Mathematical Notes, vol. 80, no. 6, 2006, pp. 888–892.
Translated from Matematicheskie Zametki, vol. 80, no. 6, 2006, pp. 946–949.

Local Entropy in Learning Theory

Yu. V. Malykhin

Received March 23, 2006; in final form, April 11, 2006

Key words: *entropy*, *ε-entropy*, *local entropy*, *learning theory*, *accuracy confidence function*.

1. NOTATION

If $F(\varepsilon)$ and $G(\varepsilon)$ are functions of $\varepsilon \in (0, \varepsilon_0)$, then, by definition, $F(\varepsilon) \sim G(\varepsilon)$ if

$$\lim_{\varepsilon \to +0} \frac{F(\varepsilon)}{G(\varepsilon)} = 1,$$

and $F(\varepsilon) \asymp G(\varepsilon)$, if $F(\varepsilon) \le aG(\varepsilon)$ and $G(\varepsilon) \le bF(\varepsilon)$ for some positive constants a and b. By log we denote the base 2 logarithm function.

2. SETTING OF THE LEARNING THEORY PROBLEM

Suppose that ρ is an unknown probability measure on $X \times Y$ and the data set

$$\boldsymbol{z} = \boldsymbol{z}^{(m)} = ((x_1, y_1), \dots, (x_m, y_m))$$

is a sample of independent random variables whose probability measure is ρ. The purpose of our paper is to approximate the regression function

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

where $\rho(y|x)$ is the corresponding conditional probability measure. The regression function minimizes the error

$$\mathscr{E}(f) = \int_{X \times Y} (f(x) - y)^2 d\rho$$

with respect to the function f; this fact follows from the equality $\mathscr{E}(f) - \mathscr{E}(f_\rho) = \|f - f_\rho\|^2_{L_2(\rho_X)}$ (the measure ρ_X is the projection of the measure ρ on the set X, i.e., $\rho_X(S) := \rho(S \times Y)$). Thus, f_ρ provides the best characterization of the dependence of y on x.

In this paper, we consider the following setting of the problem. First, we assume that the distribution of X is given, i.e., the marginal measure ρ_X is known. Second, it is assumed that a fixed class Θ of f_ρ is defined *a priori*. Third, the regression function must be approximated with respect the norm on the $L_2(\rho_X)$-space.

Now let us consider the formal setting of the problem. Let us set $X = \mathbb{R}^d$, $Y = [-M, M]$, and $Z = X \times Y$. By an *estimator* we mean an arbitrary map $E_m \colon \boldsymbol{z} \mapsto f_{\boldsymbol{z}}$ which reconstructs the Borel function $f_{\boldsymbol{z}} \colon X \to Y$ corresponding to the data set $\boldsymbol{z} = \boldsymbol{z}^{(m)} \in Z^m$. Suppose that $\mathscr{M}$ is a certain class of Borel probability measures on Z. The *accuracy confidence function* is defined according to Ref. [1]:

$$\mathrm{AC}_m(\mathscr{M}, \eta) = \inf_{E_m} \sup_{\rho \in \mathscr{M}} \rho^m\{\boldsymbol{z} : \|f_\rho - f_{\boldsymbol{z}}\|_{L_2(\rho_X)} \ge \eta\}.$$

 0001-4346/2006/8056-0888

By a probability measure we mean the outer probability measure, because, generally speaking, the set $\{\boldsymbol{z} : \|f_\rho - f_{\boldsymbol{z}}\| \geq \eta\}$ does not need to be measurable with respect to the measure ρ^m. Next, let the Borel probability measure μ on X be fixed, and let a certain set Θ of Borel functions $X \to Y$ be given. By $\mathscr{M}(\Theta, \mu)$ we denote the class of measures ρ such that $\rho_X = \mu$ and $f_\rho \in \Theta$. Thus, our purpose is to study the function $\mathrm{AC}_m(\mathscr{M}(\Theta, \mu), \eta)$.

