Московский государственный университет
имени М.В. Ломоносова
Механико-математический факультет

На правах рукописи

УДК 519.21, 519.234

Рафиков Евгений Геннадьевич

ЭФФЕКТ КОНЦЕНТРАЦИИ МЕРЫ В СТАТИСТИЧЕСКИХ ЗАДАЧАХ НЕПАРАМЕТРИЧЕСКОГО ОЦЕНИВАНИЯ.

01.01.05 — теория вероятностей и математическая статистика

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва 2007

Работа выполнена на кафедре теории вероятностей Механико-математического факультета Московского государственного университета имени М. В. Ломоносова.

Научный руководитель: доктор физико-математических наук, профессор Ю. Н. Тюрин

Официальные оппоненты: доктор физико-математических наук, доцент В. В. Ульянов

кандидат физико-математических наук, доцент С. А. Пирогов

Ведущая организация: Математический институт Стеклова РАН

Защита диссертации состоится 16 марта 2007 года в 16 часов 15 минут на заседании диссертационного совета Д.501.001.85 в Московском государственном университете имени М. В. Ломоносова по адресу: 119992, ГСП-2, г. Москва, Ленинские горы, МГУ, Механико-математический факультет, аудитория 16-24.

С диссертацией можно ознакомиться в библиотеке Механико-математического факультета (Главное здание, 14 этаж).

Автореферат разослан 16 февраля 2007 года.

Учёный секретарь диссертационного
совета Д.501.001.85 в МГУ,
доктор физико-математических
наук, профессор Т. П. Лукашенко

Общая характеристика работы

Актуальность темы.

Во многих задачах математической статистики наряду с асимптотическим поведением стохастического объекта важно также знать и его неасимптотическое поведение.

Известно, что если $X_1, \ldots, X_m$ — одномерные независимые и одинаково распределённые случайные величины, где $F(x)$ и $F_m(x)$ — соответствующие им настоящая и эмпирическая функции распределения, то имеет место предельное асимптотическое поведение:

Теорема (Колмогоров).

$$\mathsf{Pr}\Big\{\sup_{x\in\mathbb{R}}|F_m(x)-F(x)|>t/\sqrt{m}\Big\}\to 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2t^2},\quad \textit{при } m\to\infty.$$

Из теоремы Колмогорова следует также и неасимптотическая оценка:

Утверждение.

$$\mathsf{Pr}\Big\{\sup_{x\in\mathbb{R}}|F_m(x)-F(x)|>t\Big\}\le C\cdot e^{-2mt^2},$$

аналог которой в случае распределений в $\mathbb{R}^d$ был получен Кифером (Kiefer):

Теорема (Кифер). *Для произвольного $\alpha>0$ существует положительная константа $K=K(\alpha,d)$, такая что для всех $m\in\mathbb{N}$ верно:*

$$\mathsf{Pr}\Big\{\sup_{x\in\mathbb{R}^d}|F_m(x)-F(x)|>t\Big\}\le K\cdot e^{-(2-\alpha)\cdot mt^2}.$$

Эмпирический процесс, который фигурирует в указанных выше результатах, является частным случаем более общего объекта:

$$\sup_{A\in\mathcal{A}}|\mathsf{P}_m(A)-\mathsf{P}(A)|, \tag{1}$$

Получение точных оценок для вероятностных хвостов (1) тесно связано с задачей о *вероятностях больших уклонений*, подробно изучавшейся Сановым[1]. См. также монографию[2]. Большая часть *статистической теории обучений* также

[1] И.Санов, *О вероятностях больших уклонений случайных величин*, Мат. сборник, **42** (1957), p.11-44.

[2] A.Dembo, O.Zeitouni, *Large Deviations Techniques and Applications*, Springer (1998).

посвящена изучению этой задачи при определённых условиях на вероятностную меру $\mathsf{P}(\cdot)$ и набор множеств $\mathcal{A}$.[3,4,5] Для фиксированного множества A и вероятностной меры $\mathsf{P}(\cdot)$ по закону больших чисел имеет место сходимость почти наверное $\mathsf{P}_m(A) - \mathsf{P}(A) \to 0$ при $m \to \infty$. Более того, известное неравенство Хёфдинга даёт следующую (неасимптотическую) оценку на скорость сходимости:

$$\mathsf{Pr}\big\{|\,\mathsf{P}_m(A) - \mathsf{P}(A)| > t\big\} \le 2e^{-2mt^2}.$$

Если набор $\mathcal{A}$ конечен, то естественно получаем обобщение предыдущего неравенства:

$$\mathsf{Pr}\big\{\sup_{A\in\mathcal{A}} |\,\mathsf{P}_m(A) - \mathsf{P}(A)| > t\big\} \le 2|\mathcal{A}| \cdot e^{-2m\cdot t^2}.$$

Однако если семейство $\mathcal{A}$ имеет бесконечную мощность, как это и бывает во многих интересных случаях, имеющих прикладное значение, задача становится гораздо более сложной. Оказывается, что ключевую роль в оценках экспоненциального типа для этой задачи играет ряд чисто комбинаторных характеристик семейства $\mathcal{A}$, таких как *размерность Вапника-Червоненкиса* и *энтропия Вапника-Червоненкиса*.

Часть результатов в этой области посвящена тому, что на неизвестную меру $\mathsf{P}(\cdot)$ накладываются дополнительные естественные условия. Так, важным является случай, когда мера $\mathsf{P}(\cdot)$ и семейство $\mathcal{A}$ таковы, что *множество значений мер* $\big\{\mathsf{P}(A),\, A \in \mathcal{A}\big\}$ отделено от некоторой окрестности $1/2$. Улучшению оценок на поведение вероятностных хвостов для (1) в этом случае посвящена первая часть нашей диссертации. В качестве числовой характеристики такой отделимости мы используем разновидность *информации Кульбака-Лейблера*, играющую важную роль в *теории информации*.[6,7] Близкие характеристики, такие как *расстояние Кульбака-Лейблера*, используются для решения задачи *оценивания плотности*[8], а также в *асимптотической теории оценивания*.[9]

Для получения экспоненциально быстрых оценок в изучаемом нами случае используется техника из *теории эмпирических процессов*, что также делает те-

[3] В.Вапник, А.Червоненкис, *Теория распознавания образов*, Наука (1974).
[4] В.Вапник, А.Червоненкис, *Восстановление зависимостей по эмпирическим данным*, Наука (1979).
[5] L.Devroye, L.Gyorfi, G.Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer (1996).
[6] А.Колмогоров, *Теория информации и теория алгоритмов*, Наука (1987).
[7] T.Cover, J.Thomas, *Elements of Information Theory*, Wiley (1991).
[8] Л.Деврой, Л.Дьёрфи, *Непараметрическое оценивание плотности. L1-подход.* Пер. с англ., Мир (1988).
[9] И.Ибрагимов, Р.Хасьминский, *Асимптотическая теория оценивания,* Наука (1979).

