APPLICATIONES MATHEMATICAE
41,1 (2014), pp. 1–12

E. Gordienko, J. Ruiz de Chávez and E. Zaitseva (Mexico City)

ON CONVERGENCE OF THE EMPIRICAL MEAN METHOD FOR NON-IDENTICALLY DISTRIBUTED RANDOM VECTORS

Abstract. We consider the following version of the standard problem of empirical estimates in stochastic optimization. We assume that the underlying random vectors are independent and not necessarily identically distributed but that they satisfy a "slow variation" condition in the sense of the definition given in this paper. We show that these assumptions along with the usual restrictions (boundedness and equicontinuity) on a class of functions allow one to use the empirical mean method to obtain a consistent sequence of estimates of infimums of the functional to be minimized. Also, we provide certain estimates of the rate of convergence.

1. Motivation. In this paper we consider the following non-stationary version of the usual one-stage stochastic optimization problem.

Let $\mathcal{A}$ be a given set, $S \subset \mathbb{R}^m$ be a given closed set, and $f : \mathcal{A} \times \mathbb{R}^m \to \mathbb{R}$ be a fixed function measurable with respect to the second argument.

Suppose that on some probability space $(\Omega, \mathcal{F}, P)$ a sequence $\xi_1, \xi_2, \ldots$ of independent random vectors with values in S is defined. The distribution (on $(S, \mathcal{B}_S)$) of ξ_t is denoted by μ_t, $t \geq 1$. These distributions are not supposed to be identical.

The "original" problem of stochastic optimization considered in this paper consists of evaluating the following sequence of "minimal risks" (or "minimal losses")

$$R_t^* := \inf_{\alpha \in \mathcal{A}} Ef(\alpha, \xi_t) \equiv \inf_{\alpha \in \mathcal{A}} \int_S f(\alpha, x)\, \mu_t(dx). \tag{1.1}$$

As is frequently the case in such optimization problems, it is assumed that

2010 *Mathematics Subject Classification*: Primary 90C15; Secondary 65C05.
Key words and phrases: stochastic optimization, empirical distributions, probability metrics, slowly varying measures, consistent estimates, rate of convergence.

DOI: 10.4064/am41-1-1

μ_t, $t \geq 1$, are unknown. Instead, at each ("time") $t = 1, 2, \ldots$ the realizations of $\xi_1, \ldots, \xi_t$ (observations) are available, where ξ_k has distribution μ_k, $k = 1, \ldots, t$.

Let $\hat{\mu}_t := t^{-1} \sum_{k=1}^{t} \delta_{\xi_k}$, $t \geq 1$, be (as in the case of i.i.d. random vectors $\xi_1, \xi_2, \ldots$) *empirical measures*, and define

$$L_t^* := \inf_{\alpha \in \mathcal{A}} \int_S f(\alpha, x)\, \hat{\mu}_t(dx) \equiv \inf_{\alpha \in \mathcal{A}} \frac{1}{t} \sum_{k=1}^{t} f(\alpha, \xi_k), \qquad t \geq 1. \tag{1.2}$$

Also we set

$$\varDelta_t := |R_t^* - L_t^*|, \qquad t \geq 1. \tag{1.3}$$

The aim of the paper is to propose conditions on $\{\mu_t,\, t \geq 1\}$ and on the class of functions $\mathcal{F} := \{f(\alpha, \cdot) : \alpha \in \mathcal{A}\}$ that allow proving that $E\varDelta_t \to 0$ as $t \to \infty$.

The "classical" variant of this problem, when $\xi_1, \xi_2, \ldots$ are i.i.d. random vectors and $R_t^* \equiv R^* := \inf_{\alpha \in \mathcal{A}} Ef(\alpha, \xi_1)$, has been very well studied (under different names: "empirical estimation in stochastic optimization", "empirical mean method", "empirical risk (or losses) minimization", "Monte Carlo approximation", etc.). From a great variety of works on this topic, we just mention the books [1, 15, 18]. In many cases, under different conditions the consistency ($L_t^* - R^* \to 0$) or asymptotic normality of $L_t^* - R^*$ (along with asymptotic normality of the approximation to optimal solutions) has been proved, and also numerous exponential bounds on $P(|L_t^* - R^*| > \varepsilon)$ have been obtained. (Merely as examples we point out the works [1, 5, 7, 9–12, 14–16, 18, 19].)

Concerning the wide field of non-stationary stochastic optimization, it is often assumed that in (1.1) the function f can depend on "time" t, that is, for each $t \geq 1$ in (1.1) a function f_t appears (instead of f). Also in some works it is assumed that in (1.2) in place of $f_t(\alpha, \xi_t)$ one has $f_t(\alpha, \xi_t) + \eta_t$, where η_t is a random "measurement error".

