Задача о двуруком бандите в приложении к параллельной обработке данных

А.В.Колногоров[1]

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

Семинар «Стохастический анализ в задачах»

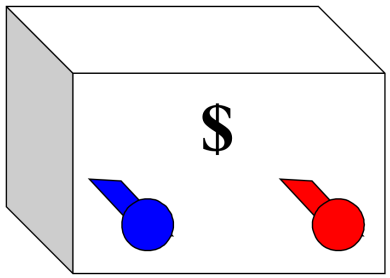
г. Москва

21 сентября 2013

r-этапное управление

r-этапное управление

Задача о двуруком бандите



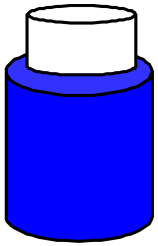
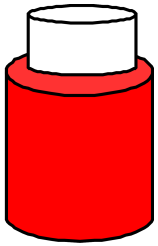
Это игральный автомат с двумя рукоятками. При нажатии ℓ-ой рукоятки доход игрока равен 1 с вероятностью p_ℓ и 0 с вероятностью q_ℓ $(p_\ell + q_\ell = 1,\ \ell = 1, 2)$.

Игрок может нажать рукоятки 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 p_1, p_2. Но чтобы определить эту рукоятку, он должен протестировать их обе, и это ведет к уменьшению его полного выигрыша.

Двухэтапный подход - 1

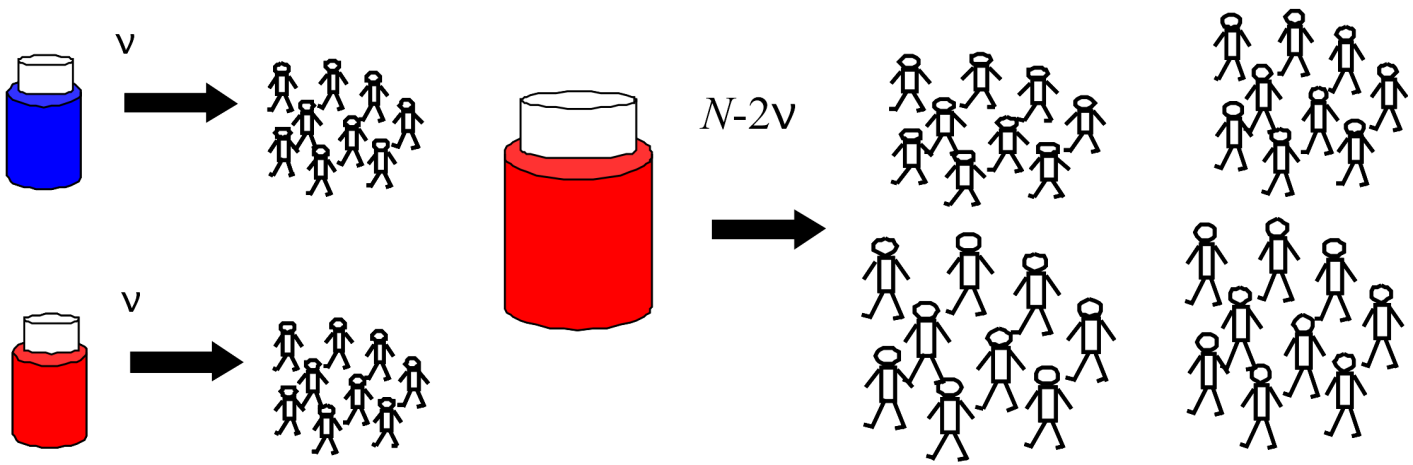
Пусть имеется очень большая группа N пациентов и 2 альтернативных лекарства с различными и неизвестными эффективностями.

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

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 $\nu \propto N^{2/3}$.



r-этапный подход - 1

Начальные Этапы $(i = 1, \ldots, r-1)$***:*** Если на $(i-1)$-ом этапе не удалось выделить лучший вариант, то на i-ом этапе оба варианта применяются по ν_i раз. Затем полные текущие доходы на i-ом этапе X_{i1}, X_{i2} сравниваются. Если выполняется неравенство $|X_{i1} - X_{i2}| \geq \Delta_i$, то вариант, соответствующий меньшему из доходов X_{i1}, X_{i2} отбрасывается; оставшийся вариант применяется на заключительном этапе. Предполагаем, что $\Delta_{r-1} = 0$, поэтому переход к ***Заключительному Этапу*** в конце $(r-1)$-ого этапа случится обязательно.

Заключительный Этап: На этом этапе применяется единственный вариант, соответствующий максимальной величине дохода на предыдущем i-ом начальном этапе $(i \leq r-1)$. Он считается лучшим по результатам сравнения.

r-этапный подход - 2

Найдены асимптотически оптимальные параметры стратегии $\{\nu_i, \Delta_i\}$ и установлено, что порядок минимаксного риска определяется величиной N^α, где $\alpha = 2^{r-1}/(2^r - 1)$. При параллельной обработке полное время работы определяется количеством этапов r, а не числом данных N.

Таблица: α как функция r

$r = 2$	$r = 3$	$r = 4$	$r = 5$	$\dots$	$r = \infty$
$\alpha = \frac{2}{3}$	$\alpha = \frac{4}{7}$	$\alpha = \frac{8}{15}$	$\alpha = \frac{16}{31}$	$\dots$	$\alpha = \frac{1}{2}$

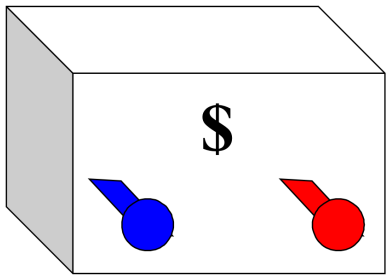
Публикации результатов

1. Колногоров А.В. Об оптимальном априорном времени обучения в задаче о «двуруком бандите» // Пробл. передачи информ. 2000. Т. 36, № 4. С. 117-127.
2. Kolnogorov A.V., Melnikova S.V. Minimax R-Stage Strategy for the Multi-Armed Bandit Problem // Proceedings of the 9-th IFAC Workshop on Adaptation and Learning in Control and Signal Processing, ALCOSP'07. Available online at http://www.ifac-papersonline.net.
3. Колногоров А.В. Задача о двуруком бандите для систем с параллельной обработкой данных // Пробл. передачи информ. 2012. Т. 48, № 1, С. 83–95.

Параллельное управление

Параллельное управление

Нормальный двурукий бандит



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

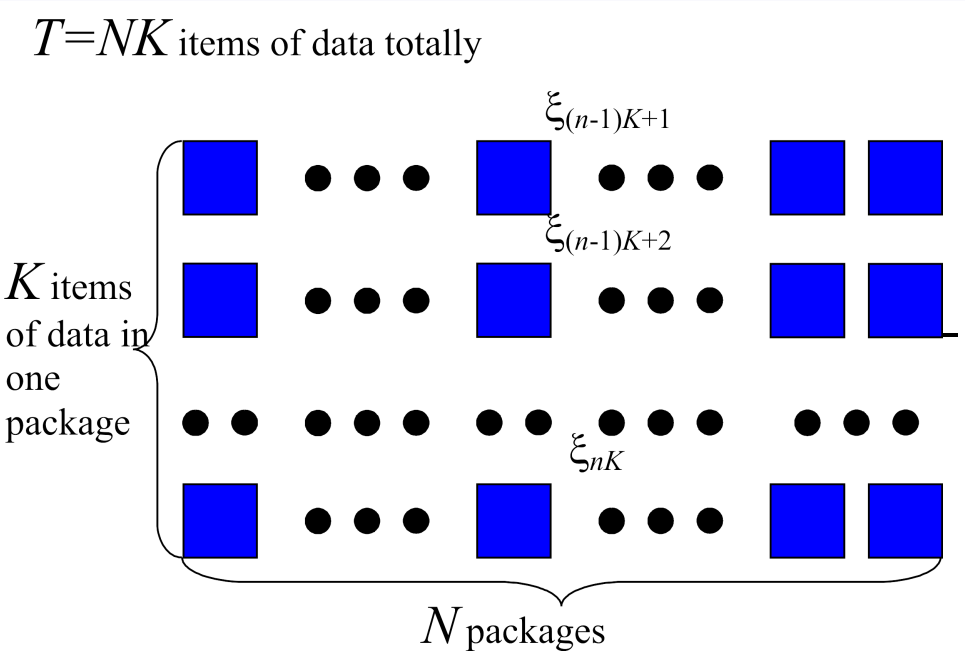
Формально выигрыши можно рассматривать как управляемый случайный процесс $\xi_1, \xi_2, \ldots, \xi_n$, значения которого зависят только от выбираемых вариантов $y_1, y_2, \ldots, y_n$ и имеют нормальную плотность распределения с единичной дисперсией и математическим ожиданием m_ℓ, если выбран вариант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

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

Почему *нормальный* двурукий бандит?



