказывает утверждение теоремы.

Случай 2: $M \geqslant 3$, $N = 2$. Матрицу Z можно записать в виде

$$\begin{bmatrix} z_1^1 & z_1^2 \\ z_2^1 & z_2^2 \\ \vdots & \vdots \\ z_M^1 & z_M^2 \end{bmatrix}.$$

Элементы матрицы удовлетворяют соотношению $z_k^2 = 1 - z_k^1$, $k = 1, \ldots, M$. Возьмем любые $k, j \in \{1, \ldots, M\}$ такие, что $z_k^1 > z_j^1$ (или, эквивалентно, $z_j^2 > z_k^2$). Положим $p := z_k^1$, $q := z_j^1$ и через $z_{-j,k}^i$ обозначим $(N-2)$-кортеж, состоящий из элементов $\{z_1^i, \ldots, z_M^i\} \setminus \{z_j^i, z_k^i\}$. Из (2.1) получаем следующие неравенства:

$$\begin{aligned} &pF_1(p, q, z_{-(j,k)}^1) + (1-p)F_2(1-p, 1-q, z_{-(j,k)}^2) \\ &\quad > pF_1(q, p, z_{-(j,k)}^1) + (1-p)F_2(1-q, 1-p, z_{-(j,k)}^2), \\ &qF_1(q, p, z_{-(j,k)}^1) + (1-q)F_2(1-q, 1-p, z_{-(j,k)}^2) \\ &\quad > qF_1(p, q, z_{-(j,k)}^1) + (1-q)F_2(1-p, 1-q, z_{-(j,k)}^2). \end{aligned}$$

Таким образом, если определить A и B посредством равенств

$$\begin{aligned} A &= F_1(p, q, z_{-(j,k)}^1) - F_1(q, p, z_{-(j,k)}^1) = F_1(z_k^1, z_{-k}^1) - F_1(z_j^1, z_{-j}^1), \\ B &= F_2(1-p, 1-q, z_{-(j,k)}^2) - F_2(1-q, 1-p, z_{-(j,k)}^2) \\ &= F_2(z_k^2, z_{-k}^2) - F_2(z_j^2, z_{-j}^2), \end{aligned}$$

то из леммы 6.1 будет следовать, что $A > 0$ и $B < 0$. Это доказывает (3.9) для $i = 1$ и $i = 2$.

Случай 3: $M \geqslant 2$, $N \geqslant 3$. Определим

$$Z = \begin{bmatrix} z_1^1 & z_1^2 & \ldots & z_1^N \\ z_2^1 & z_2^2 & \ldots & z_2^N \\ \vdots & \vdots & \vdots & \vdots \\ z_M^1 & z_M^2 & \ldots & z_M^N \end{bmatrix}.$$

Докажем (3.9) для i-го столбца матрицы Z. Выберем любые две строки (два типа) $j, k \in \{1, \ldots, M\}$, где $j \neq k$. Поскольку все, что требуется от матрицы Z, — это чтобы ее строки были невырожденными вероятностными распределениями, и поскольку значения F_i зависят лишь от значений в i-м столбце Z, то, чтобы завершить доказательство теоремы, нам нужно показать только, что для любых $p := z_k^i$ и $q := z_j^i$ таких, что $1 > p > q > 0$, а также для любых $z_{-(k,j)}^i \in (0,1)^{M-2}$ (если $M = 2$, то последнее требование излишне) справедливо неравенство

$$F_i(p, q, z_{-(k,j)}^i) > F_i(q, p, z_{-(k,j)}^i). \tag{6.6}$$

Чтобы применить (2.1) и лемму 6.1, преобразуем матрицу Z без изменения ее i-го столбца и найдем матрицу $\widetilde{Q}$ с теми же вероятностями типов s_ℓ, что и у исходной матрицы Q. Как отмечено выше, мы всегда можем это сделать, взяв $q_{\ell i} = z^i_\ell s_\ell$. Более того, при рассмотрении случаев 1 и 2 мы применяли лемму 6.1 к разности платежных функций в двух различных состояниях, соответствующих двум различным столбцам матрицы Z. Без ограничения общности предположим, что $i \neq 1$. Вместо исходной матрицы Z будем работать со следующей Z-матрицей, в которой i-й столбец остался неизменным: для $0 < \varepsilon < 1$ и $a := 1 - \varepsilon$ определим матрицу $\widetilde{Z}$ посредством равенств

