

УДК 519.865

НАХОЖДЕНИЕ МИНИМАКСНЫХ СТРАТЕГИИ И РИСКА В БИНАРНОЙ СЛУЧАЙНОЙ СРЕДЕ

А.В.Колногоров

FINDING THE MINIMAX RISK AND MINIMAX STRATEGY IN A BINARY RANDOM ENVIRONMENT

A.V.Kolnogorov

Институт электронных и информационных систем НовГУ, Alexander.Kolnogorov@novsu.ru

Предложен алгоритм нахождения минимаксных стратегии и риска в бинарной стационарной случайной среде. Алгоритм сводится к поиску глобального максимума функции, численно равной значению минимаксного риска для среды, характеризуемой некоторым конечным подмножеством исходного множества параметров.

Ключевые слова: управление в случайной среде, задача о двуруком бандите, минимаксный подход, системы линейных неравенств, теорема Хелли

We propose an algorithm of finding the minimax risk and minimax strategy in a binary stationary random environment. The algorithm reduces to searching the global maximum of the function being equal to the value of minimax risk for the environment described by some finite subset of the initial parameter set.

Keywords: control in random environment, two-armed bandit problem, minimax approach, systems of linear inequalities, Helly's theorem

1. Введение

В работе развиваются исследования по нахождению минимаксных стратегии и риска в бинарной стационарной случайной среде, начатые в [1,2]. Бинарная стационарная случайная среда — это управляемый случайный процесс ξ_t , $t=1, \dots, T$, значения которого $\{\xi_t\}$ интерпретируются как доходы и зависят только от текущих выбираемых действий $\{y_t\}$ следующим образом: $\Pr\{\xi_t=1|y_t=\ell\}=p_\ell$, $\Pr\{\xi_t=0|y_t=\ell\}=q_\ell$, $\ell=1, \dots, K$ ($K \geq 2$).

Таким образом, процесс полностью описывается векторным параметром $\theta=(p_1, \dots, p_K)$, который предполагается неизвестным. Известным предполагается множество Θ его допустимых значений, которое описывает класс рассматриваемых процессов и может быть произвольным подмножеством единичного K -мерного куба $\{\theta: 0 \leq p_\ell \leq 1, \ell=1, \dots, K\}$, который соответствует классу всех управляемых процессов рассматриваемого вида.

Стратегия σ описывается массивом вероятностей, $\Pr\{y_t=\ell|y^{t-1}; \xi^{t-1}\}=\sigma_\ell(y^{t-1}; \xi^{t-1})$, $\ell=1, \dots, K$, определяющих выбор действия y_t в зависимости от известной предыстории — примененных действий $y^{t-1}=y_1, \dots, y_{t-1}$ и полученных в ответ на них доходов $\xi^{t-1}=\xi_1, \dots, \xi_{t-1}$. При $t=1$ предыстория отсутствует, поэтому зависимость от нее может быть опущена в обозначениях, т.е. $\sigma_\ell(y^{t-1}; \xi^{t-1})=\sigma_\ell$, $\ell=1, \dots, K$ при $t=1$. Ясно, что вероятности $\{\sigma_\ell(y^{t-1}; \xi^{t-1})\}$ являются неотрицательными числами, удовлетворяющими равенствам $\sum_{\ell=1}^K \sigma_\ell(y^{t-1}; \xi^{t-1})=1$ для любого $t=1, \dots, T$ и любой возможной предыстории $(y^{t-1}; \xi^{t-1})$.

Обозначим через $p^*=\max(p_1, \dots, p_K)$. Тогда

$$Losses_T(\sigma, \theta) = E_{\sigma, \theta} \left(\sum_{t=1}^T (p^* - \xi_t) \right)$$

— функция потерь дохода относительно его максимально возможного значения вследствие неполноты

информации. Здесь через $E_{\sigma, \theta}$ обозначен знак математического ожидания по мере, порожденной стратегией σ и параметром θ . Обозначим через

$$R_T(\Theta) = \inf_{\{\sigma\}} \sup_{\Theta} \text{Losses}_T(\sigma, \theta) \quad (1)$$

— минимаксный риск, определенный на множестве параметров Θ , через $\sigma(\Theta)$ — соответствующую минимаксную стратегию. В [2] установлено, что минимаксная стратегия $\sigma(\Theta)$ существует для любого множества Θ . При $K=2$ этот результат был ранее установлен в [3].

Основной результат данной статьи состоит в том, что минимаксный риск (1) на любом множестве Θ численно равен минимаксному риску на некотором конечном его подмножестве $\theta_1^0, \dots, \theta_n^0$, при этом для числа n дается оценка, зависящая от K и T .