матику и приёмы, обсуждаемые в диссертации, близкими к таким областям как классическая *непараметрическая статистика*, *семипараметрическая статистика*, *теория M-оценивания*, *теория Гауссовских процессов* и *равномерные законы больших чисел и повторного логарифма*.

Вторая часть диссертации посвящена современным проблемам регрессионного анализа. Пусть $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ — некоторые Борелевские множества. В зависимости от предположений о характере неизвестной зависимости между $x \in X$ и $y \in Y$ известен ряд классических постановок *задачи обучения по конечной выборке*.

Так в *теории приближений* считается, что имеется детерминированная неизвестная зависимость на произведении $X \times Y$. Точнее, предполагается, что для точек из известного конечного набора $\mathbf{z} = \big\{(x_1, y_1), \ldots, (x_m, y_m),\ x_i \in X, y_i \in Y\big\}$ верно соотношение $y_i = f(x_i), i = 1, \ldots, m$, где f принадлежит некоторому заранее известному функциональному классу $\mathcal{H}$. Задача состоит в нахождении *наилучшего приближения* к f внутри класса $\mathcal{H}$. Ошибка приближения обычно измеряется в L_p-норме по отношению к мере Лебега на X, где $1 \leq p < \infty$.

В случае, когда допускается какая-то случайность на характер зависимости между x-ами и y-ами, говорят о *теории оценивания функции регрессии.* Рассматривается класс моделей вида $y = f(x) + \epsilon$, где $f \in \mathcal{H}$, а ϵ — случайная величина. Например, считается, что для множества $\mathbf{z} = \big\{(x_1, y_1), \ldots, (x_m, y_m),\ x_i \in X, y_i \in Y\big\}$ выполняются равенства $y_i = f(x_i) + \epsilon_i$, где $x_1, \ldots, x_m$ фиксированы, а ϵ_i — независимые одинаково распределённые случайные величины с $\mathsf{E}\epsilon_i = 0$. Задача, как и выше, состоит в том, чтобы построить некоторое приближение $f_{\mathbf{z}}$ для f (или оценку f), оптимальное в смысле $\mathsf{E}(||f - f_{\mathbf{z}}||^2)$ в одной из стандартных норм. Также рассматривается постановка задачи, когда "входы" $x_1, \ldots, x_m$ сами случайны и как-то распределены в соответствии с неизвестным вероятностным законом на X.[10]

В своей работе мы имеем дело с задачей, когда считается, что произведение $Z = X \times Y$ снабжено некоторой вероятностной Борелевской мерой ρ, а *обучающие примеры* $\mathbf{z}$ — суть случайная ρ-выборка длины m. Задача заключается в оценивании *функции регрессии* y-ов по x-ам, $f_\rho(x) = \mathsf{E}\{y|x\}$. Пусть ρ_X —

[10] L.Gyorfi, M.Kohler, A.Krzyzak, H.Walk, *A Distribution-Free Theory of Nonparametric Regression,* Springer Series in Statistics (2002).

это проекция меры ρ на X. Для описания качества оценивания мы используем общепринятую характеристику скорости убывания вероятностных хвостов

$$\mathsf{Pr}\{\|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)} > t\}, \tag{2}$$

где $f_{\mathbf{z}}(x)$ — это некоторая *функция-оценщик* функции регрессии $f_\rho(x)$. Для такой постановки типичными условиями на неизвестную меру ρ являются предположение на принадлежность $f_\rho(x)$ некоторому классу $\mathcal{H}$, а также условия на равномерную по x скорость убывания вероятностных хвостов по y.

В ряде случаев, когда известно поведение классических в *теории приближений* характеристик класса $\mathcal{H}$, таких как *энтропийные числа* и *Колмогоровские поперечники*, нами получены экспоненциальные оценки убывания (2) при согласованном поведении m и t. Они являются обобщением результатов, полученных в ряде недавних работ, основные из которых приведены в конце страницы.[11,12]

Экспоненциальные оценки в первой части выражены в терминах *комбинаторных характеристик* семейства множеств в $\mathbb{R}^d$ или, что то же самое, семейства функций на $\mathbb{R}^d$ со значениями в $\{0,1\}$, в то время как результаты второй части формулируются в терминах *метрических характеристик* классов действительнозначных функций на $\mathbb{R}^d$. Известно, что существует тесная связь между этими характеристиками, и наши результаты лишний раз подтверждают это.

Цель работы.

Цель диссертационной работы — получить обобщения классических результатов Колмогорова, Смирнова, Кифера, Вапника и Червоненкиса о количественном поведении отклонений эмпирических частот событий от вероятностей, а также показать, что в ряде случае существуют оптимальные оценщики неизвестной функции регрессии и доказать, что имеют место оптимальные вероятностные границы для их разности в нескольких случаях, когда класс гипотез задаётся в классических терминах теории приближений.

[11] F.Cucker, S.Smale, *On the mathematical foundations of learning,* Bulletin of AMS, **39** (2001), p.1-49.
[12] V. Temlyakov. *Approximation in Learning Theory.* IMI Preprints **05** (2005), p.1-43.

Научная новизна.

В работе получены как новые результаты, так и новые более простые доказательства ряда известных ранее утверждений. Основные результаты диссертации состоят в следующем.

