Робастное параллельное управление в случайной среде (задаче о двуруком бандите)

А.В.Колногоров[1]

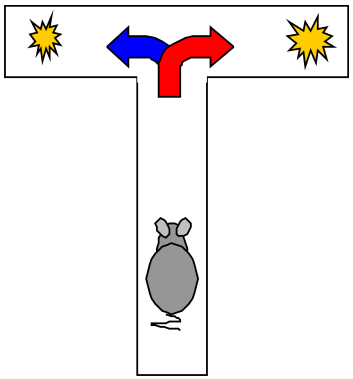
[1]Новгородский государственный университет им. Ярослава Мудрого
Alexander.Kolnogorov@novsu.ru

Большой семинар кафедры теории вероятностей МГУ

19 октября 2011 г.

Целесообразное поведение в случайной среде

1. Цетлин М.Л. Исследования по теории автоматов и моделированию биологических систем. М.: Наука, 1969
2. Tsetlin, M.L. Automation Theory and Modeling of Biological Systems. Academic Press, New York. 1973.



Животное (обычно, крыса) должно выбрать одно из 2-х направлений в Т-образном лабиринте. В конце лабиринта его ожидает удар тока с вероятностями q_1 и q_2.
Вероятности q_1, q_2 были фиксированы и крыса демонстрировала способность к обучению выбирать направление, которому соответствовала меньшая вероятность.

Формальное определение целесообразного поведения

Можно рассматривать случайный процесс реакций среды $\xi_1, \xi_2, \ldots, \xi_n$, зависящих только от текущих выбираемых вариантов (направлений) $y_1, y_2, \ldots, y_n$ следующим образом:

$$\Pr(\xi_n = 1 | y_n = \ell) = p_\ell, \quad \Pr(\xi_n = 0 | y_n = \ell) = q_\ell, \quad \ell = 1, 2.$$

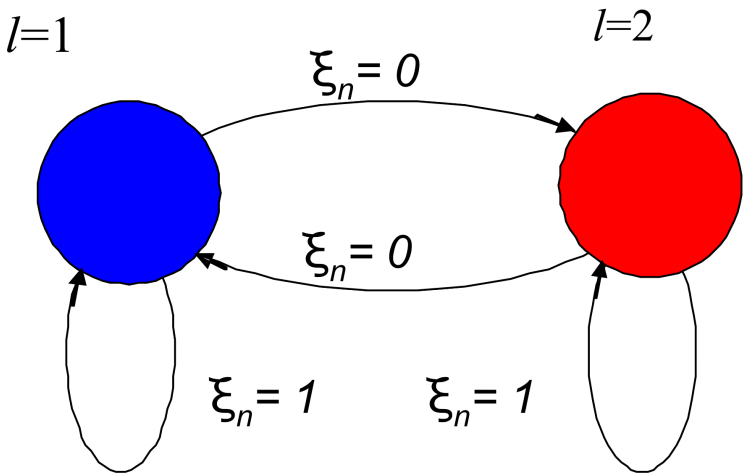
- $\xi_n = 1$ - нет удара током
- $\xi_n = 0$ - есть удар током
- $y_n = 1$ - поворот налево
- $y_n = 2$ - поворот направо

Вероятности p_1, p_2 фиксированы в процессе управления, но неизвестны. Значения процесса интерпретируются как доходы и цель состоит в максимизации среднего ожидаемого дохода. При этом *целесообразное поведение* обеспечивает во всех средах ожидаемый доход выше, чем при равновероятном применении обоих вариантов.

Простейший автомат

Возможная стратегия крысы в лабиринте:

- Изменить направление, если получен удар тока
- Не менять направление, если удар тока не получен

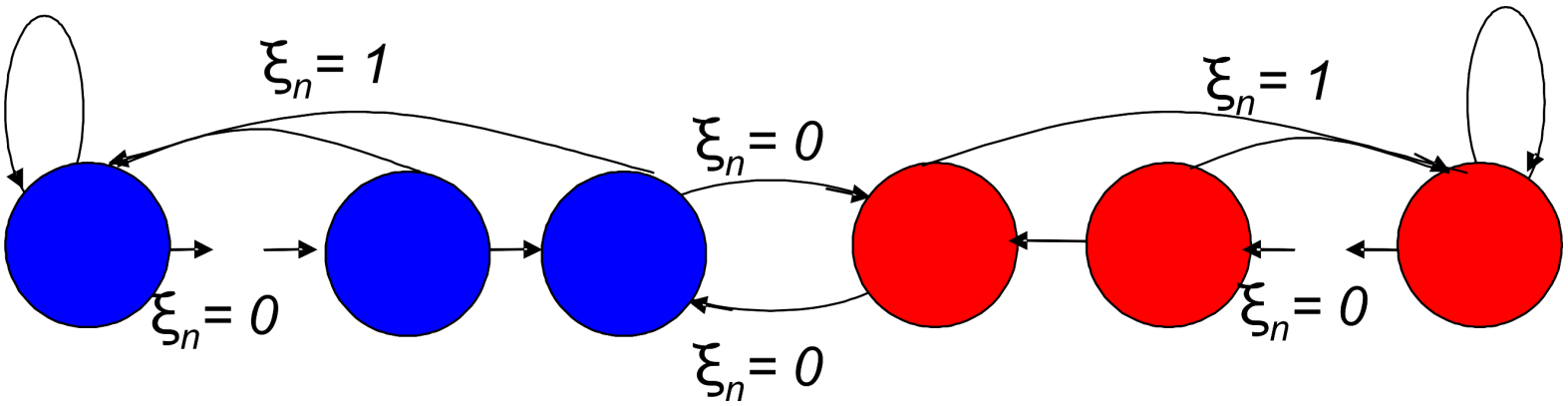


Стратегия может быть представлена симметричным автоматом с 2 состояниями.

Автомат Роббинса - Кринского

Более сложная автоматная стратегия:

- Перейти в самое глубокое состояние, если удар тока не получен
- Перейти в менее глубокое состояние или поменять направление (вариант) в наименее глубоком состоянии, если удар тока получен



Стохастические автоматы с переменной структурой

Варшавский В.И. Коллективное поведение автоматов. М.: Наука, 1973.

Состояние автомата в момент времени n - текущие вероятности выбора вариантов $\ell = 1$ и $\ell = 2$. Их можно представить вектором $\pi(n) = (\pi_1(n), \pi_2(n))$. Такой автомат имеет бесконечную память. Переходы описываются правилом

$$\pi(n+1) = R_n(\pi(n), y_n, \xi_n).$$

Если за выбор текущего варианта $\ell(n)$ получено поощрение ($\xi_n = 1$), вероятность выбора его на следующем шаге $\pi_\ell(n+1)$ растет, если штраф ($\xi_n = 0$) — то уменьшается.

Идентификационный подход

1. Срагович В.Г. Теория адаптивных систем. М.: Наука. 1976
2. Срагович В.Г. Адаптивное управление. М.: Наука, 1981.
3. Sragovich, V.G. Mathematical Theory of Adaptive Control // Interdisciplinary Mathematical Sciences – Vol. 4. World Scientific. New Jersey, London, . . . 2006.

В процессе управления делаются оценки параметров среды $\hat{\theta}_n = (\hat{p}_{1n}, \hat{p}_{2n})$. Пересчет вектора состояния осуществляется по правилу:

$$\pi(n+1) = Q_n(\pi(n), \hat{\theta}_n).$$

Возможная стратегия. Применить варианты по очереди и найти $\hat{\theta}_2$. Затем при $n = 3, 4, \ldots$ повторять процедуру: применять вариант, соответствующий текущей большей оценке $\hat{p}_{\ell n}$ с вероятностью $1 - \delta_n$, текущей меньшей оценке $\hat{p}_{\ell n}$ – с вероятностью δ_n, после чего пересчитывать $\hat{\theta}_{n+1}$. Для оптимальности нужно, чтобы

$$\lim_{n\to\infty} \delta_n = 0, \quad \sum_{n=1}^{\infty} \delta_n = \infty.$$

Цели управления

Должны выполняться для всех сред из некоторого класса:

$$\lim_{n\to\infty} E\xi_n \geq (p_1 \vee p_2) - \varepsilon, \quad \lim_{n\to\infty} N^{-1}\sum_{n=1}^{N} E\xi_n \geq (p_1 \vee p_2) - \varepsilon,$$

$$\Pr\left(\lim_{n\to\infty} N^{-1}\sum_{n=1}^{N} \xi_n \geq (p_1 \vee p_2) - \varepsilon\right) = 1$$

– ε-оптимальность в слабом и сильном смыслах,

$$\lim_{n\to\infty} E\xi_n = p_1 \vee p_2, \quad \lim_{n\to\infty} N^{-1}\sum_{n=1}^{N} E\xi_n = p_1 \vee p_2,$$

$$\Pr\left(\lim_{n\to\infty} N^{-1}\sum_{n=1}^{N} \xi_n = p_1 \vee p_2\right) = 1$$

– асимптотическая оптимальность в слабом и сильном смыслах.

