

# ENTROPY CONDITIONS FOR $L_r$ -CONVERGENCE OF EMPIRICAL PROCESSES

A. CAPONNETTO, E. DE VITO, AND M. PONTIL

ABSTRACT. The Law of Large Numbers (LLN) over classes of functions is a classical topic of Empirical Processes Theory. The properties characterizing classes of functions on which the LLN holds uniformly (*i.e.* Glivenko-Cantelli classes) have been widely studied in the literature. An elegant sufficient condition for such a property is finiteness of the Koltchinskii-Pollard entropy integral, and other conditions have been formulated in terms of suitable combinatorial complexities (*e.g.* the Vapnik-Chervonenkis dimension). In this paper, we endow the class of functions  $\mathcal{F}$  with a probability measure and consider the LLN relative to the associated  $L_r$  metric. This framework extends the case of uniform convergence over  $\mathcal{F}$ , which is recovered when  $r$  goes to infinity. The main result is a  $L_r$ -LLN in terms of a suitable uniform entropy integral which generalizes the Koltchinskii-Pollard entropy integral.

## 1. INTRODUCTION

Uniform Laws of Large Numbers (u-LLN) are widely studied results in Statistics. In the usual setting, we are given a finite set of points  $\mathbf{x} = (x_1, \dots, x_n) \in X^n$  sampled i.i.d. from a fixed but unknown probability measure  $P$  on  $X$ , and a class  $\mathcal{F}$  of real-valued functions on  $X$ . The aim of u-LLN is to establish conditions on the class  $\mathcal{F}$  which ensure the uniform convergence of the empirical average  $P_n f = \frac{1}{n} \sum_i f(x_i)$  to the mean  $Pf = \int_X f(x) dP(x)$ , that is <sup>1</sup>

$$(1) \quad \forall P \in \mathcal{P}(X), \forall \epsilon > 0 \quad \lim_{n \rightarrow \infty} \mathbb{P}_{\mathbf{x}} \left[ \sup_{f \in \mathcal{F}} |Pf - P_n f| \geq \epsilon \right] = 0,$$

where  $\mathcal{P}(X)$  is the set of all probability measures on  $X$ . Classes of function  $\mathcal{F}$  fulfilling condition (1) are called Glivenko-Cantelli classes.

Laws of Large Numbers (LLN) over classes of functions are classical results in Empirical Processes Theory. In particular, the characterization of Glivenko-Cantelli classes has been extensively studied in this literature. A number of techniques have been introduced to capture this concept, for example through the notions of VC-dimension [15, 16, 17, 14], scale-sensitive VC-dimension [1], Koltchinskii-Pollard entropy integral [8, 9, 5], etc.

In this paper, we endow the class of functions  $\mathcal{F}$  with a probability measure and consider the LLN relative to an  $L_r$  metric. This framework extends the case

---

*Date:* March 15, 2007.

*2000 Mathematics Subject Classification.* Primary 60G15, 60G51.

*Key words and phrases.* Empirical Processes, Uniform Entropy, Rademacher Averages, Glivenko-Cantelli Classes.

<sup>1</sup>In the paper we denote by  $\mathbb{P}[E]$  the probability of the event  $E$ , and by  $\mathbb{E}[F]$  the expectation of the random variable  $F$

of uniform convergence over  $\mathcal{F}$ , which is recovered when  $r$  goes to infinity. More precisely, we introduce the pseudo-norm

$$\|P\|_{\mu,r} = \left( \int_{\mathcal{F}} |Pf|^r d\mu(f) \right)^{\frac{1}{r}},$$

where  $\mu$  is a prescribed probability measure on  $\mathcal{F}$ , and consider the convergence of the stochastic process  $(P_n - P)$  relative to this norm. To illustrate our notation, let us consider a simple example where  $X = \mathbb{R}$  and  $\mathcal{F}$  is the space of characteristic functions of half-lines, that is  $\mathcal{F} = \{f_t : t \in \mathbb{R}\}$ , where  $f_t(x) = 1$  if  $x \leq t$  and zero otherwise. In this case, the function  $t \mapsto Pf_t$  is the cumulative distribution function associated to  $P$  and  $\|P - P_n\|_{\mu,r}$  is the  $L_r$  distance between the true cumulative distribution function and the empirical distribution function, respectively.

