Теорема Эрроу о диктаторе

Савватеев А.В., Шварц Д.А.

1. Введение

Постановка задачи (в ее современном понимании) такова: пусть перед обществом стоят несколько альтернатив: $a_1, \ldots, a_l$. Это могут быть какие-либо проекты, типа «построить мост», «открыть магазин», «учредить университет» и т.п. Всех членов общества опрашивают, предлагая им упорядочить имеющиеся альтернативы в порядке убывания привлекательности. По результатам опроса определяется коллективное решение, тоже как-то упорядочивающее альтернативы в соответсвии с индивидуальными предпочтениями. Весь вопрос в том, как именно определить коллективное решение.

Пожалуй, самым исторически ранним примером такой процедуры было правило простого большинства: одна альтернатива предпочтительней другой, если так считает большинство членов общества. Проблема в том, что это правило может не приводить ни к какому результату.

Изложенный выше классический пример впервые был построен французским ученым маркизом де Кондорсе в конце 18 века.

Примерно в то же время, другой французский ученый Борда предложил следующий способ принятия решений.

Членам общества предлагается строго упорядочить альтернативы в порядке убывания предпочтительности, а именно, присвоить им номера: $1, 2, \ldots, l$. Затем каждой альтернативе присваивается её суммарный «рейтинг», равный сумме номеров во всех n списках, где n — население страны (т.е. количество членов общества). «Общественное мнение» представляет собой список альтернатив, упорядоченный по рейтингу. Возможно, некоторые альтернативы станут для общества равноценными (не будучи равноценными ни для каких членов общества).

Подобными схемами «формирования общественного мнения» пользовались долгое время и пользуются по сию пору. Вместе с тем, было замечено, выбор по правилу Борда обладает странным свойством: для того, чтобы узнать, какая из данных двух альтернатив a и b более предпочтительна для общества, приходится узнавать мнение членов общества относительно всех прочих альтернатив, что создает богатые возможности для манипулирования.

Целью многих дальнейших исследований стал поиск «истинного», самого справедливого и неманипулируемого правила выбора, которое бы обладало тем свойством, что для сравнения данной пары альтернатив не нужно было опрашивать избирателей об их отношении к остальным. Это и накладно (дополнительная работа), и как-то нелогично. Было предложено множество хитроумных правил выбора, но ни одно из них не удовлетворяло простым и, казалось бы, вполне естественным требованиям.

Каково же было замешательство в научных кругах, когда в середине XX века (1951 год) один из величайших мат-экономистов XX века Кеннет Эрроу объявил, что единственным правилом «из всех вообще мыслимых правил», удовлетворяющим вышеупомянутому требованию, является правило диктатуры, т.е. полное игнорирование предпочтений и пожеланий всех граждан, кроме одного, который во всех случаях навязывает обществу свою волю. Столь

неожиданный и нетривиальный результат вызвал некоторый застой: только в середине 70-х годов стали появляться результаты, дополняющие теорему Эрроу или как-либо иначе с ней тесно связанные.

Спустя ровно 50 лет (в 2001 г.) вышла работа, в которой было изложено существенно более короткое и прозрачное доказательство, чем то, которое представил сам Эрроу. Авторам, в свою очередь, удалось еще немного его сократить.

2. Постановка задачи

Обозначим за N множество игроков (точнее, «членов общества»), а за A — множество доступных обществу *альтернатив*.

Занумеруем игороков числами от 1 до N. и обозначим за P_i предпочтения i-го игрока.

Для формализации понятия «предпочтение» придется ввести несколько вспомогательных понятий.

Бинарным отношением P на множестве A называется подмножество упорядоченных пар элементов A, т.е. $P \subset A \times A$.

Факт $(a, b) \in P$ также записывается, как aPb (по аналогии с бинарными отношениями «$>$», «$=$»,...).

Предпочтения игроков и общественное мнение записываются бинарными отношениями по следующему правилу: пара (a, b) входит в бинарное отношение в том и только в том случае, когда игрок (соответственно, общественное мнение) считает альтернативу a не хуже альтернативы b (в частности, пара (a, a) всегда входит в отношение).