Рекуррентные алгоритмы

1. Назин А.В., Позняк А.С. Адаптивный выбор вариантов. М.: Наука, 1986.
2. Poznyak, A.S. and Najim, K. Learning Automata and Stochastic Optimization. Lecture Notes in Control and Information Sciences 225. Springer-Verlag. Berlin, Heidelberg, New York. 1997.

Проанализированы известные и предложены новые алгоритмы типа САПС. Для анализа использовался главным образом метод стохастической аппроксимации.
Наряду с асимптотической оптимальностью алгоритмов исследовалась гарантированная скорость сходимости в среднеквадратическом

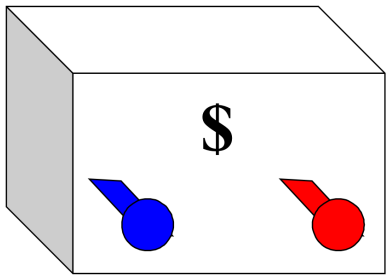
$$E\left((p_1 \vee p_2) - N^{-1}\sum_{n=1}^{N}\xi_n\right)^2 .$$

Последовательное управление по неполным данным

1. Пресман Э. Л., Сонин И.М. Последовательное управление по неполным данным. – М.: Наука, 1982.
2. Presman, E.L. and Sonin, I.M. Sequential Control with Incomplete Information. Academic Press. New York. 1990.

Рассмотрена байесовская постановка задачи управления с конечным множеством параметров в дискретном и непрерывном времени. В дискретном времени управляемый процесс бинарный, в непрерывном – пуассоновский. Целью является максимизация ожидаемого полного дохода.
В непрерывном времени предложена схема одновременного применения вариантов (с разделением ресурса). Решена задача синтеза оптимального управления на конечном и бесконечном времени.

Задача о двуруком бандите



Это игральный автомат с двумя рукоятками. При нажатии ℓ-ой рукоятки доход игрока равен 1 с вероятностью p_ℓ и 0 с вероятностью q_ℓ $(p_\ell + q_\ell = 1,\ \ell = 1, 2)$.

Игрок может нажать рукоятки в общей сложности N раз. Его целью является максимизация математического ожидания полного дохода. Вероятности p_1, p_2 фиксированы в процессе управления, но неизвестны игроку.

Дилемма "Информация или управление"

Для игрока оптимальной стратегией было бы всегда выбирать ту рукоятку, которой соответствует максимальное значение вероятностей p_1, p_2. Но чтобы определить эту рукоятку, он должен протестировать их обе, и это ведет к уменьшению его полного выигрыша.

Байесовский подход - 1

Формально доходы рассматриваются как случайный процесс $\xi_1, \xi_2, \ldots, \xi_n$, зависящий только от текущих выбираемых вариантов $y_1, y_2, \ldots, y_n$, именно $\Pr(\xi_n = 1 | y_n = \ell) = p_\ell$, $\Pr(\xi_n = 0 | y_n = \ell) = q_\ell$, $\ell = 1, 2$. Стратегия σ определяет выбор вариантов y_n, $n = 1, \ldots, N$ и может использовать всю текущую информацию о процессе: n_1, n_2 – количества нажатий и m_1, m_2 – полные выигрыши на обеих рукоятках. Функция потерь такова

$$L_N(\sigma, \theta) = E_{\sigma,\theta} \left(\sum_{n=1}^{N} ((p_1 \vee p_2) - \xi_n) \right),$$

где $\theta = (p_1, p_2)$ -– параметр процесса. Пусть $\Lambda(d\theta)$ есть априорное распределение на множестве параметров Θ. Байесовский риск равен

$$R_N^B(\Lambda) = \inf_{\{\sigma\}} \int_\Theta L_N(\sigma, \theta) \Lambda(d\theta),$$

соответствующая оптимальная стратегия σ^B называется байесовской.

Байесовский подход - 2

Berry, D.A. and Fristedt, B. Bandit Problems: Sequential Allocation of Experiments. Chapman and Hall. London, New York.,1985.

Известен простой рекуррентный алгоритм определения байесовских стратегии и риска. Как пишут Берри и Фристедт: "...it is not that researchers in bandit problems tend to "Bayesians"; rather Bayes's theorem provides a convenient mathematical formalism that allows for adaptive learning and so is an ideal tool in sequential decision problems". Для вычисления риска надо последовательно решать уравнение

$$R_n^B(\Lambda) = \min_{\ell=1,2} E_\Lambda \left((p_1 \vee p_2 - x) + R_{n-1}^B(\Lambda(y_1 = \ell, \xi_1 = x) | y_1 = \ell) \right),$$

при этом байесовская стратегия – та, которая обеспечивает решение этого уравнения. Цветом выделены минимальные полные потери при условии выбора на 1-ом шаге ℓ-ого варианта.

Адаптивное обучение и байесовский формализм

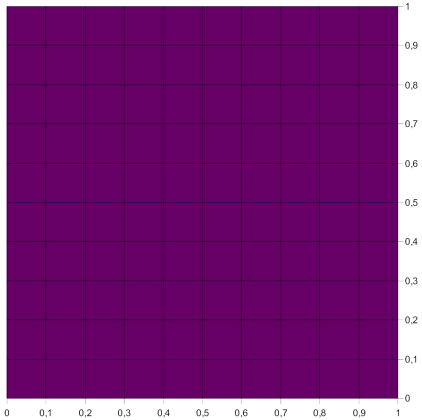
Адаптация = Идентификация + Управление

$$R_n^B(\Lambda) = \min_{\ell=1,2} E_\Lambda \left((p_1 \vee p_2 - x) + R_{n-1}^B(\Lambda(y_1 = \ell, \xi_1 = x) | y_1 = \ell) \right).$$

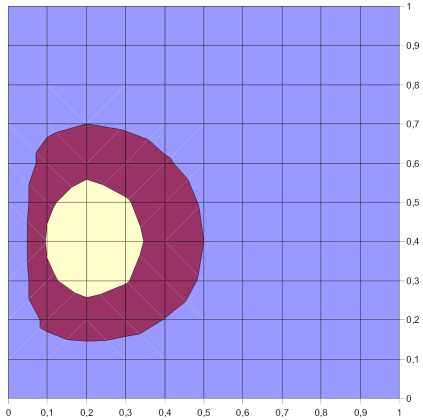
Цветом выделены части уравнения, обеспечивающие идентификацию и управление.

Априорное и апостериорное распределения

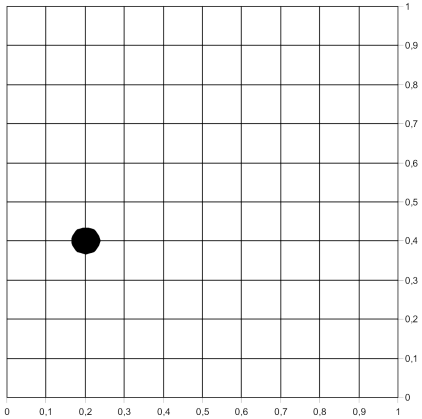
Равномерное априорное распределение



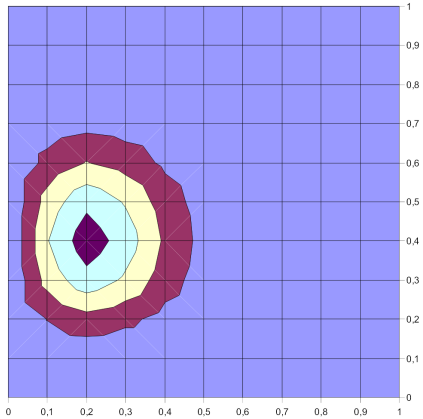
Апостериорные распределения



Фактическое значение параметра



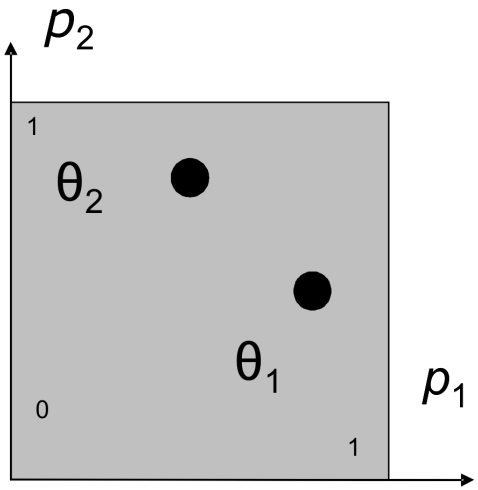
$m_1 = 1,\ n_1 = 5,\ m_2 = 2,\ n_2 = 5$



$m_1 = 2,\ n_1 = 10,\ m_2 = 4,\ n_2 = 10$

Близорукая стратегия Фельдмана

Feldman, D. Contributions to the "Two–Armed Bandit" Problem. Ann. Math. Stat. 1962. V. 33. P. 847–856.