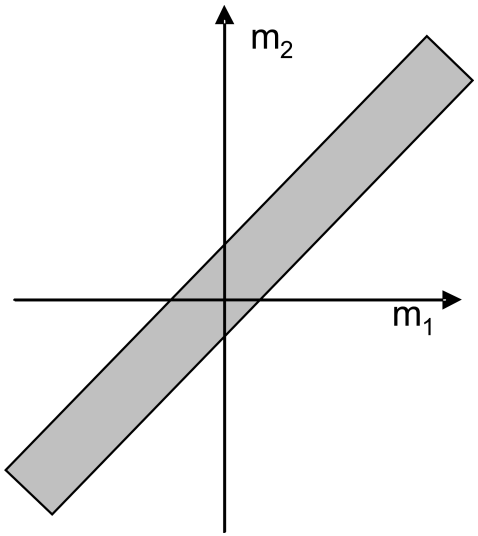
Предположим, что надо обработать большое число данных $T = NK$, причем для обработки доступны два альтернативных метода. Обработка может быть успешной $(\xi_t = 1)$ или неуспешной $(\xi_t = 0)$.

Вероятности успешной и неуспешной обработки зависят только от выбранных методов (действий) и равны p_ℓ и q_ℓ соответственно $(\ell = 1, 2)$. Известно, что $p_1 = p$, а p_2 близка к p. Определим процесс

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

Распределения ξ'_n близки к нормальным, дисперсии близки к 1, а первое математическое ожидание равно 0.

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)

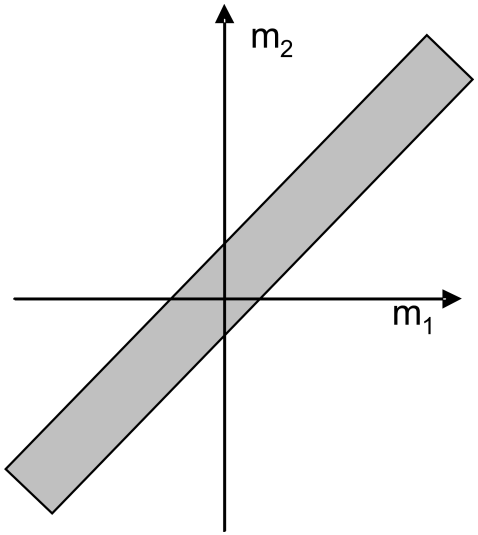
При $N \to \infty$ выполнено неравенство:

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

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:

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

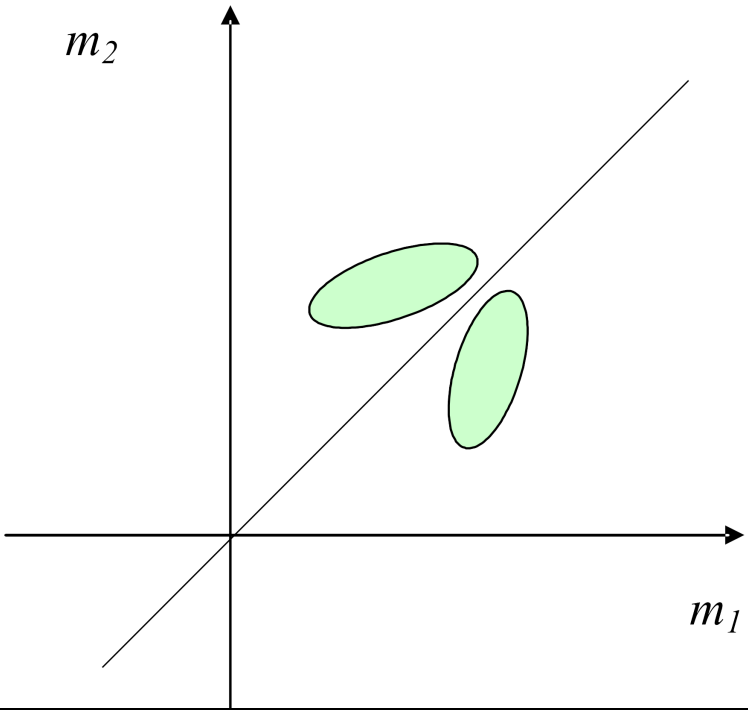
Минимаксная стратегия совпадает с некоторой байесовской, вычисленной для наихудшего априорного распределения.

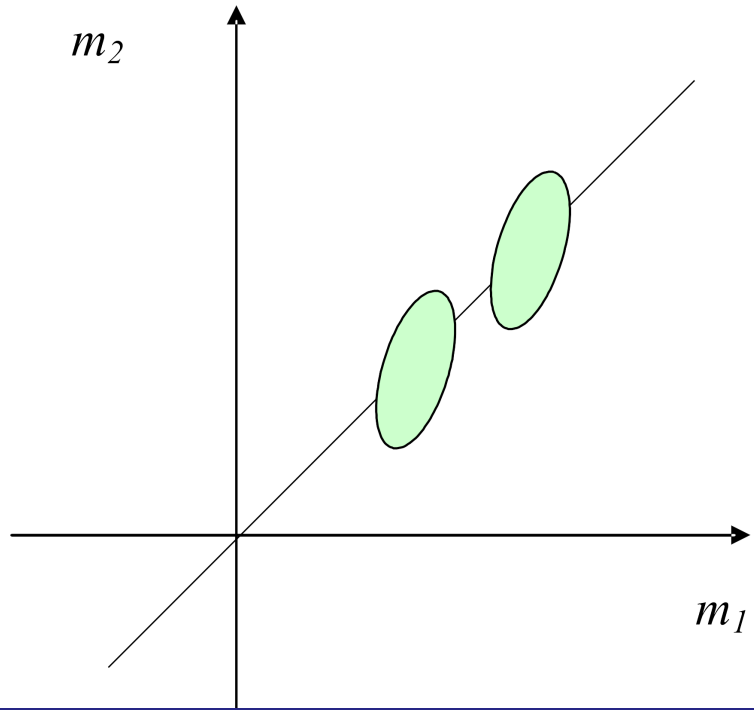
Свойства априорного распределения

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.



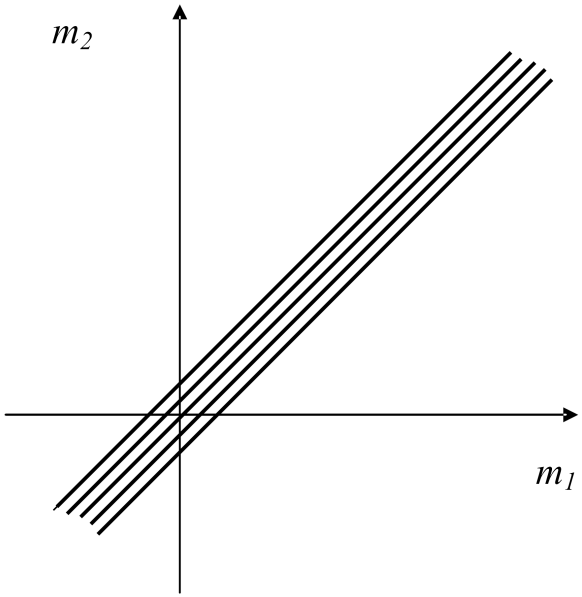


Асимптотически наихудшее априорное распределение

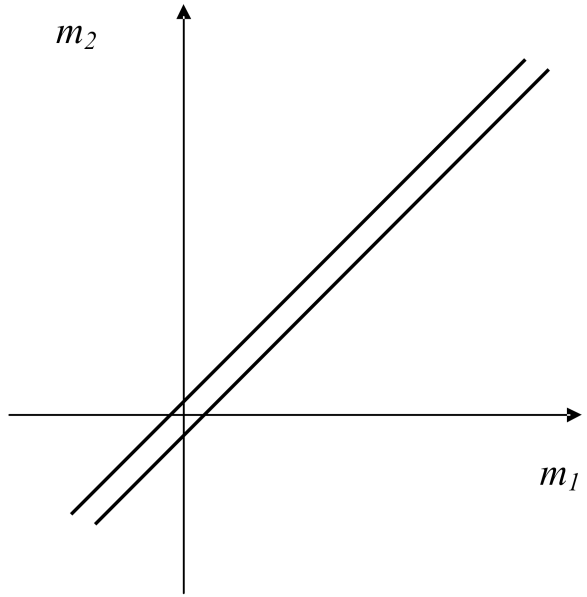
Может быть установлено с использованием свойства ***вогнутости байесовского риска***. Пусть λ_1, λ_2 — априорные плотности распределения, неотрицательные α_1, α_2 таковы, что $\alpha_1 + \alpha_2 = 1$. Тогда