Бинарное отношение P называется линейным порядком ($\mathcal{LO}$), если оно транзитивно (из aPb и bPc следует aPc), антисимметрично (если при $a \neq b$ верно aPb, то bPa верным быть не может) и полно ($\forall a, b$ или aPb или bPa имеет место).

(Менее формально, это «строгое предпочтение» — альтернативы пронумерованы и первая альтернатива предпочтительней всех остальных, вторая — всех, кроме первой, и т.д. последняя из альтернатив менее предпочтительна.

Бинарное отношение P называется слабым порядком ($\mathcal{WO}$), если оно полно и транзитивно, однако не обязательно антисимметрично — то есть, при $a \neq b$ возможна ситуация, что aPb и bPa одновременно: альтернативы a и b равнозначны между собой.

(Менее формально, это «нестрогое предпочтение» — т.е. упорядочиваются не сами альтернативы, а классы эквивалентных альтернатив.)

Упражнение 1. *Придайте точный смысл последнему утверждению и убедитесь тем самым, что определение через бинарные отношения совпадает с интуитивным пониманием строгого и нестрогого упорядочения.*

Мы предполагаем, что индивидуальные предпочтения P_i являются линейными порядками, а коллективное предпочтение может быть и слабым порядком, то есть некоторые альтернативы для общества могут быть. равноценны.

Набор отношений $(P_1, \ldots, P_n)$ называется *профилем участников* и обозначается как $\vec{P} = \{P_i\}_{i=1}^n$.

Задача общественного выбора состоит в построении по любому мыслимому профилю участников $\vec{P}$ отношения P, отражающего мнение коллектива как единого целого. Более формально, *функционалом общественного выбора* F называется правило построения коллективного предпочтения по личным, то есть произвольное отображение $F : \mathcal{LO}^N \to \mathcal{WO}$. Обозначим для краткости $P = F(\vec{P})$.

Для удобства дальнейшего введем в рассмотрение множество участников, которое включает пару (a, b) в свои отношения P_i, т.е.

$$V(a, b; \vec{P}) = \{i \in N \mid (a, b) \in P_i\}.$$

Мы накладываем на функционал F следующие два ограничения:

(1) Паретовость, или эффективность (скажем, если все члены общества считают, что новый кабак предпочтительнее новому детсаду, то функционал общественного выбора тоже должен признавать открытие кабака строго более насущной задачей, нежели открытие детсада).

Формально это условие можно записать так

$$\forall(x, y), \forall \vec{P}\ \{V(x, y; \vec{P}) = N\} \Rightarrow xPy \text{ и не } yPx).$$

(2) Независимость от посторонних альтернатив: для любой пары вариантов x и y из A решение о том, какой вариант предпочтительнее другого в коллективном решении P, зависит от информации относительно только этих вариантов в индивидуальных отношениях P_i. То есть, скажем, при изучении вопроса о том, что предпочтительнее: строить кабак или детский сад, мнение населения выявляется только относительно этих двух альтернатив и не выясняется, например, как люди оценивают строительство моста, по сравнению со строительством кабака или детского садика.

Формально же эта аксиома записывается следующим образом: для любой пары (x, y) и любых двух профилей $\vec{R}$ и $\vec{R}'$

$$V(x, y; \vec{R}) = V(x, y; \vec{R}') \Rightarrow (xRy \Leftrightarrow xR'y),$$

Оба упомянутых правила нашим условиям не удовлетворяют: Правило простого большинства не дает решений в классе слабых порядков, правило Борда зависит от посторонних альтернатив.

Существует пример «нечестного» функционала, заведомо удовлетворяющего этим двум требованиям. А именно, представим себе, что среди членов общества есть человек, во всех случаях навязывающий коллективу своё мнение. (в разных ситуациях таким человеком может быть диктатор, руководитель похода, отец семейства, царь-сатрап...). Примем это мнение за функционал. Формально, этот функционал определяется так: $F(P_1, \ldots, P_n) = P_d$ при некотором d, то есть, функционал общественного мнения есть проекция профиля участников $\vec{P}$ на d-ю координату. Построенный функционал носит название *диктаторского*, или попросту *диктатуры d-го участника*.