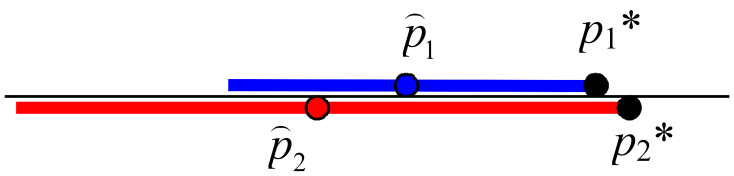


В этом случае $\Theta = \{\theta_1, \theta_2\}$, $\theta_1 = (p_1, p_2)$, $\theta_2 = (p_2, p_1)$, $p_1 > p_2$, $\lambda(\theta_1) = \lambda_1$, $\lambda(\theta_2) = \lambda_2$, $\lambda_1 + \lambda_2 = 1$.

- Выбрать 1-ый вариант, если больше текущая апостериорная вероятность θ_1
- Выбрать 2-ой вариант, если больше текущая апостериорная вероятность θ_2

Асимптотическая байесовская теорема

1. Lai, T.L. and Robbins, H. Asymptotically Efficient Adaptive Allocation Rules. Advances in Applied Mathematics, 1985, V. 6, P. 4-22.
2. Lai, T.L. Adaptive treatment allocation and the multi-armed bandit problem. The Annals of Statist., 1987, V. 25, P.1091-1114.



Предложены стратегии, выбирающие на каждом шаге вариант, которому соответствует бóльшее значение верхней границы доверительного интервала оценки параметра. Ширина интервала тем больше, чем меньше применялся данный вариант.

При широких предположениях установлены оценки (при $N \to \infty$)

$$L_N(\sigma^B, \theta) \propto \ln N, \quad R_N^B(\Lambda) \propto \ln^2 N.$$

Указаны точные значения множителей, которые зависят от информационного числа Куллбака-Лайблера (Kullback-Leibler).

A problem of two populations

Robbins, H. Some aspects of the sequential design of experiments. Bulletin of Amer. Math. Soc., 1952, V. 58, P.527-535.

"... In what follows we shall discuss a few simple problems in sequential design which are now under investigation and which are different from those usually met with in statistical literature. Optimum solutions to these problems are not known. Still, it is often better to have reasonably good solutions of the proper problems than optimum solutions of the wrong problems. In the present state of statistical theory this principle applies with particular force to problems in sequential design."

"... It would be interesting to know the value of $\phi(N)$[1] and the explicit description of any "minimax" rule R for which the value $\phi(N)$ is attained."

[1] $\phi(N) = N^{-1} R_N^M(\Theta)$, $R_N^M(\Theta)$ – minimax risk.

Минимаксные риск и стратегия

1. Vogel, W. An asymptotic minimax theorem for the two-armed bandit problem. Ann. Math. Stat., 1961, V. 31, P.444-451
2. Fabius, J., and van Zwet, W.R. Some remarks on the two-armed bandit. Ann. Math. Stat., 1970, V. 41, 1906 -1916.

Минимаксный риск равен:

$$R_N^M(\Theta) = \inf_{\{\sigma\}} \sup_{\Theta} L_N(\sigma, \theta),$$

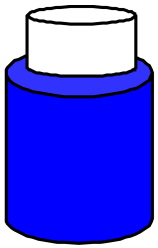
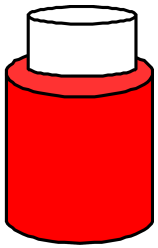
соответствующая оптимальная стратегия σ^M называется минимаксной[2]. Прямое нахождение минимаксных стратегии и риска практически невозможны. Как пишут Фабиус и Ван Цвет: “the algebra involved becomes progressively more complicated with increasing N and seems to remain prohibitive already for N as small as 5”. Однако известна асимптотическая минимаксная теорема Фогеля, гласящая, что при $N \to \infty$, $D = 0,25$:

$$0,530 \leq (DN)^{-1/2} R_N^M(\Theta) \leq 0,752.$$

[2] $\phi(N) = N^{-1} R_N^M(\Theta)$

Двухэтапный подход - 1

1. Cheng, Y. (1994). Multistage decision problems. Sequential Analysis. V. 13, 329-350.
2. Witmer, J.A. (1986). Bayesian multistage decision problems. Ann. Stat. V. 14, 283-297.

Пусть имеется очень большая группа N пациентов и 2 альтернативных лекарства с различными и неизвестными эффективностями.

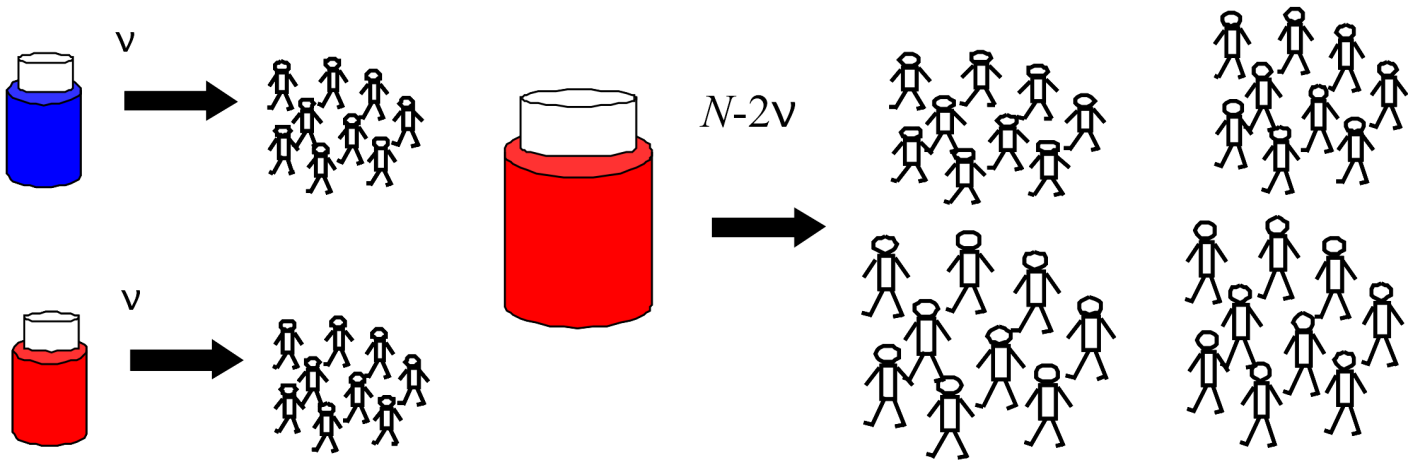
Эти лекарства могут рассматриваться как варианты, причем p_ℓ, q_ℓ -– вероятности успешного и неуспешного лечения, $\ell = 1, 2$. Значения процесса определяются так: $\xi_n = 1$, если пациент номер n поправился и $\xi_n = 0$, если нет.

Варианты в данном случае нельзя применять последовательно один за другим, так как лечение пациента требует значительного времени. В этом случае требуется использовать *параллельную обработку*!

Двухэтапный подход - 2

Колногоров А.В. Об оптимальном априорном времени обучения в задаче о “двуруком бандите” // Пробл. передачи информ. 2000. т. 36. № 4. С. 117-127.

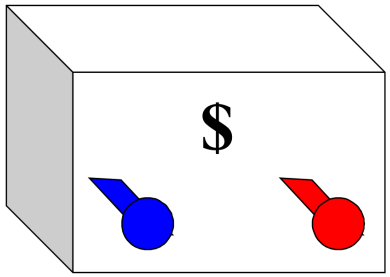
На 1-ом этапе оба лекарства даются равным достаточно большим группам ν пациентов. В конце 1-ого этапа подсчитываются количества выздоровевших пациентов в обеих группах. Затем более эффективное лекарство дается оставшимся $N - 2\nu$ пациентам на 2-ом этапе. При $N \to \infty$ оптимально выбирать $\nu \propto N^{2/3}$.