3. SOME DEFINITIONS

It is obvious that the more "massive" is the class Θ, the larger is the error of the estimators. In order to describe this dependence, we recall the necessary definitions.

Suppose that (S, τ) is a metric space and $A \subset S$. Denote by $N_\varepsilon(A, S)$ the number of elements of the minimal ε-net for A in S. The base 2 logarithm of this quantity will be denoted by $\mathscr{H}_\varepsilon(A, S)$; it is called the *$\varepsilon$-entropy* of A in S. The quantity

$$\varepsilon_n(A, S) := \inf\{\varepsilon > 0 : \mathscr{H}_\varepsilon(A, S) \leq n\}$$

is called the *entropy width* of the set A in S. By P_ε we denote the maximal size of an ε-set-packing in A, i.e.,

$$P_\varepsilon(A) = P_\varepsilon(A, S) := \sup\{n : \exists\, x_1, \ldots, x_n \in A \;\; \tau(x_i, x_j) \geq \varepsilon\}.$$

It is clear that P_ε is independent of the ambient space; sometimes, S will be used in order to specify the metric on A.

Moreover, we will use a slightly different kind of entropy, a "local " entropy. Let us fix $c > 1$. Define a *local set packing number* of the set A:

$$\overline{P}_\varepsilon(c, A) = \overline{P}_\varepsilon(c, A, S) := \sup\{n : \exists x_1, \ldots, x_n \in A \;\; \varepsilon \leq \tau(x_i, x_j) \leq c\varepsilon\}. \tag{1}$$

This definition was introduced in [1]. A similar definition of the entropy was used earlier in [2].

4. PRELIMINARY FACTS

Suppose that the situation is as described in Sec. 2. Let Θ be a compact subset in $L_2(\mu)$; hence all the entropy characteristics of Θ are finite. Set

$$\overline{P}(\varepsilon) := \overline{P}_\varepsilon(c, \Theta, L_2(\mu)).$$

In the paper [1], a lower bound was established in terms of the local entropy for the accuracy confidence function. Before passing to the statement of the theorem, let us consider the following construction. Suppose that $Y = [-1, 1]$. Consider the collection of functions $\{f_i\}_{i=1}^{\overline{P}(\eta)}$ introduced in Definition (1). One can associate a measure ρ_i to each function f_i:

$$d\rho_i(x, y) = \left(\frac{1 + f_i(x)}{2}\, d\delta_1(y) + \frac{1 - f_i(x)}{2}\, d\delta_{-1}(y)\right) d\mu(x),$$

where $d\delta_\xi$ is the Dirac measure of unit mass in ξ. Note that the measures defined in this way belong to $\mathscr{M}(\Theta, \mu)$, because $(\rho_i)_X = \mu$ and $f_{\rho_i} = f_i$.

Theorem [1, Theorem 3.1]. *Let $Y = [-1, 1]$. Suppose that, for some $\eta > 0$, the functions $\{f_i\}$, $i = 1, \ldots, \overline{P}(2\eta)$, from (1) satisfy the condition $\|f_i\|_{C(X)} \leq 1/4$. Then, for any estimator $f_{\boldsymbol{z}}$, there exists an $i \in \{1, \ldots, \overline{P}(2\eta)\}$ such that*

$$\rho_i\{\boldsymbol{z} : \|f_{\boldsymbol{z}} - f_i\|_{L_2(\mu)} \geq \eta\} \geq \min\Big(\frac{1}{2}, (\overline{P}(2\eta) - 1)^{1/2} e^{-8c^2 m\eta^2 - 3/e}\Big), \qquad m = 1, 2, \ldots\,.$$

The upper bound for the accuracy confidence function was obtained in terms of the standard entropy. The following theorem from [3] provides the sharpest estimate.