Упражнение 2. *Убедитесь в том, что вы понимаете, почему диктаторский функционал удовлетворяет вышеуказанным требованиям.*

К.Эрроу (Kenneth Arrow) в 1951 году установил, что других функционалов, удовлетворяющих двум указанным ограничениям, не существует. Точнее:

Теорема 1. *Пусть число альтернатив больше 2. Тогда функционал, удовлетворяющий условиям эффективности и независимости от посторонних альтернатив, является диктаторским.*

Перед тем, как доказывать эту основную теорему, давайте заметим, что ограничение на число альтернатив является существенным.

Упражнение 3. *Покажите, что если альтернатив всего две, то голосование по простому большинству удовлетворяет условиям эффективности и независимости, но не является диктаторским, если, конечно, членов общества больше, чем один. (При равенстве голосов, альтернативы объявляются равнозначными с точки зрения коллектива.)*

Для доказательства теоремы Эрроу нам понадобится несколько лемм.

Лемма 1. (о нейтральности) *Пусть $V(x, y, \vec{P}) = V(z, t, \vec{P})$, тогда из xPy следует zPt.*

Переставив местами пары (x, y) и (z, t), получаем, что на самом деле в этом случае наблюдается эквивалентность: $xPy \Leftrightarrow zPt$.

Эта лемма утверждает об «отсутсвии двойных стандартов» — если, например, одни и те же избиратели считают строительство кабака более важной задачей, нежели строительство детского сада, и строительство школы предпочитают строительству ясель, то, коль скоро принято решение, что строить кабак важнее, чем детский сад, то аналогичное решение должно быть принято относительно школы и ясель.

Доказательство. Его мы разделим на два этапа — сначала заменим y на t, а потом x на z.

1) Докажем, что при $y \neq t$, если $V(x, y, \vec{P}) = V(x, t, \vec{P})$, и xPy, то xPt. При доказательстве этого, а также многих последующих утверждений, будет использоваться один и тот же приём. А именно, по условию независимости от посторонних альтернатив, ответ на вопрос «xPt или tPx» зависит только от множества $W_{x,t} = V(x, t, \vec{P})$; помимо этой, никакой информации о профиле P не требуется.

А следовательно, если доказать, что $xP't$ для какого-то одного профиля с тем же самым $W_{x,t} = V(x, t, \vec{P'})$, то это будет верно и для всех остальных таких профилей, в частности, и для исходного профиля P.

Поэтому рассмотрим следующий «удобный» профиль P':

$V(x, y, \vec{P'})$	$V(y, x, \vec{P'})$
$\vdots$	$\vdots$
x	y
y	t
t	x
$\vdots$	$\vdots$

Так как $V(x, y, \vec{P}) = V(x, y, \vec{P'})$, то имеем $xP'y$ (по условию, xPy, а по аксиоме независимости $xPy \Leftrightarrow xP'y$); далее, $yP't$ по условию эффективности (все избиратели считают y предпочтительней t).

Поскольку P' транзитивно, получаем $xP't$. И опять применяем аксиому независимости, получая, что xPt.

По аналогии с предыдущим, что если $V(x, t, \vec{P}) = V(z, t, \vec{P})$, и xPt, то zPt.

