APPLICATIONES MATHEMATICAE
34,1 (2007), pp. 47–52

Karol Dziedziul and Barbara Wolnik (Gdańsk)

# NOTE ON UNIVERSAL ALGORITHMS FOR LEARNING THEORY

*Abstract.* We study the universal estimator for the regression problem in learning theory considered by Binev *et al.* This new approach allows us to improve their results.

**1. Introduction.** S. Cucker and S. Smale [1] determined the scope of learning theory. We present a general approach which corresponds to [2] and [3]. The problem is the following. Let $X = [0,1]^d$ and $Y = [-A, A]$. On the product space $Z = X \times Y$ there is an unknown probability Borel measure $\varrho$. We shall assume that the marginal probability measure $\varrho_X(S) = \varrho(S \times Y)$ on $X$ is a Borel measure. We have

$$d\varrho(x,y) = d\varrho(y|x)d\varrho_X(x).$$

We are given the data $\mathbf{z} \subset Z$ of $m$ independent random observations $z_j = (x_j, y_j)$, $j = 1, \dots, m$, identically distributed according to $\varrho$. We are interested in estimating the *regression function*

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

in $L^2(X, \varrho_X)$ norm which will be denoted by $\|\cdot\|$.

To do it let $\mathbf{M} = \{M_v\}_{v \in T}$ denote any family of measurable functions on $X$ such that for all $v \in T$,

$$0 \le M_v(x) \le 1, \quad x \in X, \tag{1}$$

and

$$\sum_{v \in T} M_v(x) = 1, \quad x \in X. \tag{2}$$

2000 *Mathematics Subject Classification*: 68T05, 41A36, 41A45, 62G05.
*Key words and phrases*: nonparametric regression, learning theory.

 

An example is the family $\{\chi_I\}_{I\in T}$, where $\chi_I$ denotes the indicator function of $I$ and $\{I : I \in T\}$ is any partition of $X$ (in [2] the sets $I$ are dyadic cubes). Another example is obtained if we consider a triangulation $T$ of $X$ with vertices $\{v\}_{v\in T}$ and the corresponding system of functions $\{M_v\}_{v\in T}$ which are continuous on $X$, linear on each component of this triangulation and

$$M_v(w) = \begin{cases} 1 & \text{for vertices } w = v, \\ 0 & \text{for } w \neq v. \end{cases}$$

It is not hard to check that the family $\{M_v\}_{v\in T}$ satisfies (1) and (2).

Now for a given family $\mathbf{M}$ we define the operator

$$Q_{\mathbf{M}} f(x) = \sum_{v\in T} c_v(f) M_v(x),$$

where

$$c_v(f) = \frac{\alpha_v(f)}{\varrho_v}, \qquad \alpha_v(f) = \int\limits_X f M_v \, d\varrho_X, \qquad \varrho_v = \int\limits_X M_v \, d\varrho_X,$$

and the estimator

$$f_{\mathbf{z}}(x) = \sum_{v\in T} c_v(\mathbf{z}) M_v(x),$$

where

$$c_v(\mathbf{z}) = \frac{\alpha_v(\mathbf{z})}{\varrho_v(\mathbf{z})}, \qquad \alpha_v(\mathbf{z}) = \frac{1}{m}\sum_{j=1}^{m} y_j M_v(x_j), \qquad \varrho_v(\mathbf{z}) = \frac{1}{m}\sum_{j=1}^{m} M_v(x_j).$$

If $\varrho_v = 0$ then we define $c_v = 0$, and if $\varrho_v(\mathbf{z}) = 0$ then we put $c_v(\mathbf{z}) = 0$. Note also that $E\alpha_v(\mathbf{z}) = \alpha_v$ (here and subsequently, $\alpha_v := \alpha_v(f_\varrho)$, $c_v := c_v(f_\varrho)$) and $E\varrho_v(\mathbf{z}) = \varrho_v$. Moreover

$$\operatorname{Var}(yM_v(x)) \le \int\limits_Z y^2 M_v^2(x) \, d\varrho(x,y) \le A^2 \int\limits_X M_v^2(x) \, d\varrho_X(x),$$

hence

$$\operatorname{Var}(yM_v(x)) \le A^2 \int\limits_X M_v(x) \, d\varrho_X(x) = A^2 \varrho_v, \tag{3}$$

$$\operatorname{Var}(M_v(x)) \le E(M_v(x))^2 \le E(M_v(x)) = \varrho_v. \tag{4}$$