Theorem [3, Theorem 1.3]. *Suppose that the set Θ satisfies the following condition*:

$$\varepsilon_n(\Theta, L_2(\mu)) \le Dn^{-r}, \quad n = 1, 2, \dots, \qquad \|f\|_{L_2(\mu)} \le D \quad \forall\, f \in \Theta.$$

Then there exists an estimator $f_{\boldsymbol{z}}$ such that, for any $\eta \ge \varepsilon_0 = C(M, D, r)m^{-r/(1+2r)}$, any measure $\rho \in \mathscr{M}(\Theta, \mu)$, and $m \ge 240(M/D)^2$, the following inequality holds:

$$\rho^m\{\boldsymbol{z} : \|f_{\boldsymbol{z}} - f_\rho\|_{L_2(\mu)} \ge \eta\} \le \exp\left(-\frac{m\eta^2}{800M^2}\right).$$

In accordance with this theorem, the estimator $f_{\boldsymbol{z}}$, can be constructed as follows. The minimal $(\varepsilon_0/49)$-set packing $\mathscr{N}$ in $L_2(\mu)$ is considered for the set Θ. Further, we set

$$f_{\boldsymbol{z}} := \arg\min_{f \in \mathscr{N}} \mathscr{E}_{\boldsymbol{z}}(f), \qquad \text{where} \quad \mathscr{E}_{\boldsymbol{z}}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2.$$

In fact, the following lemma was proved in [3]; this assertion will be used further in our paper.

Lemma. *Suppose that $\mathscr{N}$ is a collection of functions $X \to Y$, $\#\mathscr{N} = N$. Suppose that*

$$\min_{f \in \mathscr{N}} \|f - f_\rho\|_{L_2(\rho_X)} \le \frac{\eta}{2}.$$

Then the following inequality holds:

$$\rho^m\{\boldsymbol{z} : \|f_{\boldsymbol{z}} - f_\rho\|_{L_2(\rho_X)} \ge \eta\} \le (N+1)\exp\left(-\frac{m\eta^2}{640M^2}\right)$$

for the estimator $f_{\boldsymbol{z}} = \arg\min_{f \in \mathscr{N}} \mathscr{E}_{\boldsymbol{z}}(f)$.

Remark. If the minimum of $\mathscr{E}_{\boldsymbol{z}}(f)$ is attained for several functions f, we choose an arbitrary one. Thus, the lemma states that the corresponding inequality holds for any function which minimizes $\mathscr{E}_{\boldsymbol{z}}(f)$.

5. MAIN RESULTS

In this paper, we establish upper bounds for the target function in terms of the local entropy.

Suppose that $c = 20$. Thus, $\overline{P}(\eta) := \overline{P}_\eta(20, \Theta, L_2(\mu))$. In the proof of this theorem, $\|\cdot\|$ denotes the norm on the space $L_2(\mu)$.

Theorem. *Suppose that*

$$\overline{P}(\eta) \le \varphi(\eta) \qquad \forall\, \eta > 0,$$

and φ does not increase. There exists an estimator $f_{\boldsymbol{z}}$ such that, for any measure $\rho \in \mathscr{M}(\Theta, \mu)$ and any $\eta > 0$, the following inequality holds:

$$\rho^m\{\boldsymbol{z} : \|f_{\boldsymbol{z}} - f_\rho\| > \eta\} \le c_1\varphi\left(\frac{\eta}{16}\right)\exp(-c_2 m\eta^2), \tag{2}$$

where $c_i = c_i(M)$.

Proof. Set $\eta_0 = M$ and $\eta_j = 2^{-j}\eta_0$. Choose the minimal number k so that $m\eta_k^2 \le 1$. Suppose that A_j are the maximal $(\eta_j/2)$-pack in Θ. Set $\overline{A}_j := A_j \cap B(f_\rho, 5\eta_j)$, where $B(x, R)$ is the closed ball of radius R centered at the point x. Note that, for arbitrary distinct functions $f, g \in \overline{A}_j$, the estimate $\eta_j/2 \le \|f - g\| \le 10\eta_j$ is valid; therefore, $\#\overline{A}_j \le \overline{P}(\eta_j/2)$.

