Lebesgue type inequalities in approximation and estimation

Vladimir Temlyakov

University of South Carolina

(Texas A&M, October, 2007)

The Lebesgue inequality

A. Lebesgue proved the following inequality: for any 2π-periodic continuous function f one has

$$\|f - S_n(f)\|_\infty \leq (4 + \frac{4}{\pi^2} \ln n) E_n(f)_\infty,$$

where $S_n(f)$ is the nth partial sum of the Fourier series of f and $E_n(f)_\infty$ is the error of the best approximation of f by the trigonometric polynomials of order n in the uniform norm $\| \cdot \|_\infty$.

1. Approximation. Redundant systems

We say a set of functions $\mathcal{D}$ from a Hilbert space H is a dictionary if each $g \in \mathcal{D}$ has norm one $(\|g\| := \|g\|_H = 1)$ and the closure of span$\mathcal{D}$ coincides with H. We let $\Sigma_m(\mathcal{D})$ denote the collection of all functions (elements) in H which can be expressed as a linear combination of at most m elements of $\mathcal{D}$. Thus each function $s \in \Sigma_m(\mathcal{D})$ can be written in the form

$$s = \sum_{g \in \Lambda} c_g g, \quad \Lambda \subset \mathcal{D}, \quad \#\Lambda \le m,$$

where the c_g are real or complex numbers. For a function $f \in H$ we define its best m-term approximation error

$$\sigma_m(f) := \sigma_m(f, \mathcal{D}) := \inf_{s \in \Sigma_m(\mathcal{D})} \|f - s\|.$$

Orthogonal Greedy Algorithm

If H_0 is a finite dimensional subspace of H, we let P_{H_0} be the orthogonal projector from H onto H_0. That is $P_{H_0}(f)$ is the best approximation to f from H_0.

Orthogonal Greedy Algorithm (OGA). We define $f_0 := f$. Then for each $m \geq 1$ we inductively define:

1). $\varphi_m \in \mathcal{D}$ is any element satisfying (we assume existence)

$$|\langle f_{m-1}, \varphi_m \rangle| = \sup_{g \in \mathcal{D}} |\langle f_{m-1}, g \rangle|;$$

2).

$$G_m(f, \mathcal{D}) := P_{H_m}(f), \quad \text{where} \quad H_m := \operatorname{span}(\varphi_1, \dots, \varphi_m);$$

3).

$$f_m := f - G_m(f, \mathcal{D}).$$

Examples

It is clear that for an orthonormal basis $\mathcal{B}$ of a Hilbert space H we have for each f

$$\|f - G_m(f, \mathcal{B})\| = \sigma_m(f, \mathcal{B}).$$

There is a nontrivial classical example of a redundant dictionary, having the same property: OGA realizes the best m-term approximation for each individual function. We describe that dictionary now. Let Π be a set of functions from $L_2([0,1]^2)$ of the form $u(x_1)v(x_2)$ with the unit L_2-norm. Then for this dictionary and $H = L_2([0,1]^2)$ we have for each $f \in H$

$$\|f - G_m(f, \Pi)\| = \sigma_m(f, \Pi).$$

Incoherent dictionaries

We consider dictionaries that have become popular in signal processing. Denote

$$M(\mathcal{D}) := \sup_{g \neq h; g,h \in \mathcal{D}} |\langle g, h \rangle|$$

the coherence parameter of a dictionary $\mathcal{D}$. For an orthonormal basis $\mathcal{B}$ we have $M(\mathcal{B}) = 0$. It is clear that the smaller the $M(\mathcal{D})$ the more the $\mathcal{D}$ resembles an orthonormal basis. However, we should note that in the case $M(\mathcal{D}) > 0$ the $\mathcal{D}$ can be a redundant dictionary.