$$\widetilde{Z}^t_l := \begin{cases} z^t_l, & \text{если } l \in \{1,\dots,M\} \setminus \{j,k\}, \\ p, & \text{если } l = k,\ t = i, \\ q, & \text{если } l = j,\ t = i, \\ a - p, & \text{если } l = k,\ t = 1, \\ a - q, & \text{если } l = j,\ t = 1, \\ \dfrac{\varepsilon}{N-2} & \text{в противном случае,} \end{cases}$$

где p, q — любые числа из интервала $(0, a)$ такие, что $p > q$. Тогда при любом выборе ε, p и q получаем, что $\widetilde{Z}$ является Z-матрицей, которая отличается от Z только j-й и k-й строками, и эти строки имеют вид

$$\begin{bmatrix} a-p & \dfrac{\varepsilon}{N-2} & \cdots & \dfrac{\varepsilon}{N-2} & p & \dfrac{\varepsilon}{N-2} & \cdots & \dfrac{\varepsilon}{N-2} \\ a-q & \dfrac{\varepsilon}{N-2} & \cdots & \dfrac{\varepsilon}{N-2} & q & \dfrac{\varepsilon}{N-2} & \cdots & \dfrac{\varepsilon}{N-2} \end{bmatrix}.$$

Применяя IC-неравенство (2.1) к строкам j и k, получаем

$$\sum_{t=1}^N \widetilde{z}^t_k F_t(\widetilde{z}^t_k, \widetilde{z}^t_{-k}) > \sum_{t=1}^N \widetilde{z}^t_k F_t(\widetilde{z}^t_j, \widetilde{z}^t_{-j}),$$
$$\sum_{t=1}^N \widetilde{z}^t_j F_t(\widetilde{z}^t_j, \widetilde{z}^t_{-j}) > \sum_{t=1}^N \widetilde{z}^t_j F_t(\widetilde{z}^t_k, \widetilde{z}^t_{-k}).$$

Заметим, что для $t \in \{1,\dots,N\} \setminus \{1, i\}$ имеют место равенства $\widetilde{z}^t_j = \widetilde{z}^t_k = \varepsilon/(N-2)$. Поэтому в обеих частях полученных неравенств присутствуют слагаемые вида $[\varepsilon/(N-2)]F_t\big(\varepsilon/(N-2), \varepsilon/(N-2), z^t_{-(j,k)}\big)$, которые взаимно сокращаются. Следовательно, эти неравенства принимают вид

$$\begin{aligned} &(a-p)F_1(a-p, a-q, z^1_{-(j,k)}) + pF_i(p, q, z^i_{-(j,k)}) \\ &\qquad > (a-p)F_1(a-q, a-p, z^1_{-(j,k)}) + pF_i(q, p, z^i_{-(j,k)}), \\ &(a-q)F_1(a-q, a-p, z^1_{-(j,k)}) + qF_i(q, p, z^i_{-(j,k)}) \\ &\qquad > (a-q)F_1(a-p, a-q, z^1_{-(j,k)}) + qF_i(p, q, z^i_{-(j,k)}). \end{aligned}$$

Если определить величины A и B посредством равенств

$$A = F_i(p, q, z^i_{-(j,k)}) - F_i(q, p, z^i_{-(j,k)}),$$
$$B = F_1(a-p, a-q, z^1_{-(j,k)}) - F_1(a-q, a-p, z^1_{-(j,k)})$$

и применить лемму 6.1, получим, что $A > 0$. Это доказывает неравенство (6.6) для $a > p > q > 0$. Устремляя ε к нулю, приходим к (6.6) для $1 > p > q > 0$.

