On the lower bounds
for the rates of convergence in estimation at a point under multi-index constraint

Nora Serdyukova [1]

[1] *Departamento de Estadística, Facultad de Ciencias Físicas y Matemáticas*
Universidad de Concepción, Avda. Esteban Iturra s/n - Barrio Universitario
Concepción, Región VIII, CHILE. E-mail : Nora.Serdyukova@gmail.com

Résumé. Dans le cadre de l'estimation non paramétrique d'une fonction multidimensionnelle on cherche à obtenir la borne inférieure minimax. On suppose que la fonction à estimer possède la structure « multi-index » dans lequel ni fonction de lien et ni vecteurs d'indice ne sont connus. Par exemple, en régression, ce hypothèse signifie que l'espérance de la variable réponse est défini par celle sachant uniquement une projection du vecteur de covariables sur un sous-espace de dimension plus petite. Par conséquent, cette manière de réduire la dimension est un compromis convenable entre les approches paramétrique et purement non paramétrique. D'après les résultats obtenus pour les pertes ponctuelle, sous l'hypothèse structurelle, on a un nouveau type de bornes inférieures minimax.

Mots-clés. Estimation non paramétrique, Modèle de multi-index, Borne inférieure, Vitesse minimax.

Abstract. In the framework of multivariate function estimation one seeks lower bounds for the minimax risk. One assumes that the function to be estimated possesses a multi-index structure where neither the link function nor the index vectors are known. For example, in regression this assumption means that the expectation of the response is defined by the response given the projection of the covariate vector onto a low-dimensional subspace. Therefore, this convenient dimension reduction approach is a compromise between the parametric and fully nonparametric models. The obtained results show that under pointwise losses imposing the structural constraints leads to new types of minimax lower bounds for "standard" nonparametric models.

Keywords. Nonparametric estimation, Multi-index model, Lower bounds, Minimax rate.

1 Introduction

The present talk adresses lower bounds for the minimax risk under structural assumptions on the function to be estimated. The obtained results show that imposing structural constraints leads to new types of minimax lower bounds for “standard” nonparametric models.

The lower bounds for the minimax risk, apart from being a challenging mathematical problem, serve as a benchmark for the best obtainable quality of an arbitrary estimator. Let us briefly review the main building blocks of the problem. Let $\widehat{Q}$ be some estimator of a parameter $Q(\theta)$ with θ in some parameter space Θ. Consider a risk determined by a polynomial loss function

$$\mathcal{R}_{r,d}^{(n)}\left(\widehat{Q},Q\right)=\left\{\mathbb{E}_{\theta}^{(n)}\left[d^{r}\left(\widehat{Q},Q(\theta)\right)\right]\right\}^{1/r},\qquad \theta\in\Theta,\ r\in[1,\infty), \tag{1}$$

where d is some semi-metric and $\mathbb{E}_{\theta}^{(n)}$ denotes the mathematical expectation with respect to $\mathbb{P}_{\theta}^{(n)}$, a family of probability measures generated by observations. Clearly, it is preferable to have a uniform in θ upper bound on the risk (1), that is, to bound the maximum risk $\sup_{\theta\in\Theta}\mathcal{R}_{r,d}^{(n)}\left(\widehat{Q},Q\right)$ where Θ may be either a subset of a finite-dimensional space (parametric setup) or an infinite-dimensional space (nonparametric setup). In the latter case usually $\Theta=\mathbb{F}$, a sufficiently rich set of functions and the target of estimation is a function $\theta=F\in\mathbb{F}$ which may be, for instance, a regression function in a regression model or a signal in the Gaussian white noise (GWN) model. (These models will be sometimes referred to as “standard” statistical models.) In what follows we will consider a problem of nonparametric estimation at a given point, that is, when $Q(F)=F(t)$ with t in some bounded interval of $\mathbb{R}^{d}$. The corresponding risk of some estimator $\widehat{F}(t)$ of $F(t)$ is then given by

$$\mathcal{R}_{r,t}^{(n)}\left(\widehat{F},F\right)=\left(\mathbb{E}_{F}^{(n)}|\widehat{F}(t)-F(t)|^{r}\right)^{1/r},\qquad t\in\mathcal{D}\subset\mathbb{R}^{d}.$$

One aims at obtaining as great as possible lower bound $\psi_n(\mathbb{F})$ on the minimax risk usually called lower rate of convergence or minimax lower bound,

$$\inf_{\widetilde{F}}\sup_{F\in\mathbb{F}}\mathcal{R}_{r,t}^{(n)}\left(\widetilde{F},F\right)\gtrsim\psi_n(\mathbb{F}),\qquad n\to\infty,$$

where $\mathbb{F}$ is some class of functions. The latter inequality says that, on the class $\mathbb{F}$, the estimators of $F(t)$ cannot converge to $f(t)$ faster than $\psi_n(\mathbb{F})$.

We see that a great advantage of the described above approach is that it allows us to judge the accuracy of arbitrary estimators. However, there are some “subjective” components : the choice of the loss function and the functional class $\mathbb{F}$. In what follows we fix the loss function to be the pointwise semi-norm and investigate the effect of selection of $\mathbb{F}$.