First results

The first general Lebesgue type inequality for the OGA for the M-coherent dictionary has been obtained in [Gilbert, Muthukrishnan, Strauss, 2003]. They proved that

$$\|f_m\| \le 8m^{1/2}\sigma_m(f) \quad \text{for} \quad m < 1/(32M).$$

The constants in this inequality were improved in [Tropp, 2004] (see also [Donoho, Elad, Temlyakov, 2004]):

$$\|f_m\| \le (1+6m)^{1/2}\sigma_m(f) \quad \text{for} \quad m < 1/(3M).$$

New results

The following inequalities has been obtained in [Donoho, Elad, Temlyakov, 2007].

Theorem 1.1 Let a dictionary $\mathcal{D}$ have the mutual coherence $M = M(\mathcal{D})$. Assume $m \le 0.05 M^{-2/3}$. Then for $l \ge 1$ satisfying $2^l \le \log m$ we have

$$\|f_{m(2^l-1)}\| \le 6 m^{2^{-l}} \sigma_m(f).$$

Corollary 1.1 Let a dictionary $\mathcal{D}$ have the mutual coherence $M = M(\mathcal{D})$. Assume $m \le 0.05 M^{-2/3}$. Then we have

$$\|f_{[m \log m]}\| \le 24 \sigma_m(f).$$

2. Learning Theory

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 $f : X \to Y$ define the error

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

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 $f_\rho(x)$ to be the conditional expectation of y with respect to measure $\rho(\cdot|x)$. The function f_ρ is known in statistics as the regression function of ρ.

Setting

It is clear that if $f_\rho \in L_2(\rho_X)$ then it minimizes the error $\mathcal{E}(f)$ over all $f \in L_2(\rho_X)$: $\mathcal{E}(f_\rho) \le \mathcal{E}(f)$, $f \in L_2(\rho_X)$. Thus, in the sense of error $\mathcal{E}(\cdot)$ the regression function f_ρ is the best to describe the relation between inputs $x \in X$ and outputs $y \in Y$. Now, 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 well with high probability. We assume that (x_i, y_i), $i = 1, \dots, m$ are independent and distributed accoding to ρ. We note that it is easy to see that for any $f \in L_2(\rho_X)$

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L_2(\rho_X)}.$$

Least Squares Estimator

It is well known in statistics that the following way of building $f_{\mathbf{z}}$ provides a near optimal estimator in many cases. First, choose a right hypothesis space $\mathcal{H}$. Second, construct $f_{\mathbf{z},\mathcal{H}} \in \mathcal{H}$ as the empirical optimum (least squares estimator). We explain this in more detail. We define

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

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 (LSE).

Sparse approximants

Let $\mathcal{D}(n,q) := \{g_l^n\}_{l=1}^{N_n}$, $n \in \mathbb{N}$, $N_n \le n^q$, $q \ge 1$, be a system of bounded functions defined on X. We will consider a sequence $\{\mathcal{D}(n,q)\}_{n=1}^{\infty}$ of such systems. In building an estimator, based on $\mathcal{D}(n,q)$, we are going to use n-term approximations with regard to $\mathcal{D}(n,q)$:

$$G_\Lambda := \sum_{l \in \Lambda} c_l g_l^n, \quad |\Lambda| = n. \tag{2.1}$$

A standard assumption that we make in supervised learning theory is that $|y| \le M$ almost surely. This implies that we always assume that $|f_\rho| \le M$.

Boundedness assumption

Denoting $\|f\|_{B(X)} := \sup_{x \in X} |f(x)|$, we rewrite the above assumption in the form $\|f_\rho\|_{B(X)} \leq M$. It is natural to restrict our search to estimators $f_{\mathbf{z}}$ satisfying the same inequality $\|f_{\mathbf{z}}\|_{B(X)} \leq M$. Now, there are two standard in learning theory ways to go. In the first way (I) we are looking for an estimator of the form (2.1) with an extra condition

$$\|G_\Lambda\|_{B(X)} \leq M. \tag{2.2}$$

In the second way (II) we take an approximant G_Λ of the form (2.1) and truncate it, i.e. consider $T_M(G_\Lambda)$, where T_M is a truncation operator: $T_M(u) = u$ if $|u| \leq M$ and $T_M(u) = M \operatorname{sign} u$ if $|u| \geq M$. Then automatically $\|T_M(G_\Lambda)\|_{B(X)} \leq M$.