1. Получена новая (неасимптотическая) оценка скорости сходимости эмпирических частот к вероятностям в терминах функции информации и размерности Вапника-Червоненкиса, обобщающая результаты Колмогорова, Смирнова, Кифера, Хёфдинга, Чернова, Вапника, Червоненкиса и Девроя.

2. Получено более простое доказательство результата Талагранда, обобщающее одномерное неравенство Чернова, а также дан частичный ответ к гипотезе Талагранда о поведении супремумов важного класса эмпирических процессов.

3. Получен ряд оценок на поведение вероятностных хвостов отклонения функции-оценщика от функции регрессии в случае, когда класс гипотез задаётся в терминах энтропийных чисел и Колмогоровских поперечников.

4. Построен класс универсальных оценщиков функции регрессии и приведены оценки вероятностных хвостов их отклонения.

Методы исследования.

В диссертации используются методы теории эмпирических процессов, теории приближений и результаты о концентрации вероятностной меры. А именно, используется техника симметризации, неравенства Чернова, Бернштейна, Талагранда, неравенства Карла, формула Стирлинга, моментные неравенства и лемма Варадана, информация Кульбака-Лейблера и ряд неравенств, связывающих метрические и комбинаторные характеристики функционального пространства.

Теоретическая и практическая ценность.

Диссертация носит теоретический характер. Её методы и результаты могут найти применение в вероятностной теории распознавания образов и в теории оценивания функции регрессии.

Апробация работы.

Результаты диссертации в разное время докладывались и обсуждались на следующих семинарах механико-математического факультета МГУ: «Непараметрическая статистика и временные ряды» (руководители - д.ф.-м.н., профессор Ю.Н. Тюрин, д.ф.-м.н., профессор В.Н. Тутубалин, к.ф.-м.н., доцент М.В. Болдин); «Большой Кафедральный Семинар кафедры теории вероятностей»; «Статистическая теория обучений и её применения» (руководитель - д.ф.-м.н., профессор С.В. Конягин); «Теория приближений и приложения» (руководители - д.ф.-м.н., профессор, чл.-корр. РАН Б.С. Кашин, д.ф.-м.н., профессор С.В. Конягин);

а также представлялись на следующих международных конференциях: «25-я Европейская Конференция по Математической Статистике» (Осло, 2005); «Математические основы теории обучений II» (Париж, 2006); «26-я Европейская Конференция по Математической Статистике» (Торунь, 2006); Российско-Скандинавский Симпозиум «Теоретическая и Прикладная Теория Вероятностей» (Петрозаводск, 2006).

Публикации.

Результаты диссертации опубликованы в 6 работах, список которых приведён в конце автореферата, см. [1-6].

Структура диссертации.

Диссертация состоит из введения, двух глав (разбитых на разделы), заключения и списка литературы, насчитывающего 34 наименования. Общий объём диссертации — 73 страницы.

Краткое содержание диссертации

Во **введении** приведён краткий исторический обзор по тематике работы, изложены цели и методы исследования, а также структура диссертации.

Пусть $X_1, \ldots, X_m$ — независимые одинаково распределённые в сответствии с некторой вероятностной мерой $\mathsf{P}(\cdot)$ случайные величины в Евклидовом про-

странстве $\mathbb{R}^d$. Обозначим через m *длину выборки*, а через $\mathsf{P}_m(\cdot)$ — соответствующую *эмпирическую меру*. **Первая глава** посвящена изучению поведения вероятностных хвостов эмпирического процесса

$$\sup_{A\in\mathcal{A}} |\,\mathsf{P}_m(A) - \mathsf{P}(A)|,$$

для случая, когда $\mathcal{A}$ — семейство Борелевских множеств в $\mathbb{R}^d$ с конечной *размерностью Вапника-Червоненкиса* $V = V(\mathcal{A})$. В предположении, что множество значений мер $\{\mathsf{P}(A),\, A \in \mathcal{A}\}$ отделено от *некоторой окрестности* $1/2$, мы улучшаем ряд известных классических оценок. Введём необходимые определения.

Определение. *Наибольшее целое число V, для которого существует набор из V точек $x_1, \ldots, x_V$, удовлетворяющих свойству*

$$\#\Big\{\{x_1, \ldots, x_V\} \cap A :\ \ A \in \mathcal{A}\Big\} = 2^V,$$

называется размерностью Вапника-Червоненкиса семейства $\mathcal{A}$.

Если не оговаривается противное, везде ниже под $V = V(\mathcal{A})$ мы понимаем *конечную размерность Вапника-Червоненкиса* набора множеств $\mathcal{A}$.

Определение. *Для вещественных чисел $p, q \in (0,1)$ расходимостью Кульбака-Лейблера $H(p,q)$ называют следующее выражение:*

$$H(p,q) = p \cdot \ln\frac{p}{q} + (1-p)\cdot \ln\frac{1-p}{1-q}.$$

Определение. *Критическим значением $p_{\mathcal{A}}$ для системы множеств $\mathcal{A}$ и меры $\mathsf{P}(\cdot)$ назовём ближайшую к $1/2$ из двух величин p_+ и p_-, где*

$$p_+ = \inf_{A\in\mathcal{A},\, \mathsf{P}(A)>1/2} \mathsf{P}(A), \quad p_- = \sup_{A\in\mathcal{A},\, \mathsf{P}(A)<1/2} \mathsf{P}(A).$$

Неформально можно рассматривать $p_{\mathcal{A}}$ как ближайшее к $1/2$ значение меры $\mathsf{P}(A)$ на множествах $A \in \mathcal{A}$.