In general, to obtain consistent estimates of the infimums in (1.1) in the non-stationary case (using, for example, empirical measures as in (1.2), stochastic approximation involving the Kiefer–Wolfowitz procedure, methods of stochastic quasigradient (see, e.g., [6]), or other algorithms of stochastic optimization, learning and control (see, e.g., [1, 2, 6, 15, 18]), in many works the presence of certain "parametric models" of a drift $\{f_t,\, t \geq 1\}$ (or "regression"), or some asymptotic properties of $\{f_t,\, t \geq 1\}$ are assumed.

To ensure some consistency of estimators (for example, the convergence to zero of the sequence $\varDelta_t$, $t \geq 1$, from (1.3) in some sense), the aforementioned random sequences $\{\xi_t\}$ and $\{\eta_t\}$ should either vanish as $t \to \infty$, or satisfy some "stationary-like" and "independence-like" conditions which allow one to make use of laws of large numbers (for martingales, for instance) and their uniform versions.

There are many different settings of non-stationary problems in stochastic approximation, estimation and control theory, etc., where the above mentioned "regularity properties" of random disturbances are not assumed. In such cases, as a rule, instead of consistency only some upper bounds on the asymptotic error of estimators can be obtained (for example, in applications of algorithms of stochastic approximation in problems such as tracking or stabilizing control; see e.g. [2] and [17]).

In this paper we are not concerned with methods of finding the infimum in (1.2). Rather we put forward a simpler question: Under what conditions the random variables $\varDelta_t$, $t = 1, 2, \ldots$, in (1.3) converge to zero (say, in probability)? It is clear that this is not the case in general (even when $\mathcal{A}$ consists of a unique element). Moreover, if μ_t varies with $t = 1, 2, \ldots$ in an arbitrary and unpredictable way, then it is impossible to estimate (by any method) R_t^* in (1.1) consistently (as $t \to \infty$) using the observations $\xi_1, \xi_2, \ldots$.

In the next section we define *slowly varying sequences* $\{\mu_t,\, t \geq 1\}$ of probability measures. For such sequences, additionally assuming tightness of $\{\mu_t,\, t \geq 1\}$, and boundedness and equicontinuity of the set $\mathcal{F}$ of functions, we prove that in (1.3), $E\varDelta_t \to 0$ as $t \to \infty$.

In Section 3 we establish some estimates of the rate of vanishing of $\kappa(0, \varDelta_t)$, where κ is the Ky Fan metric (that metrizes convergence in probability).

2. Convergence in $\mathbb{L}_1$. It is well-known that even in the case of i.i.d. vectors ξ_t, $t \geq 1$, the values of L_t^* in (1.2) converge (as $t \to \infty$) to $R^* \equiv R_t^*$ in (1.1) only under certain restrictions on the class of functions $f(\alpha, \cdot) : S \to \mathbb{R}$ in (1.1), which in what follows is denoted by

$$\mathcal{F} := \{f(\alpha, \cdot) : \alpha \in \mathcal{A}\}. \tag{2.1}$$

(See [18] for necessary and sufficient conditions of convergence in probability and with probability one.)

In this section we consider the following class $\mathcal{F}$.

Assumption 1. Functions f from $\mathcal{F}$ are uniformly bounded and equicontinuous at each point $x \in S$.

Let $\mathcal{M}$ denote the set of all probability measures on $(S, \mathcal{B})$ (where $\mathcal{B} \equiv \mathcal{B}_S$ is the Borel σ-algebra), and ρ be any given semimetric on $\mathcal{M}$.

Definition 1. We say that a sequence $\{\mu_t\} \subset \mathcal{M}$ is *slowly varying with respect to ρ* if

$$\lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{t} \rho(\mu_k, \mu_t) = 0. \tag{2.2}$$

Let us set

$$\rho_{\mathcal{F}}(\mu,\nu) := \sup\Big\{\Big|\int_S f\,d\mu - \int_S f\,d\nu\Big| : f \in \mathcal{F}\Big\}, \qquad \mu,\nu \in \mathcal{M}, \tag{2.3}$$

where $\mathcal{F}$ is from (2.1), and let $|\cdot|$ denote the Euclidean norm in $\mathbb{R}^m$.

Theorem 1. *Let Assumption* 1 *hold, let $\{\mu_t\}$ be slowly varying with respect to $\rho_{\mathcal{F}}$, and let $\{\mu_t\}$ be tight, i.e.*

$$\lim_{n\to\infty}\sup_{t\ge1}\mu_t(x\in S : |x| > n) = 0. \tag{2.4}$$

Then

$$\lim_{t\to\infty} E\Delta_t = 0. \tag{2.5}$$

Remark 1. (a) Measurability of Δ_t (defined in (1.3)) follows from the assumptions of Theorem 1 and its proof.