Упражнение 4. *Докажите по аналогии с предыдущим, что если $V(x, t, \vec{P}) = V(z, t, \vec{P})$, и xPt, то zPt.*

Утверждение леммы напрямую следует из установленных фактов. ■

Упражнение 5. *Пользуясь аксиомой эффективности и тем, что во вспомогательном профиле P' мы наблюдаем $V(y,t;P') = N$, докажите, что на самом деле область значений функционала F лежит строго внутри множества линейных порядков.*

Определение 1. *Пусть дан профиль $\vec{P}$ индивидуальных предпочтений. Назовем альтернативу a экстремальной, если все избиратели считают ее либо лучшей, либо худшей, то есть если $\forall i$ либо $\forall b \neq a$ aP_ib, либо $\forall b \neq a$ bP_ia.*

Лемма 2. (об экстремальной альтернативе) *Экстремальная альтернатива a и в коллективном предпочтении тоже займет либо первое, либо последнее место, то есть, если a экстремальна, то либо $\forall b \neq a$ aPb, либо $\forall b \neq a$ bPa.*

Надо сказать, что в этой лемме уже содержится шокирующий факт. В самом деле, если половина общества считает данную альтернативу самой лучшей, а другая половина — самой худшей, то коллективное решение должно, вроде бы, поставить эту альтернативу куда-нибудь в середину, таким образом, находя определённый компромисс. Этого, однако, не происходит.

Доказательство. Обозначим за M множество избирателей, считающих a лучшей альтернативой. Рассмотрим произвольную альтернативу $b \neq a$. Заметим, что избиратели, считающие a лучшей, в частности ставят a над b, а считающие a худшей — ставят b над a, то есть $V(a,b,\vec{P}) = M$ и никак не зависит от b. Следовательно, для произвольных альтернатив b и c, не совпадающих с a, мы имеем $V(a,b,\vec{P}) = V(a,c,\vec{P})$.

По лемме о нейтральности либо одновременно aPb и aPc, либо bPa и cPa, либо $b \sim a$ и $c \sim a$, то есть в коллективном решении все альтернативы либо одновременно лучше a, либо одновременно хуже a, либо одновременно равнозначны с a.

В первом случае a — худшая альтернатива, во втором — лучшая. В силу упражнения 5 последний случай невозможен. ■

Теперь приступим к доказательству теоремы.

Доказательство.

Нахождение диктатора.

Рассмотрим серию профилей, отличающихся друг от друга только мнением избирателей об альтернативе a: в первом профиле все считают a лучшей альтернативой, во втором профиле участник 1 считает a худшей, остальные — лучшей, в третьем a — худшая альтернатива уже по мнению участников 1 и 2, по мнению остальных a лучшая. И так далее. В последнем, $(n+1)$-м профиле все будут считать a худшей альтернативой.

P_1	$\dots$	P_i	$\dots$	P_n	P
a	$\dots$	a	$\dots$	a	a
$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

P_1	$\dots$	P_i	$\dots$	P_n	P
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a	$\dots$	a	$\dots$	a	a

P_1	$\dots$	P_i	P_{i+1}	$\dots$	P_n	P
$\dots$	$\dots$	$\dots$	a	$\dots$	a	a?
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a	$\dots$	a	$\dots$	$\dots$	$\dots$	a?

По аксимоме единогласия a будет худшей альтернативой в первом случае и лучшей в последнем, а по лемме об экстремальной альтернативе — или лучшей или худшей альтернативой

во всех остальных случаях. Поэтому в какой-то момент a за один шаг из худшей альтернативы станет лучшей. (Возможно, потом она станет снова худшей, и потом снова лучшей, и т.д., но для нас важно лишь то, что существует момент, когда наша альтернатива из худшей становится лучшей. Возьмём, например, и рассмотрим первый такой момент.)

Случится это при смене предпочтений ровно одного участника. Обозначим его d. Приведенное рассуждение показывает, что d может навязать свое мнение всему коллективу, но пока что только по поводу конкретной экстремальной альтернативы и в конкретных обстоятельствах.

Ситуацию иллюстрируют следующие профили:

P_1	$\dots$	P_{d-1}	P_d	P_{d+1}	$\dots$	P_n	P
$\dots$	$\dots$	$\dots$	$\dots$	a	$\dots$	a	$\dots$
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$
a	$\dots$	a	a	$\dots$	$\dots$	$\dots$	a

и

P_1	$\dots$	P_{d-1}	P_d	P_{d+1}	$\dots$	P_n	P
$\dots$	$\dots$	$\dots$	a	a	$\dots$	a	a
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\vdots$
a	$\dots$	a	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$

Далее мы покажем, что именно d является диктатором, то есть может навязать свое мнение по поводу любой пары альтернатив независимо от мнений других участников, то есть для любых альтернатив b и c из bP_dc следует bPc.