The main result of the paper is a  $L_r$ -LLN involving a finiteness condition for a suitable generalization of the Koltchinskii-Pollard entropy integral.

We note that  $u$ -LLN play also an important role in the foundations of Learning Theory. In particular, the notion of Glivenko-Cantelli class introduced in the former context is equivalent to the *learnability* notion of a class of functions  $\mathcal{F}$ , see, for example [1] and references therein. Hence, our results can also be seen as a relaxation of the learnability results in Learning Theory.

The paper is organized as follows. In Section 2, we introduce our framework and in particular we give the definition of *touchstone class* and induced  $L_r$  metric. In Section 3, we collect some known results about the convergence of empirical measures  $P_n$  to the unknown measure  $P$  relative to the uniform semi-norm  $\|\cdot\|_{\mathcal{F}}$ . In particular, we define the Koltchinskii-Pollard entropy integral  $I(\mathcal{F})$  of the class  $\mathcal{F}$ , which is used in Theorem 1 to bound the uniform deviation of the process  $P_n - P$ . In Section 4, we study the  $L_r$ -LLN in terms of a suitable uniform entropy integral which generalizes the Koltchinskii-Pollard entropy integral. This section contains the main results of the paper. In Subsection 4.1, we define the uniform entropy integral relative to the  $L_r$  metric, and show its relation to the Koltchinskii-Pollard entropy integral (Theorem 2). In Subsection 4.2, we generalize the results of Section 3 to the  $L_r$  setting (Theorem 3). Proofs of the results given in Section 4 are postponed to Appendices A and B.

## 2. TOUCHSTONE CLASSES AND $L_r$ SEMI-NORMS

Let  $X$  be a locally compact separable metric space, for example any closed subset of  $\mathbb{R}^k$ . The space of (signed) bounded measures  $\mathcal{M}(X)$  over  $X$  is defined as the dual of the Banach space  $C_0(X)$  of continuous functions on  $X$  which vanish at infinity<sup>2</sup> (see, for example, [2, Appendix C.18]).

The set of probability measures on  $X$  is denoted by  $\mathcal{P}(X)$ . It is well-known that the Banach space structure of  $\mathcal{M}(X)$  induces the following metric over  $\mathcal{P}(X)$ ,

$$(2) \quad d(P, P') = \sup_{f \in \mathcal{F}} |Pf - P'f|,$$

where we use the notation  $Pf = \int_X f(x)dP(x)$  and  $\mathcal{F}$  is the unit ball in  $C_0(X)$ , that is,

$$(3) \quad \mathcal{F} = \{f \in C_0(X) \mid \sup_{x \in X} |f(x)| \leq 1\}.$$

---

<sup>2</sup>This means that, for every  $\epsilon > 0$  there is a compact subset  $K_\epsilon$  of  $X$  with  $|f(x)| \leq \epsilon$  for every  $x \in X \setminus K_\epsilon$ .

According to the definition (2)–(3) two probability measures are  $\epsilon$ -close to each other whenever for every  $f \in \mathcal{F}$  they have  $\epsilon$ -close pairings with  $f$ . In a sense, approximating a probability measure  $P$ , relative to the metric  $d$  is equivalent to simultaneously approximating as many linear functionals as the functions in  $\mathcal{F}$ . However, in various situations this notion of distance may often be excessively strong. In fact, in many applications (*e.g.* density estimation) it is interesting to estimate just a very limited class of linear functionals of the unknown probability measure. It is therefore natural to look for weaker distances than (2)–(3).