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

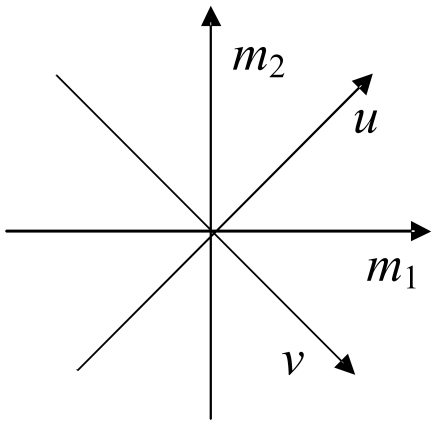
У асимптотически наихудшего распределения на линиях плотность постоянна:



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

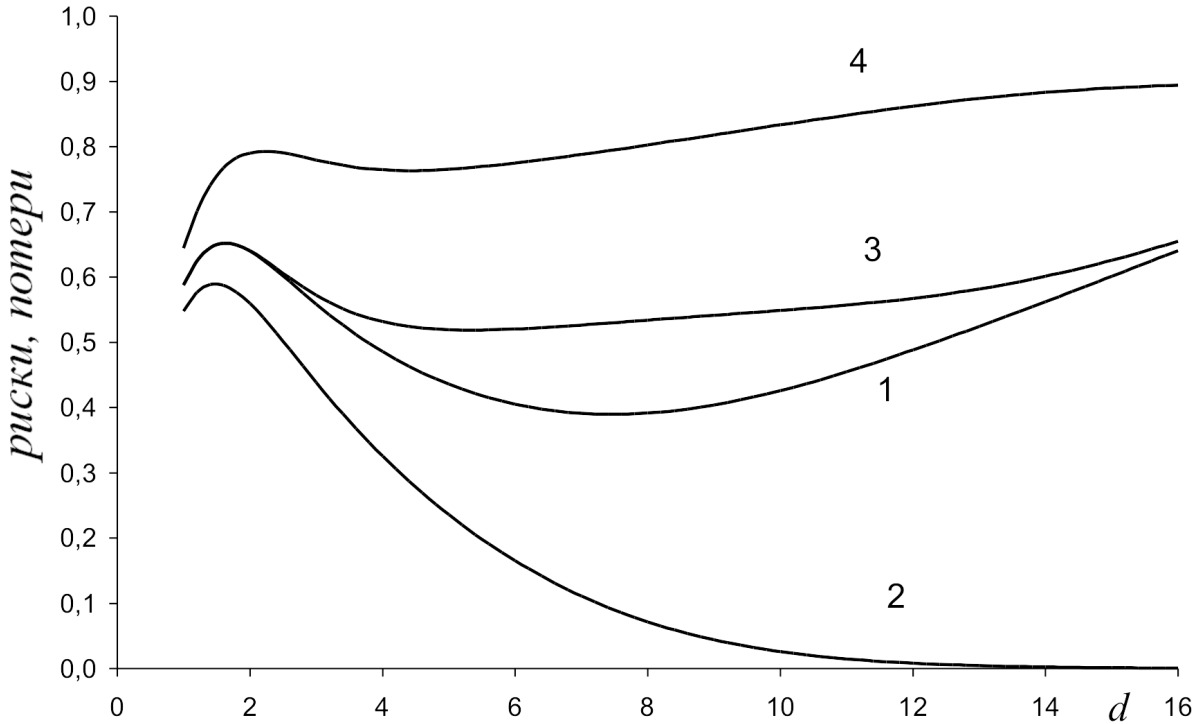
$$\lim_{a\to\infty} R_N^B(\nu_a(u,v)) = 4\int_0^\infty v\rho(v)dv + \int_{-\infty}^\infty R_{1,1}(z)dz.$$

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

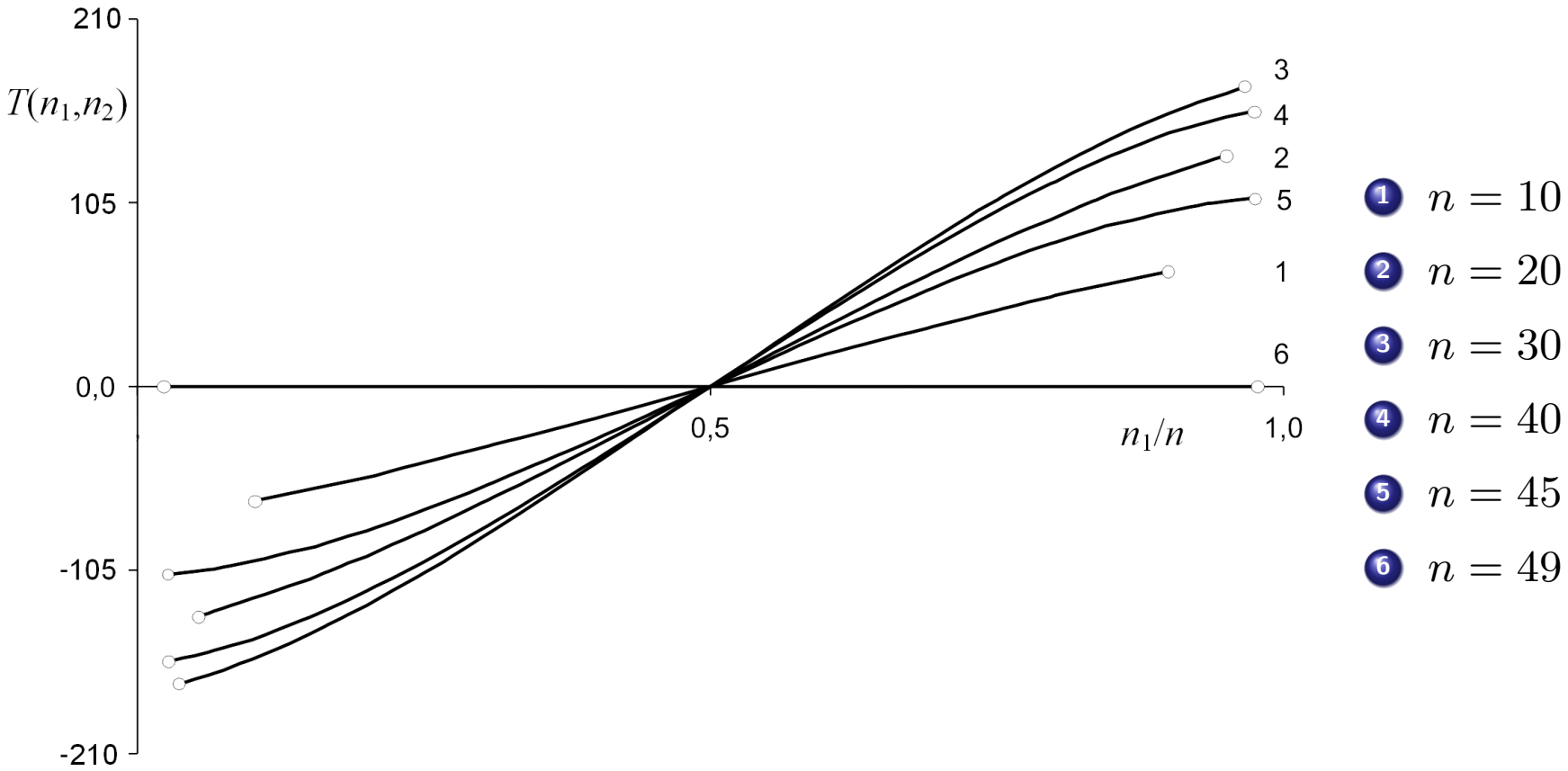
Предположим, что $\rho(v)$ сосредоточена в точках $v = \pm dN^{-1/2}$ с вероятностями 1/2. Тогда d соответствует максимуму приведенного байесовского риска. Этот максимум при $N = 50$ приблизительно равен 0,65 при $d \approx 1,7$. Для подтверждения предположения вычислялись приведенные потери и их значение 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 $Z > T(n_1, n_2)$ и 2-ой вариант, если $Z < T(n_1, n_2)$. При $Z = T(n_1, n_2)$ выбор может быть произволен.