2. Стратегия и вариация стратегии

Рассмотрим достаточную статистику $\zeta_{t-1} = ((n_1, m_1), \dots, (n_K, m_K))$, где (n_ℓ, m_ℓ) равны соответственно количеству применений варианта с номером ℓ и количеству полученных при этом доходов равных 1 ($\ell=1, \dots, K$). Тогда (см. [2]) функцию потерь можно записать в виде

$$L_T(\pi, \theta) = \sum_{t=1}^T \sum_{\{\zeta_{t-1}\}} \sum_{\ell=1}^K \pi_\ell(\zeta_{t-1}) a(\zeta_{t-1}; \theta), \quad (2)$$

где все $a(\zeta_{t-1}; \theta)$ — непрерывные ограниченные функции θ . Совокупность неотрицательных чисел $\{\pi_\ell(\zeta_{t-1})\}$ образует выпуклый симплекс вида

$$\sum_{\ell=1}^K \pi_\ell = 1, \quad (3)$$

$$\sum_{\ell=1}^K \pi_\ell(\zeta_{t-1}) = \sum_{k=1}^K \sum_{x=0}^1 \sum_{\zeta_{t-2}: S(\zeta_{t-2}, k, x) = \zeta_{t-1}} \pi_k(\zeta_{t-2}) \quad (4)$$

для всех возможных достаточных статистик $\{\zeta_{t-1}\}$, $t=2, \dots, T$. Здесь через $S(\zeta_{t-1}, y_t, \xi_t) \rightarrow \zeta_t$ обозначен оператор, формирующий статистику ζ_t по $(\zeta_{t-1}, y_t, \xi_t)$. В [2] установлено, что система вероятностей π , удовлетворяющая условиям (3), (4), позволяет однозначно определить стратегию σ , обеспечивающую выполнение равенства $L_T(\pi, \theta) = \text{Losses}_T(\sigma, \theta)$ при всех $\theta \in \Theta$. Соответствующую минимаксную стратегию обозначим $\pi(\Theta)$.

Определим вариацию стратегии δ совокупностью величин $\{\delta_\ell(\zeta_{t-1})\}$, удовлетворяющих условиям

$$\sum_{\ell=1}^K \delta_\ell = 0, \quad (5)$$

$$\sum_{\ell=1}^K \delta_\ell(\zeta_{t-1}) = \sum_{k=1}^K \sum_{x=0}^1 \sum_{\zeta_{t-2}: S(\zeta_{t-2}, k, x) = \zeta_{t-1}} \delta_k(\zeta_{t-2}) \quad (6)$$

для всех возможных предысторий $\{\zeta_{t-1}\}$, $t=2, \dots, T$.

Если заданы некоторая стратегия π и вариация δ , то $\pi + \delta$, определенная совокупностью величин $\{\pi_\ell(\zeta_{t-1}) + \delta_\ell(\zeta_{t-1})\}$, удовлетворяет условиям (3), (4) и, следовательно, также является стратегией в случае неотрицательности всех этих величин. В дальнейшем будем называть допустимыми вариации δ , для которых $\pi + \varepsilon_0 \delta$ является стратегией для некоторого $\varepsilon_0 > 0$. В этом случае стратегиями, очевидно, будут и $\pi + \varepsilon \delta$ при $0 < \varepsilon < \varepsilon_0$. Запрещенными будем называть такие вариации δ , для которых $\pi + \varepsilon \delta$ не является стратегией ни при каком $\varepsilon > 0$ в силу отрицательности какой-либо из ее величин. Ясно, что допустимые и запрещенные вариации зависят от варьируемой стратегии π . Следующая лемма позволяет различать допустимые и запрещенные вариации.

Лемма 1. Пусть даны некоторая стратегия π и вариация δ . Для того чтобы вариация δ была допустимой для стратегии π , необходимо и достаточно выполнения условия

$$\delta_\ell(\zeta_{t-1}) \geq 0, \text{ если } \pi_\ell(\zeta_{t-1}) = 0$$

для всех $\ell=1, \dots, K$ и всех ζ_{t-1} , $t=1, \dots, T$.

Доказательство. Необходимость следует из того, что в противном случае при любом $\varepsilon > 0$ появятся отрицательные величины $\pi_\ell(\zeta_{t-1}) + \varepsilon \delta_\ell(\zeta_{t-1})$.