A natural way to weaken the distance (2)–(3) is to suitably restrict the class of functions  $\mathcal{F}$ . Inspired by [12] we name *touchstone class* a class of functions  $\mathcal{F}$  inducing a metric over  $\mathcal{P}(X)$  through equation (2). A classical example of metric of type (2) is the Kolmogorov-Smirnov distance, which is obtained when  $X = \mathbb{R}$  and  $\mathcal{F}$  is the class of step functions on  $\mathbb{R}$  (see Example 1 below).

However, in many applications the metric (2) may be still too strong. In fact, we would like two probability measures to be  $\epsilon$ -close even if they do not have  $\epsilon$ -close evaluations over a tiny fraction of the functionals induced by  $\mathcal{F}$ . The formalization of this idea can be accomplished by suitably endowing  $\mathcal{F}$  with a probability measure  $\mu$ , and considering, for some  $r \geq 1$ , the pseudo-distance

$$(4) \quad d_r(P, P') = \left( \int_{\mathcal{F}} |Pf - P'f|^r d\mu(f) \right)^{\frac{1}{r}}.$$

Since the measure  $\mu$  is finite and equation (4) has the form of the distance between the functionals  $f \mapsto Pf$  and  $f \mapsto P'f$  in the Banach space  $L_r(\mathcal{F}, \mu)$ , from Hölder's inequality  $d_r(P, P')$  is non-decreasing in  $r$ . Moreover, as  $r \rightarrow \infty$ ,  $d_r(P, P')$  converges to the right hand side of equation (2) with the supremum replaced by an essential supremum. At least for countable classes  $\mathcal{F}$  (as in Example 2 below) this expression is equal to the right hand side of equation (2) itself, whenever the condition  $\text{supp } \mu = \mathcal{F}$  is fulfilled. However, establishing a rigorous link between equations (2) and (4) for more general classes  $\mathcal{F}$ , requires some additional technical assumptions and is the goal of Proposition 1 later in this section. The definition below formalizes the notion of touchstone class. In order to avoid technical problems endowing a touchstone class  $\mathcal{F}$  with a probability measure, we regard  $\mathcal{F}$  as a locally compact separable metric space with respect to a given metric. In most applications (see for example Examples 1 and 2 later in section) the metric space structure over  $\mathcal{F}$  is naturally induced by a suitable space of parameters  $\mathcal{T}$  through a parametrization function  $t \mapsto f_t$ .

**Definition 1.** *A touchstone class over  $X$  is a family  $\mathcal{F}$  of functions from  $X$  to  $[-1, 1]$  equipped with a structure of locally compact separable metric space.  $\mathcal{F}$  is endowed with a probability measure  $\mu$ , satisfying the properties*

- (a) *the map  $(f, x) \mapsto f(x)$  is measurable from  $\mathcal{F} \times X$  into  $[-1, 1]$ ;*
- (b) *for every  $f \in \mathcal{F}$  there exists a measurable subset  $A_f \subset \mathcal{F}$  with<sup>3</sup>*

$$\mu(A_f \cap B(f, \delta)) > 0 \quad \forall \delta > 0$$

*and, for all  $x \in X$  and  $\epsilon > 0$ , there is  $\delta > 0$  such that*

$$|f'(x) - f(x)| \leq \epsilon \quad \forall f' \in A_f \cap B(f, \delta).$$

<sup>3</sup>Here  $B(f, \delta)$  is the open ball in  $\mathcal{F}$ , with center  $f$  and radius  $\delta$

Here and in the following, *measurability* is always relative to the  $\sigma$ -algebra induced by metric introduced in Definition 1. Therefore, when we say that a subset of  $\mathcal{F}$  (or a function over  $\mathcal{F}$ ) is measurable, we mean that it is measurable with respect to (w.r.t.) this  $\sigma$ -field.