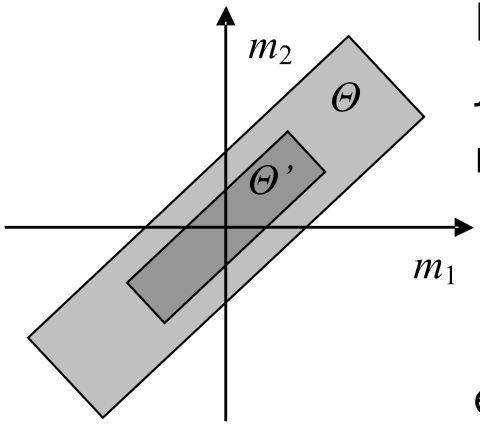


Параллельная обработка - 1

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

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

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

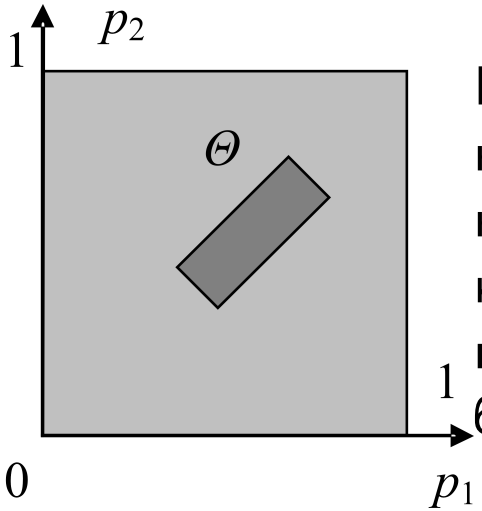


Потребуем, чтобы новая стратегия σ' так же управляла процессом ξ'_n как стратегия σ управляет процессом ξ_n. Тогда приведенные потери равны, т.е.

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

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 $(\xi_t = 0)$.

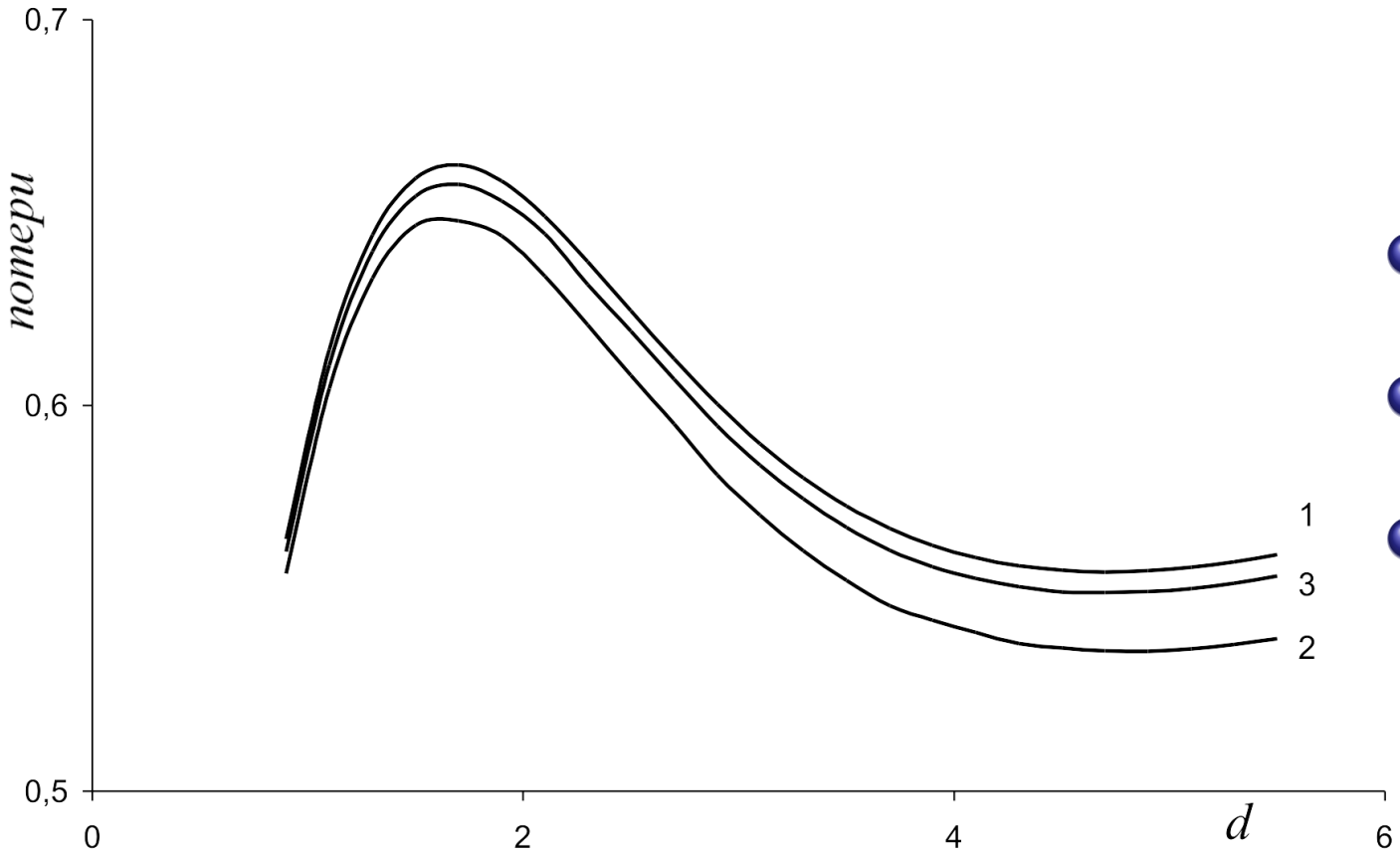
Вероятности успешной и неуспешной обработки зависят только от выбранных методов (вариантов) и равны p_ℓ и q_ℓ соответственно $(\ell = 1, 2)$. Пусть известно, что p_1, p_2 близки к $p = 0,5$. Разобьем все данные на $N = 30$ блоков по $K = 20$ каждый и определим процесс

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

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 $T = 3000$

Предельный переход. Оценки минимаксного риска

Предельный переход. 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

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

Дифференциальное уравнение

При некоторых дополнительных ограничениях $r(S, t_1, t_2)$ является обобщенным решением дифференциального уравнения в частных производных

$$\min_{\ell=1,2}\left(r'_{t_\ell} + \frac{1}{t_\ell}r + \frac{s}{t_\ell}r'_s + \frac{t_\ell^2}{2}r''_{ss} + g^{(\ell)}(s, t_1, t_2)\right) = 0$$

с начальными условиями

$$\lim_{t_1+t_2\to 1} r(s, t_1, t_2) = 0 \quad \text{при} \quad t_1 > \varepsilon_0,\ t_2 > \varepsilon_0,$$