(b) Since $\{\Delta_t\}$ is a bounded sequence, under the above conditions the sequence Δ_t, $t \ge 1$, converges to zero in $\mathbb{L}_p$, for every $p > 0$.

It is not difficult to give examples showing that, in general, all assumptions of Theorem 1 (i.e. Assumption 1, slow variation and tightness) are essential for (2.5) to hold.

Under Assumption 1 the sequence $\{\mu_t\}$ is slowly varying in $\rho_{\mathcal{F}}$ if, for example, $\{\mu_t\}$ is slowly varying with respect to the total variation metric $\rho = V$. If $\mathcal{F}$ consists of bounded Lipschitzian functions (with the same Lipschitz constant for all $f \in \mathcal{F}$), then $\rho_{\mathcal{F}} \le cd$ for some $c < \infty$, where

$$\begin{aligned} d &= d(\mu,\nu) \\ &:= \sup\Big\{\Big|\int_S f\,d\mu - \int_S f\,d\nu\Big| : f \text{ with } \|f\|_\infty + \sup_{x\ne y}\frac{|f(x)-f(y)|}{|x-y|} \le 1\Big\}, \end{aligned}$$

is the Dudley metric, [4]. Thus, the second condition in Theorem 1 is satisfied if $\{\mu_t\}$ is slowly varying with respect to d (or with respect to the Lévy–Prokhorov metric, see [4]).

Note that if $\{\mu_t\}$ converges in ρ, then it is slowly varying with respect to ρ. The following example provides slowly varying sequences which may not be convergent.

Example 1. Let $S = \mathbb{R}$, and suppose that for every $t \ge 1$ the distribution μ_t has a density $g_t(x) = g(x - a_t)$, $x \in \mathbb{R}$, where g is a fixed density and $a_t \in \mathbb{R}$ is a shift parameter. Assume that

$$a_t = h(\log^p(t)), \quad t = 1,2,\ldots, \quad \text{where } p \in (0,1),$$

and that the function h satisfies the Lipschitz condition (say, with a constant L).

It follows that $\{\mu_t\}$ is slowly varying with respect to the Dudley metric d. Moreover, if h is bounded, then $\{\mu_t\}$ is tight. (The simplest example: $\mu_t \sim \mathrm{Norm}(a_t, \sigma)$ where $a_t = \sin(\log^p(t))$.)

Indeed, let $1 < m < t$. Then for every function $\varphi : \mathbb{R} \to \mathbb{R}$ with

$$\|\varphi\|_{BL} := \|\varphi\|_\infty + \|\varphi\|_L := \|\varphi\|_\infty + \sup_{x \neq y} \frac{|\varphi(x) - \varphi(y)|}{|x - y|} \leq 1 \tag{2.6}$$

we have

$$\begin{aligned} &\Big|\int \varphi(x) g(x - a_m)\, dx - \int \varphi(x) g(x - a_t)\, dx\Big| \\ &\qquad \leq L|\log^p(m) - \log^p(t)| \leq \frac{Lp}{\log^{1-p}(m)} \, \frac{t - m}{m}. \end{aligned} \tag{2.7}$$

A sequence $\{\varepsilon_t\}$ of nonnegative numbers can be chosen in such a way that $\varepsilon_t \to 0$ and $\varepsilon_t \log^{1-p}(t\varepsilon_t) \to \infty$ (for example, $\varepsilon_t := 1/\log^\alpha(t)$ with small enough positive α).

We set $m = m(t) := [t\varepsilon_t] + 1$, where $[\cdot]$ stands for integer part.

Then since $d \leq 1$ we obtain in (2.2) with $\rho = d$ (using (2.7))

$$\frac{1}{t} \sum_{k=1}^{t} d(\mu_\kappa, \mu_t) \leq \frac{m(t)}{t} + \frac{Lp}{\log^{1-p}(t\varepsilon_t)} \, \frac{(t - t\varepsilon_t)^2}{t^2 \varepsilon_t} \to 0 \quad \text{as } t \to \infty.$$

3. Rate of convergence estimation.

3. Rate of convergence estimation. Estimation of the rate of vanishing of Δ_t can be important from the computational point of view, and in some problems of probabilistic pattern recognition and learning (see, e.g., [1,18]).

We will measure the distance between zero and Δ_t in terms of the Ky Fan metric:

$$\kappa(0, \Delta_t) := \inf\{\varepsilon : P(\Delta_t > \varepsilon) < \varepsilon\}. \tag{3.1}$$

The conditions of Theorem 2 below impose rather strong restrictions on the moments of μ_t, but allow unbounded functions f in the set $\mathcal{F}$.