Hypothesis spaces I

Let us look in more detail at the hypothesis spaces generated in the above two cases. In the case (I) we use the following compacts in $B(X)$ as a source of estimators

$$F_n(q) := \{f : \exists \Lambda \subset [1, N_n], |\Lambda| = n, f = \sum_{l \in \Lambda} c_l g_l^n, \|f\|_{B(X)} \leq M\}.$$

An important good feature of $F_n(q)$ is that it is a collection of sparse (at most n terms) estimators. An important drawback is that it may not be easy to check if (2.2) is satisfied for a particular G_Λ of the form (2.1).

Hypothesis spaces II

In the case (II) we use the following sets in $B(X)$ as a source of estimators

$$F_n^T(q) := \{f : \exists \Lambda \subset [1, N_n], |\Lambda| = n, f = T_M(\sum_{l \in \Lambda} c_l g_l^n)\}.$$

An obvious good feature of $F_n^T(q)$ is that by definition we have $\|f\|_{B(X)} \leq M$ for any f from $F_n^T(q)$. An important drawback of it is that $F_n^T(q)$ has (in general) a rather complex structure. In particular, applying the truncation operator T_M to G_Λ we loose (in general) the sparseness property of G_Λ.

Covering numbers

Now, when we have specified our hypothesis spaces, we can look for an existing theory that provides the corresponding error bounds. The general theory is well developed in the case (I). We will use a variant of such a general theory developed in Temlyakov (2005). This theory is based on the following property of compacts $F_n(q)$, formulated in terms of covering numbers:

$$N(F_n(q), \epsilon, B(X)) \leq (1 + 2M/\epsilon)^n n^{qn}. \qquad (2.3)$$

We now formulate the corresponding results. For a compact Θ in a Banach space B we denote $N(\Theta, \epsilon, B)$ the covering number that is the minimal number of balls of radius ϵ with centers in Θ needed for covering Θ.

Approximation tools

Let a, b, be two positive numbers. Consider a collection $\mathcal{K}(a,b)$ of compacts K_n in $B(X)$ that are contained in the M-ball of $B(X)$ and satisfy the following covering numbers condition

$$N(K_n, \epsilon, B(X)) \leq (a(1+1/\epsilon))^n n^{bn}, \quad n = 1, 2, \ldots. \tag{2.4}$$

The following theorem has been proved in Temlyakov (2005). We begin with the definition of our estimator. Let as above $\mathcal{K} := \mathcal{K}(a,b)$ be a collection of compacts K_n in $B(X)$ satisfying (2.4).

Penalized Least Squares Estimator

We take a parameter $A \geq 1$ and consider the following Penalized Least Squares Estimator (PLSE)

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

with

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

Denote for a set L of a Banach space B

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

Lebesgue type inequality for PLSE

Theorem 2.1 For $\mathcal{K} := \{K_n\}_{n=1}^{\infty}$ satisfying (2.4) and $M > 0$ there exists $A_0 := A_0(a, b, M)$ such that for any $A \geq A_0$ and any ρ such that $|y| \leq M$ a.s. we have

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

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

A particular case

It is clear from (2.3) and from definition of $F_n(q)$ that we can apply Theorem 2.1 to the sequence of compacts $\{F_n(q)\}$ and obtain the following error bound with probability $\geq 1 - m^{-c(M)A}$

$$\|f_{\mathbf{z}}^A - f_\rho\|_{L_2(\rho_X)}^2 \leq \min_{1\leq j\leq m} \left(3d(f_\rho, F_j(q))_{L_2(\rho_X)}^2 + \frac{4Aj\ln m}{m} \right). \tag{2.5}$$

Comments

We note that the inequality (2.5) is the Lebesgue type inequality. Indeed, in the left side of (2.5) we have an error of a particular estimator $f_{\mathbf{z}}^A$ built as the PLSE and in the right side of (2.5) we have $d(f_\rho, F_j(q))_{L_2(\rho_X)}$ - the best error that we can get using estimators from $F_j(q)$, $j = 1, 2, \dots$.