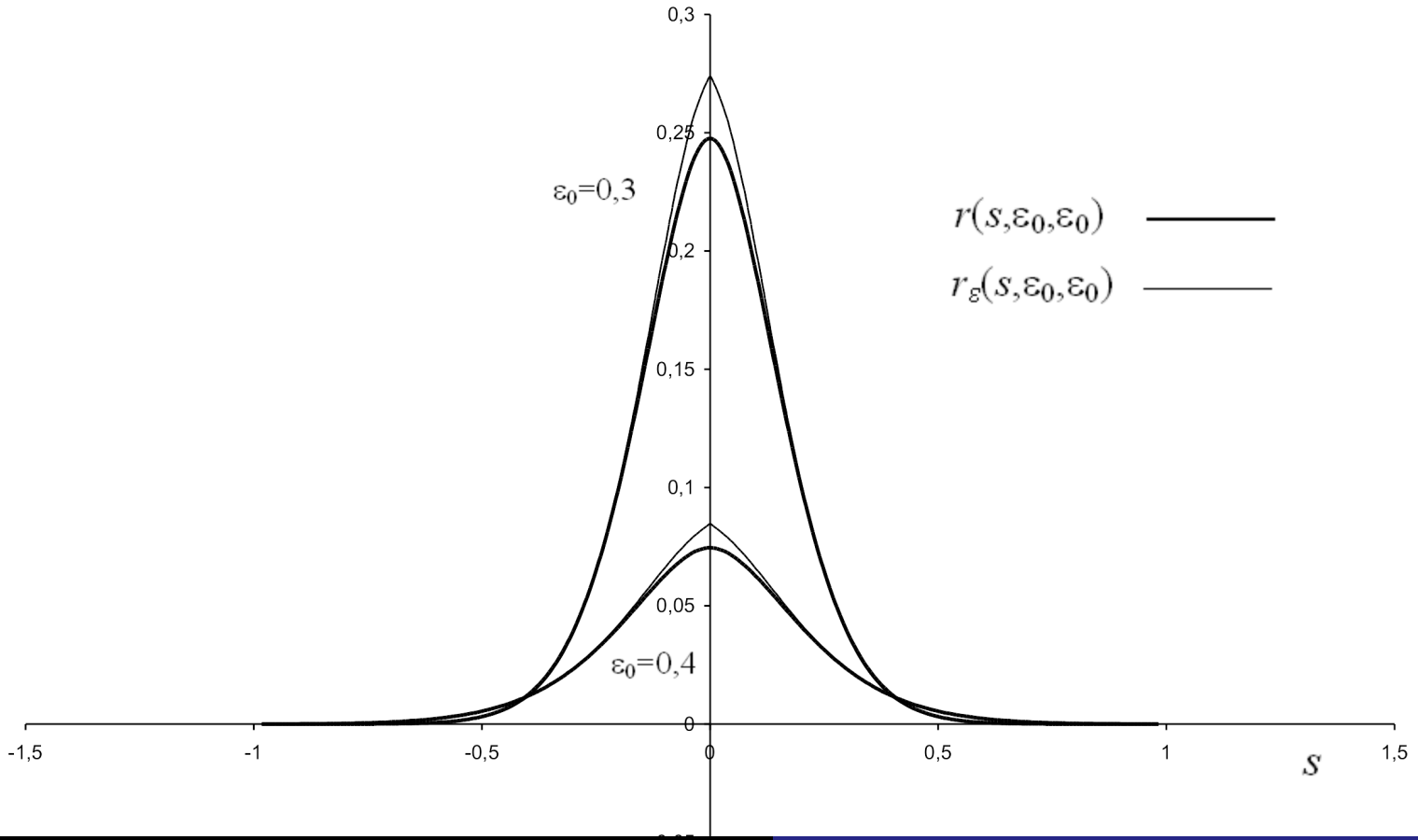
и граничными условиями

$$\lim_{s\to+\infty} r(s, t_1, t_2) = \lim_{s\to-\infty} r(s, t_1, t_2) = 0$$

при $2\varepsilon_0 < t_1 + t_2 < 1$, $t_1 > \varepsilon_0$, $t_2 > \varepsilon_0$.

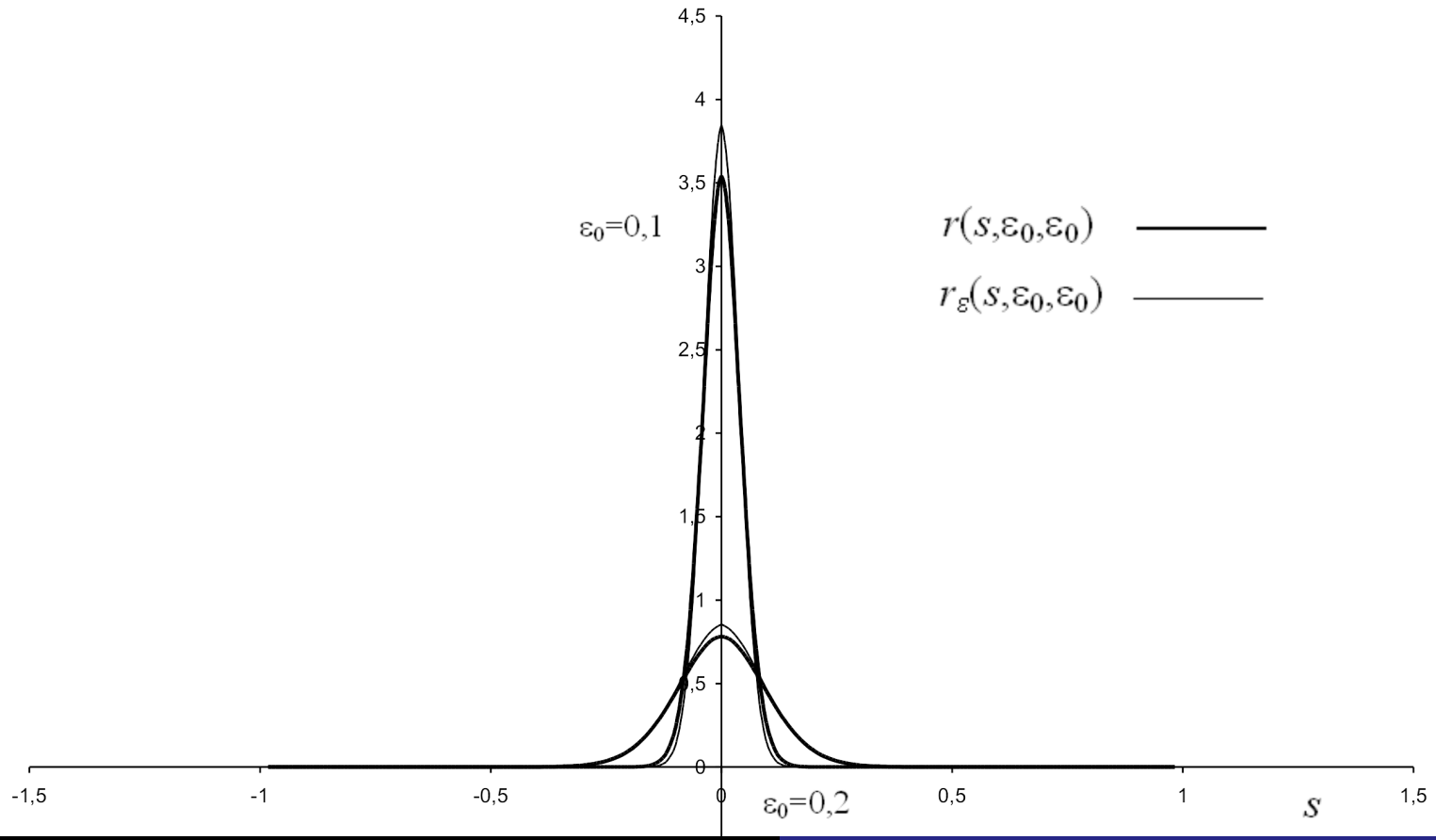
Численные эксперименты - 1

$\varepsilon = 0,05$	$r(\varrho, \varepsilon_0)$	$r_\varepsilon(\varrho, \varepsilon_0)$
$\varepsilon_0 = 0,4$	1,397	1,399
$\varepsilon_0 = 0,3$	1,114	1,118



Численные эксперименты - 2

$\varepsilon = 0,05$	$r(\varrho, \varepsilon_0)$	$r_\varepsilon(\varrho, \varepsilon_0)$
$\varepsilon_0 = 0,2$	0,869	0,875
$\varepsilon_0 = 0,1$	0,707	0,707



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

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

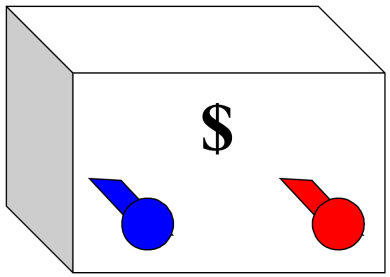
В частности, при всех N кратных 50 на $\Theta = \{|m_1 - m_2| \le 32N^{-1/2}\}$ справедлива оценка

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

Однорукий бандит

Задача об одноруком бандите

Нормальный однорукий бандит



Это игральный автомат с двумя рукоятками, причем характеристики одной из них известны. При выборе ℓ-ой рукоятки игрок получает случайный доход. Этот доход имеет нормальное распределение с единичной дисперсией и математическим ожиданием m_ℓ.

Игрок может нажать рукоятки в общей сложности N раз. Его цель – максимизация (в некотором смысле) полного ожидаемого дохода. Оба распределения фиксированы в процессе управления. Значение m_1 известно и без ограничения общности $m_1 = 0$. Значение $m_2 = m$ неизвестно.

Основная идея

Выбор первой рукоятки не меняет информацию, известную игроку. Поэтому если первая рукоятка однажды будет выбрана, то она будет выбираться до конца управления. Следовательно, оптимальная стратегия предписывает выбирать вторую рукоятку на некоторой начальной стадии управления. а затем переключается на выбор первой рукоятки до конца управления.