Assumption 2. There are nonnegative finite numbers a, b, γ and σ such that each function f from $\mathcal{F}$ in (2.1) satisfies the following conditions ($\alpha \in \mathcal{A}$; $n = 1, 2, \ldots$):

$$|f(x, \alpha)| \leq a n^\gamma, \qquad |f(x, \alpha) - f(y, \alpha)| \leq b n^\sigma |x - y| \tag{3.2}$$

for every $x, y \in S$ such that $|x|, |y| \leq n$.

Theorem 2. *Let Assumption 2 hold, and for every* $\lambda > 0$,

$$\sup_{t \geq 1} \int_S |x|^\lambda \, \mu_t(dx) \leq K_\lambda < \infty. \tag{3.3}$$

Then for each fixed $\delta \in (0, (2+m)^{-1})$ *there exists a sequence* $c(\delta, t)$, $t = 1, 2, \ldots,$ *such that*

$$\lim_{t\to\infty} c(\delta, t) = 0, \tag{3.4}$$

$$\kappa(0, \varDelta_t) \le c(\delta, t)t^{-\delta} + \min\Big\{1, t^{-1}\sum_{k=1}^{t} \rho_{\mathcal{F}}(\mu_k, \mu_t)\Big\}, \qquad t = 1, 2, \ldots. \tag{3.5}$$

In the last theorem the semimetric $\rho_{\mathcal{F}}$ is defined by (2.3) with the family of functions $\mathcal{F}$ specified in Assumption 2. If $\mathcal{F}$ contains unbounded functions, then the property of "slow variation" of $\{\mu_t\}$ in $\rho_{\mathcal{F}}$ in general does not hold when $\{\mu_t\}$ is slowly varying, say, with respect to the total variation metric. Nevertheless, the less restrictive moment conditions than in (3.3) allow one to prove the following assertion.

Proposition 1. *Suppose Assumption* 2 *holds, and for some* $\lambda > 1 + \gamma$ (*where* γ *is from* (3.2)) *inequality* (3.3) *is satisfied. Then, if* $\{\mu_t\}$ *is slowly varying with respect to the Dudley metric* d, *then it is slowly varying with respect to* $\rho_{\mathcal{F}}$.

Moreover, for each $1 \le k < t$,

$$\begin{aligned}\rho_{\mathcal{F}}(\mu_k, \mu_t) \le \inf_{N=1,2,\ldots} \Big\{&[aN^{\gamma} + bN^{\sigma}]\, d(\mu_k, \mu_t) \\ &+ 2aK_\lambda\Big[\sum_{n=N+1}^{\infty} n^{\gamma}(n-1)^{-\lambda} + N^{\gamma-\lambda}\Big]\Big\}.\end{aligned} \tag{3.6}$$

Corollary 1. *Suppose that* $\{\mu_t\}$ *is slowly varying with respect to the Dudley metric. Then under the conditions of Theorem* 2,

$$\kappa(0, \varDelta_t) \to 0 \quad \text{as } t \to \infty. \tag{3.7}$$

Remark 2. Combining (3.5) and (3.6) yields an estimate on the vanishing rate of $\kappa(0, \varDelta_t)$ (which, in particular, depends on $t^{-1}\sum_{k=1}^{t} d(\mu_k, \mu_t)$).

4. Proofs

4.1. The proof of Theorem 1. In view of Assumption 1 the set of functions $\mathcal{F}$ in (2.1) is relatively compact in the topology of uniform convergence on bounded sets from S. Let us show that under (2.4) this topology can be metrized by the following metric:

$$r(\varphi, \psi) := \sup_{t\ge 1}\sum_{n=1}^{\infty} h_{n,t} \sup_{x\in Q_n} |\varphi(x) - \psi(x)|, \qquad \varphi, \psi \in \mathcal{F}, \tag{4.1}$$

where

$$Q_n := \{x \in S : n - 1 \le |x| < n\}, \qquad n = 1, 2, \ldots, \tag{4.2}$$

$$h_{n,t} := \max\{n^{-2}, \mu_t(Q_n)\}, \qquad n, t = 1, 2, \ldots. \tag{4.3}$$

Note that for every n, t,

$$0 < h_{n,t} \le 1. \tag{4.4}$$

First, suppose $\varphi_m \to \varphi$ uniformly on each bounded set, and $\varepsilon > 0$. From (2.4) it follows that there exists N_1 such that $\sup_{t\ge 1}\sum_{n=N_1}^{\infty}\mu_t(Q_n) < \varepsilon$, or for some $N \ge N_1$, $\sup_{t\ge 1}\sum_{n=N}^{\infty} h_{n,t} < \varepsilon$. Then from (4.1)–(4.3) we find that for all large enough m, $r(\varphi_m, \varphi) < \varepsilon + 2b\varepsilon$, where b is a constant that bounds the functions from $\mathcal{F}$.