Suppose that Λ_j is the set of $\boldsymbol{z}$ such that the minimum of $\mathscr{E}_{\boldsymbol{z}}$ on the set $\overline{A}_j$ is attained in the interior of the ball $B(f_\rho, \eta_j)$. By applying the lemma (η_j plays the role of η and $\overline{A}_j$ that of $\mathscr{N}$), we obtain

$$\rho^m(\Lambda_j) \geq 1 - \left(\overline{P}\left(\frac{\eta_j}{2}\right) + 1\right) \exp\left(-\frac{m\eta_j^2}{640M^2}\right).$$

Now let us construct the estimator. Set $f_{\boldsymbol{z}}^{(0)} := \arg\min\{\mathscr{E}_{\boldsymbol{z}}(f) : f \in A_0\}$. Next, define the estimators $f_{\boldsymbol{z}}^{(s)}$ by induction using the following formula:

$$f_{\boldsymbol{z}}^{(s+1)} := \arg\min\{\mathscr{E}_{\boldsymbol{z}}(f) : f \in A_{s+1} \cap B(f_{\boldsymbol{z}}^{(s)}, 3\eta_{s+1})\}. \tag{3}$$

As the final estimator $f_{\boldsymbol{z}}$, we take $f_{\boldsymbol{z}}^{(k)}$.

Let us prove by induction that the inequality $\|f_{\boldsymbol{z}}^{(s)} - f_\rho\| \leq \eta_s$ holds for

$$\boldsymbol{z} \in \Lambda^{(s)} := \bigcap_{j=0}^{s} \Lambda_j.$$

Indeed, for $s = 0$, this statement follows from the inclusion $\boldsymbol{z} \in \Lambda_0$. Suppose that the inequality holds for s; let us prove it for $s+1$. The condition $\|f_{\boldsymbol{z}}^{(s)} - f_\rho\| \leq \eta_s = 2\eta_{s+1}$ implies that

$$B(f_\rho, \eta_{s+1}) \subset B(f_{\boldsymbol{z}}^{(s)}, 3\eta_{s+1}) \subset B(f_\rho, 5\eta_{s+1}).$$

Since $\boldsymbol{z}$ belongs to Λ_{s+1}, the minimal value of $\mathscr{E}_{\boldsymbol{z}}$ on the set $\overline{A}_{s+1} = A_{s+1} \cap B(f_\rho, 5\eta_{s+1})$ is attained in the interior of the ball $B(f_\rho, \eta_{s+1})$. Therefore, this minimum is attained exactly on $f_{\boldsymbol{z}}^{(s+1)}$. Hence $f_{\boldsymbol{z}}^{(s+1)} \in B(f_\rho, \eta_{s+1})$ as required to complete the proof by induction.

Now consider the case $\eta > 0$. For $\eta < 4\eta_k$, inequality (2) is obvious. Suppose that $\eta \geq 4\eta_k$. Consider a number s such that $\eta/8 < \eta_s \leq \eta/4$. Then if $z \in \Lambda^{(s)}$, Eq. (3) yields

$$\|f_{\boldsymbol{z}} - f_\rho\| \leq \|f_{\boldsymbol{z}}^{(s)} - f_\rho\| + \|f_{\boldsymbol{z}}^{(s)} - f_{\boldsymbol{z}}^{(k)}\| \leq \eta_s + 3\eta_s \leq \eta.$$