Формальная постановка задачи

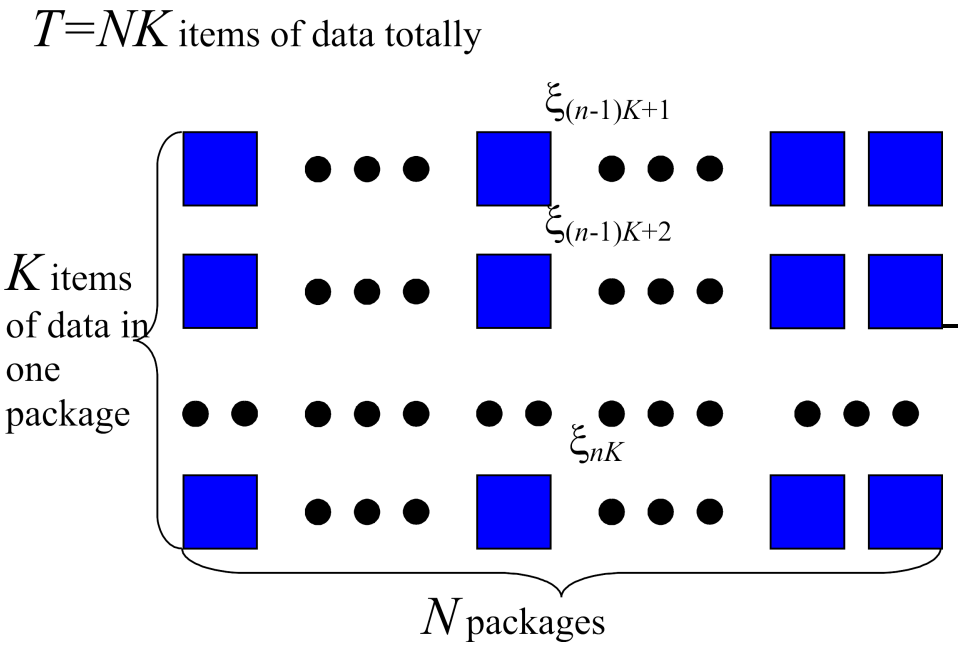
Формально доходы рассматриваются как управляемый случайный процесс $\xi_1, \xi_2, \ldots, \xi_N$, значения которого зависят только от выбираемых в текущий момент времени рукояток (в дальнейшем называемых действиями) $y_1, y_2, \ldots, y_N$ и имеют нормальные распределения с единичными дисперсиями и математическими ожиданиями m_ℓ, если выбрано ℓ-е действие

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

Такой процесс полностью описывается векторным параметром $\theta = (0, m)$. Управляющая стратегия σ зависит от всей предыстории процесса. Пусть в момент времени n все еще выбиралось второе действие. Тогда предыстория процесса описывается парой (X, n). Здесь n – полное число применений второго действия, X – полный доход. Функция потерь определена следующим образом

$$L_N(\sigma, \theta) = N(0 \vee m) - E_{\sigma,\theta}\left(\sum_{n=1}^{N} \xi_n\right).$$

Почему *нормальный* однорукий бандит?



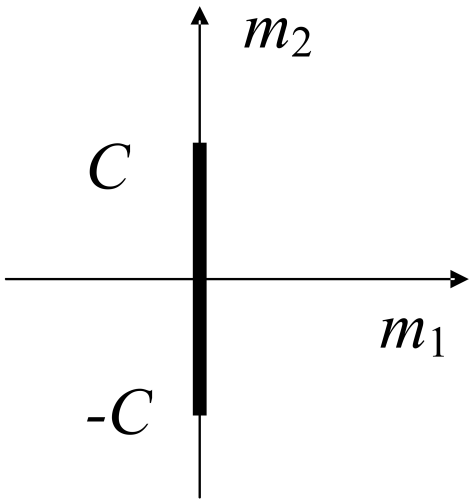
Предположим, что надо обработать большое число данных $T = NK$, причем для обработки доступны два альтернативных метода. Обработка может быть успешной $(\xi_t = 1)$ или неуспешной $(\xi_t = 0)$.

Вероятности успешной и неуспешной обработки зависят только от выбранных методов (действий) и равны p_ℓ и q_ℓ соответственно $(\ell = 1, 2)$. Известно, что $p_1 = p$, а p_2 близка к p. Определим процесс

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

Распределения ξ'_n близки к нормальным, дисперсии близки к 1, а первое математическое ожидание равно 0.

Минимаксная и байесовская постановки



Предположи, что множество параметров является следующим $\Theta = \{(0, m) : |m| \leq C\}$, with $0 < C < \infty$. Минимаксный риск определяется как

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

соответствующая оптимальная стратегия σ^M называется минимаксной стратегией.

Рассмотрим априорное распределение $\lambda(m)$ на Θ. Байесовский риск определяется как

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

соответствующая оптимальная стратегия σ^B называется байесовской стратегией.

Рекуррентное уравнение

Рассмотри уравнение, решаемое методом динамического программирования

$$R_n(\cdot) = \min(R_n^{(1)}(\cdot), R_n^{(2)}(\cdot)),$$

где $R_N^{(1)}(\lambda; X, N) = R_N^{(2)}(\lambda; X, N) = 0$ and

$$R_n^{(1)}(\lambda; X, n) = (N-n) g_n^{(1)}(\lambda; X, n),$$

$$R_n^{(2)}(\lambda; X, n) = g_n^{(2)}(\lambda; X, n) + \int_{-\infty}^{+\infty} R_{n+1}(\lambda; X+Y, n+1) h_n\left(X - nY\right) dY$$

при $0 \le n < N$. Здесь

$$h_n(Y) = \left(\tfrac{n+1}{2\pi n}\right)^{1/2} \exp\left(-\tfrac{Y^2}{2n(n+1)}\right), \ n \ge 1, \quad h_0(Y) = 1,$$

$$g_n^{(1)}(\lambda; X, n) = \int_0^C m f_n(X - nm) \lambda(m) dm,$$

$$g_n^{(2)}(\lambda; X, n) = \int_0^C m f_n(X + nm) \lambda(-m) dm,$$

$$f_n(X) = (2\pi n)^{-1/2} \exp(-X^2/(2n)).$$

Байесовский риск и оптимальная (байесовская) стратегия

Байесовский риск

Байесовский риск вычисляется как

$$R_N^B(\lambda) = \min(R_0^{(1)}(\lambda), R_0^{(2)}(\lambda)).$$

Оптимальная (байесовская) стратегия

Оптимальная (байесовская) стратегия предписывает выбирать в текущий момент времени действие, соответствующее меньшей из величин $R_n^{(1)}(\lambda; X, n)$, $R_n^{(2)}(\lambda; X, n)$, при их равенстве выбор может быть произвольным. Если первое действие было однажды выбрано, то его следует выбирать далее до конца управления.

Кусочно-постоянные стратегии

Рассмотрим стратегию, допускающую параллельное управление. Такая стратегия применяет один и тот же вариант к группе из $M = \varepsilon N$ данных. Для простоты N кратно M. Если данные поступают последовательно, такая стратегия позволяет реже переключать варианты, так как использует одинаковые варианты на интервалах времени $(1, M), (M+1, 2M), \dots (N-M+1, N)$.

При параллельной обработке M данных обрабатываются одновременно. Стратегия суммирует доходы, полученные при параллельной обработке, поэтому в силу ЦПТ их распределения могут быть близки к нормальным и в том случае, если исходные распределения нормальными не являлись.

Положим $t = N^{-1}n$, $x = N^{-1/2}X$, $y = N^{-1/2}Y$, $\varepsilon = MN^{-1}$,
$v = N^{1/2}m$, $\varrho(v) = N^{-1/2}\lambda(m)$, $c = N^{1/2}C$,
$r_\varepsilon^{(\ell)}(\varrho; x, t) = R_n^{(\ell)}(\lambda; X, n)$, $r_\varepsilon(\varrho; x, t) = R_n(\lambda; X, n)$,
$r_\varepsilon^{(\ell)}(\varrho) = N^{-1/2}R_0^{(\ell)}(\lambda)$, $r_\varepsilon(\varrho) = N^{-1/2}R_0(\lambda)$.