Second, if $r(\varphi_m, \varphi) \to 0$ then (since $h_{n,t} \ge 1/n^2$) for every n,

$$\sup_{x\in Q_n} |\varphi_m(x) - \varphi(x)| \to 0,$$

and the uniform convergence on bounded sets holds.

Now, denote by $\mathcal{L}(\mathcal{F})$ the set of all functions $g : \mathcal{F} \to \mathbb{R}$ bounded by the above defined constant b and satisfying the Lipschitz condition with respect to the metric r in (4.1).

For every integer $1 \le j \le t$, let

$$Y_{j,t} := \{\varphi(\xi_j) - E\varphi(\xi_t) : \varphi \in \mathcal{F}\}, \tag{4.5}$$

$$Z_j := \{\varphi(\xi_j) - E\varphi(\xi_j) : \varphi \in \mathcal{F}\}. \tag{4.6}$$

(That is, $Y_{j,t}$ and Z_j are real valued "random processes" indexed by $\varphi \in \mathcal{F}$.)

Let us show that for each $\omega \in \Omega$ (where $(\Omega, \mathcal{F}, P)$ is the probability space where ξ_t, $t \ge 1$, are defined), $Y_{j,t}, Z_j \in \mathcal{L}(\mathcal{F})$.

If $\varphi, \psi \in \mathcal{F}$ and $\xi_j \in Q_n$ then

$$|\varphi(\xi_j) - \psi(\xi_j)| \le h_{n,j}^{-1} h_{n,j} \sup_{x\in Q_n} |\varphi(x) - \psi(x)|,$$

or (see (4.1), (4.3))

$$|\varphi(\xi_j) - \psi(\xi_j)| \le \ell(\xi_j) r(\varphi, \psi) \tag{4.7}$$

for arbitrary ξ_j. In (4.7),

$$\ell(\xi_j) := h_{n,j}^{-1} \quad \text{when } \xi_j \in Q_n. \tag{4.8}$$

On the other hand, in view of (4.3),

$$\begin{aligned} |E\varphi(\xi_j) - E\psi(\xi_j)| &= \Big|\sum_{n=1}^{\infty} E\{[\varphi(\xi_j) - \psi(\xi_j)],\, \xi_j \in Q_n\}\Big| \\ &\le \sum_{n=1}^{\infty} h_{n,j}^{-1} P(\xi_j \in Q_n) h_{n,j} \sup_{x\in Q_n} |\varphi(x) - \psi(x)| \le r(\varphi, \psi). \end{aligned} \tag{4.9}$$

From (4.7) and (4.9) we see that the random functions $Y_{j,t}(\varphi)$, $Z_j(\varphi)$ in (4.5), (4.6) satisfy the Lipschitz condition with the random constant $L \equiv L(\xi_j) := 1 + \ell(\xi_j)$, where $\ell(\xi_j)$ was defined in (4.8).

For every function $Z \in \mathcal{L}(\mathcal{F})$ let us set

$$\|Z\| := \sup\{|Z(\varphi)| : \varphi \in \mathcal{F}\}. \tag{4.10}$$

From the definitions given in (1.1)–(1.3),

$$\begin{aligned} \varDelta_t &\le \sup_{\alpha\in\mathcal{A}} \Big| t^{-1}\sum_{j=1}^{t} f(\xi_j,\alpha) - \int_S f(x,\alpha)\,\mu_t(dx)\Big| \\ &= \sup_{\varphi\in\mathcal{F}} \Big| t^{-1}\sum_{j=1}^{t} [\varphi(\xi_j) - E(\xi_t)]\Big|. \end{aligned} \tag{4.11}$$

Thus by (4.5), (4.6), (4.10) and (4.11) we get

$$\begin{aligned} \varDelta_t \le \Big\| t^{-1}\sum_{j=1}^{t} Y_{j,t}\Big\| &= \Big\| t^{-1}\sum_{j=1}^{t} Z_j - t^{-1}\sum_{j=1}^{t}(Z_j - Y_{j,t})\Big\| \\ &\le \Big\| t^{-1}\sum_{j=1}^{t} Z_j\Big\| + t^{-1}\sum_{j=1}^{t}\|Z_j - Y_{j,t}\| \\ &= \Big\| t^{-1}\sum_{j=1}^{t} Z_j\Big\| + t^{-1}\sum_{j=1}^{t}\rho_{\mathcal{F}}(\mu_j,\mu_t). \end{aligned} \tag{4.12}$$

The last equality follows from the definition of $\rho_{\mathcal{F}}$ in (2.3).