Параллельное управление

Параллельное управление

Задача о двуруком бандите с нормально распределенными доходами



Это игральный автомат с двумя рукоятками. При нажатии ℓ-ой рукоятки доход игрока имеет нормальное распределение с единичной дисперсией и математическим ожиданием m_ℓ.

Игрок может нажать рукоятки в общей сложности N раз. Его целью является максимизация (в некотором смысле) математического ожидания полного дохода. Математические ожидания m_1, m_2 фиксированы в процессе управления, но неизвестны игроку.

Дилемма "Информация или управление"

Для игрока оптимальной стратегией было бы всегда выбирать ту рукоятку, которой соответствует максимальное значение математических ожиданий m_1, m_2. Но чтобы определить эту рукоятку, он должен протестировать обе, и это ведет к уменьшению его полного выигрыша.

Формальная постановка задачи

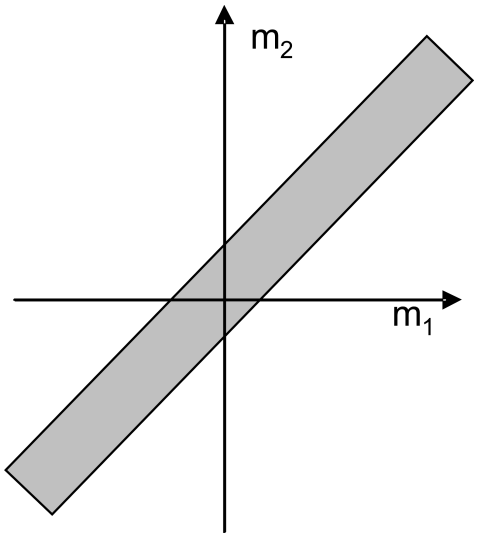
Формально выигрыши можно рассматривать как управляемый случайный процесс $\xi_1, \xi_2, \ldots, \xi_n$, значения которого зависят только от выбираемых вариантов $y_1, y_2, \ldots, y_n$ и имеют нормальную плотность распределения с единичной дисперсией и математическим ожиданием m_ℓ, если выбран вариант ℓ

$$f(x|m_\ell) = (2\pi)^{-1/2} \exp\left\{-(x - m_\ell)^2/2\right\},$$

при этом процесс полностью характеризуется векторным параметром $\theta = (m_1, m_2)$. Стратегия управления σ определяет выбор вариантов y_n, $n = 1, \ldots, N$ и в общем случае может использовать всю предысторию процесса $y_1, \xi_1, \ldots, y_{n-1}, \xi_{n-1}$. Достаточно знать 4 текущие величины: n_1, n_2 – количества выборов обоих вариантов и X_1, X_2 -– полные доходы за их выбор. Функция потерь определяется следующим образом

$$L_N(\sigma, \theta) = E_{\sigma,\theta}\left(\sum_{n=1}^{N}((m_1 \vee m_2) - \xi_n)\right).$$

Минимаксная постановка задачи



Будем предполагать, что множество параметров удовлетворяет ограничению $\Theta = \{(m_1, m_2) : |m_1 - m_2| \le 2c_1, |m_1 + m_2| \le 2c_2\}$, где $0 < c_1 < \infty$, $0 < c_2 < \infty$. Минимаксный риск определяется как

$$R_N^M(\Theta) = \inf_{\{\sigma\}} \sup_{\Theta} L_N(\sigma, \theta),$$

обеспечивающая его стратегия σ^M называется минимаксной стратегией.

Робастность минимаксного подхода

При применении стратегии σ^M следующее неравенство выполнено на всем множестве Θ:

$$L_N(\sigma^M, \theta) \le R_N^M(\Theta).$$

Асимптотическая минимаксная теорема Фогеля (W.Vogel)

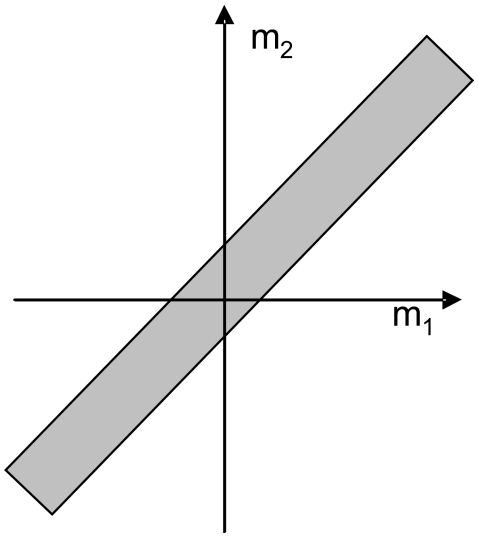
При $N \to \infty$ выполнено неравенство:

$$0,530 \leq N^{-1/2} R_N^M(\Theta) \leq 0,752.$$

Пороговая стратегия

Оценка сверху обеспечивается следующей стратегией.
Следует применять варианты по очереди до тех пор, пока абсолютная разность полных доходов за их применение не превысит величины $\alpha N^{1/2}$ или не истечет время управления. Если порог превышен, а время управления не истекло, то далее следует применять только вариант, соответствующий большему значению дохода на начальном этапе. Оценке сверху соответствуют $\alpha \approx 0,584$, $|m_1 - m_2| \approx 0.37 N^{-1/2}$.

Байесовская постановка задачи



Рассмотрим в этой области априорное распределение с плотностью $\lambda(m_1, m_2)$. Байесовский риск определяется как

$$R_N^B(\lambda) = \inf_{\{\sigma\}} \int_\Theta L_N(\sigma, \theta)\lambda(\theta)d\theta.$$

Апостериорная плотность распределения

$$\lambda(m_1, m_2|X_1, n_1, X_2, n_2) =$$

$$= \frac{f_{n_1}(X_1|n_1 m_1) f_{n_2}(X_2|n_2 m_2)\lambda(m_1, m_2)}{\iint\limits_\Theta f_{n_1}(X_1|n_1 m_1) f_{n_2}(X_2|n_2 m_2)\lambda(m_1, m_2)dm_1, dm_2},$$

где $f_D(x|M) = (2\pi D)^{-1/2} \exp\left\{-(x-M)^2/(2D)\right\}$, причем $f_n(X|nm) = 1$ при $n = 0$.

Уравнения для вычисления байесовского риска

$$R^B_{N-n}(\cdot) = \min(R^{(1)}_{N-n}(\cdot), R^{(2)}_{N-n}(\cdot)),$$

$$R^{(1)}_0(\cdot) = R^{(2)}_0(\cdot) = 0,$$

$$R^{(1)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = \iint_\Theta \Big((m_2 - m_1)^+ +$$

$$+E^{(1)}_x R^B_{N-n-1)}(\lambda; X_1 + x, n_1 + 1, X_2, n_2)\Big)\, \lambda(m_1, m_2 | X_1, n_1, X_2, n_2) dm_1 dm_2,$$

$$R^{(2)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = \iint_\Theta \Big((m_1 - m_2)^+ +$$

$$+E^{(2)}_x R^B_{N-n-1}(\lambda; X_1, n_1, X_2 + x, n_2 + 1)\Big)\, \lambda(m_1, m_2 | X_1, n_1, X_2, n_2) dm_1 dm_2,$$

при $n < N$, где $n = n_1 + n_2$,

$$E^{(\ell)}_x R(x) = \int_{-\infty}^{+\infty} R(x) f(x|m_\ell) dx, \quad \ell = 1, 2.$$

Основная теорема теории игр

При широких предположениях минимаксный риск совпадает с байесовским, вычисленным для наихудшего априорного распределения, то есть справедливо равенство:

$$R_N^M(\Theta) = \sup_{\{\Lambda\}} R_N^B(\Lambda) = R_N^B(\Lambda^0).$$

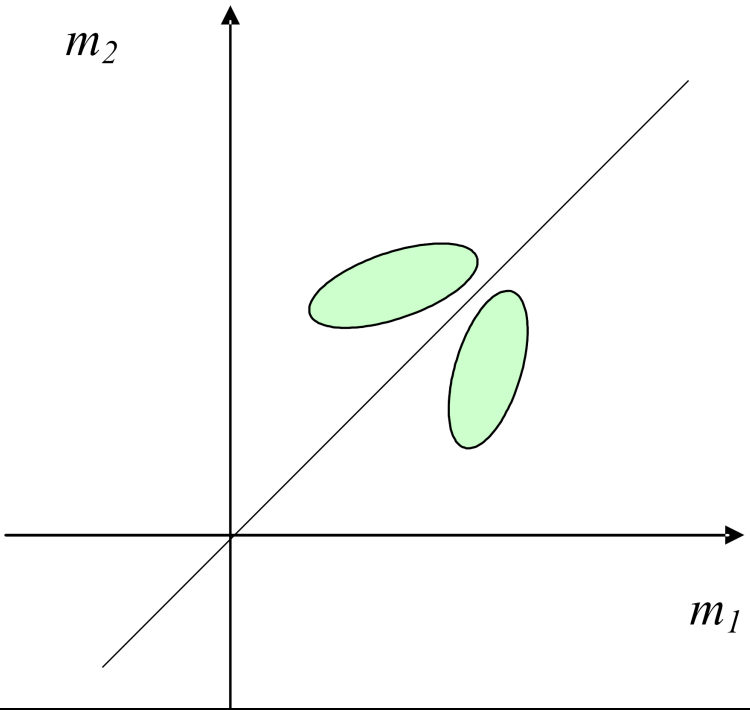
Минимаксная стратегия совпадает с некоторой байесовской, вычисленной для наихудшего априорного распределения.