Теорема 3.2 доказана.

Доказательство теоремы 4.1. Требуется получить представление скорингового балла BTS в виде логарифмов локальных апостериорных вероятностей.

Определим

$$p_{ij} = \Pr(X^\pi_i = 1,\, X^\gamma_j = 1),$$

где мы используем тот факт, что в силу свойства взаимозаменяемости правая часть не зависит от выбора $\pi \neq \gamma$. Тогда

$$\Pr(X^\pi = x^\pi \mid X^\gamma = x^\gamma) = \frac{p_{ij}}{\sum_{k=1}^M p_{kj}}. \tag{6.7}$$

Требуется установить следующие три свойства.

Свойство I: $y^\pi_j = \sum_{i=1}^M \bigl(x^\pi_i p_{ij} / \sum_{k=1}^M p_{ki}\bigr)$.

Свойство II: $\ln \Pr(X^\gamma = x^\gamma \mid X^\pi = x^\pi) = \sum_{j=1}^M x^\gamma_j \ln y^\pi_j$, где условие под знаком вероятности означает, что вероятность рассматривается при условии правдивого ответа.

Свойство III: $\ln \Pr(X^\pi = x^\pi \mid \overline{X} = \overline{x}) = \sum_{k=1}^M x^\pi_k \ln \overline{x}_k$.

Свойство I следует из предположения о байесовской игре: респонденты вычисляют условные вероятности, руководствуясь байесовским подходом. Свойство II является следствием свойства I и равенства (6.7). Для доказательства свойства III рассмотрим ℓ такое, что $x^\pi_\ell = 1$. Из теоремы де Финетти следует, что

$$\Pr(X^\pi = x^\pi \mid \overline{X} = \overline{x}) = \overline{x}_\ell = \sum_{k=1}^M x^\pi_k \overline{x}_k.$$

Сумма в правой части полученного равенства всегда имеет только одно ненулевое слагаемое. Отсюда, взяв логарифм от обеих частей, получаем свойство III.

Далее, пусть величины x^γ таковы, что

$$\overline{x}_k = \lim_n \frac{1}{n} \sum_s x^\gamma_k.$$

Заметим, что в силу свойства взаимозаменяемости мы можем изменить порядок респондентов так, что $\pi = 1$ и $\gamma = 2, \ldots, n+1$. Для данных

π и γ имеем $\ln \Pr(X^\pi = x^\pi \mid X^\gamma = x^\gamma) = \sum_{j=1}^M x_j^\pi \ln y_j^\gamma$. Мы всегда можем исключить из рассмотрения те γ, для которых $\Pr(X^\gamma = x^\gamma) = 0$. Таким образом, фактически имеется только конечное число возможных M-кортежей x^γ таких, что $0 < \Pr(X^\gamma = x^\gamma) < 1$, и существуют нижняя граница A и верхняя граница B такие, что $0 < A \leqslant \Pr(X^\gamma = x^\gamma) \leqslant B < 1$. Отсюда следует, что $A = \sqrt[n]{A^n} \leqslant \sqrt[n]{\prod_{s=1}^n \Pr(X^\gamma = x^\gamma)} \leqslant \sqrt[n]{B^n} = B$. Логарифмическая функция является непрерывной, поэтому $\ln \lim f = \lim \ln f$, если f и $\lim f$ — финитные и строго положительные функции. Следовательно, предел $\lim_{n\to\infty} \prod_{s=1}^n \Pr(X^\gamma = x^\gamma)$ существует, не равен нулю и мы можем внести логарифм под знак предела.