Итак, пусть b и c — произвольные альтернативы и bP_dc. Поскольку по условию независимости от посторонних альтернатив решение относительно b и c не зависит от третьей альтернативы a, то достаточно предъявить профиль $P^{\vec{(1)}}$ с тем же самым множеством $W = V(b, c; P^{\vec{(1)}}) = V(b, c; \vec{P})$, в котором $bP^{(1)}c$ — в этом случае мы заключим, что и bPc тоже.

А именно, рассмотрим профиль $P^{\vec{(1)}}$, в котором для всех участников, кроме d альтернатива a экстремальна, и последние ранжируют ее также, как и при выборе диктатора. «Диктатор» же (то есть, пока что просто участник d) в профиле $P^{\vec{(1)}}$ считает a промежуточной альтернативой между b и c.

Итак, достаточно доказать, что $bP^{(1)}c$ для следующего профиля $P^{\vec{(1)}}$:

$P_1^{(1)}$	$\dots$	$P_{d-1}^{(1)}$	$P_d^{(1)}$	$P_{d+1}^{(1)}$	$\dots$	$P_n^{(1)}$
$\dots$	$\dots$	$\dots$	$\dots$	a	$\dots$	a
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\ddots$
$\vdots$	$\ddots$	$\vdots$	b	$\vdots$	$\ddots$	$\ddots$
$\vdots$	$\ddots$	$\vdots$	a	$\vdots$	$\ddots$	$\ddots$
$\vdots$	$\ddots$	$\vdots$	c	$\vdots$	$\ddots$	$\ddots$
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\ddots$
a	$\dots$	a	$\dots$	$\dots$	$\dots$	$\dots$

Для доказательства рассмотрим два вспомогательных профиля $\vec{P}''$ и $\vec{P}'$, отличающихся от $P^{(1)}$ только предпочтениями d.

$$\vec{P}' = \begin{array}{ccccccc} P'_1 & \dots & P'_{d-1} & P'_d & P'_{d+1} & \dots & P'_n \\ \dots & \dots & \dots & \dots & a & \dots & a \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots \\ \vdots & \ddots & \vdots & b & \vdots & \ddots & \ddots \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots \\ \vdots & \ddots & \vdots & c & \vdots & \ddots & \ddots \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots \\ a & \dots & a & a & \dots & \dots & \dots \end{array}$$

$$\vec{P}'' = \begin{array}{ccccccc} P''_1 & \dots & P''_{d-1} & P''_d & P''_{d+1} & \dots & P''_n \\ \dots & \dots & \dots & a & a & \dots & a \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots \\ \vdots & \ddots & \vdots & b & \vdots & \ddots & \ddots \\ \vdots & \ddots & \vdots & c & \vdots & \ddots & \ddots \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \ddots \\ a & \dots & a & \dots & \dots & \dots & \dots \end{array}$$

Заметим, что $V(b, a, \vec{P}') = V(b, a, \vec{P^{(1)}}) = \{1, \dots d\}$. Кроме того, по определению d, альтернатива a в профиле $\vec{P}'$ — худшая, в частности, $bP'a$. По лемме о нейтральности имеем, что $bP^{(1)}a$.

Рассмотрим теперь профили $\vec{P}''$ и $\vec{P^{(1)}}$. Имеем: $V(a, c, \vec{P}'') = V(a, c, \vec{P^{(1)}}) = \{d, \dots n\}$. По определению d, альтернатива a в профиле $\vec{P}''$ — лучшая, в частности, $aP''c$. По лемме о нейтральности имеем, что $aP^{(1)}c$.

Итак, имеем $bP^{(1)}a$ и $aP^{(1)}c$. Поскольку коллективное предпочтение транзитивно, получаем $bP^{(1)}c$. Теорема доказана. ■