GENERAL PROCEDURE FOR SELECTING LINEAR ESTIMATORS

Alexander Goldenshluger, Oleg Lepski

Vol. 57, No. 2, pp. 1–608

GENERAL PROCEDURE FOR SELECTING LINEAR ESTIMATORS *

GOLDENSHLUGER A.V.† AND LEPSKI O.V.‡

(*Translated by*)

Аннотация. In the general statistical experiment model we propose a procedure for selecting an estimator from a given family of linear estimators. We derive an upper bound on the risk of the selected estimator and demontrate how this result can be used in order to construct minimax and adaptive minimax estimators in specific nonparametric estimation problems.

Key words. statistical experiment, linear estimators, oracle inequality, adaptive minimax estimation, majorant.

1. Introduction. Let $(\mathcal{X}^{(n)}, \mathcal{B}^{(n)}, \mathbf{P}_f^{(n)}, f \in \mathbf{F})$, $n \in \mathbf{N}^*$ be a sequence of statistical experiments generated by the observation $X^{(n)}$. Here $\mathcal{B}^{(n)}$ is the σ-algebra, generated by the random element $X^{(n)} \in \mathcal{X}^{(n)}$, and distribution of $X^{(n)}$ belongs to the family of the probability measures $(\mathbf{P}_f^{(n)}, f \in \mathbf{F})$. Let $(D, \mathfrak{D}, \nu)$ be a measurable space, and let $\mathbf{F}_0$ be a given set of measurable functions $f\colon\ D \to \mathbf{R}$. In this paper we will suppose that $\mathbf{F} \subseteq \mathbf{F}_0$; spaces of continuous and bounded functions and $\mathbf{L}_2(D, \nu)$ are classical examples of the set $\mathbf{F}$.

Consider the problem of estimating function f on the basis of the observation $X^{(n)}$. By an estimator of function f we mean any $\mathcal{B}^{(n)}$-measurable map $\widehat{f}\colon\ \mathcal{X}^{(n)} \to \mathbf{F}_0$. With any estimator $\widehat{f}$ we associated the risk

$$\mathcal{R}_\ell^{(n)}[\widehat{f}; f] = \left\{\mathbf{E}_f^{(n)}[\ell(\widehat{f} - f)]^q\right\}^{1/q}, \tag{1}$$

where $\mathbf{E}_f^{(n)}$ is the expectation with respect to the probability measure $\mathbf{P}_f^{(n)}$, $\ell\colon\ \mathbf{F}_0 \to \mathbf{R}_+$ is a semi–norm, and $q \geqq 1$ is a given real number. Our goal is to develop an estimator of function f with "small" risk $\mathcal{R}_\ell^{(n)}[\widehat{f}; f]$.

The following three models are typical examples of statistical experiments.

Example 1 (regression model). Let

$$X^{(n)} = \{(Y_1, Z_1), \ldots, (Y_n, Z_n)\}, \qquad n \in \mathbf{N}^*,$$

where

$$Y_i = f(Z_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{2}$$

and ξ_i, $i = 1, \ldots, n$, are independent identically distributed random variables with zero mean, $\mathbf{E}\,\xi_1 = 0$, and finite second moment, $\mathbf{E}\,\xi_1^2 = \sigma^2 < \infty$. The design points Z_i, $i = 1, \ldots, n$, are assumed to be either fixed points in D, or independent random elements, identically distributed on D.

*The first author is supported by ISF grant №104/11.

†Department of Statistics, University of Haifa, 31905 Haifa, Israel; e-mail: goldensh@stat.haifa.ac.il

‡Laboratoire d'Analyse, Topologie et Probabilités UMR CNRS 6632, Université Aix-Marseille, 39, rue F. Joliot Curie, 13453 Marseille, France; e-mail: lepski@cmi.univ-mrs.fr

Example 2 (Gaussian white noise model). Let W be the white noise on $\mathfrak{D}$ of intensity ν (see. [8, c. 36]). Consider the statistical experiment, generated by the observation $X^{(n)} = \{Y_n(\phi),\ \phi \in \mathbf{L}_2(D,\nu)\}$, $n \in \mathbf{N}^*$, where $Y_n(\cdot)$ satisfies the equation

$$Y_n(\phi) = \langle f, \phi\rangle + \frac{1}{\sqrt{n}} \int_D \phi(u)\, W(\mathrm{d}u) \qquad \forall \phi \in \mathbf{L}_2(D,\nu). \tag{3}$$

Here $\langle\cdot,\cdot\rangle$ is the scalar product on $\mathbf{L}_2(D,\nu)$ and $f \in \mathbf{L}_2(D,\nu)$. Recall, that

$$Y_n(\phi) \sim \mathcal{N}\big(\langle f, \phi\rangle, n^{-1}\|\phi\|_2^2\big), \tag{4}$$

where $\|\cdot\|_2$ denotes the norm in the space $\mathbf{L}_2(D,\nu)$.

Example 3 (density model). Let $\mathbf{P}$ be a probability measure on the σ-algebra $\mathfrak{D}$ with density f with respect to the measure ν. Assume that we observe vector $X^{(n)} = (X_1, \ldots, X_n)$, $n \in \mathbf{N}^*$, where X_i, $i = 1, \ldots, n$, are independent identically distributed random elements with values in D, and distributed according to $\mathbf{P}$.

Choosing semi–norm $\ell(\cdot)$ differently in (1), we come to different estimation problems.

1. *Global estimation.* If we want to estimate the whole funstion f in a given set $D_0 \subseteq D$, then it is natural to choose the semi–norm ℓ as the $\mathbf{L}_p$-norm on (D_0,ν): $\ell(g) = \|g\|_{p,\nu}$ (in the sequel, for the sake of brevity, in the norm notation we omit dependence on ν). In this case the corresponding risk is given by the formula

$$\mathcal{R}_p^{(n)}[\widehat{f}; f] = \Big[\mathbf{E}_f^{(n)} \big\|\widehat{f} - f\big\|_p^q\Big]^{1/q}, \qquad p \in [1,\infty]. \tag{5}$$

2. *Pointwise estimation.* If we are interested in estimating function f at a fixed point $x \in D_0 \subseteq D$, then the semi–norm and the corresponding risk are defined as follows: $\ell(g) = g(x)$,

$$\mathcal{R}_x^{(n)}[\widehat{f}; f] = \Big\{\mathbf{E}_f^{(n)} \big|\widehat{f}(x) - f(x)\big|^q\Big\}^{1/q}.$$

Note, that in the definition of the semi–norm we use $D_0 \subseteq D$. This one allows to avoid the discussion of boundary effects in those models where such effects can exist (Examples 1, 2 and Example 3, if D is a bounded set).

Our approach to the outlined above estimation problem is based on the idea of selecting an estimator from a given family of linear estimators. Linear methods are ubiquitous in nonparametric estimation problems. It is well known that they are "optimal" in many statistical problems. Let us discuss a typical form of the linear estimator of f in the models of Examples 1–3.

Regression model. Linear estimator of regression function f is defined by the formula

$$\widehat{f}(x) = \sum_{j=1}^n K_j(x) Y_j,$$

where the weights $K_j(\cdot)$, $j = 1, \ldots, n$, satisfy the condition $\sum_{j=1}^n K_j(x) = 1$ for any $x \in D_0$. The following important example of the weights leads to the Nadaraya–Watson kernel estimator (here $D \subseteq \mathbf{R}^d$):

$$K_j(x) = \frac{w_h(Z_j - x)}{\sum_{i=1}^n w_h(Z_i - x)}, \qquad j = 1, \ldots, n,$$

where $w\colon \mathbf{R}^d \to \mathbf{R}$ is a given function, $h \in \mathbf{R}^d_+ \setminus \{0\}$, $h = (h_1, \dots, h_d)$ and w_h is given by $w_h(\cdot) = [\prod_{i=1}^d h_i^{-1}] w(\cdot/h)$. Here and in the sequel for vectors $x, y \in \mathbf{R}^d$ we understand the division x/y in the coordinate–wise sense. Note that if variables Z_j are deterministic, then

$$\mathbf{E}_f^{(n)}\big[\widehat{f}(x)\big] = \sum_{j=1}^n K_j(x) f(Z_j).$$

Gaussain white noise model. Let $K\colon D \times D \to \mathbf{R}$ be a function satisying the condition $\int_D K(t,x)\,\nu(\mathrm{d}t) = 1$, $K(\cdot, x) \in \mathbf{L}_2(D, \nu)$ for any $x \in D$. A linear estimator of f is defined by the expression

$$\widehat{f}(x) = Y_n\big(K(\cdot, x)\big), \qquad x \in D_0.$$

In this model for any $x \in D_0$ estimator $\widehat{f}(x)$ is a Gaussian random variable with mean $\int_D K(t,x) f(t)\,\nu(\mathrm{d}t)$ and variance $\int_D K^2(t,x)\,\nu(\mathrm{d}t)$.

