

On optimal and universal estimators in learning theory

Vladimir Temlyakov

(University of South Carolina)

1. Introduction. Notations

1. Approximation theory. Recovery of functions.

Deterministic model: given

$$y_i = f(x_i), \quad i = 1, \dots, m, \quad f \in \Theta.$$

Recover $f \in \Theta$ (find an approximant of f). Error of approximation is measured in some norm $\| \cdot \|$.

2. Statistics. Regression theory.

a) Fixed design model: given

$$y_i = f(x_i) + \epsilon_i, \quad x_1, \dots, x_m \text{ — fixed,}$$

ϵ_i — i.i.d., $E\epsilon_i = 0$, $f \in \Theta$.

Find an approximant for f (estimator \hat{f}). The unknown function f is called **regression function**. Error is measured by expectation $E(\|f - \hat{f}\|^2)$.

b) Random design model: given

$$y_i = f(x_i) + \epsilon_i,$$

x_1, \dots, x_m – random, distributed according ρ_X ; ϵ_i – i.i.d. (independent of x_i), $E\epsilon_i = 0$, $f \in \Theta$. Error is measured by expectation $E(\|f - \hat{f}\|^2)$.

c) Distribution-free theory of regression.

Let $X \subset \mathbb{R}^d$, $Y \subset \mathbb{R}$ be Borel sets, ρ be a Borel probability measure on $Z = X \times Y$. For $f : X \rightarrow Y$ define **the error**

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

Consider $\rho(y|x)$ - conditional (with respect to x) probability measure on Y and ρ_X - the marginal probability measure on X (for $S \subset X$, $\rho_X(S) = \rho(S \times Y)$). Define

$$f_\rho(x) := \int_Y y d\rho(y|x).$$

The function f_ρ minimizes the error $\mathcal{E}(f)$. It is known in statistics as the **regression function** of ρ . Given: (x_i, y_i) , $i = 1, \dots, m$, i.i.d. distributed according ρ , $|y| \leq M$ a.e. Find an estimator \hat{f} for f_ρ . Error: $E(\|f_\rho - \hat{f}\|_{L_2(\rho_X)}^2)$. Assume $f_\rho \in \Theta$. For a class Θ consider

$$E(\Theta, m, \hat{f}) := \sup_{f_\rho \in \Theta} E(\|f_\rho - \hat{f}\|_{L_2(\rho_X)}^2)$$

$$E(\Theta, m) := \inf_{\hat{f}} E(\Theta, m, \hat{f}).$$

Reference: a book by L. Györfy, M. Kohler, A. Krzyzak, and H. Walk, A distribution-free theory of nonparametric regression, Springer, Berlin, 2002.

Learning theory

The setting is similar to the above 2c). Our goal is to find an estimator $f_{\mathbf{z}}$, on the base of given data

$\mathbf{z} = ((x_1, y_1), \dots, (x_m, y_m))$ that approximates f_{ρ} (or its projection) well with high probability. We assume that (x_i, y_i) , $i = 1, \dots, m$ are independent and distributed according to ρ . Study the probability distribution function

$$\rho^m \{ \mathbf{z} : \|f_{\rho} - f_{\mathbf{z}}\|_{L_2(\rho_X)}^2 \geq \eta \}.$$

Let $\Theta = W_{\infty}^s$ – Sobolev class with smoothness s (univariate). Stone, 1982, proved

$$E(W_{\infty}^s, m) \gg m^{-\frac{2s}{1+2s}}.$$

Kohler, 2000, proved the corresponding upper estimates.

The previous works in regression theory and learning theory indicate importance of a characteristic of a class Θ closely related to the concept of entropy numbers. For a compact subset Θ of a Banach space B we define the entropy numbers as follows

$$\epsilon_n(\Theta, B) := \inf \{ \epsilon : \exists f_1, \dots, f_{2^n} \in \Theta :$$

$$\Theta \subset \cup_{j=1}^{2^n} (f_j + \epsilon U(B)) \}$$

where $U(B)$ is the unit ball of Banach space B . We denote $N(\Theta, \epsilon, B)$ the covering number that is the minimal number of balls of radius ϵ needed for covering Θ .

