

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 ϱ . 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 ϱ . 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 $v \in T$,

$$(1) \quad 0 \leq M_v(x) \leq 1, \quad x \in X,$$

and

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

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 χ_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

$$c_v(f) = \frac{\alpha_v(f)}{\varrho_v}, \quad \alpha_v(f) = \int_X f M_v d\varrho_X, \quad \varrho_v = \int_X M_v d\varrho_X,$$

and the estimator

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

where

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

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

$$\text{Var}(yM_v(x)) \leq \int_Z y^2 M_v^2(x) d\varrho(x, y) \leq A^2 \int_X M_v^2(x) d\varrho_X(x),$$

hence

$$(3) \quad \text{Var}(yM_v(x)) \leq A^2 \int_X M_v(x) d\varrho_X(x) = A^2 \varrho_v,$$

$$(4) \quad \text{Var}(M_v(x)) \leq E(M_v(x))^2 \leq E(M_v(x)) = \varrho_v.$$

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

$$(5) \quad \text{Prob}\{|\alpha_v - \alpha_v(\mathbf{z})| \geq \varepsilon\} \leq 2 \exp\left(-\frac{3m\varepsilon^2}{6A^2\varrho_v + 4A\varepsilon}\right),$$

$$(6) \quad \text{Prob}\{|\varrho_v - \varrho_v(\mathbf{z})| \geq \varepsilon\} \leq 2 \exp\left(-\frac{3m\varepsilon^2}{6\varrho_v + 2\varepsilon}\right).$$

The main result of this paper is



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

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

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

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

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

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

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

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

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

$$E|c_v - c_v(\mathbf{z})|^2 = \int_{\varrho_v(\mathbf{z}) > 0} |c_v - c_v(\mathbf{z})|^2 + \int_{\varrho_v(\mathbf{z}) = 0} \left| \frac{\alpha_v - \alpha_v(\mathbf{z})}{\varrho_v} \right|^2.$$

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

$$(7) \quad \left| \frac{a}{b} - \frac{s}{t} \right| \leq \frac{1}{|b|} |a - s| + \frac{|s|}{|bt|} |t - b|$$

to get

$$(8) \quad \left| \frac{a}{b} - \frac{s}{t} \right|^2 \leq 2 \frac{|a - s|^2}{b^2} + 2 \frac{1}{b^2} \frac{s^2}{t^2} |t - b|^2,$$

which in particular gives

$$\left| \frac{a_v}{\varrho_v} - \frac{a_v(\mathbf{z})}{\varrho_v(\mathbf{z})} \right|^2 \leq 2 \frac{|a_v - a_v(\mathbf{z})|^2}{\varrho_v^2} + 2 \left(\frac{a_v(\mathbf{z})}{\varrho_v(\mathbf{z})} \right)^2 \frac{|\varrho_v - \varrho_v(\mathbf{z})|^2}{\varrho_v^2}.$$

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

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

thus

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



Consequently,

$$E\|Q_T f_\varrho - f_{\mathbf{z}}\|^2 \leq C \sum_{v \in T} \frac{1}{m \varrho_v^2} (\text{Var}(y M_v(x)) + \text{Var}(M_v(x))) \varrho_v.$$

By (3) and (4) we get

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

and this finishes the proof.

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

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

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 $0 < s \leq 2$) 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

$$(10) \quad \|f - Q_{\mathbf{M}} f\| \leq CN^{-s}.$$

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

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

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 $N = |T| = \lceil m^{1/(1+2s)} \rceil$. Then*

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

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 $\eta > 0$,*

$$(11) \quad \text{Prob}\{\|Q_{\mathbf{M}} f_\varrho - f_{\mathbf{z}}\| > \eta\} \leq 4N e^{-c m \eta^2 / N},$$

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

Proof. By the convexity of the square function we have

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



This gives

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

Note that

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

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

Therefore we can write

$$\text{Prob}\{\|Q_{\mathbf{M}}f_{\varrho} - f_{\mathbf{z}}\| > \eta\} \leq \sum_{v: \varrho_v > \eta^2/4A^2N} \text{Prob}\left\{|c_v - c_v(\mathbf{z})| > \frac{\eta}{\sqrt{N\varrho_v}}\right\}.$$

To estimate the last sum, note that if

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

and

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

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

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

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

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

Therefore

$$\begin{aligned} &\text{Prob}\left\{|c_v - c_v(\mathbf{z})| > \frac{\eta}{\sqrt{N\varrho_v}}\right\} \\ &\leq \text{Prob}\left\{|\alpha_v(\mathbf{z}) - \alpha_v| > \frac{\varrho_v \eta}{4\sqrt{N\varrho_v}}\right\} + \text{Prob}\left\{|\varrho_v(\mathbf{z}) - \varrho_v| > \frac{\varrho_v \eta}{4A\sqrt{N\varrho_v}}\right\}. \end{aligned}$$

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



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

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. Kerkycharian, 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)