Let $D \subseteq \mathbf{R}^d$ and ν be the Lebesgue measure. If, for example, function $w\colon D \to \mathbf{R}$ is such that $\int w(t)\,\mathrm{d}t = 1$, and $h \in \mathbf{R}^d_+ \setminus \{0\}$, then, setting $K(t,x) = \big(\prod_{i=1}^d h_i^{-1}\big) w((t-x)/h)$, we come to the *kernel estimator*.

Density model. A linear estimator in the density model is given by the formula $\widehat{f}(x) = (1/n)\sum_{j=1}^n K\big(X_j, x\big)$, $x \in D_0$, and the Parzen–Rosenblatt kernel estimator (here again $D \subseteq \mathbf{R}^d$) is

$$\widehat{f}(x) = \frac{1}{n}\Big(\prod_{i=1}^d h_i^{-1}\Big) \sum_{j=1}^n w\Big(\frac{X_j - x}{h}\Big).$$

Here $\mathbf{E}_f^{(n)}[\widehat{f}(x)] = \int K(t,x) f(t)\,\nu(\mathrm{d}t)$.

It should emphasized that linear estimators not necessarily depend linearly on the observations (Example 3). The common feature, however, is that for each fixed x the expectation of $\widehat{f}(x)$ is a linear functional of f. We take this property as the definition of a linear estimator in the context of the general statistical experiment (see Definition 1 in Section 2).

Thus, in different models linear estimators are characterized by some *weight function* K. It is known that accuracy of linear estimators is determined by the weight K, or, for kernel estimators, by kernel w and bandwidth h. Therefore, irrespectively of the considered model, while constructing a kernel estimator we naturally face the crucial problem: *How to select the kernel and the bandwidth of the kernel estimator*? *Or , more generally, how to select the weight function while constructing the lienar estimator*? Note that the risk (1) (or any reasonable upper bound on it) depends on the function to be estimated f; therefore, direct optimization of the risk with respect to the weight function is impossible.

In this paper we propose a solution the the described problem which is based on selection an estimator from a fixed family of linear estimators. In order to clarify our approach, suppose for a moment that we have a family $\mathcal{F}(\mathcal{K})$ of linear estimator, indexed by a set of the weight function, i.e. $\mathcal{F}(\mathcal{K}) = \{\widehat{f}_K, K \in \mathcal{K}\}$. Here $\widehat{f}_K$ is the linear estimator associated with the weight function K (the precise definition will be given below). For example, we can view $\mathcal{K}$ as the set of all weights of the type $[\prod_{i=1}^d h_i^{-1}] w_h(\cdot/h)$, where kernel w is fixed, and badwidth $h \in \mathbf{R}^d_+ \setminus \{0\}$ takes values in a given interval $[h_{\min}, h_{\max}] \subset \mathbf{R}^d$; other examples of families $\mathcal{K}$ will be

considered below. We propose a fullt data–driven selection rule $\widehat{K}$, and, under rather general conditions on the family $\mathcal{F}(\mathcal{K})$, we establish an upper bound on the risk of the selected estimator $\widehat{f}_{\widehat{K}}$. The obtained upper bound has the following form: for any function $f \in \mathbf{F}$

$$\mathcal{R}_\ell^{(n)}[\widehat{f}_{\widehat{K}}; f] \leqq \inf_{K\in\mathcal{K}} U_\ell^{(n)}(K, f) + \Delta_\ell^{(n)}(\mathcal{K}), \tag{6}$$

where quantity $U_\ell^{(n)}(K, f)$ depends on K, f, while the remainder $\Delta_\ell^{(n)}(\mathcal{K})$ is independent of of f and completely determined by the family $\mathcal{K}$. Note that inequality (6) itself does not guarantee that the selected estimator is "good". However analysis of the expression on the right hand side of (6) allows to derive many useful *minimax*, *adaptive minimax* и *oracle* results in different estimation settings. Derivation of such results from (6) relies upon the following ideas.

Оракульные неравенства. В некоторых задачах можно показать, что существуют константы c_1 и c_2 такие, что

(i) $U_\ell^{(n)}(K, f) \leqq c_1 \mathcal{R}_\ell^{(n)}[\widehat{f}_K; f]$ для всех $f \in \mathbf{F}$ и $K \in \mathcal{K}$;

(ii) $\Delta_\ell^{(n)}(\mathcal{K}) \leqq c_2 \inf_{K\in\mathcal{K}} \mathcal{R}_\ell^{(n)}[\widehat{f}_K; f]$ для всех $f \in \mathbf{F}$.

Тогда (i) и (ii) вместе с (6) влекут *оракульное неравенство* для риска выбранной оценки:

$$\mathcal{R}_\ell^{(n)}[\widehat{f}_{\widehat{K}}; f] \leqq (c_1 + c_2) \inf_{K\in\mathcal{K}} \mathcal{R}_\ell^{(n)}[\widehat{f}_K; f].$$

Другими словами, риск выбранной оценки с точностью до мультипликативной константы совпадает с риском наилучшей оценки из $\mathcal{F}(\mathcal{K})$. Следовательно, если семейство $\mathcal{F}(\mathcal{K})$ содержит "хорошую" оценку, то правило выбора *имитирует* эту "хорошую" оценку с точностью до постоянного множителя.

Минимаксный подход. В рамках *минимаксного подхода* предполагается, что функция f принадлежит некоторому подмножеству Σ множества $\mathbf{F}$. Обычно Σ — функциональный класс, который выбирается статистиком, и этот выбор осуществляется на основе имеющейся априорной информации об оцениваемой функции. Точность оценки $\widehat{f}$ измеряется наихудшим (максимальным) риском на классе Σ:

$$\mathcal{R}_\ell^{(n)}[\widehat{f}; \Sigma] = \sup_{f\in\Sigma} \mathcal{R}_\ell^{(n)}[\widehat{f}; f],$$

и цель состоит в том, чтобы найти *оптимальную по порядку оценку* $\widehat{f}_*$ такую, что

$$\mathcal{R}_\ell^{(n)}[\widehat{f}_*; \Sigma] \asymp \varphi_n(\Sigma) := \inf_{\widehat{f}} \mathcal{R}_\ell^{(n)}[\widehat{f}; \Sigma], \qquad n \to \infty;$$

здесь инфимум берется по всем возможным оценкам. Известно, что линейные оценки являются оптимальными по порядку во многих задачах для разных полунорм ℓ и функциональных классов Σ.

Теперь предположим, что мы получили неравенство вида (6), и что выполнены следующие условия:

(i) существует весовая функция $K_* \in \mathcal{K}$ такая, что $\sup_{f\in\Sigma} U_\ell^{(n)}(K_*, f) \asymp \varphi_n(\Sigma)$ при $n \to \infty$;

(ii) $\Delta_\ell^{(n)}(\mathcal{K}) = O(\varphi_n(\Sigma))$ при $n \to \infty$.

Тогда оракульное неравенство (6) влечет $\mathcal{R}_\ell^{(n)}[\widehat{f}_{\widehat{K}};\Sigma] \asymp \varphi_n(\Sigma)$, т.е. выбранная оценка $\widehat{f}_{\widehat{K}}$ является оптимальной по порядку на Σ.

Минимаксное адаптивное оценивание. Основной недостаток минимаксного подхода состоит в том, что выбор весовой функции определяется только функциональным классом Σ и линейная оценка, которая оптимальна по порядку на классе Σ, обычно неоптимальна на другом классе Σ'.

Этот факт мотивирует разработку *адаптивных оценок*, которые оптимальны по порядку на некоторой шкале функциональных классов $\{\Sigma_s, s \in S\}$, а не только на одном классе Σ. Здесь s — так называемый мешаюший параметр и S — множество мешающих параметров. Другими словами, мы хотим найти оценку $\widehat{f}_*$ такую, что для любого $s \in S$ имеет место

$$\mathcal{R}_\ell^{(n)}[\widehat{f}_*;\Sigma_s] \asymp \varphi(\Sigma_s), \qquad n \to \infty. \tag{7}$$

Впервые оценки, удовлетворяющие (7), были получены в работе [1] в модели гауссовского белого шума, когда $\{\Sigma_s, s \in S\}$ — шкала классов Соболева и $\ell(g) = \|g\|_2$. В задачах глобального оценивания ($\ell(g) = \|g\|_p$, $1 \leqq p \leqq \infty$), адаптивные минимаксные оценки были предложены в статье [7] для случая, когда $\{\Sigma_s, s \in S\}$ — шкала классов Гёльдера.

Используя неравенство (6), легко показать, что оценка $\widehat{f}_{\widehat{K}}$ удовлетворяет (7), если для всех $s \in S$ выполняются следующие условия:

(i) для любого $s \in S$ существует весовая функция $K_s \in \mathcal{K}$ такая, что $\sup_{f\in\Sigma_s} U_\ell^{(n)}(K_s, f) \asymp \varphi_n(\Sigma_s)$ при $n \to \infty$;

(ii) $\Delta_\ell^{(n)}(\mathcal{K}) = O(\inf_{s\in S} \varphi_n(\Sigma_s))$ при $n \to \infty$.