The best obtainable rates for classical statistical models are well known [see, for instance, Ibragimov and Has’minskii (1981)]. Let $F:\mathbb{R}^{d}\to\mathbb{R}$ be either an unknown signal in the GWN model or a regression function. Consider $\mathbb{H}_d(\beta,L)$, $\beta>0,L>0$, the isotropic Hölder class, or, more generally, the anisotropic Hölder classes $\mathbb{H}_d(\boldsymbol{\beta}, L)$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)$, [see Definition 1]. Then (when necessary) under regularity conditions the minimax rates are

$$\psi_n(\beta, L) \;=\; L^{d/(2\beta+d)} n^{-\beta/(2\beta+d)}, \tag{2}$$

$$\psi_n(\gamma, L) \;=\; L^{1/(2\gamma+1)} n^{-\gamma/(2\gamma+1)}, \qquad \gamma^{-1} = \sum_{k=1}^{d} \beta_k^{-1}, \tag{3}$$

for $\mathbb{H}_d(\beta, L)$ and $\mathbb{H}_d(\boldsymbol{\beta}, L)$, respectively. In the anisotropic case (3) the dimension d is hidden in the harmonic mean.

What this story is really about. Nevertheless, being quite common, the choice of $\mathbb{F} = \mathbb{H}(\beta, L)$ or $\mathbb{F} = \mathbb{H}(\boldsymbol{\beta}, L)$ is rather subjective. Suppose that in a "standard" statistical model one seeks an estimator of the value $F(t)$, $t \in \mathcal{D}$, of a function $F : \mathbb{R}^d \to \mathbb{R}$ under a *structural constraint* that there exist an unknown function $f : \mathbb{R}^m \to \mathbb{R}$, $m \le d$, and some unknown linearly independent unit vectors $\theta_k \in \mathbb{S}^{d-1}$, $k = 1, \dots, m$, such that

$$F(x) = f\Big(\theta_1^\top x, \dots, \theta_m^\top x\Big). \tag{4}$$

This model assumption is called "multi-index" and appears, for instance, in semiparametric estimation and dimension reduction problems [see Stone (1985) and Hristache et al. (2001)]. A natural question then arises : whether the rate appearing in the lower bounds on the minimax risk for the smoothness classes of such "structured" functions [for the precise definition see Definition 2] coincide with the rates in (2) and (3) ? In what follows, it will be shown that the answer is negative : the lower bounds contain an additional logarithmic factor.

2 Main results

We start this section with definitions of smoothness classes of functions.

Let $D_j^l g$ denote the l th order partial derivative of $g : \mathbb{R}^m \to \mathbb{R}$ with respect to the variable z_j ; and let $\lfloor \beta_k \rfloor$ be the largest integer strictly less than β_k .

Definition 1 *Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)$, $\beta_k > 0,\ k = 1, \dots, m$, and $L > 0$. A function $g : \mathbb{R}^m \to \mathbb{R}$ belongs to the anisotropic Hölder class $\mathbb{H}_m(\boldsymbol{\beta}, L)$ if g has continuous partial derivatives of all orders $l \le \lfloor \beta_k \rfloor$, $k = 1, \dots, m$, and for all $k = 1, \dots, m$*

$$\|D_k^l f\|_\infty \le L \qquad \forall l \le \lfloor \beta_k \rfloor$$

$$\Big| D_k^{\lfloor \beta_k \rfloor} g(z_1, \dots, z_{k-1}, z_k + \tau, z_{k+1} \dots, z_m) - D_k^{\lfloor \beta_k \rfloor} g(z_1, \dots, z_k, \dots, z_m) \Big|$$
$$\le L \tau^{\beta_k - \lfloor \beta_k \rfloor} \qquad \forall z \in \mathbb{R}^m, \tau \in \mathbb{R}.$$

Definition 2 . *Let* $\Theta_m = (\theta_1, \dots, \theta_m)$ *be a* $d \times m$ *matrix with linearly independent rows* $\theta_k \in \mathbb{S}^{d-1}$ *,* $k = 1, \dots, m$ *. Define the class of (anisotropic) multi-index functions :*

$$\mathbb{F}_{m,d}(\boldsymbol{\beta}, L) = \Big\{ F : \mathbb{R}^d \to \mathbb{R} \ \Big| \ F(x) = f\Big(\Theta_m x\Big),\ f \in \mathbb{H}_m(\boldsymbol{\beta}, L), 1 \le m \le d \Big\}.$$

For $m = 1$ this class consists of the single-index functions $F(x) = f(\theta^\top x)$. The adaptive estimation of such functions was studied in Lepski and Serdyukova (2014).

When $\boldsymbol{\beta} = (\beta, \dots, \beta)$, we will write $\mathbb{F}_{m,d}(\beta, L)$ (isotropic case). In addition, denote by $\mathbb{F}^{\mathrm{anis}}_{m,d}(\boldsymbol{\beta}, L) = \mathbb{F}_{m,d}(\boldsymbol{\beta}, L) \backslash \mathbb{F}_{m,d}(\beta, L)$ the class of multi-index functions with purely anisotropic link functions.