Let us now briefly discuss the points in Definition 1. Assumption (a) ensures that every function in  $\mathcal{F}$  is measurable and bounded on  $X$ . Hence, for every probability measure  $P \in \mathcal{P}(X)$ , we have that  $\mathcal{F} \subset L_2(X, P)$ . Furthermore, as a consequence of Fubini's Theorem, for every  $M \in \mathcal{M}(X)$ , the function  $f \mapsto Mf = \int_X f(x)dM(x)$ , is integrable with respect to the measure  $\mu$ .

It is not difficult to verify that Assumption (b) implies that the support of  $\mu$  is  $\mathcal{F}$ , which was exactly the assumption we made in the previous informal discussion in the case of a countable class  $\mathcal{F}$ . Moreover, notice the two following important cases for which Assumption (b) can be easily fulfilled. In the first case, the map  $f \mapsto f(x)$  is continuous for all  $x \in X$ , then Assumption (b) holds with  $A_f = \mathcal{F}$  for every  $f \in \mathcal{F}$ . In the second case,  $\mathcal{F}$  is discrete, then for every  $f \in \mathcal{F}$ , it holds  $\mu(\{f\}) > 0$ , and Assumption (b) is satisfied when  $A_f = \{f\}$ . However, Definition 1 embraces important examples where both  $\mathcal{F}$  is not discrete and the mappings  $f \mapsto f(x)$  are not continuous (see Examples 1 and 2 at the end of this section).

**Definition 2.** Let  $(\mathcal{F}, \mu)$  be a touchstone class and  $M \in \mathcal{M}(X)$ . We define the semi-norms

$$\begin{aligned} \|M\|_{\mu, r} &= \left( \int_{\mathcal{F}} |Mf|^r d\mu(f) \right)^{\frac{1}{r}}, & r \in [1, \infty) \\ \|M\|_{\mu, \infty} &= \operatorname{ess\,sup}_{f \in \mathcal{F}} |Mf| \\ \|M\|_{\mathcal{F}} &= \sup_{f \in \mathcal{F}} |Mf|. \end{aligned}$$

The next proposition clarifies some properties of the above semi-norms and the role of Assumption (b) in Definition 1.

**Proposition 1.** With the above notation, we have for every  $M \in \mathcal{M}(X)$ , that

- (1) the map  $r \mapsto \|M\|_{\mu, r}$  is continuous on  $[1, \infty]$ , increasing and bounded from above by  $\|M\|_{\mathcal{F}}$ ;
- (2)  $\|M\|_{\mu, \infty} = \|M\|_{\mathcal{F}}$ .

*Proof.* Part (1) follows from the finiteness of  $\mu$  and well known properties of  $L_r$  norms (see, for example, [11, Theorem 5.8.35]).

We prove part (2) by contradiction. Assume that there is  $M \in \mathcal{M}(X)$  and  $f \in \mathcal{F}$  such that  $|Mf| > \|M\|_{\mu, \infty}$  and, without loss of generality,  $Mf > 0$ . Let  $A_f \subset \mathcal{F}$  as in Assumption (b) in Definition 1, and  $\epsilon = (Mf - \|M\|_{\mu, \infty})/2$ , we claim that there is  $\delta > 0$  such that

$$(5) \quad |Mf' - Mf| \leq \epsilon \quad \forall f' \in A_f \cap B(f, \delta),$$

and, hence,

$$Mf' \geq \frac{Mf + \|M\|_{\mu, \infty}}{2} > \|M\|_{\mu, \infty} \quad \forall f' \in A_f \cap B(f, \delta).$$

By assumption  $\mu(A_f \cap B(f, \delta)) > 0$ , so

$$\operatorname{ess\,sup} \{|Mf'| : f' \in A_f \cap B(f, \delta)\} > \|M\|_{\mu, \infty},$$

which is a contradiction.