Однако следует подчеркнуть, что соотношение (7) не всегда выполнено. Например, в задаче поточечного оценивания ($\ell(g) = g(x)$) не существует оценки, которая является оптимальной по порядку одновременно на двух классах Гёльдера $\mathbf{H}(\alpha_1, L_1)$ и $\mathbf{H}(\alpha_2, L_2)$ с $\alpha_1 \neq \alpha_2$ (см. [6]). Тем не менее существуют оценки, *почти оптимальные по порядку* (с точностью до логарифмического по n множителя) на шкале классов Гёльдера, и этот логарифмический множитель не может быть устранен. Здесь важно отметить, что оценки, достигающие наилучшей скорости сходимости на шкале классов, даются некоторой процедурой случайного (измеримого по наблюдению) выбора из семейства линейных оценок.

В настоящей работе мы предлагаем общую процедуру выбора оценки из заданного семейства линейных оценок. Наша процедура может быть применена к любому статистическому эксперименту и любому семейству линейных оценок, удовлетворяющему некоторому условию коммутативности (см. разд. 2). В частности, упомянутое условие коммутативности выполняется для любых ядерных оценок (если игнорировать граничные эффекты в схемах, где они возникают). Мы выводим верхнюю границу на риск выбранной оценки и показываем, как эта граница может быть использована дла получения минимаксных и адаптивных минимаксных результатов в разных задачах (см. разд. 5). Процедура выбора, представленная в этой статье, обобщает и развивает идеи, разработанные в [13], [14], [16] для моделей гауссовского белого шума и плотности.

Статья организована следующим образом. В разд. 2 мы обсуждаем свойства линейных оценок и связанных с ними весовых функций в контексте общего статистического эксперимента. Предлагаемая процедура выбора требует установления равномерных верхних границ (мажорант) для полунормы некоторых случайных процессов. Соответствующие определения и понятия вводятся в разд. 3. В разд. 4

мы определяем процедуру выбора, формулируем и доказываем основные результаты этой статьи. Раздел 5 иллюстрирует приложение основных результатов к задачам оценивания плотности в $\mathbf{L}_p$ и оценивания в модели "projection pursuit" с наблюдениями в белом шуме.

2. Линейные оценки.

2.1. Линейные оценки и связанные с ними весовые функции. Пусть $\widehat{f}$ — оценка функции f на основе наблюдения $X^{(n)}$, т.е. $\widehat{f}$ есть $\mathcal{B}^{(n)}$-измеримое отображение из $\mathcal{X}^{(n)}$ в $\mathbf{F}_0$. Предположим, что математическое ожидание $\mathbf{E}_f^{(n)}[\widehat{f}(x)]$, $x \in D$, существует для всех $f \in \mathbf{F}$ и $\mathbf{E}_f^{(n)}[\widehat{f}(\cdot)] \in \mathbf{F}_0$.

Definition 1. Оценка $\widehat{f}$ функции f называется *линейной*, если существуют функция $K\colon\ D \times D \to \mathbf{R}$ и σ-конечная мера μ на $\mathfrak{D}$ такие, что

$$\mathbf{E}_f^{(n)}\big[\widehat{f}(x)\big] = \int_D K(t,x) f(t)\,\mu(\mathrm{d}t) \qquad \forall f \in \mathbf{F}, \quad \forall x \in D. \tag{8}$$

В дальнейшем любая линейная оценка будет обозначаться $\widehat{f}_K$, с явным указанием соответствующей функции K из определения (8).

Заметим, что μ может совпадать или не совпадать с мерой ν, определенной выше. Важными примерами меры μ являются: (i) мера Лебега на $D \subset \mathbf{R}^d$; (ii) считающая мера $\sum_{i=1}^n \delta_{Z_i}$, где $Z_i \in D$, $i = 1, \ldots, n$, — заданная последовательность и δ_z — масса Дирака.

Итак, оценка $\widehat{f}$ линейна, если для всякого $x \in D$ математическое ожидание $\widehat{f}(x)$ является линейным функционалом от f. Назовем

$$S_K(x) = \int_D K(t,x) f(t)\,\mu(\mathrm{d}t)$$

линейным сглаживателем, который будем понимать как аппроксимацию функции f в точке x.

Точность линейной оценки $\widehat{f}_K$ характеризуется *смещением* (ошибкой аппроксимации f линейным сглаживателем S_K)

$$B_K(f,x) = B_K(x) = \mathbf{E}_f^{(n)}\big[\widehat{f}_K(x)\big] - f(x) = S_K(x) - f(x) \tag{9}$$

и *стохастической ошибкой*

$$\xi_K(f,x) = \xi_K(x) = \widehat{f}_K(x) - \mathbf{E}_f^{(n)}\big[\widehat{f}_K(x)\big].$$

В частности, $\widehat{f}_K(x) - f(x) = B_K(f,x) + \xi_K(f,x)$, и в силу неравенства треугольника

$$\mathcal{R}_\ell^{(n)}[\widehat{f}_K; f] \leqq \ell(B_K) + \Big\{\mathbf{E}_f^{(n)}\big[\ell(\xi_K)\big]^q\Big\}^{1/q}.$$

Последнее неравенство обычно называется в литературе *bias-variance decomposition*.

Перейдем к обсуждению некоторых естественных свойств, которыми должна обладать функция K в (8).

Definition 2. Пусть $D_0 \subseteq D$. Функция $K\colon\ D \times D \to \mathbf{R}$ называется *весовой функцией* или просто *весом*, если

$$\int_D K(t,x)\,\mu(\mathrm{d}t) = 1 \qquad \forall x \in D_0.$$

Множество всех таких весовых функций будем обозначать $\mathcal{W}(D, D_0)$. Для любого веса K мы имеем

$$B_K(f, x) = \int_D K(t, x)[f(t) - f(x)]\,\mu(\mathrm{d}t), \qquad x \in D_0;$$

поэтому смещение постоянной функции тождественно равно нулю. Весовые функции являются ключевыми элементами при построении линейных оценок в различных непараметрических моделях. Мы иллюстрируем этот факт с помощью примеров 1–3, рассмотренных в разд. 1.

Examples. 1. *Модель регрессии с детерминированным планом.* Рассмотрим модель регрессии (2), где Z_i, $i = 1, \dots, n$, — детерминированные элементы в D. Очевидно, что для любой функции $K\colon\ D \times D \to \mathbf{R}$ оценка $\widehat{f}_K(\cdot) = \sum_{i=1}^n K(Z_i, \cdot)Y_i$ является линейной. Действительно,

$$\mathbf{E}_f^{(n)}\big[\widehat{f}_K(\cdot)\big] = \sum_{i=1}^n K(Z_i, \cdot)f(Z_i) = \int_D K(t, \cdot)f(t)\,\mu(\mathrm{d}t),$$

где μ — считающая мера на $\mathfrak{D}$.

2. *Модель гауссовского белого шума.* Рассмотрим модель, введенную в примере 2 из разд. 1. В силу (4), для любой функции $K\colon\ D \times D \to \mathbf{R}$, удовлетворяющей $K(\cdot, x) \in \mathbf{L}_2(D, \nu)$, оценка $\widehat{f}_K(\cdot) = Y_n\big(K(\cdot, x)\big)$ линейна, и

$$\mathbf{E}_f^{(n)}\big[\widehat{f}_K(x)\big] = \int_D K(t, x)f(t)\,\nu(\mathrm{d}t).$$

Таким образом, в этом примере меры μ и ν совпадают.

3. *Модель плотности.* В примере 3 из разд. 1 оценка $\widehat{f}_K(\cdot) = \sum_{j=1}^n K(\cdot, X_j)$ линейна, поскольку

$$\mathbf{E}_f^{(n)}\big[\widehat{f}_K(\cdot)\big] = \int_D K(t, \cdot)f(t)\,\nu(\mathrm{d}t). \tag{10}$$

Эта оценка может быть построена для любой функции $K\colon\ D \times D \to \mathbf{R}$, для которой интеграл в (10) определен. Заметим, что здесь опять $\mu = \nu$. Этот пример демонстрирует, что линейная оценка не обязательно линейна по наблюдениям.

В следующем определении мы вводим некоторое подможество множества $\mathcal{W}(D, D_0)$. Пусть D_1 — множество такое, что $D_0 \subseteq D_1 \subseteq D$.

Definition 3. Функция $K\colon\ D \times D \to \mathbf{R}$ такая, что

$$\int_D K(t, x)\,\mu(\mathrm{d}t) = 1 \qquad \forall x \in D_1,$$
$$\operatorname{supp}\big\{K(\cdot, x)\big\} \subseteq D_1 \qquad \forall x \in D_0,$$

называется D_1-*весовой функцией* или D_1-*весом*. Множество всех D_1-весов обозначается $\mathcal{W}_{D_1}(D, D_0)$.