Далее, используя доказанное выше, из свойств I–III получим

$$\sum_{k=1}^M \overline{x}_k \ln y_k^\pi = \lim_n \frac{1}{n} \sum_s \ln \Pr(X^\gamma = x^\gamma \mid X^\pi = x^\pi),$$
$$\sum_{k=1}^M x_k^\pi \ln \overline{y}_k = \lim_n \frac{1}{n} \sum_s \ln \Pr(X^\pi = x^\pi \mid X^\gamma = x^\gamma);$$

воспользовавшись правилом Байеса, заключаем, что

$$\begin{aligned}
\mathrm{BTS}^\pi &= \sum_{k=1}^M x_k^\pi \ln \frac{\overline{x}_k}{\overline{y}_k} + \sum_{k=1}^M \overline{x}_k \ln y_k^\pi \\
&= \ln \Pr(X^\pi = x^\pi \mid \overline{X} = \overline{x}) + \lim_n \frac{1}{n} \sum_s \ln \Pr(X^\gamma = x^\gamma \mid X^\pi = x^\pi) \\
&\quad - \lim_n \frac{1}{n} \sum_s \ln \Pr(X^\pi = x^\pi \mid X^\gamma = x^\gamma) \\
&= \ln \left(\Pr(X^\pi = x^\pi \mid \overline{X} = \overline{x}) \lim_n \prod_{s=1}^n \frac{\Pr^{1/n}(X^\gamma = x^\gamma \mid X^\pi = x^\pi)}{\Pr^{1/n}(X^\pi = x^\pi \mid X^\gamma = x^\gamma)} \right) \\
&= \ln \left(\Pr(X^\pi = x^\pi \mid \overline{X} = \overline{x}) \frac{\lim_n \prod_{s=1}^n \Pr^{1/n}(X^\gamma = x^\gamma)}{\Pr(X^\pi = x^\pi)} \right) \\
&= \ln \Pr(\overline{X} = \overline{x} \mid X^\pi = x^\pi) - \ln \Pr(\overline{X} = \overline{x}) + \lim_n \frac{1}{n} \sum_s \ln \Pr(X^\gamma = x^\gamma).
\end{aligned}$$

Так как два последних слагаемых не зависят от π и $\sum_r \mathrm{BTS}^\pi = 0$, получаем равенство (4.1). Далее, для фиксированных n и $\overline{x}$ обозначим через n_j число респондентов типа j, так что

$$\sum_j n_j = n.$$

Тогда мы можем записать (4.1) в виде

$$\begin{aligned}\mathrm{BTS}^{\pi} &= \ln \Pr(\overline{X} = \overline{x} \mid X^{\pi} = x^{\pi}) - \lim_{n\to\infty} \frac{1}{n}\Bigg[\sum_{s=1}^{n_1} \ln \Pr(\overline{X} = \overline{x} \mid X^{\pi} = x^{1}) \\ &\quad + \cdots + \sum_{s=n_{M-1}+1}^{n_M} \ln \Pr(\overline{X} = \overline{x} \mid X^{\pi} = x^{M})\Bigg] \\ &= \ln \Pr(\overline{X} = \overline{x} \mid X^{\pi} = x^{\pi}) - \lim_{n\to\infty}\Bigg[\frac{n_1}{n} \ln \Pr(\overline{X} = \overline{x} \mid X^{\pi} = x^{1}) \\ &\quad + \cdots + \frac{n_M}{n} \ln \Pr(\overline{X} = \overline{x} \mid X^{\pi} = x^{M})\Bigg].\end{aligned}$$

Так как $\lim_{n\to\infty}(n_j/n) = \Pr(T^{\pi} = j \mid \overline{X} = \overline{x})$, равенство (4.2) установлено. Теорема 4.1 доказана.

Авторы выражают благодарность Джорджу Георгиадису, Мэттью Эллиоту, Катрине Лигетт, Патрику Рэю, слушателям конференций XVI Southwest Economic Theory Conference, 13th Conference on Research on Economic Theory & Econometrics, the 2017 Bayesian Crowd Workshop, а также анонимным рецензентам за замечания, предложения и комментарии к работе. Предыдущая версия статьи имела название “Mechanism design for an agnostic planner: universal mechanisms, logarithmic equilibrium payoffs and implementation”. Суждения и выводы, изложенные в настоящей работе, выражают личное мнение авторов и не должны рассматриваться как официальная позиция IARPA, DoI/NBC или Правительства США.