Finally let us prove claim (5) by contradiction, assuming that for every  $i \in \mathbb{N}$  there is  $f'_i \in A_f \cap B(f, \frac{1}{i})$  such that  $|Mf'_i - Mf| > \epsilon$ . However, by assumption, the sequence  $(f'_i(x))_{i \in \mathbb{N}}$  converges to  $f(x)$  for all  $x \in X$ . Since  $f'_i$  and  $f$  are bounded functions, the Lebesgue dominated convergent theorem implies that

$$\lim_{i \rightarrow \infty} Mf'_i = Mf,$$

which is a contradiction.  $\square$

Notice that in general  $\|\cdot\|_{\mu,r}$  is only a semi-norm on  $\mathcal{M}(X)$ . Indeed, from Proposition 1 it follows that if there exists a  $M \neq 0$  with  $\|M\|_{\mathcal{F}} = 0$ , then for every  $r \in [1, \infty]$ ,  $\|M\|_{\mu,r} = 0$ , and therefore  $\|\cdot\|_{\mu,r}$  is not a norm. Now, by definition,  $\|M\|_{\mathcal{F}} = 0$  implies that  $Mf = 0$  for all  $f \in \mathcal{F}$ , nevertheless if  $\mathcal{F}$  does not separate the points of  $\mathcal{M}(X)$ , it can happen that  $M \neq 0$ .

Since the pseudo-metric introduced in equations (2) and (4) can be expressed in the form

$$\begin{aligned} d(P, P') &= \|P - P'\|_{\mathcal{F}}, \\ d_r(P, P') &= \|P - P'\|_{\mu,r}, \end{aligned}$$

by Proposition 1 we conclude that  $d_r(P, P')$  is increasing as a function of  $r$ , and

$$(6) \quad \lim_{r \rightarrow \infty} d_r(P, P') = d_{\infty}(P, P') = d(P, P').$$

We now present two simple examples of the described construction. In the following sections they will be used to illustrate the forthcoming developments.

**Example 1.** *Characteristic functions of orthants.* We let  $X = \mathbb{R}^k$  and

$$\mathcal{F} = \{f_t : t \in \mathbb{R}^k\},$$

where  $f_t(x) = \mathbf{1}\{x_i \leq t_i, \forall i \in \{1, \dots, k\}\}$ , with  $\mathbf{1}\{a\}$  the indicator function of the predicate  $a$ , and  $x_i$  the  $i$ -th component of the vector  $x \in \mathbb{R}^k$ .

Here  $\mathcal{T} = \mathbb{R}^k$  plays the role of parameter space for  $\mathcal{F}$ , therefore we endow  $\mathcal{F}$  with the metric induced by the Euclidean structure of  $\mathbb{R}^k$ .

We let  $\mu$  be an arbitrary probability measure on the metric space  $\mathcal{F}$ , satisfying the condition  $\text{supp } \mu = \mathcal{F}$ . In this example the evaluation functionals  $f \mapsto f(x)$  are not continuous in  $t = x$ , nevertheless Assumption (b) in Definition 1 may be fulfilled thanks to the upper semi-continuity of the functions in  $\mathcal{F}$ . In fact, it is easy to verify that a suitable choice for the sets  $A_f$  is

$$A_{f_t} = \{f_{t'} : t'_i \geq t_i, \forall i \in \{1, \dots, k\}\} \quad \forall t \in \mathbb{R}^k.$$

**Example 2.** *Binary digits.* We use the binary expansion of real numbers in  $(0, 1)$ . For every  $x \in (0, 1)$  we define the sequence  $(b_i(x))_{i \in \mathbb{N}}$  of numbers in  $\{0, 1\}$ , fulfilling the equation<sup>4</sup>  $x = \sum_i b_i(x)2^{-i}$ .

We let  $X = (0, 1)$  and,

$$\mathcal{F} = \{b_t : t \in \mathbb{N}\}.$$

In this case the parameter space is  $\mathcal{T} = \mathbb{N}$ , and  $\mathcal{F}$  inherits its metric from it. Since  $\mathcal{F}$  is discrete, recalling the discussion following Definition 1, we conclude that

<sup>4</sup>For rational  $x$ , the expansion is not unique. In this case ties are broken by choosing the unique finite expansion.

for arbitrary  $\mu$  fulfilling  $\mu(\{f\}) > 0$  for every  $f \in \mathcal{F}$ , the choice  $A_f = \{f\}$  verifies the assumptions in Definition 1.