Ясно, что любой D_1-вес является весом. Действительно, первое соотношение в определении влечет, что $\mathcal{W}_{D_1}(D, D_0) \subset \mathcal{W}(D, D_1)$, и $\mathcal{W}(D, D_1) \subset \mathcal{W}(D, D_0)$, так как $D_0 \subseteq D_1$. Для $D \subseteq \mathbf{R}^d$ типичный пример D_1-веса дается функцией $K\colon\ D \times D \to \mathbf{R}$, которая удовлетворяет следующим условиям:

(i) для действительного числа $\delta > 0$ предположим, что

$$K(t, x) = 0 \quad \forall t, x \in \mathbf{R}^d\colon\ \|t - x\| \geqq \delta,$$

где $\|\cdot\|$ — евклидова норма;

(ii) $\int_{\mathbf{R}^d} K(t,x)\,\mu(\mathrm{d}t) = 1\ \forall x \in \mathbf{R}^d$.

Для фиксированных интервалов $D_0 \subset D_1 \subset D \subset \mathbf{R}^d$ найдется достаточно малое δ такое, что K является D_1-весом.

2.2. Коммутативная система весов. Снабдим множество всех D_1-весов $\mathcal{W}_{D_1}(D, D_0)$ следующей операцией: для любой пары $K, K' \in \mathcal{W}_{D_1}(D, D_0)$ определим

$$[K \otimes K'](\cdot,\cdot) = \int_{D_1} K(\cdot, y) K'(y, \cdot)\,\mu(\mathrm{d}y).$$

Definition 4. Будем говорить, что $K \in \mathcal{W}_{D_1}(D, D_0)$ и $K' \in \mathcal{W}_{D_1}(D, D_0)$ *коммутируют*, если

$$[K \otimes K'](\cdot, x) \equiv [K' \otimes K](\cdot, x) \qquad \mu\text{-п.в.}, \quad \forall x \in D.$$

Подмножество $\mathcal{K}$ множества $\mathcal{W}_{D_1}(D, D_0)$ называется *коммутативной системой весов*, если любая пара его элементов коммутирует.

Коммутативные системы весов играют важную роль при построении и в анализе нашей процедуры выбора. В частности, при довольно слабых технических условиях процедура выбора может быть применена к любому семейству линейных оценок, порожденному коммутативной системой весов.

Приведем несколько примеров коммутативных систем весов.

Example 4. Пусть $D = [-a, a]^d$, $D_0 = [-a_0, a_0]^d$ и $D_1 = [-a_1, a_1]^d$, где $a > a_1 > a_0 > 0$ — заданные действительные числа. Положим $\mu(\mathrm{d}t) = \mathrm{d}t$, и пусть $\mathcal{W}_\delta$ — множество непрерывных функций $w\colon \mathbf{R}^d \to \mathbf{R}$ таких, что $\int_{\mathbf{R}^d} w(t)\,\mathrm{d}t = 1$ и $\mathrm{supp}\{w\} \subseteq [-\delta, \delta]^d$, где $0 < \delta < \min\{a - a_1, a_1 - a_0\}$. Пусть

$$\mathcal{K}[\mathcal{W}_\delta] = \Big\{K\colon\ \mathbf{R}^d \times \mathbf{R} \to \mathbf{R}\colon\ K(t,x) = w(t-x),\ w \in \mathcal{W}_\delta\Big\}.$$

Легко видеть, что множество $\mathcal{K}[\mathcal{W}_\delta]$ является коммутативной системой весов. В самом деле, интегрирование в определениях D_1-веса и $[K \otimes K']$ может быть заменено интегрированием по всему пространству $\mathbf{R}^d$. Таким образом, операция $\otimes$ представляет собой свертку $*$, и поэтому

$$[K \otimes K'] = [K * K'] = [K' * K] = [K' \otimes K].$$

На самом деле, с точностью до граничных эффектов, любой набор весов, соответствующий ядерным оценкам, образует коммутативную систему весов. Заинтересованного читателя мы отсылаем к статье [13], где приведены различные примеры таких коллекций, соответствующих ядерным оценкам в структурных моделях.

Example 5. Пусть Λ — подмножество пространства l_2, и пусть $\{\psi_k, k \in \mathbf{N}^d\}$ — ортонормированный базис в $\mathbf{L}_2\big(D, \nu\big)$, обладающий следующими свойствами:

$$\psi_0 \equiv c \neq 0, \qquad \int_D \psi_k(t)\,\nu(\mathrm{d}t) = 0 \quad \forall k \neq 0 = (0, \ldots, 0).$$

В частности, тензорное произведение одномерных тригонометрических базисов удовлетворяет этим условиям.

Пусть $\lambda = \bigl(\lambda_k, k \in \mathbf{N}^d\bigr)$; рассмотрим семейство весов

$$\mathcal{K}[\Lambda] = \Bigl\{K_\lambda\colon\ D \times D \to \mathbf{R}\colon\ K_\lambda(t,x) = \sum_{k\in\mathbf{N}^d} \lambda_k \psi_k(t)\psi_k(x),\ \lambda \in \Lambda\Bigr\}.$$

Если $\lambda_0 \nu(D) = c^{-1}$ для всех $\lambda \in \Lambda$, то K_λ является D-весом. Тогда $\mathcal{K}\bigl[\Lambda\bigr]$ — коммутативная система весов для любого $\Lambda \subseteq l_2$, если выбрать $\mu = \nu$. В самом деле, для всех $\lambda, \lambda' \in \Lambda$ имеем

$$\begin{aligned}
[K_\lambda \otimes K_{\lambda'}](t,x) &= \int_D \Biggl[\sum_{k\in\mathbf{N}^d} \lambda_k \psi_k(t)\psi_k(y) \sum_{r\in\mathbf{N}^d} \lambda'_r \psi_r(y)\psi_r(x)\Biggr] \nu(\mathrm{d}y) \\
&= \sum_{k\in\mathbf{N}^d, r\in\mathbf{N}^d} \lambda_k \lambda'_r \psi_k(t)\psi_r(x) \int_D \Bigl[\psi_k(y)\psi_r(y)\Bigr] \nu(\mathrm{d}y) \\
&= \sum_{k\in\mathbf{N}^d} \lambda_k \lambda'_k \psi_k(t)\psi_k(x) = [K_{\lambda'} \otimes K_\lambda](t,x).
\end{aligned}$$

Example 6. Пусть $D_0 = \bigl\{Z_1, \ldots, Z_n\bigr\}$, где Z_i, $i = 1, \ldots, n$, — фиксированные точки в D. Пусть $\mu = \sum_{i=1}^n \delta_{Z_i}$ и предположим, что мы интересуемся оцениванием функции f в точках плана D_0. Это типичная задача, возникающая в модели регрессии. Пусть $\mathcal{W}$ есть некоторое заданное множество функций $w\colon\ D\times D \to \mathbf{R}$; тогда множество весов может быть выбрано как множество $(n \times n)$-матриц

$$\mathcal{K} = \Bigl\{K = \{w(Z_i, Z_j)\}_{i,j=1,\ldots,n},\ w \in \mathcal{W}\Bigr\}.$$

Следовательно, $K \otimes K' = KK'$, и поэтому $\mathcal{K}$ является коммутативной системой весов тогда и только тогда, когда любая пара матриц из $\mathcal{K}$ коммутируют. Необходимое и достаточное условие для этого состоит в том, что все матрицы в $\mathcal{K}$ одновременно диагонализуемы [19, теорема 1.3.19]. В статье [22] приведены многочисленные примеры линейных сглаживателей, которые связаны с коммутирующими матрицами.

Следующее простое утверждение является базовым для предлагаемого правила выбора.

Lemma 1. *Пусть $\mathcal{K}$ — коммутативная система весов; тогда для любой пары весов $K, K' \in \mathcal{K}$ и для любых $f \in \mathbf{F}$ и $x \in D_0$*

$$[S_{K\otimes K'}(x) - S_{K'}(x)] = \int_D K'(y,x) B_K(f,y)\, \mu(\mathrm{d}y), \tag{11}$$

$$[S_{K\otimes K'}(x) - S_K(x)] = \int_D K(y,x) B_{K'}(f,y)\, \mu(\mathrm{d}y). \tag{12}$$

Proof. Установим сначала (11). Используя теорему Фубини и определение D_1-веса, имеем для любого $x \in D_0$

$$\begin{aligned}
\int_D [K \otimes K'](t,x) f(t)\, \mu(\mathrm{d}t) &= \int_D \Biggl[\int_{D_1} K(t,y) K'(y,x)\, \mu(\mathrm{d}y)\Biggr] f(t)\, \mu(\mathrm{d}t) \\
&= \int_{D_1} K'(y,x) f(y)\, \mu(\mathrm{d}y) \\
&\quad + \int_{D_1} K'(y,x) \Biggl[\int_D K(t,y)(f(t) - f(y))\, \mu(\mathrm{d}t)\Biggr] \mu(\mathrm{d}y).
\end{aligned}$$

Остается заметить, что $\int_D K(t,y)[f(t)-f(y)]\,\mu(\mathrm{d}t) = B_K(f,y)$ и что, по определению D_1-веса, интегрирование веса $K(\cdot,x)$, $x \in D_0$, на D_1 может быть заменено интегрированием на D.