GWN model : We observe a path $\{Y_n(x),\ x \in [-1,1]^d\}$ satisfying the stochastic differential equation

$$Y_n(\mathrm{d}x) = F(x)\mathrm{d}x + \frac{1}{\sqrt{n}} W(\mathrm{d}x), \tag{5}$$

where W is a Brownian sheet, $1/\sqrt{n}$, $n \in \mathbb{N}$, is the deviation parameter and $F \in L_2([-1,1]^d)$.

Regression model : The observations $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ follow

$$Y_i = F(X_i) + \varepsilon_i, \qquad i = 1, \dots, n, \tag{6}$$

where $d \ge 2$, the noise $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. centered random variables and the design points $\{X_i\}_{i=1}^n$ are independent random vectors with common density g with respect to the Lebesgue measure. The sequences $\{\varepsilon_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ are assumed to be independent.

In the both models the target of estimation is the value $F(t)$, $t \in [-1/2, 1/2]^d$. In the case of regression model some additional assumption about the noise and the design should be imposed. In the case of regression model some additional assumption about the noise and the design should be imposed.

Assumption 1 *There exist constants* $q, Q > 0$ *such that, for any* $\upsilon_1, \upsilon_2 \in [-q, q]$*,*

$$\int_{\mathbb{R}} p(y + \upsilon_1) p(y + \upsilon_2) p^{-1}(y) \mathrm{d}y \le 1 + Q \Big| \upsilon_1 \upsilon_2 \Big|.$$

It is easy to see that the Gaussian density $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$, obeys the aforementioned assumption. In general, this assumption is fulfilled if the Fisher information corresponding to the density p is finite and the function $\int \Big[p'(y + \cdot) \Big]^2 p^{-1}(y) \mathrm{d}y$ is continuous at zero.

Assumption 2 *There exist constants* $\mathfrak{g} > 0$ *and* $\varpi > 1$ *such that, for all* $x \in \mathbb{R}^d$ *,*

$$g(x) \le \Big(1 + |x|_2^{\varpi} \Big)^{-1} \mathfrak{g}\,.$$

Here $|\cdot|_2$ *is the Euclidean vector norm on* $\mathbb{R}^d$ *.*

This assumption is very weak and holds for the majority of probability distributions used in statistical applications.

Theorem 1 *Let $F:\mathbb{R}^d\to\mathbb{R}$ satisfying* (4) *be either the unknown signal in the GWN model* (5) *or the regression function in* (6)*. In the latter case it is additionally assumed that Assumptions* (1) *and* (2) *hold. Then, for any $1\leq m<d\,, d\geq 2\,,\ \boldsymbol{\beta}=(\beta_1,\dots,\beta_m)\,,$ $\beta_k>0,\ k=1,\dots,m\,,\ L>0\,,\ r\geq 1\,,\ t\in[-1/2,1/2]^d\,,$ for n sufficiently large*

$$\inf_{\widetilde{F}}\sup_{F\in\mathbb{F}_{m,d}(\boldsymbol{\beta},L)}\mathcal{R}^{(n)}_{r,t}\Big(\widetilde{F},F\Big)\geq\varkappa L^{1/(2\gamma+1)}\Big[n^{-1}\ln(n)\Big]^{\gamma/(2\gamma+1)},\ \gamma^{-1}=\sum_{k=1}^m\beta_k^{-1},$$

where the infimum is taken over all estimators and $\varkappa$ is a constant independent of n and L.

Moreover, if $m=d\,,$

$$\inf_{\widetilde{F}}\sup_{F\in\mathbb{F}^{anis}_{m,d}(\boldsymbol{\beta},L)}\mathcal{R}^{(n)}_{r,t}\Big(\widetilde{F},F\Big)\geq\varkappa L^{1/(2\gamma+1)}\Big[n^{-1}\ln(n)\Big]^{\gamma/(2\gamma+1)}.$$

Remark 1 *It is worth mentioning that the obtained rates of convergence agree with the prominent Stone's dimensionality reduction principle [see Stone (1985)], particularly, we observe in the rate the "effective dimension" m. In the anisotropic case the dimensionality m is hidden in the harmonic mean. In the isotropic case, $1\leq m<d\,,$ the result of the theorem reads as*

$$\inf_{\widetilde{F}}\sup_{F\in\mathbb{F}_{m,d}(\beta,L)}\mathcal{R}^{(n)}_{r,t}\Big(\widetilde{F},F\Big)\geq\varkappa L^{m/(2\beta+m)}\Big[n^{-1}\ln(n)\Big]^{\beta/(2\beta+m)}.$$

Références

Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* **29 :6**, 1537–1566.

Ibragimov, I. A., Has'minskii, R. Z. (1981). *Statistical Estimation. Asymptotic Theory.* Applications of Mathematics, 16. Springer-Verlag, New York-Berlin.

Lepski, O. and Serdyukova, N. (2014). Adaptive estimation under single-index constraint in a regression model. *Ann. Statist.* **42 :1** 1–28.

Stone, C.J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13 :2** 689–705.