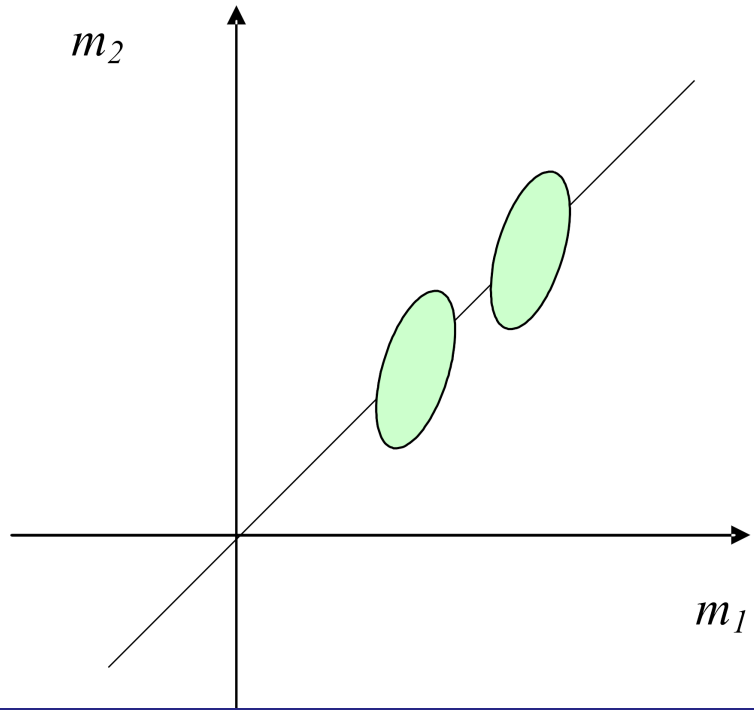
Свойства априорного распределения

Следующие преобразования $\tilde{\lambda}$ априорной плотности распределения λ не меняют байесовский риск, то есть $R_N^B(\tilde{\lambda}) = R_N^B(\lambda)$:

1. $\tilde{\lambda}^{(1)}(m_1, m_2) = \lambda(m_2, m_1)$ (для всех m_1, m_2),
2. $\tilde{\lambda}^{(2)}(m_1, m_2) = \lambda(m_1 + m, m_2 + m)$ (для всех m_1, m_2 и любого фиксированного m).

Для плотности $\tilde{\lambda}^{(2)}(m_1, m_2)$ правило выбора вариантов на первом шаге одинаково при любом m.



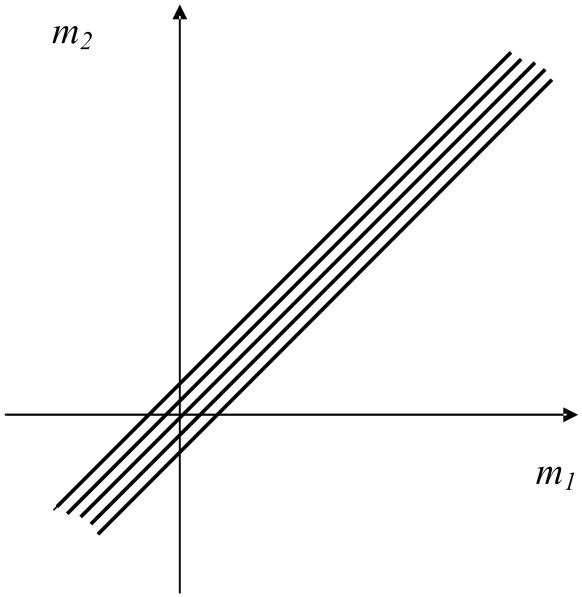


Асимптотически наихудшее априорное распределение

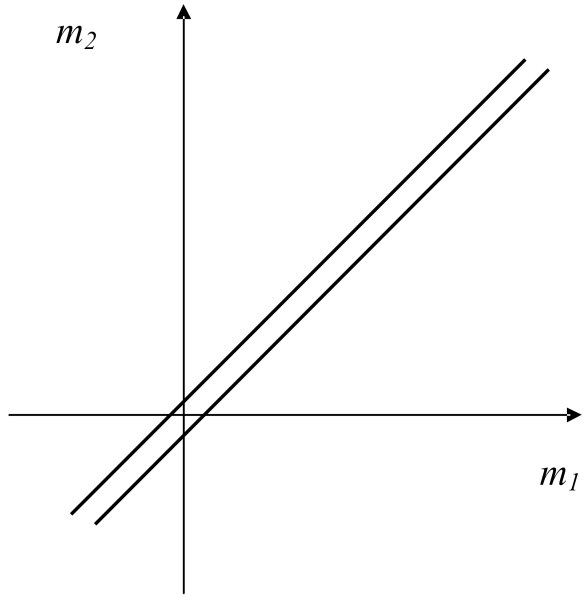
Может быть установлено с использованием свойства ***вогнутости байесовского риска***. Пусть λ_1, λ_2 — априорные плотности распределения, неотрицательные α_1, α_2 таковы, что $\alpha_1 + \alpha_2 = 1$. Тогда

$$R_N^B(\alpha_1\lambda_1 + \alpha_2\lambda_2) \geq \alpha_1 R_N^B(\lambda_1) + \alpha_2 R_N^B(\lambda_2).$$

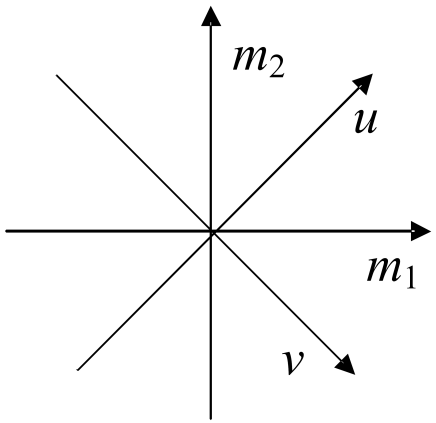
У асимптотически наихудшего распределения на линиях плотность постоянна:



Численная оптимизация позволяет предположить, что оно таково:



Рекуррентные уравнения - 1



Удобно поменять параметризацию $m_1 = u + v$, $m_2 = u - v$, тогда $\theta = (u + v, u - v)$. Наихудшее априорное распределение может быть взято в виде

$$\nu_a(u, v) = \kappa_a(u)\rho(v),$$

где $\kappa_a(u)$ – постоянная плотность на отрезке $|u| \le a$, а $\rho(-v) = \rho(v)$ при $|v| \le c_1$.