Поскольку (11) установлено для произвольной пары весов и, в силу коммутативности, $S_{K\otimes K'} = S_{K'\otimes K}$, приходим к (12). Лемма доказана.

3. Мажоранты. Наше правило выбора требует установления верхних функций (которые мы называем *мажорантами*) для семейств случайных величин, заиндексированных весами и связанных со стохастическими ошибками линейных оценок. Мажоранты могут быть детерминированными или случайными в зависимости от рассматриваемой задачи. В этом разделе мы даем необходимые определения и формулируем общие условия на мажоранты.

Пусть $\mathcal{K}$ — некоторая система весов; для любой пары $K, K' \in \mathcal{K}$ положим

$$\zeta_{K,K'} = \ell\big(\xi_{K\otimes K'} - \xi_K\big) \vee \ell\big(\xi_{K'\otimes K} - \xi_{K'}\big).$$

Введем теперь определение равномерной верхней границы для семейства случайных величин

$$\{\zeta_{K,K'},\ K, K' \in \mathcal{K}\}.$$

Definition 5. Пусть $\delta \in (0,1)$ — фиксированное число. Будем говорить, что семейство $\mathcal{B}^{(n)}$-измеримых положительных функций $\{M_{K,K'}(\delta),\ K, K' \in \mathcal{K}\}$ является δ-*мажорантой*, если для всех $f \in \mathbf{F}$ и $n \in \mathbf{N}^*$

(i) $\sup_{K,K'\in\mathcal{K}}[\zeta_{K,K'} - M_{K,K'}(\delta)]$ является $\mathcal{B}^{(n)}$-измеримым;

(ii) $\mathbf{E}_f^{(n)}\{\sup_{K,K'\in\mathcal{K}}[\zeta_{K,K'} - M_{K,K'}(\delta)]_+^q\} \leqq \delta^q$.

Здесь $[a]_+ = \max(a,0)$, $a \in \mathbf{R}$.

Заметим, что δ-мажоранта определена не единственным образом. Однако основные результаты этой статьи, представленные в п. 4.2, справедливы для любой δ-мажоранты. Конечно, в конкретных задачах мы интересуемся *минимальными* мажорантами. Основным средством для их нахождения являются экспоненциальные неравенства для полунорм гауссовских и эмпирических процессов. Мы отсылаем читателя к недавней статье [15], посвященной этой тематике. По определению, δ-мажоранта является функцией наблюдения $X^{(n)}$. Заметим, однако, что в задачах, где распределение стохастической ошибки не зависит от оцениваемой функции (модели гауссовского белого шума и регрессии), δ-мажоранта $M_{K,K'}(\delta)$ часто может быть выбрана детерминированной.

4. Правило выбора и основные результаты. В этом разделе мы определяем правило выбора из семейства линейных оценок $\mathcal{F}(\mathcal{K}) = \{\widehat{f}_K,\ K \in \mathcal{K}\}$, где $\mathcal{K}$ — коммутативная система весов.

Обозначим

$$M_K^*(\delta) = \sup_{K'\in\mathcal{K}} M_{K',K}(\delta) \qquad \forall K \in \mathcal{K}$$

и будем считать, что величина $M_K^*(\delta)$ является $\mathcal{B}^{(n)}$-измеримой для всех $K \in \mathcal{K}$.

4.1. Правило выбора. Для любого $K \in \mathcal{K}$ положим

$$\widehat{R}_K = \sup_{K'\in\mathcal{K}} \Big[\ell\big(\widehat{f}_{K\otimes K'} - \widehat{f}_{K'}\big) - M_{K,K'}(\delta)\Big] + 2M_K^*(\delta)$$

и определим правило выбора следующим образом:

$$\overline{K} = \arg \inf_{K \in \mathcal{K}} \widehat{R}_K. \tag{13}$$

Если $\overline{K} \in \mathcal{K}$ и $\overline{K}$ является $\mathcal{B}^{(n)}$-измеримым, то правило выбора (13) приводит к оценке

$$\bar{f} = \widehat{f}_{\overline{K}}. \tag{14}$$

Remark 1. Заметим, что $\widehat{R}_K \geqq M_K^*(\delta)$ для всех $K \in \mathcal{K}$. Это неравенство немедленно следует из определений $\widehat{R}_K$ и $M_K^*(\delta)$:

$$\widehat{R}_K \geqq \Big[\ell\big(\widehat{f}_{K\otimes K} - \widehat{f}_K\big) - M_{K,K}(\delta)\Big] + 2M_K^*(\delta) \geqq M_K^*(\delta).$$

В частности, это означает, что функционал, который минимизируется в (13), положителен.

Remark 2. Хотя $\mathcal{K}$ — коммутативная система весов, что влечет $S_{K\otimes K'} = S_{K'\otimes K}$, равенство $\widehat{f}_{K\otimes K'} = \widehat{f}_{K'\otimes K}$ может не выполняться. Однако если последнее равенство имеет место, то $\widehat{R}_K$ в (13) можно переопределить следующим образом:

$$\widehat{R}_K = \sup_{K' \in \mathcal{K}} \Big[\ell\big(\widehat{f}_{K\otimes K'} - \widehat{f}_{K'}\big) - M_{K,K'}(\delta)\Big] + M_K^*(\delta).$$

Легко видеть, что в этом случае $\widehat{R}_K \geqq 0$ для всех $K \in \mathcal{K}$.

Remark 3. Наши основные результаты устанавливаются для общего статистического эксперимента, и поэтому мы предполагаем, что δ-мажоранта задана. В отличие от самого правила выбора, которое *не зависит от модели*, δ-мажоранта определяется конкретной моделью, и ее нахождение может быть непростой задачей. Примеры δ-мажорант для некоторых конкретных задач представлены в разд. 5.

Подчеркнем, что доказательство измеримости и принадлежности $\overline{K}$ семейству $\mathcal{K}$ может быть упрощено, если немного переопределить правило выбора. Действительно, пусть $\widehat{K} \in \mathcal{K}$ таково, что для фиксированного числа $\tau > 0$ справедливо неравенство

$$\widehat{R}_{\widehat{K}} \leqq \inf_{K \in \mathcal{K}} \widehat{R}_K + \tau.$$

Заметим, что существование такого $\widehat{K}$ не требует доказательства, и к тому же проблема измеримого выбора также существенно упрощается. Итак, если вес $\widehat{K}$ является $\mathcal{B}^{(n)}$-измеримым, то приходим к оценке

$$\widehat{f} = \widehat{f}_{\widehat{K}},$$

для которой мы и будем доказывать наш основной результат, теорему 1.

4.2. Основные результаты. Введем следующее обозначение: для любого $K \in \mathcal{K}$ положим

$$E_K(\ell, f) = \sup_{K' \in \mathcal{K}} \ell\bigg(\int_D K'(t,\cdot) B_K(f,t)\,\mu(\mathrm{d}t)\bigg),$$

где $B_K(f,\cdot)$ — смещение линейной оценки, связанной с весом K (см. (9)).

Theorem 1. *Пусть $\mathcal{K}$ — коммутативная система весов и $\mathcal{F}(\mathcal{K})$ — соответствующее семейство линейных оценок. Предположим, что для любого $\tau > 0$ существует $\mathcal{B}^{(n)}$-измеримый вес $\widehat{K} \in \mathcal{K}$, удовлетворяющий неравенству*

$$\widehat{R}_{\widehat{K}} \leqq \inf_{K \in \mathcal{K}} \widehat{R}_K + \tau.$$

Тогда для любых $f \in \mathbf{F}$, $n \in \mathbf{N}^*$, $\delta > 0$ и $\tau > 0$

$$\mathcal{R}_\ell^{(n)}[\widehat{f}_{\widehat{K}}; f] \leqq \inf_{K \in \mathcal{K}} \Big\{ \mathcal{R}_\ell^{(n)}[\widehat{f}_K; f] + 3E_K(\ell, f) + 5\Big(\mathbf{E}_f^{(n)}[M_K^*(\delta)]^q\Big)^{1/q} \Big\} + 3\delta + 2\tau.$$

Proof. Используя неравенство треугольника, мы можем написать для любого $K \in \mathcal{K}$

$$\ell\big(\widehat{f}_{\widehat{K}} - f\big) \leqq \ell\big(\widehat{f}_{\widehat{K}} - \widehat{f}_{\widehat{K}\otimes K}\big) + \ell\big(\widehat{f}_{\widehat{K}\otimes K} - \widehat{f}_K\big) + \ell\big(\widehat{f}_K - f\big).$$

Будем оценивать отдельно каждое слагаемое в правой части последнего неравенства.