In [DeVore, Kerkyacharian, Picard, T., 2004], [Konyagin, T.1, 2004] the restrictions on a class Θ have been imposed in the following forms:

$$\epsilon_n(\Theta, \mathcal{C}) \leq Dn^{-r}, \quad n = 1, 2, \dots, \quad (1.1)$$

$$\Theta \subset DU(\mathcal{C}).$$

or

$$d_n(\Theta, \mathcal{C}) \leq Kn^{-r}, \quad n = 1, 2, \dots, \quad (1.2)$$

$$\Theta \subset KU(\mathcal{C}).$$

Here, $d_n(\Theta, B)$ is the Kolmogorov width. Kolmogorov's n -width for centrally symmetric compact set Θ in Banach space B is defined as follows

$$d_n(\Theta, B) := \inf_L \sup_{f \in \Theta} \inf_{g \in L} \|f - g\|_B$$

In [Konyagin, T.2, 2004] we impose a weaker restriction

$$\epsilon_n(\Theta, L_2(\rho_X)) \leq Dn^{-r}, \quad n = 1, 2, \dots, \quad (1.3)$$

$$\Theta \subset DU(L_2(\rho_X)).$$

We will assume that ρ and Θ satisfy the following condition.

For all $f \in \Theta$, $f : X \rightarrow Y$ is such that (1.4)

$$|f(x) - y| \leq M \quad \text{a.e.}$$

Proper function learning

For a hypothesis space \mathcal{H} denote

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

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

is the **empirical error (risk)** of f . This $f_{\mathbf{z},\mathcal{H}}$ is called the **empirical optimum**.

There are many results on the upper bounds in the following setting. Suppose that $f_{\rho} \in \Theta$. Then for $\eta \geq \eta_m(\Theta)$ for some estimator $f_{\mathbf{z}}$ one gets an upper bound for

$$\rho^m \{ \mathbf{z} : \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\rho}) \geq \eta \}.$$

We proposed to study the following function that we call the **accuracy confidence function**. Let a set \mathcal{M} of admissible measures ρ , and a sequence $\mathbb{E} := \{\mathbb{E}(m)\}_{m=1}^{\infty}$ of allowed classes $\mathbb{E}(m)$ of estimators be given. For $m \in \mathbb{N}$, $\eta > 0$ we define

$$\mathbf{AC}_m(\mathcal{M}, \mathbb{E}, \eta) := \inf_{E_m \in \mathbb{E}(m)} \sup_{\rho \in \mathcal{M}} \rho^m \{ \mathbf{z} : \|f_\rho - f_{\mathbf{z}}\|_{L_2(\rho_X)} \geq \eta \}$$

where E_m is an estimator that maps $\mathbf{z} \rightarrow f_{\mathbf{z}}$. For a example, $\mathbb{E}(m)$ could be a class of all estimators or a class of linear estimators of the form

$$f_{\mathbf{z}} = \sum_{i=1}^m w_i(x_1, \dots, x_m, x) y_i.$$

We let μ be any Borel probability measure defined on X and let $\mathcal{M}(\Theta, \mu)$ denote the set of all ρ such that $\rho_X = \mu$, $|y| \leq 1$, $f_\rho \in \Theta$. We specify $B = L_2(\mu)$ and assume that $\Theta \subset L_2(\mu)$.