Установим достаточность. Обозначим $m = \min_{\pi_\ell(\zeta_{t-1}) > 0} \pi_\ell(\zeta_{t-1})$, $M = \max_{\delta_\ell(\zeta_{t-1}) < 0} |\delta_\ell(\zeta_{t-1})|$. Из условий (3)-(6) следует, что найдется хотя бы одна строго положительная $\pi_\ell(\zeta_{t-1})$ и хотя бы одна строго отрицательная $\delta_\ell(\zeta_{t-1})$, если $\delta \neq 0$. Поэтому $m > 0$, $M > 0$ и для $\varepsilon_0 = m/M > 0$ все величины $\pi_\ell(\zeta_{t-1}) + \varepsilon_0 \delta_\ell(\zeta_{t-1})$ неотрицательны, т.е. вариация δ допустима. Лемма доказана.

Множество всех вариаций обозначим через Δ , подмножества запрещенных и разрешенных вариаций — через Δ^p и Δ^a соответственно. Ясно, что $\Delta = \Delta^a \cup \Delta^p$, $\Delta^a \cap \Delta^p = \emptyset$, причем Δ^p и Δ^a являются открытым и замкнутым множествами соответственно. Из (2) следует, что для стратегии $\pi + \varepsilon \delta$ выполнено равенство:

$$L_T(\pi + \varepsilon \delta, \theta) = L_T(\pi, \theta) + \varepsilon \cdot L_T(\delta, \theta).$$

Множество вариаций, удовлетворяющих условиям (5), (6), образует линейное пространство. При этом все вариации полностью определяются своими произвольными значениями $\{\delta_\ell(\zeta_{t-1})\}$ при $\ell < K$, а значения $\{\delta_K(\zeta_{t-1})\}$ могут быть определены из условий (5), (6). Это означает, что в качестве базиса в этом пространстве можно выбрать вариации δ , у которых есть ровно одна $\delta_\ell(\zeta_{t-1}) = 1$, а все остальные равны нулю (считаем, что $\ell < K$). Общее число таких вариаций обозначим $n(K, T)$. В [2] установлена оценка $n(K, T) \sim (K-1)T^{2K}/(2K)!$.

3. Критерий минимаксной стратегии

Для произвольной стратегии π рассмотрим множество параметров $\Theta(\pi) = \operatorname{Argmax}_{\Theta} L_T(\pi, \theta)$. Справедлива следующая теорема.

Теорема 1. Для того, чтобы стратегия π была минимаксной на множестве Θ , необходимо и достаточно, чтобы для любой ее допустимой вариации δ существовал такой параметр $\theta(\delta) \in \Theta(\pi)$, что выполнялось неравенство

$$L_T(\delta, \theta(\delta)) \geq 0. \quad (7)$$

Множество $\Theta^0 = \bigcup \theta(\delta)$ является замкнутым.

Доказательство. Установим необходимость. Пусть π — минимаксная стратегия, а δ такова, что $\pi + \varepsilon_0 \delta$ является стратегией при $\varepsilon_0 > 0$. Тогда $\pi + \varepsilon \delta$ также является стратегией при любом $0 < \varepsilon < \varepsilon_0$. Зададимся последовательностью $\{\varepsilon_n\}$, такой что $0 < \varepsilon_n < \varepsilon_0$, $\varepsilon_n \rightarrow 0$ при $n \rightarrow \infty$. Тогда для некоторой последовательности $\{\theta_n\}$ будут выполнены неравенства

$$L_T(\pi + \varepsilon_n \delta, \theta_n) \geq R_T(\Theta) \geq L_T(\pi, \theta_n), \quad n=1, 2, \dots, \quad (8)$$

причем правое неравенство следует из того, что π — минимаксная стратегия. Вычитая правые части неравенств из левых, получаем

$$L_T(\delta, \theta_n) \geq 0, \quad n=1, 2, \dots \quad (9)$$

Так как Θ — компактное множество, то из последовательности $\{\theta_n\}$ можно выбрать сходящуюся подпоследовательность; без ограничения общности можно считать, что $\{\theta_n\}$ сама есть сходящаяся последовательность, предел которой обозначим через $\theta(\delta)$. Так как $L_T(\pi, \theta)$ непрерывная функция своих аргументов, то из левого неравенства (8) при $n \rightarrow \infty$ получаем неравенство $L_T(\pi, \theta(\delta)) \geq R_T(\Theta)$, откуда следует, что $\theta(\delta) \in \Theta(\pi)$. Из (9) предельным переходом при $n \rightarrow \infty$ получаем, что для указанного $\theta(\delta)$ справедливо неравенство (7).