В силу леммы 1 имеем для любого $K \in \mathcal{K}$

$$\begin{aligned}\widehat{R}_K - 2M_K^*(\delta) &= \sup_{K' \in \mathcal{K}} \big\{ \ell\big(\widehat{f}_{K\otimes K'} - \widehat{f}_{K'}\big) - M_{K,K'}(\delta) \big\} \\ &\leqq E_K(\ell, f) + \sup_{K' \in \mathcal{K}} \big\{ \ell\big(\xi_{K\otimes K'} - \xi_{K'}\big) - M_{K,K'}(\delta) \big\} \\ &\leqq E_K(\ell, f) + \sup_{K, K' \in \mathcal{K}} \big\{ \zeta_{K,K'} - M_{K,K'}(\delta) \big\}_+.\end{aligned}$$

Таким образом,

$$\widehat{R}_K \leqq E_K(\ell, f) + 2M_K^*(\delta) + \zeta, \tag{15}$$

где для краткости мы обозначили $\zeta = \sup_{K,K' \in \mathcal{K}} \big\{ \zeta_{K,K'} - M_{K,K'}(\delta) \big\}_+$.

Кроме того, применяя лемму 1, получаем для любой пары весов $L, L' \in \mathcal{K}$

$$\ell\big(\widehat{f}_{L\otimes L'} - \widehat{f}_L\big) \leqq E_{L'}(\ell, f) + \ell\big(\xi_{L\otimes L'} - \xi_L\big),$$

и, следовательно,

$$\ell\big(\widehat{f}_{L\otimes L'} - \widehat{f}_L\big) \leqq E_{L'}(\ell, f) + M_{L',L}(\delta) + \zeta \leqq E_{L'}(\ell, f) + M_L^*(\delta) + \zeta.$$

Отсюда, полагая $L' = K$ и $L = \widehat{K}$, получим

$$\ell\big(\widehat{f}_{\widehat{K}\otimes K} - \widehat{f}_{\widehat{K}}\big) \leqq E_K(\ell, f) + M_{\widehat{K}}^*(\delta) + \zeta. \tag{16}$$

В силу замечания 4.1, определения $\widehat{K}$ и неравенства (15) имеем для любого $K \in \mathcal{K}$

$$M_{\widehat{K}}^*(\delta) \leqq \widehat{R}_{\widehat{K}} \leqq \widehat{R}_K + \tau \leqq E_K(\ell, f) + 2M_K^*(\delta) + \zeta + \tau,$$

что вместе с (16) дает

$$\ell\big(\widehat{f}_{\widehat{K}\otimes K} - \widehat{f}_{\widehat{K}}\big) \leqq 2E_K(\ell, f) + 2M_K^*(\delta) + 2\zeta + \tau. \tag{17}$$

По определению $\widehat{K}$, для любого $K \in \mathcal{K}$

$$\begin{aligned}\ell\big(\widehat{f}_{\widehat{K}\otimes K} - \widehat{f}_K\big) &= \ell\big(\widehat{f}_{\widehat{K}\otimes K} - \widehat{f}_K\big) - M_{\widehat{K},K}(\delta) + M_{\widehat{K},K}(\delta)\\ &\leqq \sup_{K'\in\mathcal{K}} \big\{\ell\big(\widehat{f}_{\widehat{K}\otimes K'} - \widehat{f}_{K'}\big) - M_{\widehat{K},K'}(\delta)\big\} + M^*_K(\delta)\\ &= \widehat{R}_{\widehat{K}} + M^*_K(\delta) - 2M^*_{\widehat{K}}(\delta) \leqq \widehat{R}_K + M^*_K(\delta) + \tau.\end{aligned}$$

Тогда, учитывая (15), получим

$$\ell\big(\widehat{f}_{\widehat{K}\otimes K} - \widehat{f}_K\big) \leqq E_K(\ell, f) + 3M^*_K(\delta) + \zeta + \tau. \tag{18}$$

Применяя неравенство треугольника, получаем из (17) и (18)

$$\ell\big(\widehat{f}_{\widehat{K}} - f\big) \leqq \ell\big(\widehat{f}_K - f\big) + 3E_K(\ell, f) + 5M^*_K(\delta) + 3\zeta + 2\tau.$$

Это, в свою очередь, влечет для любого $q \geqq 1$

$$\mathcal{R}^{(n)}_\ell[\widehat{f}_{\widehat{K}}; f] \leqq \mathcal{R}^{(n)}_\ell[\widehat{f}_K; f] + 3E_K(\ell, f) + 5\big(\mathbf{E}^{(n)}_f[M^*_K(\delta)]^q\big)^{1/q} + 3(\mathbf{E}^{(n)}_f\zeta^q)^{1/q} + 2\tau. \tag{19}$$

Принимая во внимание, что неравенство (19) справедливо для любого $K \in \mathcal{K}$ и что $(\mathbf{E}^{(n)}_f\zeta^q)^{1/q} \leqq \delta$ по определению δ-мажоранты, приходим к утверждению теоремы.

Верхняя граница на риск, установленная в теореме 1, может быть упрощена в случае, когда полунорма $\ell(\cdot)$ является $\mathbf{L}_p$-нормой, т.е. когда риск дается формулой (5).

Определим величину

$$C_{\mathcal{K}} = \sup_{K\in\mathcal{K}} \left\{ \sup_{x\in D} \int |K(t,x)|\,\mu(\mathrm{d}t)\right\} \vee \left\{ \sup_{t\in D} \int |K(t,x)|\nu(\mathrm{d}x)\right\}.$$

Corollary 1. *Пусть предположения теоремы* 1 *выполнены, и пусть* $\ell(\cdot) = \|\cdot\|_p$, $p \in [1,\infty]$. *Тогда для всех* $f \in \mathbf{F}$, $\delta > 0$ *и* $\tau > 0$ *имеем*

$$\mathcal{R}^{(n)}_p[\widehat{f}_{\widehat{K}}; f] \leqq \inf_{K\in\mathcal{K}} \left\{[3C_{\mathcal{K}} + 1]\mathcal{R}^{(n)}_p[\widehat{f}_K; f] + 5\Big(\mathbf{E}^{(n)}_f[M^*_K(\delta)]^q\Big)^{1/q}\right\} + 3\delta + 2\tau.$$

Proof. В силу леммы 1

$$\begin{aligned}\sup_{K'\in\mathcal{K}} \|S_{K'\otimes K} - S_{K'}\|_p &= \sup_{K'\in\mathcal{K}} \|S_{K\otimes K'} - S_{K'}\|_p\\ &= \sup_{K'\in\mathcal{K}} \left\| \int_D K'(y,\cdot)B_K(f,y)\mu(\mathrm{d}y)\right\|_p \leqq C_{\mathcal{K}}\|B_K(f,\cdot)\|_p.\end{aligned}$$

Здесь последнее неравенство вытекает из хорошо известных оценок для $\mathbf{L}_p$-норм интегральных операторов (см., например, [12, теорема 6.18]). Таким образом,

$$E_K(\|\cdot\|_p, f) \leqq C_{\mathcal{K}}\|B_K(f,\cdot)\|_p \qquad \forall f \in \mathbf{F}.$$

Добавим, что для любой линейной оценки $\mathcal{R}^{(n)}_p[\widehat{f}_K; f] \geqq \|B_K(f,\cdot)\|_p$. Действительно, положив $s = p/(p-1)$, приходим к следующей цепочке неравенств:

$$\begin{aligned}\mathcal{R}^{(n)}_p[\widehat{f}_K; f] &\geqq \mathbf{E}^{(n)}_f\|B_K(f,\cdot) + \xi_K(f,\cdot)\|_p\\ &= \mathbf{E}^{(n)}_f\left\{ \sup_{g:\|g\|_s=1} \int g(x)\big(B_K(f,x) + \xi_K(f,x)\big)\,\nu(\mathrm{d}x)\right\}\\ &\geqq \sup_{g:\|g\|_s=1} \mathbf{E}^{(n)}_f\left\{ \int g(x)\big(B_K(f,x) + \xi_K(f,x)\big)\,\nu(\mathrm{d}x)\right\} = \|B_K(f,\cdot)\|_p.\end{aligned}$$

Последнее равенство следует из того, что $\mathbf{E}_f^{(n)}\xi_K(f,x)=0$. Следствие доказано.

Заметим, что правило выбора требует задания параметра δ, который определяет уровень δ-мажоранты. Зависимость риска выбранной оценки от δ становится ясной из границ теоремы 1 (следствия 1) и из определения δ-мажоранты. Очевидно, что для получения разумных неравенств δ должно быть выбрано зависящим от n, $\delta=\delta_n\to 0$ при $n\to\infty$. Однако этот выбор не может быть произвольным, так как, при фиксированном n, $\mathbf{E}_f^{(n)}[M_K^*(\delta)]^q$ возрастает при $\delta\to 0$. К счастью, во многих задачах зависимость $\mathbf{E}_f^{(n)}[M_K^*(\delta)]^q$ от δ такова, что для широкого класса последовательностей δ_n найдется $c>0$ такое, что

$$\left(\mathbf{E}_f^{(n)}[M_K^*(\delta_n)]^q\right)^{1/q}\leqq c\,\mathcal{R}_\ell^{(n)}[\widehat{f}_K;f]\qquad\forall f\in\mathbf{F},\quad\forall K\in\mathcal{K}.$$

Этот факт позволяет применять неравенства теоремы 1 и следствия 1 для вывода минимаксных и адаптивных минимаксных результатов.