Theorem 2.1 [T., 2005] Let μ be a Borel probability measure on X . Assume $r > 0$ and Θ is a compact subset of $L_2(\mu)$ such that $\Theta \subset \frac{1}{2}U(\mathcal{C}(X))$ and

$$\epsilon_n(\Theta, L_2(\mu)) \asymp n^{-r}.$$

Then there exist $\delta_0 > 0$ and $\eta_m^- \leq \eta_m^+$, $\eta_m^- \asymp \eta_m^+ \asymp m^{-\frac{r}{1+2r}}$ such that

$$\mathbf{AC}_m(\mathcal{M}(\Theta, \mu), \eta) \geq \delta_0 \quad \text{for } \eta \leq \eta_m^-;$$

$$C_1 e^{-c_1(r)m\eta^2} \leq \mathbf{AC}_m(\mathcal{M}(\Theta, \mu), \eta) \leq e^{-c_2 m\eta^2}, \quad \eta \geq \eta_m^+.$$

The lower estimate in this theorem is a corollary of the corresponding lower estimates from [DKPT, 2004].

Theorem 2.1 solves the optimization problem. Let us now make some conclusions. First of all, Theorem 2.1 shows that the entropy numbers $\{\epsilon_n(\Theta, L_2(\mu))\}$ determine the behavior of the sequence $\{\mathbf{AC}_m(\mathcal{M}(\Theta, \mu), \eta)\}$ of the **AC**-functions. Secondly, it follows from the proof of Theorem 2.1 that the optimal (in the sense of order) estimator can be always constructed as a least squares estimator. Theorem 2.1 discovers a new phenomenon – **sharp phase transition**. The behavior of the accuracy confidence function changes dramatically within the **critical interval** $[\eta_m^-, \eta_m^+]$. It drops from a constant δ_0 to an exponentially small quantity $\exp(-cm^{1/(1+2r)})$. One may also call the interval $[\eta_m^-, \eta_m^+]$ **the interval of phase transition**.

Projection learning

For a compact in $L_2(\rho_X)$ set W denote by $f_W := (f_\rho)_W$ the $L_2(\rho_X)$ -projection of f_ρ onto W . In other words

$$f_W := \arg \min_{f \in W} \mathcal{E}(f).$$

Theorem 3.1 [Cucker, Smale, 2001], [DKPT, 2004] Assume that W and ρ satisfy (1.1) and (1.4). Then for $\eta \geq A_0(M, D, r)m^{-\frac{r}{1+2r}}$

$$\rho^m \{ \mathcal{E}(f_{\mathbf{z}, W}) - \mathcal{E}(f_W) \geq \eta \} \leq \exp(-c(M)m\eta^2).$$

Theorem 3.2 [KT1, 2004] Assume that ρ and W satisfy (1.1) and (1.4). Then we have the following estimates

$$\rho^m \{ \mathbf{z} : \mathcal{E}(f_{\mathbf{z},W}) - \mathcal{E}(f_W) \geq \eta \} \leq$$

$$C(M, D, r) \exp(-c(M)m\eta^2),$$

provided $r > 1/2$, $m\eta^2 \geq 1$,

$$\rho^m \{ \mathbf{z} : \mathcal{E}(f_{\mathbf{z},W}) - \mathcal{E}(f_W) \geq \eta \} \leq$$

$$C_1(M, D, r) \exp(-c(M, D, r)m\eta^{1/r}),$$

provided $r \in (0, 1/2)$, $m\eta^{1/r} \geq C_2(M, D, r)$.

Theorem 3.3 [KT1, 2004] There exist two positive constants c_1 , c_2 and a class W consisting of two functions 1 and -1 such that for every $m = 2, 3, \dots$ and $\eta \geq m^{-1/2}$ there are two measures ρ_0 and ρ_1 such that for any estimator $f_{\mathbf{z}} \in W$ for one of $\rho = \rho_0$ or $\rho = \rho_1$ we have

$$\rho^m \{ \mathbf{z} : \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_W) \geq \eta \} \geq c_1 \exp(-c_2 m \eta^2).$$

Theorem 3.4 [KT1, 2004] For any $r \in (0, 1/2]$ and for every $m \in \mathbb{N}$ there is $W \subset U(\mathcal{B}([0, 1]))$ satisfying $\epsilon_n(W, \mathcal{B}) \leq (n/2)^{-r}$ for $n \in \mathbb{N}$ such that for every estimator $f_{\mathbf{z}} \in W$ there is a ρ such that $|y| \leq 1$ and

$$\rho^m \{ \mathbf{z} : \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}((f_{\rho})_W) \geq m^{-r} \} \geq 1/7.$$