Therefore by Bernstein's inequality we have, for any $\varepsilon > 0$,

$$\operatorname{Prob}\{|\alpha_v - \alpha_v(\mathbf{z})| \ge \varepsilon\} \le 2\exp\Big(-\frac{3m\varepsilon^2}{6A^2\varrho_v + 4A\varepsilon}\Big), \tag{5}$$

$$\operatorname{Prob}\{|\varrho_v - \varrho_v(\mathbf{z})| \ge \varepsilon\} \le 2\exp\Big(-\frac{3m\varepsilon^2}{6\varrho_v + 2\varepsilon}\Big). \tag{6}$$

The main result of this paper is

THEOREM 1.1. *For any family* $\mathbf{M}$,

$$E\|Q_{\mathbf{M}} f_\varrho - f_{\mathbf{z}}\|^2 = O\Big(\frac{N}{m}\Big),$$

*where* $N = |T|$.

The new idea of the proof presented below allows us to improve the result from [2] (in Corollary 2.2 of [2] the above expectation is bounded by $O((N/m)\log N)$).

*Proof.* By (1), (2) and the convexity of the square functions we have

$$\begin{aligned} E\|Q_{\mathbf{M}} f_\varrho - f_{\mathbf{z}}\|^2 &\le \int\limits_X \sum_{v\in T} E|c_v - c_v(\mathbf{z})|^2 M_v(x)\, d\varrho_X(x) \\ &= \sum_{v\in T} E|c_v - c_v(\mathbf{z})|^2 \varrho_v. \end{aligned}$$

Note that if $\varrho_v = 0$ then $E\varrho_v(\mathbf{z}) = 0$, hence $\varrho_v(\mathbf{z}) = 0$ $\varrho^m$-a.e. Consequently,

$$E\|Q_{\mathbf{M}} f_\varrho - f_{\mathbf{z}}\|^2 \le \sum_{v\in T,\ \varrho_v>0} E|c_v - c_v(\mathbf{z})|^2 \varrho_v.$$

Fix $v$ such that $\varrho_v > 0$. We can write

$$E|c_v - c_v(\mathbf{z})|^2 = \int\limits_{\varrho_v(\mathbf{z})>0} |c_v - c_v(\mathbf{z})|^2 + \int\limits_{\varrho_v(\mathbf{z})=0} |c_v|^2.$$

Note that if $\varrho_v(\mathbf{z}) = 0$ $\varrho^m$-a.e. then $M_v(x_j) = 0$ for all $j$, hence $\alpha_v(\mathbf{z}) = 0$ $\varrho^m$-a.e. Thus

$$E|c_v - c_v(\mathbf{z})|^2 = \int\limits_{\varrho_v(\mathbf{z})>0} |c_v - c_v(\mathbf{z})|^2 + \int\limits_{\varrho_v(\mathbf{z})=0} \Big|\frac{\alpha_v - \alpha_v(\mathbf{z})}{\varrho_v}\Big|^2.$$

For $b \ne 0$ and $t \ne 0$ we use the simple inequality

$$\Big|\frac{a}{b} - \frac{s}{t}\Big| \le \frac{1}{|b|}\,|a-s| + \frac{|s|}{|bt|}\,|t-b| \tag{7}$$

to get

$$\Big|\frac{a}{b} - \frac{s}{t}\Big|^2 \le 2\,\frac{|a-s|^2}{b^2} + 2\,\frac{1}{b^2}\,\frac{s^2}{t^2}\,|t-b|^2, \tag{8}$$

which in particular gives

$$\Big|\frac{a_v}{\varrho_v} - \frac{a_v(\mathbf{z})}{\varrho_v(\mathbf{z})}\Big|^2 \le 2\,\frac{|a_v - a_v(\mathbf{z})|^2}{\varrho_v^2} + 2\Big(\frac{a_v(\mathbf{z})}{\varrho_v(\mathbf{z})}\Big)^2 \frac{|\varrho_v - \varrho_v(\mathbf{z})|^2}{\varrho_v^2}.$$

For $\varrho_v(\mathbf{z}) > 0$ we have

$$\frac{\alpha_v(\mathbf{z})^2}{\varrho_v(\mathbf{z})^2} \le A^2,$$

thus

$$E|c_v - c_v(\mathbf{z})|^2 \le \frac{3}{m\varrho_v^2}\operatorname{Var}(yM_v(x)) + \frac{2A^2}{m\varrho_v^2}\operatorname{Var}(M_v(x)).$$

Consequently,

$$E\|Q_T f_\varrho - f_{\mathbf{z}}\|^2 \le C \sum_{v\in T} \frac{1}{m\varrho_v^2} \left(\mathrm{Var}(yM_v(x)) + \mathrm{Var}(M_v(x))\right)\varrho_v.$$

By (3) and (4) we get

$$E\|Q_T f_\varrho - f_{\mathbf{z}}\|^2 \le O\Big(\sum_{v\in T} \frac{1}{m}\Big) = O\Big(\frac{N}{m}\Big),$$

and this finishes the proof.