By (4.12) we will prove (2.5) if we show that

$$\lim_{t\to\infty} E\Big\| t^{-1}\sum_{j=1}^{t} Z_j\Big\| = 0. \tag{4.13}$$

Note that $Z_1, Z_2, \ldots$ are i.i.d. bounded zero-mean random vectors with values in the normed space $(\mathcal{L}(\mathcal{F}), \|\cdot\|)$.

Let $\mathcal{G}$ be the space of all bounded continuous real-valued functions on S equipped with the topology of uniform convergence on bounded subsets of S. And let $(\overline{\mathcal{F}}, r)$ be the closure of $(\mathcal{F}, r)$ in $\mathcal{G}$. Then $(\overline{\mathcal{F}}, r)$ is a compact metric space. It is known (see e.g. [4]) that every function $Z \in \mathcal{L}(\mathcal{F})$ can be extended to a real-valued function $\overline{Z}$ defined on $\overline{\mathcal{F}}$ without changing its supremum and Lipschitz norms (see (2.6) for the definition of the last one).

Let $C(\overline{\mathcal{F}})$ be the Banach space of all bounded continuous real-valued functions on the compact set $\overline{\mathcal{F}}$ (equipped with the uniform norm), and let $\mathcal{L}(\overline{\mathcal{F}}) \subset C(\overline{\mathcal{F}})$ be the subset consisting of all the above mentioned extensions of functions from $\mathcal{L}(\mathcal{F})$.

Let also $\overline{Z}_j := \{\varphi(\xi_j) - E\varphi(\xi_j) : \varphi \in \overline{\mathcal{F}}\}$ be extensions to $\overline{\mathcal{F}}$ of the random functions from (4.6), (4.13). Then $\overline{Z}_1, \overline{Z}_2, \ldots$ are i.i.d. zero-mean bounded random vectors in $C(\overline{\mathcal{F}})$, and (4.13) will be proved if we show that

$$\lim_{t\to\infty} E\Big\|t^{-1}\sum_{j=1}^{t}\overline{Z}_j\Big\| = 0. \tag{4.14}$$

For every $c \ge 1$ the set

$$\mathcal{K}(c) := \{\overline{Z} \in C(\overline{\mathcal{F}}) : \|\overline{Z}\| \le b \text{ and } |\overline{Z}(\varphi) - \overline{Z}(\psi)| \le c r(\varphi,\psi) \text{ for all } \varphi,\psi \in \overline{\mathcal{F}}\} \tag{4.15}$$

is closed, uniformly bounded and uniformly continuous. Thus $\mathcal{K}(c)$ is compact in $C(\overline{\mathcal{F}})$.

We will show that

$$\lim_{c\to\infty}\sup_{j\ge1}\int_{\{\overline{Z}_j\notin\mathcal{K}(c)\}}\|\overline{Z}_j\|\,dP = 0. \tag{4.16}$$

Then we can apply Theorem 2.4 from [8], which states that under condition (4.16) the relation (4.14) holds true.

Because the uniform and the Lipschitzian norms of Z_j and $\overline{Z}_j$ are the same, the event $\{\overline{Z}_j \notin \mathcal{K}(c)\}$ implies (in view of (4.15) and the fact that Z_j is Lipschitzian with constant $1+\ell(\xi_j)$) the event $\{\ell(\xi_j) > c-1\}$. When $\xi_j \in Q_n$ the last event (see (4.3), (4.8)) means that $[\max\{n^{-2},\mu_j(Q_n)\}]^{-1} > c-1$, or $n > c-1$. Thus,

$$\{\ell(\xi_j) > c-1\} \subset \{|\xi_j| \ge \sqrt{c-1}-1\}.$$

By (2.4) $\lim_{c\to\infty}\sup_{j\ge1} P(|\xi_j| \ge \sqrt{c-1}-1) = 0$.

Finally, observe that according to the above argument,

$$\sup_{j\ge1}\int_{\{\overline{Z}_j\notin\mathcal{K}(c)\}}\|\overline{Z}_j\|\,dP \le b\sup_{j\ge1} P(\ell(\xi_j) > c-1).$$