We remind that by construction $f_{\mathbf{z}}^A \in F_{n(\mathbf{z})}(q)$.

Let us now discuss an application of the above theory in the case (II). We cannot apply that theory directly to the sequence of sets $\{F_n^T(q)\}$ because we don't know if these sets satisfy the covering number condition (2.4). However, we can modify the sets $F_n^T(q)$ to make them satisfy the condition (2.4).

Modified hypothesis spaces II

Let $c \geq 0$ and define

$$F_n^T(q,c) := \{f : \exists G_\Lambda := \sum_{l \in \Lambda} c_l g_l^n, \Lambda \subset [1, N_n], |\Lambda| = n,$$

$$\|G_\Lambda\|_{B(X)} \leq C_2 n^c, f = T_M(G_\Lambda)\}$$

with some fixed $C_2 \geq 1$. Then, using the inequality $|T_M(f_1(x)) - T_M(f_2(x))| \leq |f_1(x) - f_2(x)|$, $x \in X$, it is easy to get that

$$N(F_n^T(q,c), \epsilon, B(X)) \leq (2C_2(1 + 1/\epsilon)^n n^{(q+c)n}.$$

Therefore, (2.4) is satisfied with $a = 2C_2$ and $b = q + c$.

Theory versus implementation

The above estimators (built as the PLSE) are very good from the theoretical point of view. Their error bounds satisfy the Lebesgue type inequalities. However, they are not good from the point of view of implementation. For example, there is no simple algorithm to find $f_{\mathbf{z},F_n(q)}$ because $F_n(q)$ is a union of $\binom{N_n}{n}$ M-balls of n-dimensional subspaces. Thus, finding an exact LSE $f_{\mathbf{z},F_n(q)}$ is practically impossible. We now use a remark from Temlyakov (2005) that allows us to build an approximate LSE with good approximation error. We proceed to the definition of the Penalized Approximate Least Squares Estimator (PALSE).

PALSE

Let $\delta := \{\delta_{j,m}\}_{j=1}^{m}$ be a sequence of nonnegative numbers. We define $f_{\mathbf{z},\delta,K_j}$ as an estimator satisfying the relation

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\delta,K_j}) \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},K_j}) + \delta_{j,m}. \tag{2.6}$$

In other words, $f_{\mathbf{z},\delta,K_j}$ is an approximation to the least squares estimator $f_{\mathbf{z},K_j}$.
Next, we take a parameter $A \ge 1$ and define the Penalized Approximate Least Squares Estimator (PALSE)

$$f^A_{\mathbf{z},\delta} := f^A_{\mathbf{z},\delta}(\mathcal{K}) := f_{\mathbf{z},\delta,K_{n(\mathbf{z})}}$$

with

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

Lebesgue type inequality for PALSE

The theory developed in Temlyakov (2005) gives the following control of the error.

Theorem 2.2 Under the assumptions of Theorem 2.1 we have

$$\|f^A_{\mathbf{z},\delta} - f_\rho\|^2_{L_2(\rho_X)} \le \min_{1\le j\le m} \left(3d(f_\rho, K_j)^2_{L_2(\rho_X)} + \frac{4Aj\ln m}{m} + 2\delta_{j,m} \right)$$

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

Greedy algorithms in PALSE

Theorem 2.2 guarantees a good error bound for any penalized estimator built from $\{f_{\mathbf{z},\delta,K_j}\}$ satisfying (2.6). We will use greedy algorithms in building an approximate estimator. We now present results from Temlyakov (2005). We will need more specific compacts $F(n,q)$ and will impose some restrictions on g_l^n. We assume that $\|g_l^n\|_{B(X)} \le C_1$ for all n and l. We consider the following compacts instead of $F_n(q)$

$$F(n,q) := \{f : \exists \Lambda \subset [1, N_n], |\Lambda| = n, f = \sum_{l\in\Lambda} c_l g_l^n, \sum_{l\in\Lambda} |c_l| \le 1\}.$$

Then we have $\|f\|_{B(X)} \le C_1$ for any $f \in F(n,q)$ and $\|f\|_{B(X)} \le M$ if $M \ge C_1$.