Для нахождения байесовского риска относительно наихудшего априорного распределения следует вычислять риски

$$R_{n_1,n_2}(Z) = \min(R^{(1)}_{n_1,n_2}(Z), R^{(2)}_{n_1,n_2}(Z)),$$

где

$$Z = X_1 n_2 - X_2 n_1 = n_1 n_2(\hat{m}_1 - \hat{m}_2), \quad \hat{m}_\ell = \frac{X_\ell}{n_\ell}, \quad \ell = 1, 2.$$

Рекуррентные уравнения - 2

$$R^{(1)}_{n_1,n_2}(Z) = R^{(2)}_{n_1,n_2}(Z) = 0$$

при $n_1 + n_2 = N$,

$$R^{(1)}_{n_1,n_2}(Z) = g^{(1)}_{n_1,n_2}(Z) + \frac{1}{n_2}\int_{-\infty}^{+\infty} R_{n_1+1,n_2}(Z+z)h_{n_1}\left(\frac{Z-n_1 z}{n_2}\right)dz,$$
$$R^{(2)}_{n_1,n_2}(Z) = g^{(2)}_{n_1,n_2}(Z) + \frac{1}{n_1}\int_{-\infty}^{+\infty} R_{n_1,n_2+1}(Z+z)h_{n_2}\left(\frac{Z-n_2 z}{n_1}\right)dz$$

при $n_1 + n_2 < N$, $n_1 \geq 1$, $n_2 \geq 1$,

$$g^{(\ell)}_{n_1,n_2}(Z) = \int_0^\infty 2v g_{n_1,n_2}(Z,(-1)^{\ell+1}v)\rho(v)dv, \quad \ell = 1,2,$$
$$g_{n_1,n_2}(Z,v) = \frac{1}{(2\pi n_1 n_2(n_1+n_2))^{1/2}}\exp\left(-\frac{(Z+2vn_1n_2)^2}{2n_1n_2(n_1+n_2)}\right),$$
$$h_n(z) = \left(\frac{n+1}{2\pi n}\right)^{1/2}\exp\left(-\frac{z^2}{2n(n+1)}\right).$$

Байесовский риск и оптимальная стратегия

Тогда байесовский риск вычисляется по формуле

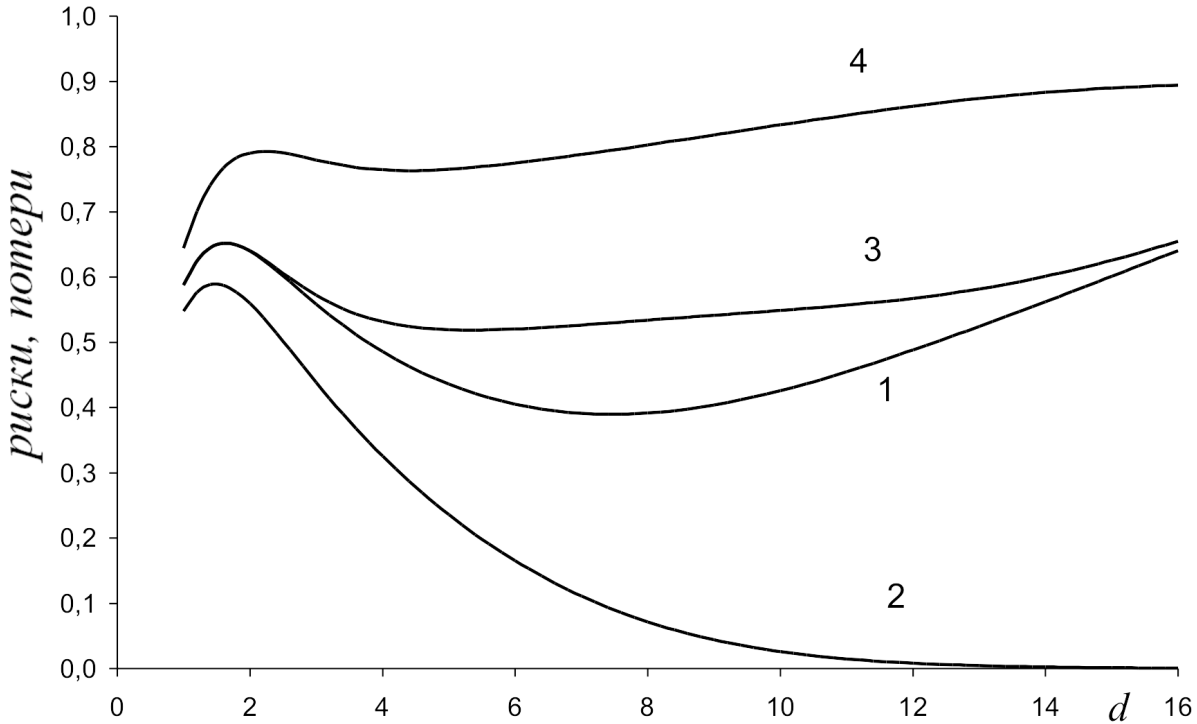
$$\lim_{a\to\infty} R_N^B(\nu_a(u,v)) = 4\int_0^\infty v\rho(v)dv + \int_{-\infty}^\infty R_{1,1}(z)dz.$$

Оптимальная стратегия

Оптимальная стратегия на первых двух шагах применяет варианты по очереди. Далее следует всегда выбирать вариант, которому соответствует меньшее значение из $R^{(1)}_{n_1,n_2}(Z)$, $R^{(2)}_{n_1,n_2}(Z)$.

Нахождение минимаксного риска

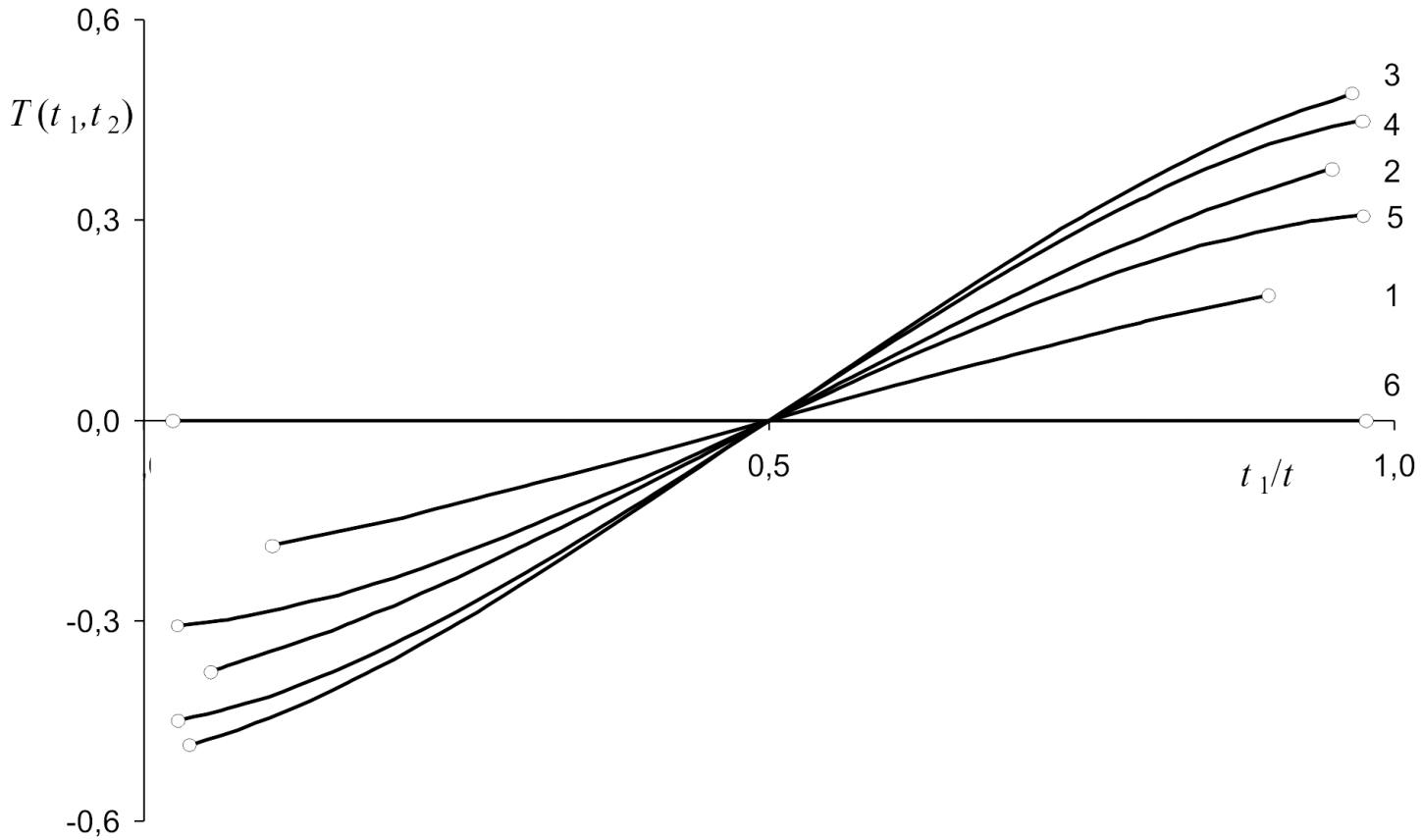
Предположим, что $\rho(v)$ сосредоточена в точках $v = \pm dN^{-1/2}$ с вероятностями 1/2. Тогда d соответствует максимуму приведенного байесовского риска. Этот максимум при $N = 50$ приблизительно равен 0,65 при $d \approx 1,7$. Для подтверждения предположения вычислялись приведенные потери и их значение оказалось не больше максимума приведенного байесовского риска.



