On optimal estimators in learning theory

Vladimir Temlyakov
(University of South Carolina)

1. Introduction. Notations

1. Approximation theory. Recovery of functions.
Deterministic model: given

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

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

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

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 $X \subset \mathbb{R}^d$, $Y \subset$ be Borel sets, ρ be a Borel probability measure on $Z = X \times Y$. For $f : X \to Y$ 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), $i = 1, \ldots, m$, i.i.d. distributed according ρ, $|y| \le M$ a.e. Find an estimator $\hat{f}$ for f_ρ. Error: $E(\|f_\rho - \hat{f}\|^2_{L_2(\rho_X)})$. Assume $f_\rho \in \Theta$. For a class Θ consider

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

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

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 accoding to ρ. Study the probability distribution function

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

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

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

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 [DKPT], [KT1] the restrictions on a class Θ have been imposed in the following forms:

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

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

or

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

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

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 [KT2] we impose a weaker restriction

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

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

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

$$\text{For all} \quad f \in \Theta, \quad f : X \to Y \quad \text{is such that} \tag{1.4}$$

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

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.

Theorem 2.1 [CS], [DKPT] Assume that ρ and Θ satisfy (1.1) and (1.4). Suppose that $f_\rho \in \Theta$. Then for $\eta \geq A_0(M, D, r)m^{-\frac{r}{1+r}}$

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

Theorem 2.2 [DKPT] Let $f_\rho \in \Theta$ and let ρ and Θ satisfy (1.2) and (1.4). Then there exists an estimator $f_{\mathbf{z}}$ such that for $\eta \geq A_0(M,r)(\ln m/m)^{\frac{2r}{1+2r}}$

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

Theorem 2.3 [KT1] Let $f_\rho \in \Theta$ and let ρ and Θ satisfy (1.1) and (1.4). Then there exists an estimator $f_{\mathbf{z}}$ such that for $\eta \geq A_0(M, D, r) m^{-\frac{2r}{1+2r}}$

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

Theorem 2.4 [KT2] Assume that ρ, Θ satisfy (1.4), and Θ satisfies (1.3) with $r > 1/2$. Suppose also $f_\rho \in \Theta$. Let $m\eta^2 \geq 1$. Then there exists an estimator $f_{\mathbf{z}}$ such that

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

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

Theorem 2.5 [KT2] Let $f_\rho \in \Theta$ and let ρ, Θ satisfy (1.4) and (1.3). Assume $r \in (0, 1/2)$ and $m\eta^{1+1/(2r)} \geq C_1(M, D, r)$. Then there exists an estimator $f_{\mathbf{z}}$ such that

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

$$C(M, D, r) \exp(-c(M, D, r) m\eta^{1+1/(2r)}).$$

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 $\mathbf{z} \to f_{\mathbf{z}}$. For a example, $\mathbb{E}(m)$ could be a class of all estimators or a class of linear estimators of the form

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

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 $\rho_X = \mu$, $|y| \le 1$, $f_\rho \in \Theta$. We specify $B = L_2(\mu)$ and assume that $\Theta \subset L_2(\mu)$.

Theorem 2.6 [T2] 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 $\eta_m^- \le \eta_m^+$, $\eta_m^- \asymp \eta_m^+ \asymp m^{-\frac{r}{1+2r}}$ such that

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

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

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 [CS], [DKPT] 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] 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 [KT2] Assume that ρ, W satisfy (1.4), (1.3) with $r > 1/2$. Let $m\eta^{1+\max(1/r,1)} \geq A_0(M, D, r) \geq 1$. Then there exists an estimator $f_{\mathbf{z}} \in W$ such that

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

$$C_1(M, D, r) \exp(-c_1(M) m\eta^2).$$

Theorem 3.4 [KT2] Assume that ρ, W satisfy (1.4), (1.3) with $r \in (0, 1/2)$. Let $m\eta^{1+1/r} \geq A_0(M, D, r) \geq 1$. Then there exists an estimator $f_{\mathbf{z}} \in W$ such that

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

$$C(M, D, r) \exp(-c(M, D, r) m\eta^{1+1/(2r)}).$$

Theorem 3.5 [KT1] 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 3.6 [KT1] 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 $\sup_\rho$ 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], [T2])

$$C_1 \exp(-c_1(M) m\eta^4) \le \sup_{W \in \mathcal{S}^r(D)} \mathbf{AC}_m^p(W, \eta)$$

$$\le C(M, D, r) \exp(-c_2(M) m\eta^4)$$

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

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

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

$$\leq C(M, D, r) \exp(-c_3(M) m \eta^4)$$

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