Теорема 1. *Пусть $p_{\mathcal{A}}$ — критическое значение для семейства множеств $\mathcal{A}$ с конечной размерностью Вапника-Червоненкиса V. Тогда для некоторых положительных констант $M = M(p_{\mathcal{A}}, V)$, $K = K(V)$ и $t_0 = t_0(p_{\mathcal{A}}, V)$ при $m\cdot t^2 > M$ и $t < t_0$ выполняется оценка*

$$\mathsf{Pr}\Big\{\sup_{A\in\mathcal{A}} |\mathsf{P}_m(A) - \mathsf{P}(A)| > t\Big\} \le K\cdot(mt^2)^{2V+4}\cdot\exp\Big\{-m\cdot\min\Big(H(p_{\mathcal{A}}+t, p_{\mathcal{A}}), H(p_{\mathcal{A}}-t, p_{\mathcal{A}})\Big)\Big\}.$$

Прокомментируем наш результат. *Размерность Вапника-Червоненкиса* V является чисто комбинаторной характеристикой системы Борелевских множеств $\mathcal{A}$ и не зависит ни от какого распределения вероятностей. Она описывает *разнообразие* $\mathcal{A}$. Другими словами, V равно наибольшему возможному числу точек в $\mathbb{R}^d$, таких что для любого (из 2^V) подмножества этих точек найдётся представитель $A \in \mathcal{A}$, пересекающийся с точками в точности по этому подмножеству. Если такое множество из V точек можно найти для любого $V \in \mathbb{N}$, то *размерность Вапника-Червоненкиса* считается равной бесконечности. Отметим, что для бесконечной системы множеств $\mathcal{A}$ её размерность $V(\mathcal{A})$ может быть как конечной, так и бесконечной. Рассмотрим несколько важных примеров.

Для набора *полупространств* в пространстве $\mathbb{R}^d$

$$\mathcal{A} = \{\{x :< w, x > +w_0 \geq 0\} : w \in \mathbb{R}^d, w_0 \in \mathbb{R}\} \text{ имеем } V(\mathcal{A}) = d + 1,$$

для семейства *отрицательных ортантов*

$$\mathcal{A} = \{(-\infty, x_1] \times \cdots \times (-\infty, x_d] : (x_1, \ldots, x_d) \in \mathbb{R}^d\} \text{ верно } V(\mathcal{A}) = d.$$

Семейство всех *выпуклых многогранников* в $\mathbb{R}^d$ (при $d \geq 2$) даёт пример набора $\mathcal{A}$ с бесконечной размерностью V.

Результат **Теоремы 1** обобщает классические неасимптотические оценки Колмогорова, Смирнова и Кифера (Kiefer) на скорость сходимости эмпирической функции распределения к настоящей, а также ряд результатов Вапника, Червоненкиса и других авторов о равномерной сходимости частот к вероятностям. Вапник и Червоненкис первыми доказали экспоненциальные вероятностные оценки для эмпирического процесса $\sup_{A\in\mathcal{A}} |\mathsf{P}_m(A) - \mathsf{P}(A)|$ в общем случае когда $V(\mathcal{A})$ конечно [13, 14].

Теорема (**Вапник-Червоненкис**). *Пусть* $\mathcal{A}$ *— класс Борелевских множеств в* $\mathbb{R}^d$ *с конечной* $V = V(\mathcal{A})$. *Тогда для* $mt^2 > 2$ *верно неравенство:*

$$\mathsf{Pr}\Big\{\sup_{A\in\mathcal{A}} |\,\mathsf{P}_m(A) - \mathsf{P}(A)| > t\Big\} \leq 4\Big(\frac{2em}{V}\Big)^V \cdot e^{-mt^2/8}.$$

[13] В. Вапник, А. Червоненкис. *О равномерной сходимости относительных частот событий к их вероятностям.* Теория вероятностей и приложения, **16**, No. 2 (1971), стр. 264-280.

[14] В. Вапник, А. Червоненкис. *Необходимые и достаточные условия для равномерной сходимости средних к математическим ожиданиям.* Теория вероятностей и приложения, **26**, No. 3 (1981), стр. 532-553.

Экспоненциальный по m множитель в этой оценке описывает вероятностный хвост отклонения для фиксированного множества $A \in \mathcal{A}$, а полиномиальный член характеризует разнообразие семейства $\mathcal{A}$. Отметим, что в упомянутых выше классических результатах и их обобщениях не делалось каких-то специальных предположений на неизвестную меру $\mathsf{P}(\cdot)$.

Такие оценки принято называть *свободными от распределения*. Они также известны в литературе как *границы Вапника-Червоненкиса.* Все такие результаты и их обобщения в общем виде можно записать так:

$$\mathsf{Pr}\{\sup_{A\in\mathcal{A}} |\mathsf{P}_m(A) - \mathsf{P}(A)| > t\} \le K(t,V)\cdot(mt^2)^v \cdot e^{-m\cdot\phi(t)}, \text{ когда } mt^2 > C, \quad (3)$$

Здесь C — некоторая константа, не зависящая от t, v — функция от V, а $\phi(t) = \gamma t^2$, где $\gamma \in (0,2]$.

Нескольким авторам удалось улучшить константы из первых оценок Вапника и Червоненкиса. Отметим здесь лишь, что Деврой (L.Devroye)[15] получил оптимальный множитель e^{-2mt^2}, а Талагранд (M. Talagrand)[16] — также и неулучшаемый показатель $v = V - 1/2$. Наше предположение о том, что множество $\{\mathsf{P}(A) : A \in \mathcal{A}\}$ отделено от некоторой окрестности $1/2$, позволяет улучшить экспоненциальный множитель в оценках вида (3) и взять в качестве $\phi(t)$ выражение

$$\phi(t) = \min\big(H(p_{\mathcal{A}} + t, p_{\mathcal{A}}), H(p_{\mathcal{A}} - t, p_{\mathcal{A}})\big). \quad (4)$$

Новый вид оценки обобщает классический результат Чернова (Chernoff) для случая отдельного множества $A \in \mathcal{A}$.

Теорема (Чернов).

$$\mathsf{Pr}\big\{(\mathsf{P}_m(A) - \mathsf{P}(A)) > t\big\} \le \exp\big\{-m\cdot H\big(\mathsf{P}(A)+t), \mathsf{P}(t)\big)\big\}, \quad \forall A \in \mathcal{A}.$$

Талагранд выдвинул гипотезу о том, что наряду с $\phi(t)$ из выражения (4) в оценке (3) может быть сохранён оптимальный *полиномиальный множитель* $(mt^2)^{V-1/2}$. Таким образом, **Теорема 1** может рассматриваться как частичное решение *гипотезы Талагранда.* Условия отделимости $\{\mathsf{P}(A),\, A \in \mathcal{A}\}$ от $1/2$

[15] L. Devroye, *Bounds for the Uniform Deviation of Empirical Measures,* Journal of Multivariate Analysis, **12** (1982), p. 72-79.