Let us make a comment on studying the accuracy confidence function for the projection learning problem. Similarly to the case of the proper function learning problem we introduce the corresponding accuracy confidence function

$$\mathbf{AC}_m^p(W, \mathbb{E}, \eta) := \inf_{E_m \in \mathbb{E}(m)} \sup_{\rho} \rho^m \{ \mathbf{z} : \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}((f_{\rho})_W) \geq \eta^2 \}$$

where \sup_{ρ} is taken over ρ such that ρ, W satisfy (1.4).

We note that the behavior of the \mathbf{AC}^p -function is well understood only in the following special cases. Let $r > 1/2$ then (see [KT1, 2004], [T., 2005])

$$\begin{aligned} C_1 \exp(-c_1(M)m\eta^4) &\leq \sup_{W \in \mathcal{S}^r(D)} \mathbf{AC}_m^p(W, \eta) \\ &\leq C(M, D, r) \exp(-c_2(M)m\eta^4) \end{aligned}$$

for $\eta \geq m^{-1/4}$.

Also for $r \geq 1$ (see [KT2, 2004])

$$\begin{aligned} C_1 \exp(-c_1(M)m\eta^4) &\leq \sup_{W \in \mathcal{S}_2^r(D)} \mathbf{AC}_m^p(W, \eta) \\ &\leq C(M, D, r) \exp(-c_3(M)m\eta^4) \end{aligned}$$

provided $\eta \gg m^{-1/4}$.

It would be interesting to find the behavior of

$$\sup_{W \in \mathcal{S}} \mathbf{AC}_m^p(W, \eta)$$

in the following cases: I. $\mathcal{S} = \mathcal{S}^r(D)$, $r \leq 1/2$; II. $\mathcal{S} = \mathcal{S}_2^r(D)$, $r < 1$; III. $\mathcal{S} = \{W : W \in \mathcal{S}_q^r(D), W \text{ is convex}\}$, $q = 1, 2, \infty$.

4. Universal Estimators

We now proceed to results on construction of universal (adaptive) estimators. Let a, β , be two positive numbers. Consider a collection $\mathcal{J}(a, \beta)$ of compacts J_n in $\mathcal{C}(X)$ satisfying

$$N(J_n, \epsilon, \mathcal{C}(X)) \leq (a(1 + 1/\epsilon))^n n^{\beta n}, \quad n = 1, 2, \dots \quad (4.1)$$

The following two theorems form a basis for construction of universal estimators. We begin with the definition of our estimator. Let as above $\mathcal{J} := \mathcal{J}(a, \beta)$ be a collection of compacts J_n in $\mathcal{C}(X)$ satisfying (4.1). We define as above

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

where

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

is the empirical error (risk) of f . This $f_{\mathbf{z}, \mathcal{H}}$ is called the empirical optimum or the least squares estimator.

We take a parameter $A \geq 1$ and consider the following estimator

$$f_{\mathbf{z}}^A := f_{\mathbf{z}}^A(\mathcal{J}) := f_{\mathbf{z}, J_{n(\mathbf{z})}}$$

with

$$n(\mathbf{z}) := \arg \min_{1 \leq j \leq m} \left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}, J_j}) + \frac{Aj \ln m}{m} \right).$$

Denote for a set L of a Banach space B

$$d(\Theta, L)_B := \sup_{f \in \Theta} \inf_{g \in L} \|f - g\|_B.$$

Theorem 4.1 [T., 2005] For $\mathcal{J} := \{J_n\}_{n=1}^{\infty}$ satisfying (4.1) and $M > 0$ there exists $A_0 := A_0(a, \beta, M)$ such that for any $A \geq A_0$ and any ρ such that $\rho, J_n, n = 1, 2, \dots$ satisfy (1.4) we have

$$\|f_{\mathbf{z}}^A - f_{\rho}\|_{L_2(\rho_X)}^2 \leq \min_{1 \leq j \leq m} (3d(f_{\rho}, J_j)_{L_2(\rho_X)}^2 + \frac{4Aj \ln m}{m})$$

with probability $\geq 1 - m^{-c(M)A}$.