Thus we have checked (4.16). ∎

4.2. On the proof of Theorem 2. Denoting in (4.12),

$$X_t := \Big\|t^{-1}\sum_{j=1}^{t} Z_j\Big\|, \qquad \beta_t := t^{-1}\sum_{j=1}^{t}\rho_{\mathcal{F}}(\mu_j,\mu_t), \tag{4.17}$$

we see that

$$\varDelta_t \le X_t + \beta_t, \qquad t \ge 1. \tag{4.18}$$

It is easy to show that

$$\kappa(0, X_t+\beta_t) \le \kappa(0,X_t) + \kappa(0,\beta_t), \tag{4.19}$$

$$\kappa(0,\beta_t) = \min\{1,\beta_t\}. \tag{4.20}$$

Thus inequality (3.5) in Theorem 2 will follow from (4.18)–(4.20), provided we show that

$$\kappa(0, X_t) \le c(\delta, t)t^{-\delta}, \qquad t \ge 1. \tag{4.21}$$

Note that by (4.6) and (4.10),

$$X_t = \sup_{\varphi\in\mathcal{F}}\Big|t^{-1}\sum_{j=1}^{t}\varphi(\xi_j) - E\varphi(\xi_j)\Big|. \tag{4.22}$$

In the case of i.i.d. random vectors $\xi_1, \xi_2, \ldots$ in (4.22), under different hypotheses about the class $\mathcal{F}$ (for instance under certain entropy conditions or boundedness and Lipschitz conditions) there are many results on bounding (often exponential in $t \ge 1$) the probabilities $P(X_t > \varepsilon)$ (see for instance [1, 5, 9–12, 14–16, 18, 19]).

Under the conditions of Theorem 2 (but with i.i.d. random vectors $\xi_1, \xi_2, \ldots$), in [7] it was proved that for each fixed $\delta \in (0, (2+m)^{-1})$,

$$\lim_{t\to 0} t^{\delta} P(X_t \ge t^{-\delta}) = 0. \tag{4.23}$$

As in many other works on this topic, to prove the results in [7] the following procedure was performed:

(a) Making use of (3.3), the functions φ from $\mathcal{F}$ are reduced (in a certain sense) to functions defined on balls of finite radius.

(b) Due to Assumption 2 on such balls these functions are bounded and satisfy the Lipschitz condition.

(c) Therefore, the corresponding function subsets are compact in the uniform metric.

(d) To estimate the size of ε-nets (for $\|\cdot\|_\infty$) the results from [13] were used.

(e) To estimate $P(\sup_{\varphi\in\Phi_m}\big|\frac{1}{t}\sum_{j=1}^{t}\varphi(\xi_j) - E\varphi(\xi_j)\big| > \varepsilon)$, where the Φ_m are certain finite sets, the Hoeffding inequality was used.

The most important point is that the proof of (4.23) given in [7] basically did not use the identity of the distributions of the random vectors $\xi_1, \xi_2, \ldots$. Thus, under the assumptions of Theorem 2 this proof can be easily adapted (with minor changes) to the case of non-identically distributed $\xi_1, \xi_2, \ldots$. For these reasons we omit this rather lengthy proof.

Finally, note that (4.23) implies (4.21) with $c(\delta, t) \to 0$ as $t \to \infty$. ■

4.3. The proof of Proposition 1. From the definitions (2.2) and inequality (3.6) it readily follows that $\{\mu_t\}$ is slowly varying with respect to $\rho_{\mathcal{F}}$ if it is slowly varying in the Dudley metric d.

Let $\varphi \in \mathcal{F}$ be arbitrary, but for now fixed.

For every $1 \le k < t$ fixed we have (see (4.2))

$$\begin{aligned} \varGamma(\varphi) &:= \Big|\int\limits_S \varphi\,[d\mu_k - d\mu_t]\Big| \\ &\le \Big|\int\limits_{S_N} \varphi\,[d\mu_k - d\mu_t]\Big| + \sum_{n=N+1}^{\infty} \int\limits_{Q_n} |\varphi|\,[d\mu_k + d\mu_t], \end{aligned} \tag{4.24}$$

where $S_N = \bigcup_{n=1}^{N} Q_n$ and N is an arbitrary natural number. By Assumption 2 on the set Q_n the function φ is bounded by aN^{γ} and φ satisfies the Lipschitz condition with constant bN^{σ}. Thus (see e.g. [4]) φ can be extended to S in such a way that the resulting extension $\widehat{\varphi}$ is a function with the same (as φ) uniform norm $\|\cdot\|_{\infty}$ and Lipschitz norm $\|\cdot\|_L$ (see (2.6)).