[16] M. Talagrand, *Sharper Bounds for Gaussian and Empirical Processes,* The Annals of Probability, **22**, No 1 (1994), p. 28-76.

означает, что $p_{\mathcal{A}} \neq 1/2$. А значит для некоторого $t_0 > 0$ множества $\{p_{\mathcal{A}} + t\}$, $\{p_{\mathcal{A}} - t\}$, $0 < t < t_0$ не пересекаются с некоторой окрестностью $1/2$. Из свойств *расходимости Кульбака-Лейблера* следует, что для таких t экспоненциальный показатель в **Теореме 1** улучшен по сравнению с оптимальным в случае предположения *свободы от распределения*.

Талагранд показал, что выполняется следующая разновидность оценки (3), которую мы приводим для частного случая:

Теорема (**Талагранд**[16]). *Пусть $\mathcal{A}$ — ограниченное семейство измеримых множеств в $\mathbb{R}^d$ с конечной $V = V(\mathcal{A})$. Пусть также $\mathsf{P}(\cdot)$ — некоторая Борелевская вероятностная мера на $\mathbb{R}^d$, а $p_{\mathcal{A}}$ — критическое значение для $\mathcal{A}$ и $\mathsf{P}(\cdot)$. Тогда если $p_{\mathcal{A}} \neq 1/2$, то для любого $\beta \in (0,1)$ существуют положительные числа $M(\beta, p_{\mathcal{A}}, V)$, $K(V)$ и $t_0(\beta, p_{\mathcal{A}}, V)$, такие что при $m \cdot t^2 > M(\beta, p, V)$ и $t < t_0(\beta, p, V)$ верно неравенство:*

$$\mathsf{Pr}\Big\{\sup_{A\in\mathcal{A}} |\mathsf{P}_m(A) - \mathsf{P}(A)| > t\Big\} \leq K(V) \cdot (mt^2)^V e^{-m(1-\beta)\cdot\min\left(H(p_{\mathcal{A}}+t,p_{\mathcal{A}}),H(p_{\mathcal{A}}-t,p_{\mathcal{A}})\right)}.$$

Эта теорема нужна нам как вспомогательная для нашего основного результата. Мы приводим новое простое её доказательство. Отметим, что наибольшую сложность здесь представляет переход от множителя $m(1-\beta)$ в показателе экспоненты просто к m, чтобы получилось прямое обобщение неравенств Чебышева и Чернова.

Поясним на примере, что объект, который мы оцениваем в **Теореме 1**, естественным образом возникает в *теории обучений*. Для этого рассмотрим *классическую задачу двухклассового распознавания* d-мерных объектов. Предположим, что заранее известен класс *решающих правил* $\mathcal{H}$, т.е. класс функций $f : \mathbb{R}^d \to \{0,1\}$, а также *обучающая выборка* $\mathbf{z} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ длины m. Здесь $x_j \in \mathbb{R}^d$, а $y_j \in \{0,1\}$. Будем понимать под *неизвестным законом*, в соответствии с которым получены или собраны наблюдения $z_j = (x_j, y_j)$, некоторое *распределение вероятностей* $\mathsf{P}(\cdot)$ на произведении $\mathbb{R}^d \times \{0,1\}$. Общая задача состоит в том, чтобы на основе выборки $\mathbf{z}$ длины m из семейства $\mathcal{H}$ выбрать в некотором смысле *наилучшее* решающее правило $f_{\mathbf{z}} \in \mathcal{H}$ с тем, чтобы использовать его для распознавания любых точек из $\mathbb{R}^d$. Например, выбрать решающее правило $f_{\mathbf{z}}$, которое по входному x-у как можно чаще выдаёт "правильный" y.

Естественный подход при выборе *оптимального* $f : \mathbb{R}^d \to \{0,1\}$, $f \in \mathcal{H}$ состоит в минимизации *эмпирической ошибки*, определяемой как

$$\frac{1}{m}\sum_{j=1}^{m} \mathbf{I}_{\{f(x_j)\neq y_j\}}.$$

Обозначим функцию, или как ещё говорят, "распознаватель", на которой достигается этот минимум, как $f_{\mathbf{z}}$. *Ошибкой правила* $f \in \mathcal{H}$ назовём P-меру множества всех таких пар $(x,y) \in \mathbb{R}^d \times \{0,1\}$, для которых $f(x) \neq y$. Возникает естественный вопрос о том, насколько *ошибка* такого "распознавателя" $f_{\mathbf{z}}$ близка к *минимально возможной ошибке* внутри класса $\mathcal{H}$. Пусть $\mathsf{P}_m(\cdot)$ — соответствующее выборке $\mathbf{z}$ эмпирическое рапределение. Тогда можно показать, что модуль интересующей нас разности оценивается сверху как

$$2 \cdot \sup_{A\in\mathcal{A}} |\mathsf{P}_m(A) - \mathsf{P}(A)|,$$

где семейство $\mathcal{A}$ представляет из себя следующий набор множеств:

$$\big\{\{x : f(x) = 1\} \times \{0\}\big\} \bigcup \big\{\{x : f(x) = 0\} \times \{1\}\big\}, \quad f \in \mathcal{H}.$$

Для фиксированного множества A и для любой непрерывной меры $\mathsf{P}(\cdot)$ по закону больших чисел имеет место сходимость почти наверное $\mathsf{P}_m(A) - \mathsf{P}(A) \to 0$ при $m \to \infty$. Более того, неравенство Чебышева и Хёфдинга даёт следующую оценку на скорость сходимости:

$$\mathsf{Pr}\big\{|\mathsf{P}_m(A) - \mathsf{P}(A)| > t\big\} \leq 2e^{-2m\cdot t^2}.$$

Если семейство $\mathcal{A}$ конечно, то естественно получаем обобщение предыдущего неравенства:

$$\mathsf{Pr}\big\{\sup_{A\in\mathcal{A}} |\mathsf{P}_m(A) - \mathsf{P}(A)| > t\big\} \leq 2|\mathcal{A}| \cdot e^{-2m\cdot t^2}.$$

Однако если набор $\mathcal{A}$ имеет бесконечную мощность, как это и бывает во многих интересных случаях, имеющих прикладное значение, задача становится гораздо более сложной.