Discretization

Let $\mathbf{z} = (z_1, \ldots, z_m)$, $z_i = (x_i, y_i)$, be given. Consider the following system of vectors in $\mathbb{R}^m$:

$$v^{j,l} := (g_l^j(x_1), \ldots, g_l^j(x_m)), \quad l \in [1, N_j].$$

We equip the $\mathbb{R}^m$ with the norm $\|v\| := (m^{-1} \sum_{i=1}^m v_i^2)^{1/2}$. Then

$$\|v^{j,l}\| \le \|g_l^j\|_{B(X)} \le C_1.$$

Consider the following system in $H = \mathbb{R}^m$ with the defined above norm $\|\cdot\|$

$$\mathcal{G} := \{v^{j,l}\}_{l=1}^{N_j}.$$

Relaxed Greedy Algorithm

Finding the estimator

$$f_{\mathbf{z},F(j,q)} = \sum_{l\in\Lambda} c_l g_l^j, \quad \sum_{l\in\Lambda} |c_l| \le 1, \quad |\Lambda| = j, \quad \Lambda \subset [1, N_j],$$

is equivalent to finding best j-term approximant of $y \in \mathbb{R}^m$ from the $A_1(\mathcal{G})$ in the space H. We apply the Relaxed Greedy Algorithm with respect to $\mathcal{G}$ to y and find, after j steps, an approximant

$$v^j := \sum_{l\in\Lambda'} a_l v^{j,l}, \quad \sum_{l\in\Lambda'} |a_l| \le 1, \quad |\Lambda'| = j, \quad \Lambda' \subset [1, N_j],$$

such that

$$\|y - v^j\|^2 \le d(y, A_1(\mathcal{G}))^2 + Cj^{-1}, \quad C = C(M, C_1).$$

Estimator

We define an estimator

$$\hat{f}_{\mathbf{z}} := \hat{f}_{\mathbf{z},F(j,q)} := \sum_{l \in \Lambda'} a_l g_l^j.$$

Then $\hat{f}_{\mathbf{z}} \in F(j,q)$ and

$$\mathcal{E}_{\mathbf{z}}(\hat{f}_{\mathbf{z},F(j,q)}) \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},F(j,q)}) + Cj^{-1}.$$

We denote $\delta := \{Cj^{-1}\}_{j=1}^m$ and define for $A \ge 1$

$$f_{\mathbf{z},\delta}^A := \hat{f}_{\mathbf{z},F(n(\mathbf{z}),q)}$$

with

$$n(\mathbf{z}) := \arg \min_{1 \le j \le m} (\mathcal{E}_{\mathbf{z}}(\hat{f}_{\mathbf{z},F(j,q)}) + \frac{Aj \ln m}{m}).$$

Error bound

By Theorem 2.2 we have for $A \geq A_0(M)$

$$\|f^A_{\mathbf{z},\delta} - f_\rho\|^2_{L_2(\rho_X)} \leq \min_{1\leq j\leq m}(3d(f_\rho, F(j,q))^2 + \frac{4Aj\ln m}{m} + 2Cj^{-1}) \tag{2.7}$$

with probability $\geq 1 - m^{-c(M)A}$.
In particular, (2.7) means that the estimator $f^A_{\mathbf{z},\delta}$ is an estimator that provides the error

$$\|f^A_{\mathbf{z},\delta} - f_\rho\|^2_{L_2(\rho_X)} \ll (\frac{\ln m}{m})^{\frac{2r}{1+2r}}$$

for f_ρ such that $d(f_\rho, F(j,q))_{L_2(\rho_X)} \ll j^{-r}$, $r \leq 1/2$. We note that the estimator $f^A_{\mathbf{z},\delta}$ is based on the greedy algorithm and it can easily be implemented.