Set $c_0 = 1/(640M^2)$ for brevity. Let us estimate $\rho^m(\Lambda^{(s)})$:

$$\begin{aligned}
1 - \rho^m(\Lambda^{(s)}) &\leq \left(\overline{P}\left(\frac{\eta_s}{2}\right) + 1\right) \exp(-c_0 m\eta_s^2) + (\overline{P}(\eta_s) + 1) \exp(-c_0 m(2\eta_s)^2) + \cdots \\
&\leq 2\varphi\Big(\frac{\eta}{16}\Big) \exp(-c_2 m\eta^2) + 2\varphi\Big(\frac{\eta}{8}\Big) \exp(-2c_2 m\eta^2) + \cdots \\
&\leq 2\varphi\Big(\frac{\eta}{16}\Big) \exp(-c_2 m\eta^2)\big(1 + \exp(-c_2 m\eta^2) + \exp(-2c_2 m\eta^2) + \cdots\big).
\end{aligned}$$

It remains to note that, for $\eta \geq 4\eta_k$, the last factor is bounded. $\square$

6. EXAMPLES

For many classes, a lower bound for the local entropy is given by the usual entropy number. As an example, consider the following lemma from [4].

Lemma [4, Lemma 2.1]. *Let Θ be a compact subset of the Banach space B. Suppose that the estimate*

$$C_1\varphi(\varepsilon) \leq \log P_\varepsilon(\Theta) \leq C_2\varphi(\varepsilon), \qquad \varepsilon \in (0, \varepsilon_1],$$

holds with the function $\varphi(\varepsilon)$ satisfying the following condition: for any $\gamma > 0$, there exists an A_γ such that the inequality $\varphi(A_\gamma\varepsilon) \leq \gamma\varphi(\varepsilon)$ is valid for all $\varepsilon > 0$. Then there exist $c_1 \geq 1$ and $\varepsilon_2 > 0$ such that

$$\log \overline{P}_\varepsilon(c_1, \Theta) \geq C_3 \log P_\varepsilon(\Theta), \qquad \varepsilon \in (0, \varepsilon_2].$$

Let us consider two examples of sets for which the value of $\overline{P}_\varepsilon$ is substantially smaller than P_ε.

The first example is a ball in any finite-dimensional space. For a ball, $\overline{P}_\varepsilon$ is bounded; at the same time, P_ε tends to infinity as $\varepsilon \to 0$.

The second example is as follows. Choose a fixed number $h > 0$. Suppose that A_h is the set of functions $f(z)$ which are analytic in the strip $|\operatorname{Im} z| < h$ and periodic with period 2π, and, for any $u \in (-h, h)$, the inequality

$$\frac{1}{2\pi}\int_0^{2\pi} |f(t+iu)|^2 dt \le 1$$

holds for these functions. Consider this set in the metric space $L_2[0, 2\pi]$, where

$$\|f\|^2 = \frac{1}{2\pi}\int_0^{2\pi} |f(t)|^2 dt.$$

(By abuse of language, one can regard A as a subset of $L_2[0, 2\pi]$.)

As is well known,

$$\log P_\varepsilon(A_h, L_2[0, 2\pi]) \sim \frac{2(\log(1/\varepsilon))^2}{h \log e}.$$

One can prove that the following weak asymptotic estimate holds:

$$\log \overline{P}_\varepsilon(c, A_h, L_2[0, 2\pi]) \asymp \log \frac{1}{\varepsilon}$$

for any fixed $c > 1$.

ACKNOWLEDGMENTS

This research was supported by the Russian Foundation for Basic Research no. 05-01-00066 and by the program "Leading Scientific Schools" under grant no. NSh-3004.2003.1.

BIBLIOGRAPHY

1. R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov, "Mathematical methods for supervised learning," *IMI Preprints*, **22** (2004), 1–51.
2. Y. Yang and A. Barron *Ann. of Statist.*, **27** (1999), no. 5, 1564–1599.
3. V. Temlyakov, "Approximation in learning theory," *IMI Preprints*, **5** (2005), 1–44.
4. V. Temlyakov, "Optimal estimators in learning theory," *IMI Preprints*, **23** (2004), 1–29.

Yu. V. Malykhin
M. V. Lomonosov Moscow State University