### 3. UNIFORM ENTROPY CONDITION AND GLIVENKO-CANTELLI PROPERTY

In this section, for a prescribed touchstone class  $\mathcal{F}$  and samples  $\mathbf{x} = (x_1, \dots, x_n)$  drawn i.i.d. from a probability measure  $P \in \mathcal{P}(X)$ , we study the convergence of the empirical measure  $P_n = \frac{1}{n} \sum_i \delta_{x_i}$  to  $P$  in the pseudo-metric  $d$  defined in equation (2). By definition (recall equation (1) in Section 1) establishing this convergence result for arbitrary  $P$  is equivalent to prove that  $\mathcal{F}$  is a Glivenko-Cantelli class. In fact, the main result of this section, Theorem 1, gives an explicit non-asymptotic upper bound on  $d(P, P_n)$  in terms of a suitable invariant of  $\mathcal{F}$ : the Koltchinskii-Pollard entropy integral  $I(\mathcal{F})$ . All the definitions and results of this section are well-known in the literature (see for instance [5, 13, 6]), and are collected here as a preliminary step toward the generalization presented in Section 4.

Let us begin by introducing the notion of Rademacher averages, which play a central role in our subsequent analysis.

**Definition 3.** *The empirical Rademacher averages of a touchstone class  $\mathcal{F}$ , relative to the samples  $\mathbf{x} = (x_1, \dots, x_n)$  are defined by<sup>5</sup>*

$$R_n(\mathcal{F}) = \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right|$$

where  $\sigma = (\sigma_1, \dots, \sigma_n)$  is a  $n$ -tuple of Rademacher variables<sup>6</sup>.

The following proposition states a fundamental bound for  $d(P, P_n)$ , the Symmetrization Lemma, in terms of Rademacher averages.

**Proposition 2.** *Let  $P$  be in  $\mathcal{P}(X)$  and  $\mathbf{x} = (x_1, \dots, x_n)$  be i.i.d. samples drawn from  $P$ . For every  $\delta \in (0, 1)$ , with probability at least  $1 - \delta$ , it holds*

$$d(P, P_n) \leq 2\mathbb{E}_{\mathbf{x}} R_n(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{n}}.$$

*Proof.* We appeal to [13, Lemma 2.3.1] to assert that  $\mathbb{E}_{\mathbf{x}} d(P, P_n) \leq 2\mathbb{E}_{\mathbf{x}} R_n(\mathcal{F})$ . The result follows by McDiarmid's inequality (see, for example, [4, Theorem 9.2]) recalling that the functions in  $\mathcal{F}$  take values in  $[-1, 1]$ .  $\square$

To proceed further in our analysis and define the Koltchinskii-Pollard entropy of  $\mathcal{F}$ , we need the notion of covering number.

<sup>5</sup>Often, in the literature the absolute value in the definition of the empirical Rademacher averages is removed, that is, one consider the quantity

$$\bar{R}_n(\mathcal{F}) = \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i).$$

This definition is equivalent to Definition 3. Specifically, one can show that  $\frac{1}{2}R_n(\mathcal{F}) \leq \bar{R}_n(\mathcal{F}) \leq R_n(\mathcal{F})$ .