5. Некоторые приложения. В этом разделе мы проиллюстрируем применимость предложенного правила выбора для построения адаптивных минимаксных оценок на примере двух задач: оценивание плотности в $\mathbf{L}_p$ и оценивание сигнала в модели гауссовского белого шума при структурных предположениях. Приведенный ниже материал основан на результатах, полученных в статьях [14], [16].

5.1. Адаптивное минимаксное оценивание плотности в $\mathbf{L}_p$. Рассмотрим задачу оценивания плотности, описанную в примере 3 из разд. 1. Мы хотим оценить f на пространстве $\mathbf{R}^d$ с малым $\mathbf{L}_p$-риском, $p\in[1,\infty)$ (см. (5)). Итак, здесь $D=D_0=\mathbf{R}^d$ и $\mu=\nu$ — мера Лебега.

Задача минимаксного оценивания плотности в $\mathbf{L}_p$ была рассмотрена в многих работах, среди которых упомянем [9], [4], [5], [18], [10], [21], [20]. Что касается адаптивного минимаксного оценивания плотности в $\mathbf{L}_p$, то эта задача гораздо менее изучена. В частности, она была рассмотрена только для одномерного случая в трех последних статьях из списка, приведенного выше.

Пусть $K\colon\mathbf{R}^d\to\mathbf{R}$ — фиксированная функция, удовлетворяющая следующим условиям: $\int K(x)\,\mathrm{d}x=1$, $\|K\|_\infty\leqq k<\infty$ и $|K(x)-K(y)|\leqq L_K|x-y|$ для всех $x,y\in\mathbf{R}^d$. Пусть $\widehat{f}_{K_h}(x)$ — ядерная оценка плотности:

$$\widehat{f}_{K_h}(x)=\frac{1}{nV_h}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)=\frac{1}{n}\sum_{i=1}^n K_h(x-X_i),$$

связанная с весом $K_h(\cdot):=V_h^{-1}K(\cdot/h)$. Здесь $h=(h_1,\dots,h_d)$ — сглаживающий параметр и $V_h:=\prod_{i=1}^d h_i$. Для двух заданных векторов $h^{\min}=(h_1^{\min},\dots,h_d^{\min})$ и $h^{\max}=(h_1^{\max},\dots,h_d^{\max})$, удовлетворяющих $0<h_i^{\min}\leqq h_i^{\max}\leqq 1$ для любого $i=1,\dots,d$, положим $\mathcal{H}=[h_1^{\min},h_1^{\max}]\times\dots\times[h_d^{\min},h_d^{\max}]$. Рассмотрим множество весов $\mathcal{K}_\mathcal{H}=\{K_h,h\in\mathcal{H}\}$ и соответствующее ему семейство ядерных оценок

$$\mathcal{F}(\mathcal{K}_\mathcal{H})=\{\widehat{f}_L,\ L\in\mathcal{K}_\mathcal{H}\}.$$

Мы применим разработанную выше процедуру выбора к семейству $\mathcal{F}(\mathcal{K}_\mathcal{H})$.

Для ядерных оценок операция $\otimes$ представляет собой свертку $*$ на $\mathbf{R}^d$, так что

$$\widehat{f}_{K_h\otimes K_\eta}(x)=\widehat{f}_{K_h*K_\eta}(x)=\frac{1}{n}\sum_{i=1}^n[K_h*K_\eta](x-X_i).$$

Следуя [16], рассмотрим семейство функций $\{M_{K_h,K_\eta}(\delta),\ h,\eta\in\mathcal{H}\}$, определенное следующим образом:

$$M_{K_h,K_\eta}(\delta)=C_p[g_p(K_\eta)+g_p(K_h*K_\eta)].$$

Здесь C_p — константа, зависящая только от p (ее точное значение дано в цитируемой статье), а функционал g_p определен так: для любой функции $U\colon\ \mathbf{R}^d\to\mathbf{R}$

— если $p\in[1,2]$, то $g_p(U)=n^{1/p-1}\|U\|_p$;

— если $p>2$, то

$$g_p(U)=\left\{\frac{1}{\sqrt{n}}\left(\int\left[\frac{1}{n}\sum_{i=1}^n U^2(x-X_i)\right]^{p/2}\mathrm{d}x\right)^{1/p}+\frac{\|U\|_p}{n^{1-1/p}}\right\}\vee\frac{\|U\|_2}{\sqrt{n}}.$$

В [16] показано, что определенное выше семейство является δ-мажорантой, где уровень δ и соответствующее математическое ожидание

$$\left[\mathbf{E}_f^{(n)}\bigl(M^*_{K_h}(\delta)\bigr)^q\right]^{1/q}=\left[\mathbf{E}_f^{(n)}\Bigl(\sup_{\eta\in\mathcal{H}}M_{K_\eta,K_h}(\delta)\Bigr)^q\right]^{1/q}$$

даются выражениями, приведенными ниже. Если для некоторой константы $c>0$ выполнено неравенство $nV_{h^{\min}}\geqq c$, то, как следует из [16, теоремы 1 и 2],

— в случае $p\in[1,2)$ для любой плотности f

$$\delta=O\bigl(n^{1/p}(\ln n)^{4d}\exp\{-cn^{2/p-1}\}\bigr),\qquad n\to\infty;$$

кроме того, $M_{K_h,K_\eta}(\delta)$ — детерминированная для $p\in[1,2]$ и $M^*_{K_h}(\delta)\leqq c(nV_h)^{1/p-1}$;

— в случае $p\in[2,\infty)$ для любой плотности f, удовлетворяющей $\|f\|_\infty\leqq\mathsf{f}_\infty$,

$$\delta=O\Bigl((\ln n)^{4d+1}\sqrt{n}\exp\bigl\{-cV_{h^{\max}}^{-2/p}\bigr\}\Bigr),\qquad n\to\infty;$$

заметим, что условие $V_{h^{\max}}\to0$ при $n\to\infty$ необходимо для состоятельности оценки $\widehat{f}_{K_{h^{\max}}}$; кроме того, мы можем гарантировать $\delta=\delta_n=O(n^{-1/2})$, выбирая $V_{h^{\max}}$ стремящимся логарифмически к нулю. Также, если $p\in(2,\infty)$, то $[\mathbf{E}_f^{(n)}(M^*_{K_h}(\delta))^q]^{1/q}\leqq c(nV_h)^{-1/2}$.

Итак, процедура выбора сводится к минимизации по $h\in\mathcal{H}$ следующего функционала:

$$\widehat{R}_{K_h}=\sup_{\eta\in\mathcal{H}}\Bigl\{\|\widehat{f}_{K_h*K_\eta}-\widehat{f}_{K_\eta}\|_p-C_p[g_p(K_\eta)+g_p(K_h*K_\eta)]\Bigr\}$$
$$+2C_p\Bigl[g_p(K_h)+\sup_{\eta\in\mathcal{H}}g_p(K_h*K_\eta)\Bigr].$$

Эти результаты вместе с границей теоремы 1, ведут к построению оценки, адаптивной на шкале функциональных классов Никольского. Мы отсылаем читателя к недавней статье [16], где представлено детальное исследование этой задачи.

5.2. Адаптивное минимаксное оценивание в модели "projection pursuit".

Рассмотрим теперь модель гауссовского белого шума, где на основе наблюдений (3) требуется оценить функцию f с малым $\mathbf{L}_\infty$-риском. Итак, здесь D и $D_0\subset D$ — некоторые интервалы в $\mathbf{R}^d$ и $\mu=\nu$ — мера Лебега.

Хорошо известно, что в задачах оценивания функций многих переменных существует эффект "проклятья размерности", который выражается в значительном ухудшении достижимой точности оценивания с ростом размерности. Следует отметить, что, даже для очень умеренной размерности, при стандартных предположениях гладкости на оцениваемую функцию достижимая скорость сходимости риска становится очень медленной. Один из способов преодоления "проклятья размерности" состоит в рассмотрении *структурных моделей*, где предполагается, что оцениваемая функция "обладает" некоторой структурой. Эта структура часто позволяет свести исходную задачу к задаче, соответствующей меньшей размерности (так называемая "эффективная размерность"). Этот подход был предложен в статье [24] и развит в многочисленных последующих работах (см., например, [11], [2], [23], [17], [3]). Однако аспекты адаптивного оценивания в структурных моделях практически не изучались. Исключением является статья [2], где адаптивная по классам Соболева оценка функции в $\mathbf{L}_2$ была построена в модели "projection pursuit". Ниже мы покажем, что в этой модели наша процедура выбора приводит к оценке функции в $\mathbf{L}_\infty$, адаптивной по классам Гёльдера.