Вторая часть диссертации посвящена задаче оценивания функции регрессии. Пусть $X \subset \mathbb{R}^d$ и $Y \subset \mathbb{R}$ — Борелевские множества, и на их произведении $Z = X \times Y$ определена Борелевская вероятностная мера ρ. Обозначим через $f_\rho : X \to Y$ *функцию регрессии* y на x, т.е. $f_\rho(x) = \mathsf{E}\{y|x\}$.

Мы изучаем задачу оценивания функции $f_\rho(x)$ по *конечной* ρ-выборке $\mathbf{z} = \{z_1, \ldots, z_m\} : z_i = (x_i, y_i), i = 1, \ldots, m$. Иногда эту задачу называют *обучением функции регрессии.*

Функцию, построенную по выборке $\mathbf{z}$ и оценивающую $f_\rho(x)$, будем обозначать как $f_\mathbf{z} : X \to Y$ и называть её *функцией-оценщиком.* Пусть ρ_X — проекция меры ρ на X. Естественной характеристикой *качества оценивания* функции f_ρ является скорость убывания вероятностных хвостов (при стремлении $\epsilon \to 0$ и $m \to \infty$)

$$\mathsf{Pr}\{\|f_\mathbf{z} - f_\rho\|^2_{L_2(X,\,\rho_X)} > \epsilon\}. \tag{5}$$

Здесь и везде ниже под $\mathsf{Pr}(\cdot)$ понимается ρ^m-вероятность на произведении Z^m, где $Z = X \times Y$. На неизвестную меру ρ мы накладываем следующие условия (6) и (7):

$$\exists\, C_1, C_2 > 0, \text{ такие что } \forall t > 0 : \ \rho\{|y| > t\} < C_1 \cdot \exp\{-C_2 \cdot t^2\}, \tag{6}$$

$$f_\rho(x) \in \mathcal{H}, \text{ где } \mathcal{H} \text{ — компактное подмножество в пространстве } C(X). \tag{7}$$

Свойство убывания вероятностных хвостов (6) также известно как *равномерная сабгауссовость* по y. Для функции $f : X \to Y$, $f \in L_2(X, \rho_X)$ её *ошибкой* назовём выражение

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 \cdot d\rho.$$

Известно, что минимум *ошибки* достигается на функции регрессии $f_\rho(x)$:

$$\mathcal{E}(f_\rho) = \inf_{f \in L_2(\rho_X)} \mathcal{E}(f), \text{ при этом для } f \in L_2(X, \rho_X) : \mathcal{E}(f) - \mathcal{E}(f_\rho) = ||f - f_\rho||^2_{L_2(X,\rho_X)},$$

а значит для любого оценщика $f_\mathbf{z} \in L_2(X, \rho_X)$ вероятностные хвосты (5) можно записать как

$$\mathsf{Pr}\{\|f_\mathbf{z} - f_\rho\|^2_{L_2(X,\,\rho_X)} > \epsilon\} = \mathsf{Pr}\{\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho) > \epsilon\}.$$

Определим *эмпирическую ошибку* $\mathcal{E}_\mathbf{z}(f)$ как

$$\mathcal{E}_\mathbf{z}(f) = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

и обозначим через $f_{\mathcal{H},\mathbf{z}}$ *минимум эмпирической ошибки* для пространства $\mathcal{H}$:

$$f_{\mathcal{H},\mathbf{z}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}_\mathbf{z}(f).$$

Обозначим также через $N(\mathcal{H}, \epsilon)$ мощность ϵ-сети для $\mathcal{H}$ в пространстве $\mathcal{C}(X)$.

Теорема 2. *Пусть мера ρ удовлетворяет условиям* (6) *и* (7). *Тогда найдутся положительные константы $C_1(\mathcal{H},\rho)$ и $C_2(\mathcal{H},\rho)$, такие что для всех $\epsilon > 0$ справедливо неравенство:*

$$\mathsf{Pr}\{\mathcal{E}(f_{\mathcal{H},\boldsymbol{z}}) - \mathcal{E}(f_\rho) > \epsilon\} \le 2\,N(\mathcal{H}, \epsilon/C_1(\mathcal{H},\rho)) \cdot \exp\{-C_2(\mathcal{H},\rho)\cdot m\epsilon^2\}.$$

Определение. *Энтропийным числом $\epsilon_n(\mathcal{H})$ порядка n для функционального класса $\mathcal{H}$ в пространстве $\mathcal{C}(X)$ называется следующая характеристика:*

$$\epsilon_n(\mathcal{H}) := \inf\big\{\epsilon : \exists f_1,\dots,f_{2^n} \in \mathcal{H} : \mathcal{H} \subset \cup_{j=1}^{2^n}(f_j + \epsilon U(\mathcal{C}(X)))\big\},$$

где $U(\mathcal{C}(X))$ — единичный шар в пространстве $\mathcal{C}(X)$.

Для случая, когда $\mathcal{H}$ задаётся скоростью убывания своих *энтропийных чисел*, мы имеем следующую оценку для вероятностных хвостов (5):

Теорема 3. *Пусть мера ρ удовлетворяет условиям* (6) *и* (7) *и для некоторых $r > 0$, $C > 0$ и всех натуральных чисел $n \in \mathbb{N}$ выполняются неравенства $\epsilon_n(\mathcal{H}) \le Cn^{-r}$. Тогда существуют положительные константы $C_1(r,\rho), C_2(r,\rho) > 0$, такие что*

$$\mathsf{Pr}\{\mathcal{E}(f_{\mathcal{H},\boldsymbol{z}}) - \mathcal{E}(f_\rho) \ge \epsilon\} \le e^{-C_1\cdot m\epsilon^2}, \quad \text{как только} \quad \epsilon \cdot m^{\frac{r}{1+2r}} > C_2.$$