Инвариантное рекуррентное уравнение

Рассмотрим рекуррентное уравнение

$$r_\varepsilon(\varrho; x, t) = \min(r_\varepsilon^{(1)}(\varrho; x, t), r_\varepsilon^{(2)}(\varrho; x, t)),$$

где $r_\varepsilon^{(1)}(\varrho; x, 1) = r_\varepsilon^{(2)}(\varrho; x, 1) = 0$,

$$r_\varepsilon^{(1)}(\varrho; x, t) = (1-t)g^{(1}(\varrho; x, t),$$
$$r_\varepsilon^{(2)}(\varrho; x, t) = \varepsilon g^{(2}(\varrho; x, t) + \int\limits_{-\infty}^{+\infty} r_\varepsilon(\varrho; x+y, t+\varepsilon) h_{t,\varepsilon}\left(x\varepsilon - ty\right) dy,$$

при $0 \le t < 1$. Здесь

$$h_{t,\varepsilon}(y) = \left(\tfrac{t+\varepsilon}{2\pi\varepsilon t}\right)^{1/2} \exp\left(-\tfrac{y^2}{2\varepsilon t(t+\varepsilon)}\right), \ t > 0, \quad h_{0,\varepsilon}(y) = 1,$$
$$g^{(1}(\varrho; x, t) = \int\limits_0^c v f_t(x - tv)\varrho(v)dv,$$
$$g^{(2}(\varrho; x, t) = \int\limits_0^c v f_t(x + tv)\varrho(-v)dv,$$
$$f_t(x) = (2\pi t)^{-1/2} \exp(-x^2/(2t)).$$

Байесовский риск и оптимальная (байесовская) стратегия

Байесовский риск

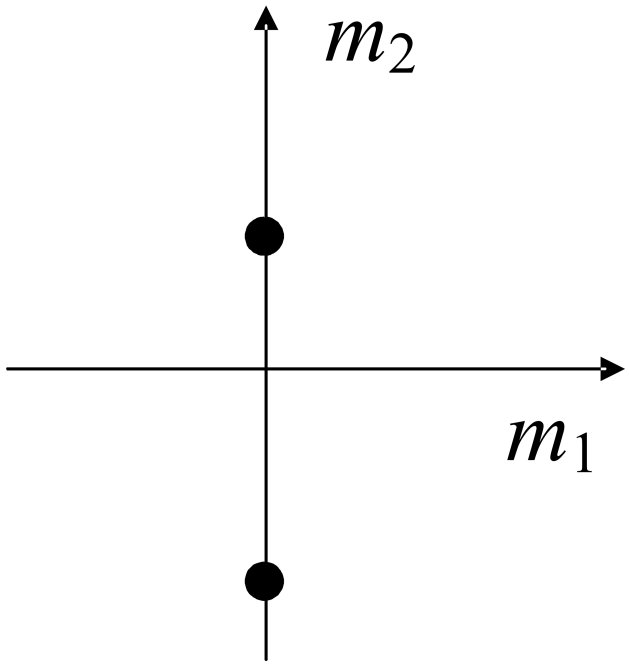
Байесовский риск вычисляется как

$$R_N^B(\lambda) = N^{1/2} \min(r_\varepsilon^{(1)}(\varrho), r_\varepsilon^{(2)}(\varrho)).$$

Оптимальная (байесовская) стратегия

Байесовская стратегия предписывает выбирать в текущий момент времени то действие, которое соответствует меньшей из величин $r_\varepsilon^{(1)}(\varrho; x, t)$, $r_\varepsilon^{(2)}(\varrho; x, t)$, при их равенстве выбор может быть произвольным. Если первое действие будет однажды выбрано, то оно применяется далее до конца управления.

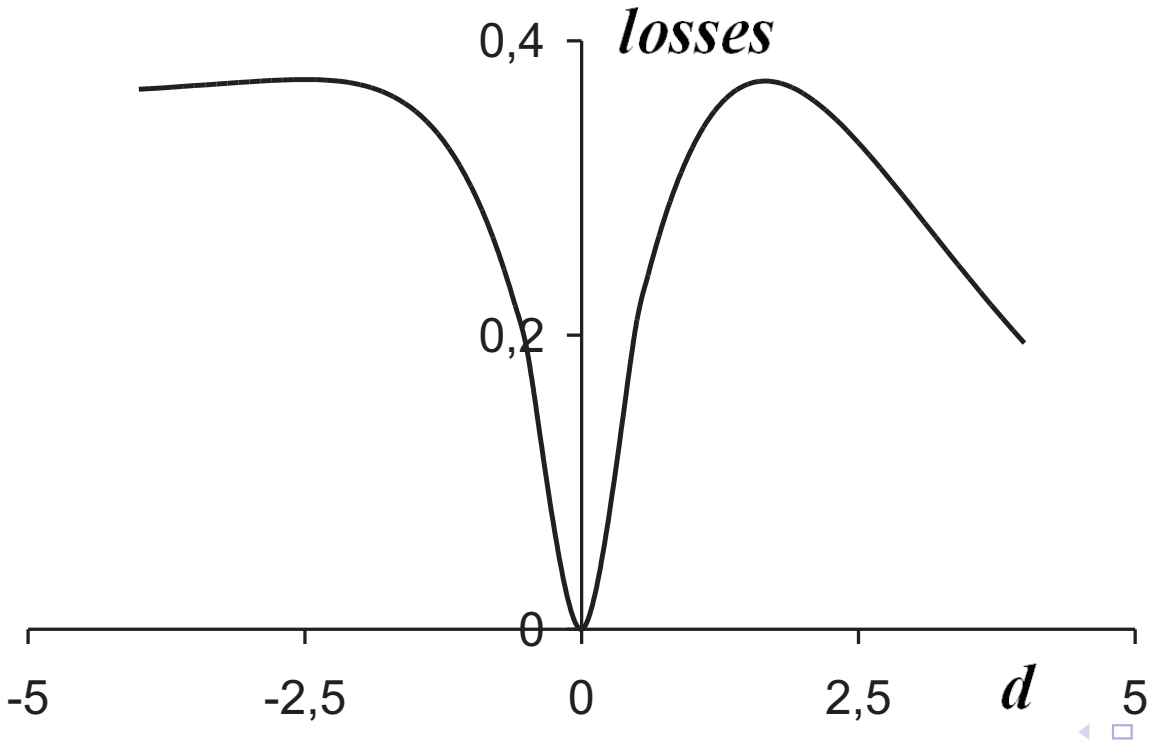
Наихудшее априорное распределение



Предполагается, что наихудшее априорное распределение сосредоточено в двух точках