1. Риски
2. Риски на последних $N - 2$ этапах
3. Потери
4. Потери за стратегию Фогеля

Нахождение минимаксной стратегии

При $n_1 \geq 1$, $n_2 \geq 1$ оптимальная стратегия предписывает выбирать 1-ый вариант, если $Z > N^{3/2}T(t_1, t_2)$ и 2-ой вариант, если $Z < N^{3/2}T(t_1, t_2)$, где $t_1 = n_1/N$, $t_2 = n_2/N$, $t = n/N$. При $Z = N^{3/2}T(t_1, t_2)$ выбор может быть произволен.



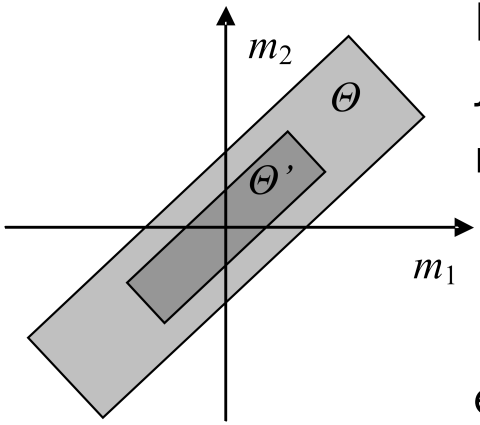
1. $t = 0,2$
2. $t = 0,4$
3. $t = 0,6$
4. $t = 0,8$
5. $t = 0,9$
6. $t = 0,98$

Параллельная обработка - 1

Стратегия может быть использована при N кратных рассмотренным. Пусть надо обработать $T = NK$ данных. Будем применять один и тот же вариант в моменты времени $t = (n-1)K+1, \ldots, nK$, а затем определим доход

$$\xi'_n = K^{-1/2} \sum_{t=(n-1)K+1}^{nK} \xi_t, \quad n = 1, \ldots, N,$$

$$D(\xi'_n | y_n = \ell) = 1, E(\xi'_n | y_n = \ell) = m_\ell, \text{ если } E(\xi_n | y_n = \ell) = K^{-1/2} m_\ell.$$

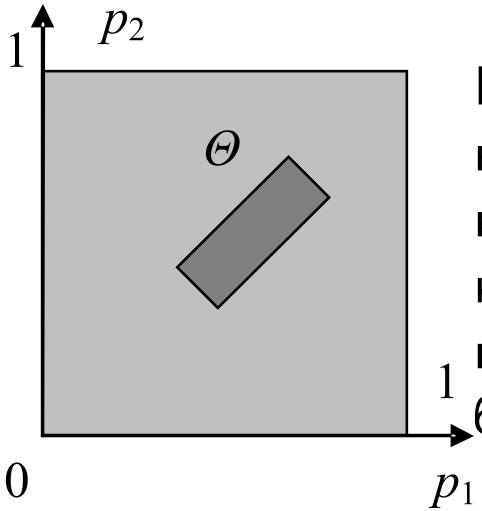


Потребуем, чтобы новая стратегия σ' так же управляла процессом ξ'_n как стратегия σ управляет процессом ξ_n. Тогда приведенные потери равны, т.е.

$$(NK)^{-1/2} L_{NK}(\sigma', \theta') = N^{-1/2} L_N(\sigma, \theta),$$

если $\theta' = (m_1 K^{-1/2}, m_2 K^{-1/2})$ и $\theta = (m_1, m_2)$.

Параллельная обработка - 2



В соответствии с ЦПТ ξ'_n могут иметь близкие к нормальным распределения, даже если распределения таковыми не являются. Пусть даны $T = 600$ пакетов данных, которые могут быть обработаны двумя альтернативными способами. Обработка может быть успешной $(\xi_t = 1)$ или неуспешной $(\xi_t = 0)$.

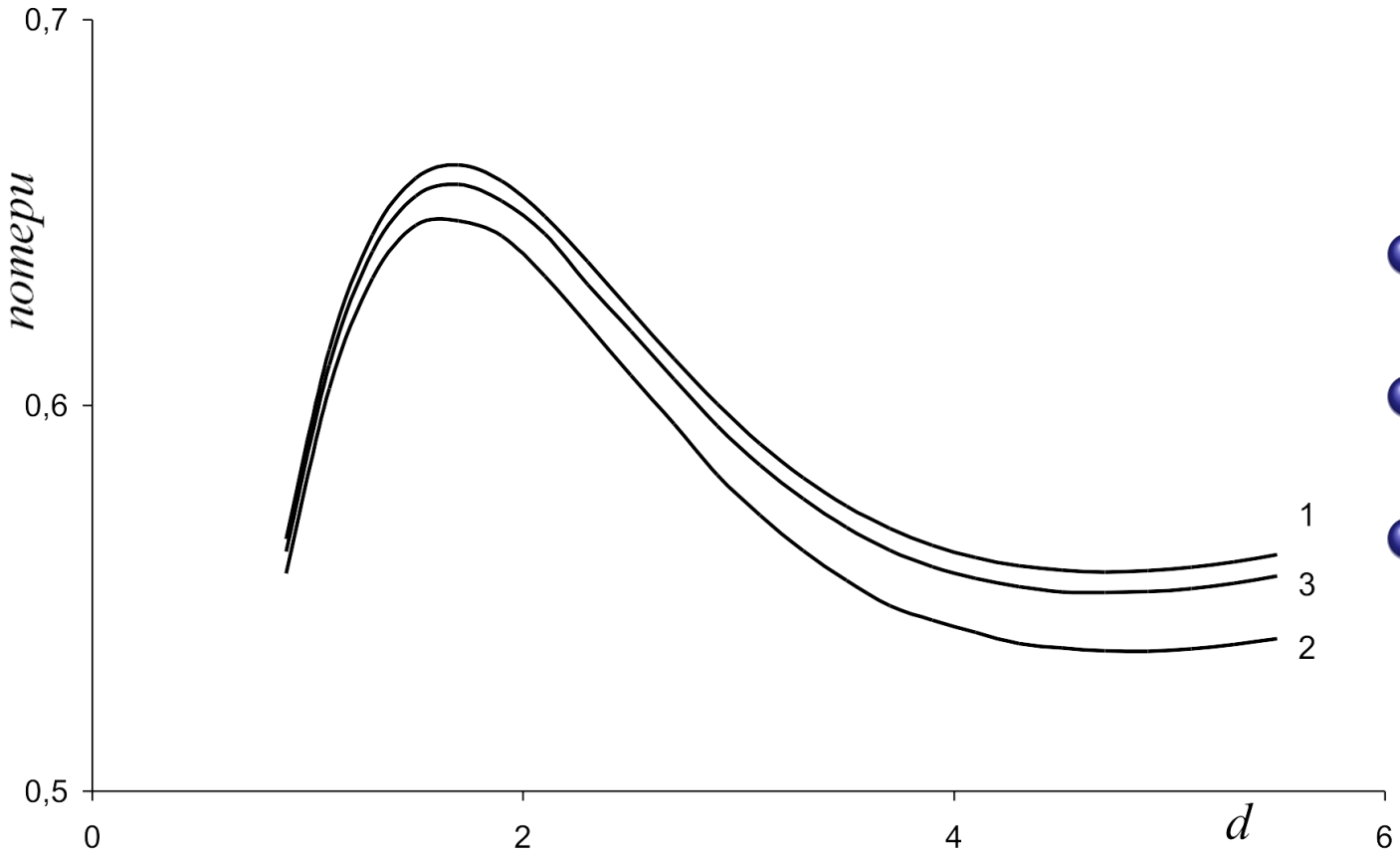
Вероятности успешной и неуспешной обработки зависят только от выбранных методов (вариантов) и равны p_ℓ и q_ℓ соответственно $(\ell = 1, 2)$. Пусть известно, что p_1, p_2 близки к $p = 0,5$. Разобьем все данные на $N = 30$ блоков по $K = 20$ каждый и определим процесс

$$\xi'_n = (DK)^{-1/2} \sum_{t=(n-1)K+1}^{nK} \xi_t, \quad n = 1, \ldots, N, \quad D = p(1-p) = 0,25.$$

Распределения ξ'_n близки к нормальным, а их дисперсии близки к 1.