Theorem 4.2 [T., 2005] Let compacts $\{J_n\}$ satisfy (4.1) and $M > 0$ be given. There exists $A_0 := A_0(a, \beta, M) \geq 1$ such that for any $A \geq A_0$ and any ρ satisfying

$$d(f_\rho, J_n)_{L_2(\rho_X)} \leq A^{1/2} n^{-r}, \quad n = 1, 2, \dots,$$

and such that $\rho, J_n, n = 1, 2, \dots$, satisfy (1.4) we have for $\eta \geq A^{1/2} \left(\frac{\ln m}{m}\right)^{\frac{r}{1+2r}}$

$$\rho^m \{ \mathbf{z} : \|f_{\mathbf{z}}^A - f_\rho\|_{L_2(\rho_X)} \geq 4A^{1/2}\eta \} \leq C e^{-c(M)m\eta^2}.$$

Corollary 4.1 Let $\mathcal{L} = \{L_n\}_{n=1}^{\infty}$ be a sequence of n -dimensional subspaces of $\mathcal{C}(X)$. For given positive numbers $D, M_1, M := M_1 + D$ there exists $A_0 := A_0(D, M)$ with the following property. For any $A \geq A_0$ there exists an estimator $f_{\mathbf{z}}^A$ such that for any ρ with the properties: $|y| \leq M_1$ a.e. with respect to ρ and

$$d(f_{\rho}, L_n)_{\mathcal{C}(X)} \leq Dn^{-r}, \quad n = 1, 2, \dots, \quad f_{\rho} \in DU(\mathcal{C}(X))$$

we have for $\eta \geq A^{1/2} \left(\frac{\ln m}{m}\right)^{\frac{r}{1+2r}}$

$$\rho^m \{ \mathbf{z} : \|f_{\mathbf{z}}^A - f_{\rho}\|_{L_2(\rho_X)} \geq 4A^{1/2}\eta \} \leq C e^{-c(M)m\eta^2}. \quad (4.2)$$

Corollary 4.1 is an extension of Theorem 4.10 from [DKPT2]. Theorem 4.10 from [DKPT2] gives (4.2) with $e^{-c(M)m\eta^2}$ replaced by $e^{-c(M)m\eta^4}$ under an extra restriction $r \leq 1/2$.

Corollary 4.2 Let $\mathbb{L} := \{\mathcal{L}_n\}_{n=1}^{\infty}$ be a sequence of collections $\mathcal{L}_n := \{L_n^j\}_{j=1}^{N_n}$ of n -dimensional subspaces L_n^j of $\mathcal{C}(X)$. Assume $N_n \leq n^{bn}$. For given positive numbers $D, M_1, M := M_1 + D$ there exists $A_0 := A_0(b, D, M)$ with the following property. For any $A \geq A_0$ there exists an estimator $f_{\mathbf{z}}^A$ such that for any ρ with the properties: $|y| \leq M_1$ a.e. with respect to ρ and

$$\min_{1 \leq j \leq N_n} d(f_{\rho}, L_n^j)_{\mathcal{C}(X)} \leq Dn^{-r}, \quad n = 1, 2, \dots, \quad f_{\rho} \in DU(\mathcal{C}(X))$$

we have for $\eta \geq A^{1/2} \left(\frac{\ln m}{m}\right)^{\frac{r}{1+2r}}$

$$\rho^m \{ \mathbf{z} : \|f_{\mathbf{z}}^A - f_{\rho}\|_{L_2(\rho_X)} \geq 4A^{1/2}\eta \} \leq Ce^{-c(M)m\eta^2}.$$