Note that if we take $N = m^{1/(1+2s)}$ for fixed $s > 0$ then

$$E\|Q_{\mathbf{M}} f_\varrho - f_{\mathbf{z}}\|^2 = O\Big(\frac{1}{m}\Big)^{2s/(1+2s)}. \tag{9}$$

To unify the linear and nonlinear approach in estimation let us introduce the sets $\mathcal{A}^s$ similar to the definition given in [2]. We have $f \in \mathcal{A}^s$, $s > 0$ (in fact it makes sense to consider $0 < s \le 2$) if $f \in L^2(\varrho_X)$ and there is $C$ such that for all $N$ there is a family $\mathbf{M} = \{M_v\}_{v\in T}$ with properties (1) and (2) such that $N = |T|$ and

$$\|f - Q_{\mathbf{M}} f\| \le CN^{-s}. \tag{10}$$

By Theorem 1.2, (9) and (10), and since

$$E\|f_\varrho - f_{\mathbf{z}}\|^2 \le 2E\|f_\varrho - Q_{\mathbf{M}} f_\varrho\|^2 + 2E\|Q_{\mathbf{M}} f_\varrho - f_{\mathbf{z}}\|^2,$$

we get the optimal rate of estimation (see [4]). This approach improves the rate of estimation in [2].

THEOREM 1.2. *Let $f_\varrho \in \mathcal{A}^s$ and let $\mathbf{M}$ be the family from the definition of the space $\mathcal{A}^s$ such that $N = |T| = [m^{1/(1+2s)}]$. Then*

$$E\|f_\varrho - f_{\mathbf{z}}\|^2 = O\Big(\frac{1}{m}\Big)^{2s/(1+2s)}.$$

Finally, we will give a general version of Theorem 2.1 in [2]. Our proof is analogous but partially simplified, so we present it for the sake of completeness. We improve the constant in estimation.

THEOREM 1.3. *For any family $\mathbf{M}$ and any $\eta > 0$,*

$$\mathrm{Prob}\{\|Q_{\mathbf{M}} f_\varrho - f_{\mathbf{z}}\| > \eta\} \le 4Ne^{-cm\eta^2/N}, \tag{11}$$

*where $N := |T|$ and $c$ depends only on $A$.*

*Proof.* By the convexity of the square function we have

$$\begin{aligned} \|Q_{\mathbf{M}} f_\varrho - f_{\mathbf{z}}\|^2 &\le \int_X \sum_{v\in T} |c_v - c_v(\mathbf{z})|^2 M_v(x)\, d\varrho_X(x) \\ &= \sum_{v\in T} |c_v - c_v(\mathbf{z})|^2 \varrho_v. \end{aligned} \tag{12}$$

This gives

$$\begin{aligned}\mathrm{Prob}\{\|Q_{\mathbf{M}}f_\varrho - f_{\mathbf{z}}\| > \eta\} &\le \mathrm{Prob}\Big\{\sum_{v\in T} |c_v - c_v(\mathbf{z})|^2\varrho_v > \eta^2\Big\}\\ &\le \sum_{v\in T} \mathrm{Prob}\Big\{|c_v - c_v(\mathbf{z})| > \frac{\eta}{\sqrt{N\varrho_v}}\Big\}.\end{aligned}$$

Note that

$$\mathrm{Prob}\Big\{|c_v - c_v(\mathbf{z})| > \frac{\eta}{\sqrt{N\varrho_v}}\Big\} = 0$$

provided $\varrho_v \le \eta^2/4A^2N$. To see this it is enough to transform this assumption to the form $\eta/\sqrt{N\varrho_v} \ge 2A$ and recall that $|c_v|$ and $|c_v(\mathbf{z})|$ are less than $A$.

Therefore we can write

$$\mathrm{Prob}\{\|Q_{\mathbf{M}}f_\varrho - f_{\mathbf{z}}\| > \eta\} \le \sum_{v\,:\,\varrho_v > \eta^2/4A^2N} \mathrm{Prob}\Big\{|c_v - c_v(\mathbf{z})| > \frac{\eta}{\sqrt{N\varrho_v}}\Big\}.$$