Параллельная обработка - 3

$$l_T(d) = (DT)^{-1/2} E_{\sigma,\theta} \left(\sum_{t=1}^{T} ((p_1 \vee p_2) - \xi_t) \right), \; 2d = |p_1 - p_2|(T/D)^{1/2}.$$



1. Расчетная кривая потерь
2. Монте-Карло при $T = 600$
3. Монте-Карло при $T = 3000$

Оценки минимаксного риска

Оценки минимаксного риска

Кусочно-постоянные стратегии

Далее рассматриваются кусочно-постоянные стратегии. Вначале оба варианта применяются по M_0 раз, а затем смена вариантов разрешается только после их применения M раз подряд. При этом предполагается, что $N - 2M_0$ кратно M.

Вместо применения варианта M раз подряд можно осуществлять параллельную обработку M данных. Стратегия суммирует доходы, полученные при параллельной обработке, поэтому их распределения могут быть близки к нормальным и тогда, когда исходные распределения среды не были таковыми.

Положим $S = ZN^{-3/2}$, $s = zN^{-3/2}$, $t_1 = n_1N^{-1}$, $t_2 = n_2N^{-1}$, $w = vN^{1/2}$, $\varepsilon = MN^{-1}$, $\varepsilon_0 = M_0N^{-1}$, $\varrho(w) = N^{-1/2}\rho(v)$, $r_\varepsilon(S, t_1, t_2) = NR_{n_1,n_2}(Z)$, $r_\varepsilon^{(\ell)}(S, t_1, t_2) = NR_{n_1,n_2}^{(\ell)}(Z)$, $\ell = 1, 2$.

Инвариантные рекуррентные уравнения

Вычисление рисков выполняется рекуррентно "с конца", т.е. следует решать уравнение:

$$r_\varepsilon(S,t_1,t_2)=\min_{\ell=1,2} r_\varepsilon^{(\ell)}(S,t_1,t_2),$$

где $r_\varepsilon^{(1)}(S,t_1,t_2)=r_\varepsilon^{(2)}(S,t_1,t_2)=0$ при $t_1+t_2=1$,

$$r_\varepsilon^{(1)}(S,t_1,t_2)=\varepsilon g^{(1)}(S,t_1,t_2)+\frac{1}{t_2}\int\limits_{-\infty}^{+\infty} r_\varepsilon(S+s,t_1+\varepsilon,t_2)h_\varepsilon\left(\frac{S\varepsilon-t_1 s}{t_2},t_1\right)ds,$$

$$r_\varepsilon^{(2)}(S,t_1,t_2)=\varepsilon g^{(2)}(S,t_1,t_2)+\frac{1}{t_1}\int\limits_{-\infty}^{+\infty} r_\varepsilon(S+s,t_1,t_2+\varepsilon)h_\varepsilon\left(\frac{S\varepsilon-t_2 s}{t_1},t_2\right)ds$$

при $t_1+t_2<1$, $t_1\geq\varepsilon_0$ и $t_2\geq\varepsilon_0$.

$$g^{(\ell)}(S,t_1,t_2)=\int\limits_0^\infty 2wg(S,(-1)^{\ell+1}w,t_1,t_2)\varrho(w)dw,\quad \ell=1,2,$$

$$g(S,w,t_1,t_2)=(2\pi t_1t_2(t_1+t_2))^{-1/2}\exp\left(-\frac{(S+2wt_1t_2)^2}{2t_1t_2(t_1+t_2)}\right),$$

$$h_\varepsilon(s,t)=\left(\frac{t+\varepsilon}{2\pi t\varepsilon}\right)^{1/2}\exp\left(-\frac{s^2}{2t\varepsilon(t+\varepsilon)}\right).$$

Байесовские стратегия и риск

Оптимальная стратегия на первых двух шагах применяет варианты по очереди. Далее текущим оптимальным является ℓ-ый вариант, если меньшее значение имеет $r_\varepsilon^{(\ell)}(S, t_1, t_2)$, $\ell = 1, 2$.

Байесовский риск, соответствующий асимптотически наихудшему распределению, вычисляется по формуле

$$\lim_{a\to\infty} R_N^B(\nu_a(u, v)) = r_\varepsilon(\varrho, \varepsilon_0)N^{1/2},$$

где

$$r_\varepsilon(\varrho, \varepsilon_0) = 4\varepsilon_0 \int_0^\infty w\varrho(w)dw + \hat{r}_\varepsilon(\varrho, \varepsilon_0), \quad \hat{r}_\varepsilon(\varrho, \varepsilon_0) = \int_{-\infty}^\infty r_\varepsilon(s, \varepsilon_0, \varepsilon_0)ds.$$

Предельный переход

Зафиксируем $\varepsilon_0 > 0$ и устремим ε к нулю. Тогда при всех S и всех t_1, t_2, для которых определены решения уравнений, существуют пределы

$$r(S, t_1, t_2) = \lim_{\varepsilon \to +0} r_\varepsilon(S, t_1, t_2) = \lim_{\varepsilon \to +0} r_\varepsilon^{(\ell)}(S, t_1, t_2), \quad \ell = 1, 2,$$

удовлетворяющие условиям Липшица по всем переменным. Это позволяет доопределить $r(S, t_1, t_2)$ по непрерывности на все допустимые S, t_1, t_2. Также при всех ϱ существуют пределы

$$\hat{r}(\varrho, \varepsilon_0) = \lim_{\varepsilon \to +0} \hat{r}_\varepsilon(\varrho, \varepsilon_0), \quad r(\varrho, \varepsilon_0) = \lim_{\varepsilon \to +0} r_\varepsilon(\varrho, \varepsilon_0),$$

которые вычисляются по формулам

$$\hat{r}(\varrho, \varepsilon_0) = \int_{-\infty}^{\infty} r(s, \varepsilon_0, \varepsilon_0) ds, \quad r(\varrho, \varepsilon_0) = 4\varepsilon_0 \int_0^{\infty} w\varrho(w) dw + \hat{r}(\varrho, \varepsilon_0).$$

Оценки минимаксного риска

Для минимаксного риска на $\Theta = \{|m_1 - m_2| \le 2cN^{-1/2}\}$ при $N \to \infty$ справедливы асимптотические оценки

$$\sup_{\varrho} \hat{r}(\varrho, \varepsilon_0) \le N^{-1/2} R_N^M(\Theta) \le \sup_{\varrho} r(\varrho, \varepsilon_0).$$

Так как $r(\varrho, \varepsilon_0) \le r_\varepsilon(\varrho, \varepsilon_0)$, то на $\Theta = \{|m_1 - m_2| \le 2cN^{-1/2}\}$ справедлива оценка сверху

$$N^{-1/2} R_N^M(\Theta) \le \sup_{\varrho} r_\varepsilon(\varrho, \varepsilon_0).$$

В частности, при всех N кратных 50 на $\Theta = \{|m_1 - m_2| \le 32N^{-1/2}\}$ справедлива оценка

$$N^{-1/2} R_N^M(\Theta) \le 0,65.$$

Выводы

- Предложена робастная стратегия параллельного управления в случайной среде
- Стратегия позволяет осуществлять управление агрегированными данными в средах, распределения которых отличны от нормальных
- Стратегия имеет пороговый характер, определяется численными методами и легко табулируется
- Минимаксный риск ищется с помощью инвариантного рекуррентного уравнения как байесовский, соответствующий наихудшему априорному распределению
- Установлено существование непрерывного предела решения инвариантного рекуррентного уравнения, если относительная продолжительность применения одного и того же варианта ε стремится к нулю
- Улучшены асимптотические оценки минимаксного риска

Публикации результатов

1. Колногоров А.В. Асимптотические оценки байесовского риска для одного класса стационарных сред // Третья междунар. конф. по проблемам управления. Пленарные доклады и избранные тр. М.: ИПУ им. В.А.Трапезникова РАН, 2006. С. 241 - 248.
2. Колногоров А.В. Нахождение минимаксных стратегии и риска в случайной среде (задаче о двуруком бандите) // АиТ. 2011. № 5. С. 127–138.
3. Kolnogorov A.V. Determination of the Minimax Risk for the Normal Two-Armed Bandit // Proceedings of the IFAC Workshop “Adaptation and Learning in Control and Signal Processing ALCOSP 2010”, Antalya, Turkey, August 26–28, 2010. http://www.ifac-papersonline.net.
4. Колногоров А.В. Робастное параллельное управление в случайной среде (задаче о двуруком бандите) // Принята к печати в АиТ.

Спасибо за внимание

Спасибо за внимание