Установим достаточность. Предположим, напротив, что условия теоремы выполнены, но π — не минимаксная стратегия и рассмотрим минимаксную стратегию $\pi(\Theta)$. Так как в этом случае выполнено неравенство

$$\max_{\Theta} L_T(\pi(\Theta), \theta) < M = \max_{\Theta} L_T(\pi, \theta),$$

то для всех $\theta \in \Theta(\pi)$ выполнено неравенство

$$L_T(\pi(\Theta), \theta) - L_T(\pi, \theta) = L_T(\pi(\Theta) - \pi, \theta) < 0.$$

Поэтому для допустимой вариации $\delta = \pi(\Theta) - \pi$ условие (9) не выполнено ни для какого $\theta \in \Theta(\pi)$. Полученное противоречие доказывает утверждение теоремы.

4. Сведение задачи к нахождению минимаксной стратегии для конечного множества параметров

Для конечного множества параметров $\{\theta_1, \dots, \theta_m\}$ определим минимаксный риск

$$R_T(\theta_1, \dots, \theta_m) = \min_{\{\pi\}} \max_{i=1, \dots, m} L_T(\pi, \theta_i),$$

соответствующую минимаксную стратегию обозначим через $\pi(\theta_1, \dots, \theta_m)$. Для дальнейшего потребуется следующий известный результат. Пусть имеется, возможно, бесконечная система линейных однородных неравенств

$$\sum_{j=1}^n a_j(\theta) \delta_j \geq 0, \quad \theta \in \Theta, \quad (10)$$

где Θ — произвольное множество параметров, а $a_j(\theta)$, $j=1, \dots, n$ — фиксированные для данного θ произвольные числа — коэффициенты неравенств, не

все равные нулю. Положим $c_i(\theta) = a_i(\theta) \left(\sum_{j=1}^n a_j^2(\theta) \right)^{-\frac{1}{2}}$,

$i=1, \dots, n$ и определим вектор $c(\theta) = (c_1(\theta), \dots, c_n(\theta))$. Справедлива следующая лемма.

Лемма 2. Пусть система линейных однородных неравенств (10) такова, что множество $C = \bigcup_{\theta \in \Theta} c(\theta)$

замкнуто, и для любого набора действительных переменных δ_j , $j=1, \dots, n$, не всех равных нулю, найдется хотя бы одно значение θ , для которого соответствующее неравенство выполняется. Тогда из системы (10) можно выделить конечную подсистему неравенств

$$\sum_{j=1}^n a_j(\theta_i) \delta_j \geq 0, \quad \theta_i \in \Theta, \quad i=1, \dots, N,$$

обладающую тем же свойством, причем для N справедлива оценка $N \leq n+1$, которая является неулучшаемой. Лемма является одним из следствий теоремы Хелли о покрытиях выпуклыми множествами и следует из результатов [4].

Теорема 2. Минимаксная стратегия $\pi(\Theta)$ совпадает с некоторой минимаксной стратегией $\pi(\theta_1^0, \dots, \theta_n^0)$ на конечном множестве параметров. На этом множестве параметров минимаксный риск достигает своего максимального значения, т.е.

$$R_T(\theta_1^0, \dots, \theta_n^0) = \max_{\{\theta_1, \dots, \theta_m\}} R_T(\theta_1, \dots, \theta_m), \quad (11)$$

а для числа n справедлива оценка $n \leq n(K, T) + 1$.

Доказательство. Рассмотрим всевозможные вариации δ , заданные наборами действительных чисел, удовлетворяющих условиям (5), (6). При доказательстве теоремы будем считать, что вариации заданы своими произвольными значениями $\{\delta_\ell(\zeta_{\ell-1})\}$, $\ell=1, \dots, K-1$; в этом случае величины $\{\delta_K(\zeta_{K-1})\}$ будут их линейными комбинациями и могут быть определены из условий (5), (6). Таким образом, множество вариаций образует линейное пространство размерности $n(K, T)$.

В силу леммы 1 все вариации δ делятся на два типа: такие, что $\varepsilon \delta$ является запрещенной для данной стратегии π при любом $\varepsilon > 0$, и такие, что $\varepsilon_0 \delta$ является допустимой для нее при некотором $\varepsilon_0 > 0$. Вся

кая запрещенная вариация, в силу леммы 1, удовлетворяет какому-либо неравенству $-\delta_\ell(\zeta_{t-1}) > 0$, где $\pi_\ell(\zeta_{t-1}) = 0$. С учетом предыдущего замечания, все неравенства вида $-\delta_K(\zeta_{t-1}) > 0$ заменятся на линейные однородные неравенства от величин $\{\delta_\ell(\zeta_{t-1})\}$, $\ell = 1, \dots, K-1$.