To estimate the last sum, note that if

$$|\alpha_v(\mathbf{z}) - \alpha_v| \le \frac{\varrho_v\eta}{4\sqrt{N\varrho_v}}$$

and

$$|\varrho_v(\mathbf{z}) - \varrho_v| \le \frac{\varrho_v\eta}{4A\sqrt{N\varrho_v}}$$

then (we know that $\varrho_v > \eta^2/4A^2N$)

$$|\varrho_v(\mathbf{z}) - \varrho_v| \le \frac{\varrho_v\eta}{4A\sqrt{N\frac{\eta^2}{4A^2N}}} = \frac{1}{2}\,\varrho_v$$

(this gives in particular $|\varrho_v(\mathbf{z})| \ge \frac{1}{2}\varrho_v$), and using (7) we get

$$\begin{aligned}|c_v(\mathbf{z}) - c_v| &= \Big|\frac{\alpha_v(\mathbf{z})}{\varrho_v(\mathbf{z})} - \frac{\alpha_v}{\varrho_v}\Big|\\ &\le \frac{1}{|\varrho_v(\mathbf{z})|}\,|\alpha_v(\mathbf{z}) - \alpha_v| + \frac{|\alpha_v|}{|\varrho_v(\mathbf{z})|\varrho_v}\,|\varrho_v(\mathbf{z}) - \varrho_v|\\ &\le \frac{1}{\frac{1}{2}\varrho_v}\cdot\frac{\varrho_v\eta}{4\sqrt{N\varrho_v}} + \frac{A}{\frac{1}{2}\varrho_v}\cdot\frac{\varrho_v\eta}{4A\sqrt{N\varrho_v}} = \frac{\eta}{\sqrt{N\varrho_v}}.\end{aligned}$$

Therefore

$$\begin{aligned}&\mathrm{Prob}\Big\{|c_v - c_v(\mathbf{z})| > \frac{\eta}{\sqrt{N\varrho_v}}\Big\}\\ &\quad\le \mathrm{Prob}\Big\{|\alpha_v(\mathbf{z}) - \alpha_v| > \frac{\varrho_v\eta}{4\sqrt{N\varrho_v}}\Big\} + \mathrm{Prob}\Big\{|\varrho_v(\mathbf{z}) - \varrho_v| > \frac{\varrho_v\eta}{4A\sqrt{N\varrho_v}}\Big\}.\end{aligned}$$

If we first use (5), (6) and then the fact that $\eta/\sqrt{N\varrho_v} \le 2A$, we finally get

$$
\begin{aligned}
&\operatorname{Prob}\{\|Q_{\mathbf{M}} f_\varrho - f_{\mathbf{z}}\| > \eta\} \\
&\qquad \le \sum_{v\,:\,\varrho_v > \eta^2/4A^2N} \Big( 2\exp\Big( -\frac{3m\eta^2}{16N\Big(6A^2 + A\frac{\eta}{\sqrt{N\varrho_v}}\Big)} \Big) \\
&\qquad\quad + 2\exp\Big( -\frac{3m\eta^2}{16A^2N\Big(6 + \frac{1}{2A}\cdot\frac{\eta}{\sqrt{N\varrho_v}}\Big)} \Big) \Big) \\
&\qquad \le \sum_{v\,:\,\varrho_v > \eta^2/4A^2N} 2\Big( \exp\Big( -\frac{3}{128}\cdot\frac{m\eta^2}{NA^2} \Big) + \exp\Big( -\frac{3}{112}\cdot\frac{m\eta^2}{NA^2} \Big) \Big) \\
&\qquad \le 4N\exp\Big( -\frac{3}{128A^2}\cdot\frac{m\eta^2}{N} \Big),
\end{aligned}
$$

which completes the proof of (11) with $c = 3/128A^2$.

## References


[1] S. Cucker and S. Smale, *On the mathematical foundations of learning*, Bull. Amer. Math. Soc. 39 (2001), 1–49.

[2] P. Binev, A. Cohen, W. Dahmen, R. DeVore and V. Temlyakov, *Universal algorithms for learning theory. Part I: piecewise constant functions*, J. Machine Learning Res. 6 (2005), 1297–1321.

[3] —, —, —, —, —, *Universal algorithms for learning theory. Part II: piecewise constant functions*, preprint.

[4] R. DeVore, G. Kerkyacharian, D. Picard and V. Temlyakov, *Approximation methods for supervised learning*, Found. Comput. Math. 1 (2006), 3–58.



Karol Dziedziul
Faculty of Applied Mathematics
Gdańsk University of Technology
Narutowicza 11/12
80-952 Gdańsk, Poland
E-mail: kdz@mifgate.pg.gda.pl

Barbara Wolnik
Institute of Mathematics
Gdańsk University
Wita Stwosza 57
80-952 Gdańsk, Poland
E-mail: Barbara.Wolnik@math.univ.gda.pl



*Received on 16.10.2006;*
*revised version on 15.2.2007* (1839)