Для доказательства предыдущих теорем мы используем несколько промежуточных результатов. Пусть для меры ρ выполняются условия (6) и (7). Тогда верны утверждения:

Лемма 4. *Пусть функция $f \in \mathcal{H}$ имеет конечную ошибку $\mathcal{E}(f)$. Тогда для некоторой константы $C = C(\mathcal{H},\rho) > 0$ и для всех $\epsilon > 0$ имеет место оценка:*

$$\mathsf{Pr}\{\mathcal{E}(f) - \mathcal{E}_{\boldsymbol{z}}(f) > \epsilon\} \le 2\exp\{-C\cdot m\epsilon^2\}.$$

Лемма 5. *Существуют константы $C_j(\mathcal{H},\rho) > 0$, $j = 1,2$, такие что для $\epsilon > 0$ верно:*

$$\mathsf{Pr}\{\sup_{f\in\mathcal{H}} |\mathcal{E}(f) - \mathcal{E}_{\boldsymbol{z}}(f)| > \epsilon\} \le 2\,N(\mathcal{H}, \epsilon/C_1(\mathcal{H},\rho)) \cdot \exp\{-C_2(\mathcal{H},\rho)\cdot m\epsilon^2\}.$$

Многие авторы изучали вопросы качества и оптимальности оценивания f_ρ в случае, когда на меру ρ наложены более простые условия ограниченности вида

$|y| < M$ ρ-п.н. для некоторого фиксированного $M > 0$, см.[17,18,19]. Наш основной вклад заключается в перенесении части их результатов на случай неограниченных y-ов.

Определение. *Колмогоровским поперечником порядка n для семейства функций $\mathcal{H} \subset \mathcal{C}(X)$ называется число $d_n(\mathcal{H}, \mathcal{C}(X))$, определяемое как*

$$d_n(\mathcal{H}, \mathcal{C}(X)) = \inf_{L_n} \sup_{f \in \mathcal{H}} \inf_{g \in L_n} \|f - g\|_{\mathcal{C}(X)},$$

где $\inf_{L_n}$ берётся по всем n-мерным линейными подпространствам в $\mathcal{C}(X)$.

Колмогоровские поперечники часто используются в *теории приближений* и описывают наилучшие приближения n-мерными линейными подпространствами. Известно, что если $\mathcal{H} \subset \mathcal{C}(X)$ — компактное множество, то из условия на убывание *Колмогоровских поперечников* вида

$$d_n(\mathcal{H}, C(X)) \leq C_1 \cdot n^{-r}, \quad n = 1, 2, \ldots \tag{8}$$

следуют неравенства для *энтропийных чисел*[20]:

$$\epsilon_n(\mathcal{H}, C(X)) \leq C_2 \cdot n^{-r}, \quad n = 1, 2, \ldots \tag{9}$$

Наш следующий результат усиливает оценки **Теоремы 3**:

Теорема 6. *Пусть мера ρ удовлетворяет условиям* (6) *и* (7). *Предположим также, что существуют такие $C, r > 0$, что выполняются неравенства:*

$$d_n(\mathcal{H}, \mathcal{C}(X)) \leq C \cdot n^{-r}, \quad n = 1, 2, \ldots$$

Тогда существует такой оценщик $f_{\boldsymbol{z}}$, что для некоторых констант $C_1(\mathcal{H}, \rho)$, $C_2(\mathcal{H}, \rho)$ выполняются неравенства:

$$\mathsf{Pr}\{\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(f_\rho) \geq \epsilon\} \leq e^{-C_1 \cdot m\epsilon^2}, \quad \textit{как только} \quad \epsilon \cdot m^{\frac{r}{1+2r}} \cdot \left(\frac{m}{\ln\, m}\right)^{\frac{r}{1+2r}} > C_2.$$

[17] F.Cucker, S.Smale, *On the mathematical foundations of learning,* Bulletin of AMS, **39** (2001), p.1-49.

[18] R.DeVore, G.Kerkyacharian, D.Picard, V.Temlyakov, *On Mathematical Methods of Learning.* IMI Preprints, **10** (2004), p.1-24.

[19] L.Gyorfi, M.Kohler, A.Krzyzak, H.Walk, *A Distribution-Free Theory of Nonparametric Regression,* Springer Series in Statistics (2002).

[20] B.Carl, *Entropy numbers, s-numbers, and eigenvalue problems*, J. Funct. Anal., **41** (1981), p.290-306.

Опишем следствие из этой теоремы. Пусть $p > 0$ действительное число. Определим класс *гладких порядка p действительнозначных функций* $\mathcal{H}_p^d$ на единичном кубе в $\mathbb{R}^d$ следующим образом.

Определение. *Представим действительное число* $p > 0$ *в виде* $p = k + \beta$, *где* $k \in \mathbb{N}$, *и* $0 < \beta \leq 1$ *и определим класс* $\mathcal{H}_p^d$ *функций* $f : [0,1]^d \to \mathbb{R}$ *следующими условиями. Скажем, что* $f \in \mathcal{H}_p^d$ *если и только если у* f *существуют все частные производные порядка* k, *все они удолетворяют условию Гёльдера с параметром* β *на множестве* $[0,1]^d$, *и* $\|f\|_{\mathcal{C}(X)} \leq 1$.

Известно, что для пространства $\mathcal{H} = \mathcal{H}_p^d$ *энтропийные числа* и *Колмогоровские поперечники* удовлетворяют неравенствам (8) и (9) с показателем $r = p/d$, см.[21], что даёт нам экспоненциальные оценки для случая, когда *пространство гипотез* для $f_\rho(x)$ представляет хорошо изученный *класс гладких функций* $\mathcal{H}_p^d$. Отметим, что оценщик из **Теоремы 6** представляет из себя более сложную конструкцию, чем рассматриваемый ранее $f_{\mathcal{H},z}$, на котором достигается *минимум эмпирической ошибки* для класса $\mathcal{H}$.

Последняя часть наших результатов касается построения *универсальных оценщиков*.