BCDD. Relaxed Greedy Algorithm

We now describe an application of greedy algorithms in learning theory from Barron, Cohen, Dahmen, and DeVore (2005). In this application one can use the Orthogonal Greedy Algorithm or the following variant of the Relaxed Greedy Algorithm.

Let $\alpha_1 := 0$ and $\alpha_m := 1 - 2/m$, $m \geq 2$. We set $f_0 := f$, $G_0 := 0$ and inductively define two sequences $\{\beta_m\}_{m=1}^{\infty}$, $\{\varphi_m\}_{m=1}^{\infty}$ as follows

$$(\beta_m, \varphi_m) := \arg \min_{\beta \in \mathbb{R}, g \in \mathcal{D}} \|f - (\alpha_m G_{m-1} + \beta g)\|.$$

Then we set

$$f_m := f_{m-1} - \beta_m \varphi_m, \quad G_m := G_{m-1} + \beta_m \varphi_m.$$

Discretization

For systems $\mathcal{D}(n,q)$ the following estimator is considered in Barron, Cohen, Dahmen, and DeVore (2005). Let as above $\mathbf{z} = (z_1, \dots, z_m)$, $z_i = (x_i, y_i)$, be given. Consider the following system of vectors in $\mathbb{R}^m$:

$$v^{j,l} := (g_l^j(x_1), \dots, g_l^j(x_m)), \quad l \in [1, N_j].$$

We equip the $\mathbb{R}^m$ with the norm $\|v\| := (m^{-1}\sum_{i=1}^m v_i^2)^{1/2}$ and normalize the above system of vectors. Denote the new system of vectors by $\mathcal{G}_j$. Now we apply either the OGA or the above defined version of the RGA to the vector $y \in \mathbb{R}$ with respect to the system $\mathcal{G}_j$. Similar to the above discussed case of the system $\mathcal{G}$ we obtain an estimator $\hat{f}_j$. Next, we look for the penalized estimator built from the estimators $\{\hat{f}_j\}$ in the following way.

BCDD. Estimator

Let

$$n(\mathbf{z}) := \arg \min_{1 \le j \le m} (\mathcal{E}_{\mathbf{z}}(T_M(\hat{f}_j)) + \frac{Aj \log m}{m}).$$

Define

$$\hat{f} := T_M(\hat{f}_{n(\mathbf{z})}).$$

Assuming that the systems $\mathcal{D}(n, q)$ are normalized in $L_2(\rho_X)$ Barron, Cohen, Dahmen, and DeVore (2005) proved the following error estimate.

BCDD. Theorem

Theorem 2.3 There exists $A_0(M)$ such that for $A \geq A_0$ one has the following bound for the expectation of the error

$$E(\|f_\rho - \hat{f}\|^2_{L_2(\rho_X)}) \leq \min_{1\leq j\leq m}(C(A,M,q)j\log m/m$$

$$+ \inf_{h\in \mathrm{span}\mathcal{D}(j,q)}(2\|f_\rho - h\|^2_{L_2(\rho_X)} + 8\|h\|^2_{\mathcal{A}_1(\mathcal{D}(j,q))}/j)). \qquad (2.8)$$

Let us make a comparison of (2.8) with (2.7). First of all, the (2.8) gives an error bound for the expectation and (2.7) gives an error bound with high probability. In this sense (2.7) is better than (2.8). However, the condition $\|g^n_l\|_{B(X)} \leq C_1$ imposed on the systems $\mathcal{D}(n,q)$ in order to obtain (2.7) is more restrictive than the corresponding assumption for (2.8).