СПИСОК ЛИТЕРАТУРЫ

1. D. J. Aldous, “Exchangeability and related topics”, *École d’été de probabilités de Saint-Flour* XIII — 1983, Lecture Notes in Math., **1117**, Springer, Berlin, 1985, 1–198.
2. A. Baillon, “Bayesian markets to elicit private information”, *Proc. Natl. Acad. Sci. USA*, **114**:30 (2017), 7958–7962.
3. D. Bergemann, S. Morris, “An introduction to robust mechanism design”, *Foundations and Trends in Microeconomics*, **8**:3 (2012), 169–230.
4. J. M. Bernardo, “Expected information as expected utility”, *Ann. Statist.*, **7**:3 (1979), 686–690.
5. T. Börgers, *An introduction to the theory of mechanism design*, With a chapter by D. Krähmer, R. Strausz, Oxford Univ. Press, New York, 2015, xv+246 pp.
6. Yuan Shih Chow, H. Teicher, *Probability Theory. Independence, interchangeability, martingales*, 3rd ed., Springer Texts Statist., Springer-Verlag, New York, 1997, xxii+488 pp.
7. J. Cvitanić, D. Prelec, S. Radas, H. Šikić, “Game of duels: information-theoretic axiomatization of scoring rules”, *IEEE Trans. Inform. Theory*, **65**:1 (2019), 530–537.

8. J. Cvitanić, D. Prelec, B. Riley, B. Tereick, "Honesty via choice-matching", AER: *Insights*, **1**:2 (2019), 179–192.
9. A. Dasgupta, A. Ghosh, "Crowdsourced judgement elicitation with endogenous proficiency", WWW'13: *Proceedings of the* 22*nd* ACM *international World Wide Web conference* (Rio de Janeiro, 2013), ACM, New York, 2013, 319–330.
10. R. Frongillo, J. Witkowski, "A geometric perspective on minimal peer prediction", *ACM Trans. Econ. Comput.*, **5**:3 (2017), 17, 27 pp.
11. T. Gneiting, A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation", *J. Amer. Statist. Assoc.*, **102**:477 (2007), 359–378.
12. P. J. Howie, Ying Wang, J. Tsai, "Predicting new product adoption using Bayesian truth serum", *J. Med. Market.*, **11**:1 (2011), 6–16.
13. R. Jurca, B. Faltings, "Robust incentive-compatible feedback payments", TADA 2006, AMEC 2006: *Agent-mediated electronic commerce. Automated negotiation and strategy design for electronic markets* (Hakodate, Japan, 2006), Lecture Notes in Comput. Sci., **4452**, Lecture Notes in Artificial Intelligence, Springer, Berlin, 2007, 204–218.
14. Yuqing Kong, G. Schoenebeck, "An information theoretic framework for designing information elicitation mechanisms that reward truth-telling", ACM *Trans. Econ. Comput.*, **7**:1 (2019), 2, 33 pp.
15. A. Kukla-Gryz, J. Tyrowicz, M. Krawczyk, K. Siwiński, "We all do it, but are we willing to admit? Incentivizing digital pirates' confessions", *Appl. Econ. Lett.*, **22**:3 (2015), 184–188.
16. Yang Liu, Juntao Wang, Yiling Chen, *Surrogate scoring rules*, 2020, arXiv: 1802.09158.
17. T. A. Loughran, R. Paternoster, K. J. Thomas, "Incentivizing responses to self-report questions in perceptual deterrence studies: an investigation of the validity of deterrence theory using Bayesian truth serum", *J. Quant. Criminol.*, **30**:4 (2014), 677–707.
18. E. Maskin, T. Sjöström, "Implementation theory", *Handbook of social choice and welfare*, v. 1, Handbooks in Econom., **19**, Elsevier/North-Holland, Amsterdam, 2002, 237–288.
19. J. McCarthy, "Measures of the value of information", *Proc. Natl. Acad. Sci. USA*, **42**:9 (1956), 654–655.
20. S. R. Miller, B. P. Bailey, A. Kirlik, "Exploring the utility of Bayesian truth serum for assessing design knowledge", *Human–Computer Interaction*, **29**:5-6 (2014), 487–515.
21. N. Miller, P. Resnick, R. Zeckhauser, "Eliciting informative feedback: the peer-prediction method", *Manag. Sci.*, **51**:9 (2005), 1359–1373.
22. R. B. Myerson, "Optimal auction design", *Math. Oper. Res.*, **6**:1 (1981), 58–73.
23. T. Offerman, J. Sonnemans, G. Van De Kuilen, P. P. Wakker, "A truth serum for non-Bayesians: correcting proper scoring rules for risk attitudes", *Rev. Econom. Stud.*, **76**:4 (2009), 1461–1489.
24. D. Prelec, "A Bayesian truth serum for subjective data", *Science*, **306**:5695 (2004), 462–466.
25. D. Prelec, H. S. Seung, J. McCoy, "A solution to the single-question crowd wisdom problem", *Nature*, **541** (2017), 532–535.
26. G. Radanovic, B. Faltings, "A robust Bayesian truth serum for non-binary signals", AAAI'13: *Proceedings of the* 27*th* AAAI *conference on artificial intelligence*, AAAI Press, 2013, 833–839.