Модель "projection pursuit" предполагает, что оцениваемая функция представима в следующем виде:

$$f(x) = \sum_{i=1}^{d} f_i(e_i^T x), \tag{20}$$

где $f_i\colon \mathbf{R} \to \mathbf{R}$, $i = 1, \dots, d$, — неизвестные функциональные компоненты, e_i, $i = 1, \dots, d$, — неизвестные линейно независимые векторы, лежащие на единичной сфере $\mathbf{S}^{d-1}$ в $\mathbf{R}^d$.

Пусть $g\colon [-1/2, 1/2] \to \mathbf{R}$ — ядро, удовлетворяющее следующим стандартным условиям: $\int g(x)\,dx = 1$, $\int g(x)x^k\,dx = 0$, $k = 1, \dots, m$, $g \in \mathbf{C}^1$. Для фиксированного сглаживающего параметра $h = (h_1, \dots, h_d)$ с компонентами, удовлетворяющими $h_{\min} \leqq h_i \leqq h_{\max}$, пусть

$$G_0(x) = \prod_{i=1}^{d} g(x_i), \qquad G_{i,h}(x) = \frac{1}{h_i}\, g\biggl(\frac{x_i}{h_i}\biggr) \prod_{j \neq i} g(x_j), \quad i = 1, \dots, d.$$

Пусть $\mathcal{E}_\eta$ — множество всех $(d \times d)$-матриц со столбцами единичной длины и определителем, по абсолютной величине большим заданного числа $\eta > 0$:

$$\mathcal{E}_\eta = \Bigl\{E\colon\ E = (e_1, \dots, e_d), e_i \in \mathbf{S}^{d-1}, |\det(E)| \geqq \eta\Bigr\}.$$

Пусть также $\mathcal{H} = [h_{\min}, h_{\max}]^d$ для некоторых заданных чисел $0 < h_{\min} \leqq h_{\max} \leqq 1$. Определим теперь ядро, связанное с параметром $\theta = (E, h) \in \Theta = \mathcal{E}_\eta \times \mathcal{H}$:

$$K_\theta(x) = |\det(E)| \sum_{i=1}^{d} G_{i,h}(E^T x) - (d-1)|\det(E)| G_0(E^T x),$$

семейство ядер $\mathcal{K}_\Theta = \{K_\theta\colon\ \theta = (E, h) \in \Theta = \mathcal{E}_\eta \times \mathcal{H}\}$ и соответствующее семейство ядерных оценок

$$\mathcal{F}(\mathcal{K}_\Theta) = \Bigl\{\widehat{f}_{K_\theta}(x) = Y_n\bigl(K_\theta(\,\cdot - x)\bigr),\ \theta \in \Theta\Bigr\}.$$

Применим разработанную процедуру выбора к семейству оценок $\mathcal{F}(\mathcal{K}_\Theta)$.

Рассмотрим семейство функций

$$\{M_{K_\theta,K_{\theta'}}(\delta), \theta=(E,h),\ \theta'=(E',h')\in\Theta\},$$

заданное формулой

$$M_{K_\theta,K_{\theta'}}(\delta)=\kappa\sqrt{\frac{\ln n}{n}}\,(\|g\|_1\|g\|_2)^d\Big[\sum_{i=1}^d h_i^{-1/2}\wedge\sum_{i=1}^d (h_i')^{-1/2}\Big],$$

где κ — некоторая положительная константа, зависящая только от d. Используя результаты статьи [14], нетрудно показать, что существует константа $\kappa=\kappa(d)$ такая, что это семейство функций является δ-мажорантой уровня $n^{-1/2}$. Заметим, что эта δ-мажоранта не зависит от E и E'.

Итак, правило выбора сводится к минимизации по $\theta=(E,h)\in\Theta$ следующего функционала:

$$\widehat{R}_{K_\theta}=\sup_{\theta'\in\Theta}\left\{\|\widehat{f}_{K_\theta*K_{\theta'}}-\widehat{f}_{K_{\theta'}}\|_\infty-\kappa\sqrt{\frac{\ln n}{n}}\,(\|g\|_1\|g\|_2)^d\sum_{i=1}^d(h_i')^{-1/2}\right\}$$
$$+2\kappa\sqrt{\frac{\ln n}{n}}\,(\|g\|_1\|g\|_2)^d\sum_{i=1}^d h_i^{-1/2}.$$

Пусть $\Sigma(\beta,L)$, $\beta=(\beta_1,\dots,\beta_d)$, $L>0$, есть класс всех функций, удовлетворяющих представлению (20) с неизвестными векторами $e_i\in\mathbf{S}^{d-1}$ и неизвестными функциональными компонентами f_i, принадлежащими классам Гёльдера $\mathbf{H}(\beta_i,L)$, $i=1,\dots,d$. Тогда наше правило выбора приводит к оценке, которая является оптимальной по порядку скорости сходимости $\mathbf{L}_\infty$-риска на любом классе $\Sigma(\beta,L)$ при условии, что $\max_{i=1,\dots,d}\beta_i\leqq m+1$.

Список литературы

[1] *Ефроймович С. Ю., Пинскер М. С.* Самообучающийся алгоритм непараметрической фильтрации. — Автоматика и телемеханика, 1984, т. 11, с. 58–65.

[2] *Голубев Г. К.* Асимптотически минимаксное оценивание функции регрессии в аддитивной модели. — Пробл. передачи информ., 1992, т. 28, с. 3–15.

[3] *Ибрагимов И. А.* Об оценке многомерной регрессии. — Теория вероятн. и ее примен., 2003, т. 48, в. 2, с. 301–320.

[4] *Ибрагимов И. А., Хасьминский Р. З.* Об оценке плотности распределения. — Записки науч. сем. ЛОМИ, 1980, т. 98, с. 61–85.

[5] *Ибрагимов И. А., Хасьминский Р. З.* Еще об оценке плотности распределения. — Записки науч. сем. ЛОМИ, 1980, т. 108, с. 72–88.

[6] *Лепский О. В.* Об одной задаче адаптивного оценивания в гауссовском белом шуме. — Теория вероятн. и ее примен., 1990, т. 35, в. 4, с. 459–470.

[7] *Лепский О. В.* Асимптотически минимаксное адаптивное оценивание. I. Верхние границы. Оптимально-адаптивные оценки. — Теория вероятн. и ее примен., 1991, т. 36, в. 4, с. 645–659.

[8] *Лифшиц М. А.* Гауссовские случайные функции. Киев: TBiMC, 1995, 246 с.

[9] *Bretagnolle J., Huber C.* Estimation des densités: risque minimax. — Z. Wahrscheinlichkeitstheor. verw. Geb., 1979, v. 47, No 2, p. 119–137.

[10] *Donoho D. L., Johnstone I. M., Kerkyacharian G., Picard D.* Density estimation by wavelet thresholding. — Ann. Statist., 1996, v. 24, p. 508–539.

[11] *Chen H.* Estimation of a projection-pursuit type regression model. — Ann. Statist., 1991, v. 19, p. 142–157.

[12] *Folland G. B.* Real Analysis. New York: Wiley, 1999, 386 p.

[13] *Goldenshluger A., Lepski O.* Universal pointwise selection rule in multivariable function estimation. — Bernoulli, 2008, v. 14, No 4, p. 1150–1190.
[14] *Goldenshluger A., Lepski O.* Structural adaptation via $\mathbf{L}_p$-norm oracle inequalities. — Probab. Theory Related Fields, 2009, v. 143, No 1–2, p. 41–71.
[15] *Goldenshluger A., Lepski O.* Uniform bounds for norms of sums of independent random functions. — Ann. Probab., 2010, v. 39, No 6, p. 2318–2384.
[16] *Goldenshluger A., Lepski O.* Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. — Ann. Statist., 2011, v. 39, No 3, p. 1608–1632.
[17] *Györfi L., Kohler M., Krzyzak A., Walk H.* A Distribution-Free Theory of Nonparametric Regression. New York: Springer, 2002, 647 p.
[18] *Hasminskii R., Ibragimov I.* On density estimation in the view of Kolmogorov's ideas in approximation theory. — Ann. Statist., 1990, v. 18, No 3, p. 999–1010.
[19] *Хорн Р., Джонсон Ч.* Матричный анализ. М.: Мир, 1989, 655 с.
[20] *Juditsky A., Lambert-Lacroix S.* On minimax density estimation on $\mathbf{R}$. — Bernoulli, 2004, v. 10, No 2, p. 187–220.
[21] *Kerkyacharian G., Picard D., Tribouley K.* L^p adaptive density estimation. — Bernoulli, 1996, v. 2, No 3, p. 229–247.
[22] *Kneip A.* Ordered linear smoothers. — Ann. Statist., 1994, v. 22, No 6, p. 835–866.
[23] *Nicoleris T., Yatracos Y.* Rates of convergence of estimators, Kolmogorov's entropy and the dimensionality reduction principle in regression. — Ann. Statist., 1997, v. 25, No 6, p. 2493–2511.
[24] *Stone C. J.* Additive regression and other nonparametric models. — Ann. Statist., 1985, v. 13, No 2, p. 689–705.