<sup>6</sup>The Rademacher variables  $(\sigma_1, \dots, \sigma_n)$  are  $\{-1, 1\}$ -valued and independent, with  $\mathbb{P}[\sigma_i = \frac{1}{2}] = \mathbb{P}[\sigma_i = -\frac{1}{2}] = \frac{1}{2}$

**Definition 4.** For every  $P \in \mathcal{P}(X)$  and  $\epsilon > 0$  we define  $\mathcal{C}(\epsilon, \mathcal{F}, P)$  as the set of all covers of  $\mathcal{F}$  by sets of the form

$$c_{\bar{f}} = \{f \in \mathcal{F} : \|f - \bar{f}\|_{L_2(X, P)} < \epsilon\} \quad \bar{f} \in \mathcal{F},$$

and the covering number of  $\mathcal{F}$  as <sup>7</sup>

$$N(\epsilon, \mathcal{F}, L_2(X, P)) = \inf\{|C| : C \in \mathcal{C}(\epsilon, \mathcal{F}, P)\}.$$

We refer to [13, Definition 2.2.3] for information on covering numbers.

The notion of uniform entropy defined below is central in Empirical Processes Theory (see for example [13, Chapter 2.5]).

**Definition 5.** For every  $\epsilon > 0$  we define the uniform entropy of a touchstone class  $\mathcal{F}$  as

$$H(\epsilon, \mathcal{F}) = \sup_n \sup_{P_n} \log N(\epsilon, \mathcal{F}, L_2(X, P_n)),$$

where the supremum is over measures of the form  $P_n = \frac{1}{n} \sum_i \delta_{x_i}$ .

The following theorem gives an upper bound on  $d(P, P_n)$  in terms of the Koltchinskii-Pollard entropy integral  $I(\mathcal{F})$ .

**Theorem 1.** Let  $P$  be in  $\mathcal{P}(X)$  and  $\mathbf{x} = (x_1, \dots, x_n)$  be i.i.d. samples drawn from  $P$ . For every  $\delta \in (0, \frac{1}{2})$ , with probability at least  $1 - \delta$ , it holds

$$(7) \quad d(P, P_n) \leq \frac{C}{\sqrt{n}} \left( I(\mathcal{F}) + \sqrt{\log \frac{1}{\delta}} \right),$$

where  $C$  is a universal constant and  $I(\mathcal{F})$  is the Koltchinskii-Pollard entropy integral of  $\mathcal{F}$  defined as

$$I(\mathcal{F}) = \int_0^\infty \sqrt{H(\epsilon, \mathcal{F})} d\epsilon.$$

*Proof.* We first note that the Koltchinskii-Pollard entropy integral is well defined since  $H(\epsilon, \mathcal{F})$  is monotone with respect to  $\epsilon$ . The inequality follows from Proposition 2 and [13, Corollary 2.2.8].  $\square$

From Theorem 1 it follows that finiteness of the Koltchinskii-Pollard entropy integral (the *uniform entropy condition*) is a sufficient condition for the Glivenko-Cantelli property of  $\mathcal{F}$ . That is, we have the following corollary.

**Corollary 1.** If  $I(\mathcal{F}) < \infty$  then  $\mathcal{F}$  is a Glivenko-Cantelli class.

Notice that in general the converse result does not hold, that is, it is not true that the Glivenko-Cantelli property implies finiteness of  $I(\mathcal{F})$ . However the equivalence holds for classes  $\mathcal{F}$  of binary-valued functions (see [6]).

Finally let us consider our examples.

**Example 1 (cont.)** We estimate the covering number of the binary-valued class of function  $\mathcal{F}$  by the standard VC-bound (see [13, Theorem 2.6.4] and [13, Example 2.6.1])

$$N(\epsilon, \mathcal{F}, L_2(X, P)) \leq \left( \frac{K}{\epsilon} \right)^{2k},$$

which holds for some constant  $K$  and every  $\epsilon \in (0, 1)$  and  $P \in \mathcal{P}(X)$ .

<sup>7</sup>We denote by  $|C|$  the cardinality of the set  $C$ .

By direct integration and noting that the covering number is exactly equal to 1 for  $\epsilon \geq 1$ , we get  $I(\mathcal{F}) \leq C'\sqrt{k}$ , for a suitable constant  $C'$ . Hence, by Corollary 1  $\mathcal{F}$  is Glivenko-Cantelli.