27. G. Radanovic, B. Faltings, “Incentives for truthful information elicitation of continuous signals”, AAAI’14: *Proceedings of the* 28*th* AAAI *conference on artificial intelligence*, AAAI Press, 2014, 770–776.

28. L. J. Savage, “Elicitation of personal probabilities and expectations”, *J. Amer. Statist. Assoc.*, **66**:336 (1971), 783–801.

29. M. J. Schervish, “A general method for comparing probability assessors”, *Ann. Statist.*, **17**:4 (1989), 1856–1879.

30. K. H. Schlag, J. J. van der Weele, “Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality”, *Theor. Econ. Lett.*, **3**:1 (2013), 38–42.

31. R. V. Vohra, *Paths, cycles and mechanism design*, Tech. rep., Northwestern Univ., Evanston, IL, 2007, 83 pp.

32. B. Waggoner, Yiling Chen, “Information elicitation sans verification”, *Proceedings of the* 3*rd workshop on social computing and user generated content* (SC-13) (Philadelphia, PA, 2013), 2013, http://yiling.seas.harvard.edu/publications/.

33. J. Witkowski, *Robust peer prediction mechanisms*, Ph.D. Diss., Albert-Ludwigs-Universität, Freiburg, 2014, 135 pp., https://freidok.uni-freiburg.de/data/10054.

34. J. Witkowski, D. C. Parkes, “A robust Bayesian truth serum for small populations”, AAAI’12: *Proceedings of the* 26*th* AAAI *conference on artificial intelligence*, AAAI Press, 2012, 1492–1498.

35. J. Witkowski, D. C. Parkes, “Learning the prior in minimal peer prediction”, *Proceedings of the* 3*rd workshop on social computing and user generated content* (SC-13) (Philadelphia, PA, 2013), 2013, http://econcs.seas.harvard.edu/files/econcs/files/witkowski_ec13.pdf.

36. P. Zhang, Yiling Chen, “Elicitability and knowledge-free elicitation with peer prediction”, AAMAS’14: *Proceedings of the* 2014 *international conference on autonomous agents and multi-agent systems* (Paris, 2014), IFAAMAS, Richland, SC, 2014, 245–252, http://yiling.seas.harvard.edu/publications/.

Поступила в редакцию
07.X.2018