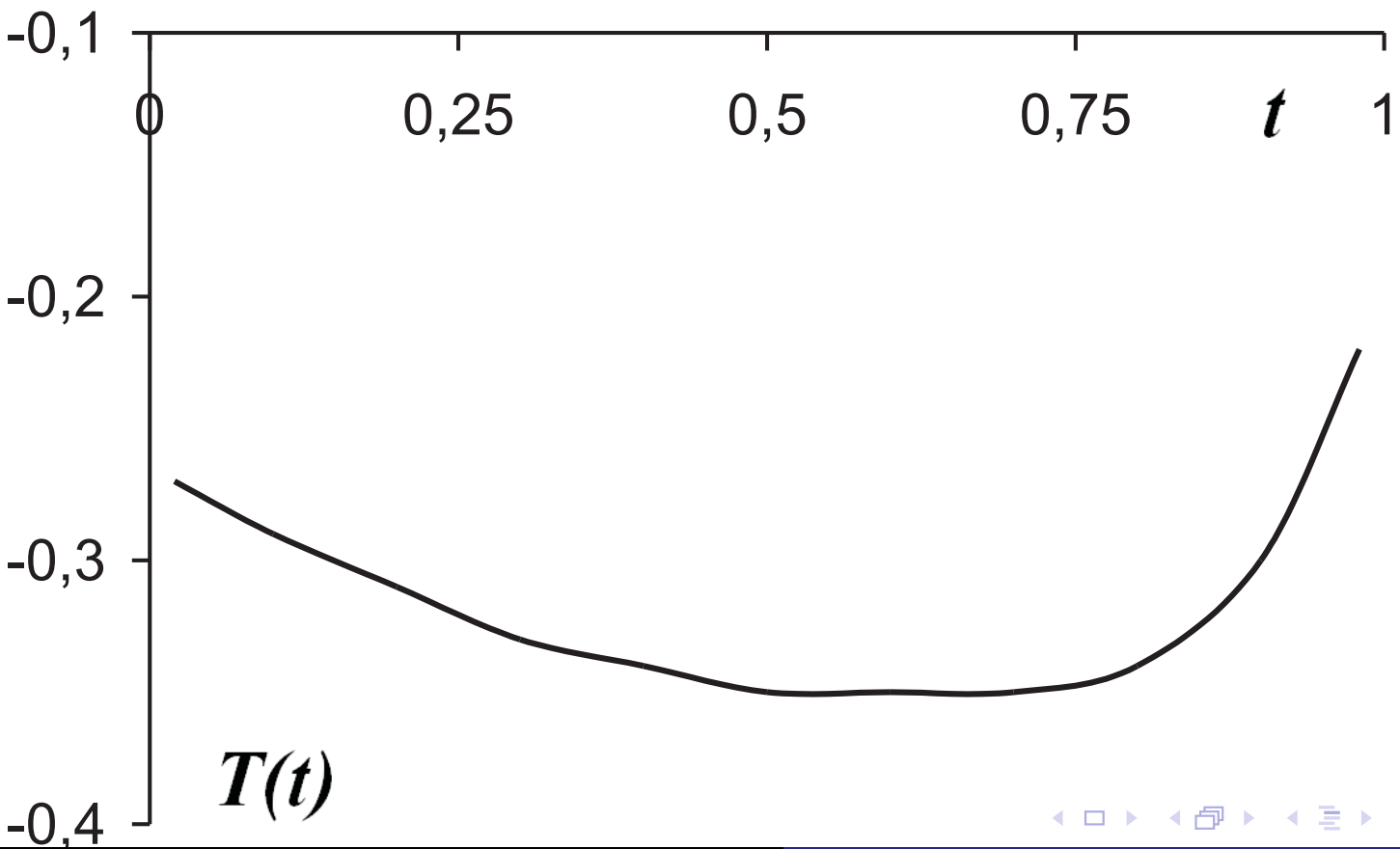
Нахождение минимаксного риска

Минимаксный риск искался в предположении, что наихудшее априорное распределение $\lambda(m)$, $m = vN^{-1/2}$ при достаточно больших N сосредоточено в двух точках $v \approx d_1$ и $v \approx -d_2$ с вероятностями $\Pr(v = d_1) = \varrho$, $\Pr(v = -d_2) = 1 - \varrho$. Численная оптимизация позволила установить, что $d_1 \approx 1,65$, $d_2 \approx 2,52$, $\varrho \approx 0,38$, $R_N^B(\lambda) \approx 0,37N^{1/2}$ и, следовательно, $R_N^M(\Theta) \approx 0,37N^{1/2}$.



Нахождение минимаксной стратегии

Найденная стратегия имеет пороговый характер. Она применяет второе действие, если $x > T(t)$ и переключается на применение первого действия до конца управления, если $x < T(t)$, где $\{T(t)\}$ – множество пороговых значений.



Параллельная обработка – 1

Пусть даны $T = 600$ данных, для обработки которых можно использовать два альтернативных метода. Обработка может быть успешной ($\xi_t = 1$) или неуспешной ($\xi_t = 0$). Вероятности успешной и неуспешной обработки равны $\Pr(\xi_t = 1 | y_t = \ell) = p_\ell$, $\Pr(\xi_t = 0 | y_t = \ell) = 1 - p_\ell$ ($\ell = 1, 2$). Пусть известно, что $p_1 = 0.5$, а p_2 близка к 0.5. Разобьем данные на $N = 50$ блоков по $K = 12$ данных в блоке и определим процесс

$$\xi'_n = (DK)^{-1/2} \sum_{t=(n-1)K+1}^{nK} (\xi_t - 0.5),$$

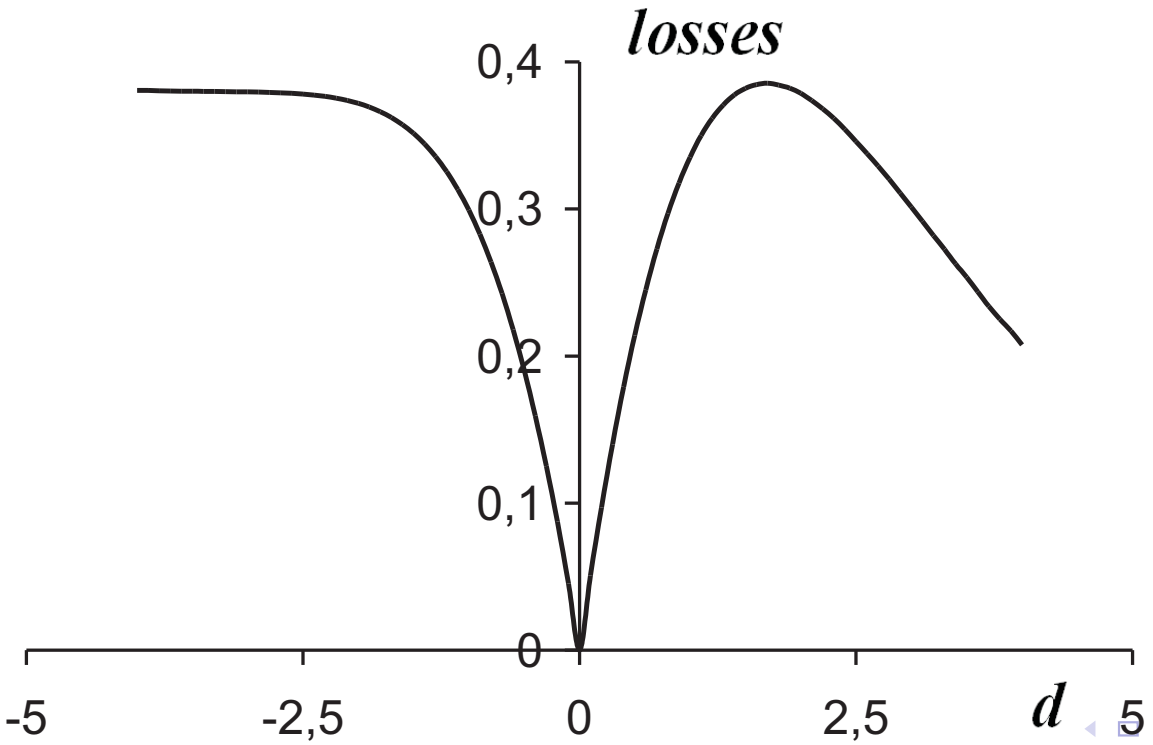
$n = 1, \ldots, N$ with $D = 0.25$. Распределения $\{\xi'_n\}$ близки к нормальным, их дисперсии близки к 1, первое математическое ожидание равно 0 и N соответствует $\varepsilon = 0.02$.

Параллельная обработка – 2

Применим минимаксную стратегию к $\{\xi'_n\}$. Результаты моделирования Монте-Карло для потерь

$$l_T(d) = (DT)^{-1/2} E_{\sigma,\theta} \left(\sum_{t=1}^{T} ((0.5 \vee p_2) - \xi_t) \right)$$

представлены ниже. Они соответствуют рассчитанным теоретически.



Некоторые публикации

1. Колногоров А.В. Нахождение минимаксных стратегии и риска в случайной среде (задаче о двуруком бандите) // АиТ. 2011. № 5. С. 127–138.
2. Колногоров А.В. Робастное параллельное управление в случайной среде (задаче о двуруком бандите) // АиТ. 2012. № 4. С. 114–130.
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. Минимаксные стратегия и риск в многоальтернативной случайной среде (задаче о многоруком бандите) // Труды IX Международной конференции "Идентификация систем и задачи управления"SICPRO’12 (3 января - 2 февраля 2012 года) - М.: Институт Проблем Управления им. В.А.Трапезникова РАН, 2012. ISBN - 978-5-91450-098-3 С. 1061-1076

Некоторые публикации

1. Колногоров А.В. Предельное описание робастного параллельного управления в задаче о двуруком бандите // Обозрение прикладной и промышленной математики, т. 19, вып. 2, 2012, с. 209-210
2. Kolnogorov, A.V. A Limiting Description of the Minimax Control in the Two-Armed Bandit Problem // Proceedings of the International Conference "Probability Theory and its Applications"in Commemoration of the Centennial of B.V.Gnedenko, Moscow, June 26-30, 2012, P.112-113
3. Kolnogorov, A.V. Robust Normal Two-Armed Bandit, One Arm Known, and Parallel Data Processing // 11th IFAC International Workshop on Adaptation and Learning in Control and Signal Processing, University of Caen Basse-Normandie, Caen, France, July 3-5, 2013. p. 263-268, DOI 10.3182/20130703-3-FR-4038.00085. Available online at http://www.ifac-papersonline.net

Спасибо за внимание

Спасибо за внимание