By the definition of the metric d (see [4]),

$$\Big|\int\limits_S \widehat{\varphi}\,[d\mu_k - d\mu_t]\Big| \le C_N d(\mu_k, \mu_t),$$

where $C_N := aN^{\gamma} + bN^{\sigma}$. Thus,

$$\begin{aligned} \int\limits_S \widehat{\varphi}\,[d\mu_k - d\mu_t] &= \int\limits_{S_N} \varphi\,[d\mu_k - d\mu_t] + \int\limits_{S\backslash S_N} \widehat{\varphi}\,[d\mu_k - d\mu_t] \\ &\le C_N d(\mu_k, \mu_t), \end{aligned} \tag{4.25}$$

or

$$\int\limits_{S_N} \varphi\,[d\mu_k - d\mu_t] \le C_N d(\mu_k, \mu_t) - \int\limits_{S\backslash S_N} \widehat{\varphi}\,[d\mu_k - d\mu_t].$$

Applying a similar inequality with opposite sign we obtain

$$\Big|\int\limits_{S_N} \varphi[d\mu_k - d\mu_t]\Big| \le C_N d(\mu_k, \mu_t) + \int\limits_{S\backslash S_N} |\widehat{\varphi}|\,[d\mu_k + d\mu_t]. \tag{4.26}$$

Since $\|\widehat{\varphi}\|_{\infty} \le aN^{\gamma}$, by (3.3) and the Chebyshev inequality the last integral in (4.26) is less than

$$2aK_{\lambda}N^{\gamma}N^{-\lambda}. \tag{4.27}$$

Applying (3.3) and the Chebyshev inequality (together with Assumption 2) to the second term on the right-hand side of (4.24), and gathering together (4.24)–(4.27), we obtain the desired inequality (3.6). ∎

Acknowledgements. The authors wish to thank the anonymous reviewers for suggestions on an earlier draft. The authors are grateful to Professor G. Escarela for language editing.

References

[1] L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.

[2] M. Duflo, *Random Iterative Models*, Springer, Berlin, 1997.

[3] J. Dupačová and R. Wets, *Asymptotic behavior of statistical estimators and of optimal solution of stochastic optimization problems*, Ann. Statist. 16 (1988), 1517–1549.

[4] R. M. Dudley, *Real Analysis and Probability*, Chapman & Hall, 1989.

[5] Yu. Ermoliev and V. I. Norkin, *Normalized convergence in stochastic optimization*, Ann. Oper. Res. 30 (1991), 187–198.

[6] Yu. Ermoliev and R. J.-B. Wets (eds.), *Numerical Techniques for Stochastic Optimization*, Springer, Berlin, 1988.

[7] E. Gordienko, *The speed of uniform convergence of the empirical risk*, Engrg. Cybernetics 1982, no. 2, 80–85.

[8] J. Hoffmann-Jørgensen and G. Pisier, *The law of large numbers and the central limit theorem in Banach spaces*, Ann. Probab. 4 (1976), 587–599.

[9] T. Homem-de-Mello, *On rates of convergence for stochastic optimization problems under non-independent and identically distributed sampling*, SIAM J. Optim. 19 (2008), 524–551.

[10] Yu. M. Kaniovski, A. J. King and R. J.-B. Wets, *Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems*, Ann. Oper. Res. 56 (1995), 189–208.

[11] V. Kaňková, *Empirical estimates in stochastic optimization via distributions tails*, Kybernetika 46 (2010), 459–471.

[12] V. Kaňková, *An approximative solution of stochastic optimization problem*, in: Trans. Eighth Prague Conference, Academia, Prague, 1978, 349–353.

[13] A. Kolmogorov and V. Tikhomirov, *ε-entropy and ε-capacity of sets in function spaces*, Amer. Math. Soc. Transl. 17 (1961), 277–364.

[14] G. Ch. Pflug, *Stochastic programs and statistical data*, Ann. Oper. Res. 85 (1999), 59–78.

[15] A. Shapiro, D. Dentcheva and A. Ruszczyński, *Lectures on Stochastic Programming. Modeling and Theory*, SIAM, 2009.

[16] A. Shapiro and T. Homem-de-Mello, *On the rate of convergence of optimal solutions of Monte Carlo approximation of stochastic programs*, SIAM J. Optim. 11 (2000), 70–86.

[17] A. T. Vakhitov, O. N. Granichin and L. S. Gurevich, *Algorithm for stochastic approximation with trial input perturbation in the non-stationary problem of optimization*, Automation and Remote Control 70 (2009), 1827–1835.

[18] V. N. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[19] V. N. Vapnik and A. Chervonenkis, *Theory of Pattern Recognition*, Nauka, Moscow, 1974 (in Russian).

E. Gordienko, J. Ruiz de Chávez (corresponding author)
Department of Mathematics
Universidad Autónoma
Metropolitana-Iztapalapa
San Rafael Atlixco 186
Col. Vicentina
C.P. 09340, Mexico City, México
E-mail: gord@xanum.uam.mx
jrch@xanum.uam.mx

E. Zaitseva
Instituto Tecnológico Autónomo
de México
Rio Hondo 1
Col. Progreso Tizapan
Del. Alvaro Obregón
C.P. 01080, Mexico City, México
E-mail: lenagordi@hotmail.com

Received on 17.6.2013;
revised version on 25.9.2013 (2186)