Определение. *Для подпространства* L *в пространстве непрерывных функций* $\mathcal{C}(X)$. *определим расстояние до функционального класса* $\mathcal{H}$ *как*

$$d(\mathcal{H}, L) := \sup_{f \in \mathcal{H}} \inf_{g \in L} \|f - g\|_\infty. \tag{10}$$

Пусть $\mathcal{L} = \{L_n\}_{n=1}^\infty$ — последовательность n-мерных ($n = 1, 2, \dots$) подпространств в $\mathcal{C}(X)$, и $0 < \alpha \leq \beta < \infty$ — действительные числа. Пусть также C, D — некоторые фиксированные положительные константы и $r \in [\alpha, \beta]$. Обозначим через $\mathcal{H}_r$ произвольное *ограниченное семейство* функций в пространстве $\mathcal{C}(X)$, удовлетоворяющее свойствам:

$$d(\mathcal{H}_r, L_n) \leq C \cdot n^{-r}, n = 1, 2, \dots; \quad \sup_{f \in \mathcal{H}_r} \|f\|_{\mathcal{C}(X)} \leq D.$$

Наконец обозначим

$$\mathcal{H}(\alpha, \beta) = \{\mathcal{H}_r,\, r \in [\alpha, \beta]\},$$

[21] В.Тихомиров, *Теория приближений.* Совр. проблемы математики. Фундаментальные направления, **14** (1987), (Итоги науки и техн.), ВИНИТИ АН СССР.

Такое условие возникло в работе[18] и является близким к естественным в *теории приближений* условиям на убывание *Колмогоровских поперечников*. Справедлива

Теорема 7. *Предположим, что для меры ρ выполняются условия* (6) *и* (7), *где $\mathcal{H}$ имеет вид $\mathcal{H} = \mathcal{H}(\alpha, \beta)$ и $0 < \alpha \le \beta < \infty$. Тогда существует универсальный оценщик $f_{\boldsymbol{z}}$ и такие положительные константы $C_1(\mathcal{H}, \rho, \alpha), C_2(\mathcal{H}, \rho, \alpha)$, что если $f_\rho \in \mathcal{H}_r \subset \mathcal{H}(\alpha, \beta)$, для некоторого $r \in [\alpha, \beta]$, то верна оценка*

$$\mathsf{Pr}\{\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(f_\rho) \ge \epsilon\} \le e^{-C_1 \cdot m\epsilon^2}, \quad \textit{как только} \quad \epsilon \cdot m^{\frac{2r}{1+2r}} \cdot \left(\frac{m}{\ln\, m}\right)^{\frac{r}{1+2r}} > C_2.$$

Оценка этой теоремы слабее, чем предыдущий результат, однако и пространство $\mathcal{H}$ теперь гораздо шире. Существование универсального семейства конечномерных подпространств $\mathcal{L} = \{L_n\}$, оптимального в смысле расстояния (10) к *Колмогоровским поперечникам* для всех $\mathcal{H}_r$, $r \in [\alpha, \beta]$ — непростой вопрос из *теории приближений*. Однако известно, что для случая $d = 1$ и пространств $\mathcal{H}_p^1$ в качестве такого семейства $\mathcal{L} = \{L_n\}$ можно взять конечномерные *пространства тригонометрических полиномов*, см.[21] Отсюда получаем такое следствие:

Следствие. *Предположим, что для меры ρ выполняются условия* (6) *и* (7). *Пусть также $0 < \alpha \le \beta < \infty$ — фиксированные числа и $\mathcal{H}$ имеет вид $\mathcal{H}(\alpha, \beta) = \{\mathcal{H}_r^1, \alpha \le r \le \beta\}$. Тогда существует универсальный для всех r оценщик $f_{\boldsymbol{z}}$ и такие положительные константы $C_1(\mathcal{H}, \rho, \alpha), C_2(\mathcal{H}, \rho, \alpha)$, что если $f_\rho \in \mathcal{H}_r$, то верно неравенство:*

$$\mathsf{Pr}\{\mathcal{E}(f_{\boldsymbol{z}}) - \mathcal{E}(f_\rho) \ge \epsilon\} \le e^{-C_1 \cdot m\epsilon^2}, \quad \textit{как только} \quad \epsilon \cdot m^{\frac{2r}{1+2r}} \cdot \left(\frac{m}{\ln\, m}\right)^{\frac{r}{1+2r}} > C_2.$$

Известно, что $\mathcal{H}_r^1 \subset \mathcal{H}_s^1$ при $0 < s < r$. Отсюда ясно, что последнее условие на ϵ и m в предыдущем утверждении можно всегда заменить на такое: $\epsilon \cdot m^{\frac{2\alpha}{1+2\alpha}} \cdot \left(\frac{m}{\ln\, m}\right)^{\frac{\alpha}{1+2\alpha}} > C_2$.

Автор глубоко благодарен своему научному руководителю профессору Ю.Н. Тюрину за постоянное внимание к работе, а также Е.Д. Лившицу, профессору В.Н. Темлякову и профессору С.В. Конягину за многочисленные полезные и плодотворные обсуждения.

Список литературы

[1] *Рафиков Е. Г.* Оценивание функции регрессии в случае неограниченных откликов. // Математические Заметки, 79 (2006), No 6, стр. 231-235.

[2] *Рафиков Е. Г.* О скорости сходимости частот к вероятностям. // Обозрение прикладной и промышленной математики, 13 (2006), No 3, стр. 632-643.

[3] *Рафиков Е. Г.* Обучение функции регрессии для неограниченных откликов. // Депонировано в ВИНИТИ РАН. 2006. N1621-B2006.

[4] *Rafikov E.* About Regression Function Estimation in the Case of Unbounded Responses. // Тезисы докладов конференции «26-я Европейская Конференция по Математической Статистике», Торунь, 2006, стр. 345-346.

[5] *Rafikov E., Livshitz E.* About Universal Estimators in Learning Theory. // Тезисы докладов конференции «25-я Европейская Конференция по Математической Статистике», Осло, 2005, стр. 237-238.

[6] *Rafikov E.* About Rates of Convergence of Empirical Frequencies to the True Probabilities. // Тезисы докладов Российско-Скандинавский Симпозиума «Теоретическая и Прикладная Теория Вероятностей», Петрозаводск, 2006, стр. 41-43.