Пусть $\pi(\Theta)$ — минимаксная стратегия, а вариация δ такова, что $\varepsilon_0 \delta$ допустима для нее при некотором $\varepsilon_0 > 0$. Из теоремы 1 следует, что для δ существует такой параметр $\theta(\delta) \in \Theta(\pi)$, что выполнено неравенство

$$L_T(\delta, \theta(\delta)) \geq 0. \quad (12)$$

Соответствующее множество C_1 является замкнутым. Множество запрещенных вариаций заменим его замыканием:

$$-\delta_\ell(\zeta_{t-1}) \geq 0, \quad (13)$$

при $\pi_\ell(\zeta_{t-1}) = 0$. Соответствующее множество C_2 конечно и, следовательно, тоже замкнуто. Очевидно, объединенная система неравенств (12), (13), для которой $C = C_1 \cup C_2$ — тоже замкнутое множество, удовлетворяет условиям леммы 2. Поэтому из системы (12), (13) может быть выделена конечная подсистема линейных однородных неравенств

$$L_T(\delta, \theta_j^0) \geq 0, \quad i = 1, \dots, m, \quad (14)$$

$$-\delta_{\ell_i}(\zeta_{t-1}) \geq 0, \quad i = m+1, \dots, n, \quad (15)$$

где $n \leq n(K, T)$. Поскольку неравенства (15) описывают только множество вариаций $\bar{\Delta}^p$, то для любого $\delta \in \Delta \setminus \bar{\Delta}^p$ выполнится хотя бы одно из неравенств (14). В силу замкнутости C_1 для любого $\delta \in \Delta \setminus \bar{\Delta}^p = \Delta^a$ также выполнится хотя бы одно из неравенств (14). Поэтому в силу теоремы 1 рассматриваемая стратегия является минимаксной на множестве $(\theta_1^0, \dots, \theta_m^0)$. Очевидно, в этом случае выполняется равенство (11). Теорема доказана.

5. Заключение

Полученные результаты позволяют свести задачу о поиске минимаксных стратегии и риска в бинарной стационарной случайной среде к задаче о поиске глобального максимума функции, численно равной минимаксному риску на конечном подмножестве параметров. Методы нахождения минимаксных стратегии и риска на конечном множестве параметров рассмотрены в [1, 2].

Работа выполнена при финансовой поддержке РФФИ, проект № 13-01-00334а.

1. Колногоров А.В. О минимаксном подходе к оптимальному целесообразному поведению в стационарных средах на конечном времени // Изв. АН СССР. Техническая кибернетика. 1988. №6. С.143-146.
2. Колногоров А.В. Алгоритм вычисления минимаксных стратегии и риска для бинарных стационарных случайных сред // Четвертая международная конференция по проблемам управления (26–30 января 2009 г.): Сб. тр. М.: Ин-т проблем управления им. В.А.Трапезникова РАН, 2009. С. 273-299.
3. Fabius J., van Zwet W.R. Some Remarks on the Two-Armed Bandit // Ann. Math. Statist. 1970. V.41. P.1906-1916.
4. Blumental L.M. Metric methods in linear inequalities // Duke Math. J. 1948. V.15. P.955-966.

References

1. Kolnogorov A.V. O minimaksnom podkhode k optimal'nomu tselesoobraznomu povedeniiu v statsionarnykh sredakh na konechnom vremeni [On minimax approach to the optimal rational behavior in stationary environments in finite amount of time]. Izv. AN SSSR. Tekhnicheskaja kibernetika, 1988, no. 6, pp. 143–146.
2. Kolnogorov A.V. Algoritm vychisleniia minimaksnikh strategii i riska dlia binarnykh stacionarnykh sluchainykh sred [An algorithm for calculating the minimax risk and minimax strategy in a binary stationary random environment]. Chetvertaia mezhdunarodnaia konferentsiia po problemam upravleniia (26 –30 ianvaria 2009 goda): Sbornik trudov [Proc. 4th Int. Conf. on the Control Problems]. Moscow, V.A.Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, 2009, pp. 273-299.
3. Fabius J., van Zwet W.R. Some Remarks on the Two-Armed Bandit. Ann. Math. Statist., 1970, vol. 41, pp. 1906–1916.
4. Blumental L.M. Metric methods in linear inequalities. Duke Math. J., 1948, vol. 15, pp. 955–966.