The pseudo-metric  $d$  is named Kolmogorov-Smirnov distance, and has been widely studied in statistics literature (*e.g.* [7, 4, 10]).

**Example 2 (cont.)** In this case  $I(\mathcal{F})$  is infinite. This fact can be proved first showing, by reasoning as in [14, Example 4.11.4], that the VC-dimension of  $\mathcal{F}$  is infinite. Hence since finiteness of VC-dimension is a necessary condition for the Glivenko-Cantelli property (over binary-valued classes), by Corollary 1 we conclude that  $I(\mathcal{F}) = \infty$ .

#### 4. $L_r$ CONVERGENCE RESULTS

In this section we present the main result of the paper, Theorem 3, which generalizes to the  $L_r$  metric the uniform convergence result given in Theorem 1.

The central concept in this analysis is a suitable generalization  $I_r(\mathcal{F}, \mu)$  of the Koltchinskii-Pollard uniform entropy integral  $I(\mathcal{F})$  defined in the previous section. This quantity and its properties are described in Subsection 4.1, while the generalization of the results from Section 3 is given in Subsection 4.2. For sake of clarity we postpone all the proofs to Appendices A and B.

**4.1. Uniform entropies.** Let us begin with some preliminary definitions.

**Definition 6.** Let  $p : I \rightarrow [0, 1]$  be a probability distribution over a denumerable set <sup>8</sup>  $I$ . For every  $r \in [1, \infty]$ , we define the quantity

$$(8) \quad h_r(p) = \inf \left\{ \left\| \sqrt{-\log q} \right\|_{L_r(I,p)}^2 : q(i) \geq 0, \sum_i q(i) = 1 \right\}.$$

Recall, for  $r \in [1, \infty)$ , adopting the convention  $(\log \frac{1}{0})^{\frac{r}{2}} 0 = 0$ , that the  $L_r$  norm appearing in the equation (8) is given by

$$\left\| \sqrt{-\log q} \right\|_{L_r(I,p)}^2 = \left( \sum_i \left( \log \frac{1}{q(i)} \right)^{\frac{r}{2}} p(i) \right)^{\frac{2}{r}},$$

and for  $r = \infty$  we have

$$\left\| \sqrt{-\log q} \right\|_{L_\infty(I,p)}^2 = \sup \left\{ \log \frac{1}{q(i)} : p(i) \neq 0 \right\}.$$

The function  $h_r$  has some nice properties collected in the following proposition.

**Proposition 3.** The function  $h_r$  fulfills the following properties.

- (a) For every  $r, r' \in [1, \infty]$ ,  $r \leq r'$  it holds  $h_r(p) \leq h_{r'}(p)$ ;
- (b)  $h_\infty(p) = \log |\{i : p(i) \neq 0\}|$ ;
- (c)  $h_2(p) = -\sum_i p(i) \log p(i)$ , the Shannon entropy of  $p$ ;
- (d) For every  $r \in [1, \infty]$ , denumerable index sets  $I$  and  $J$ , and probability distribution  $p$  over  $I \times J$

$$h_r(p) \leq 2(h_r(p_1) + h_r(p_2)),$$

$$\text{where } p_1(i) = \sum_j p(i, j) \text{ and } p_2(j) = \sum_i p(i, j).$$

<sup>8</sup>A set is denumerable if and only if it is finite or countably infinite.

The second step of our construction is to define the quantity  $H_r(\epsilon, \mathcal{F}, \mu)$  which generalizes the uniform entropy  $H_r(\epsilon, \mathcal{F})$ . To this end, we first define suitable classes of partitions of  $\mathcal{F}$ , which play a role analogous to that of the covers  $\mathcal{C}(\epsilon, \mathcal{F}, P)$ .

**Definition 7.** *Let  $(\mathcal{F}, \mu)$  be a touchstone class and  $P$  belong to  $\mathcal{P}(X)$ . For every  $\epsilon > 0$  we define  $\mathcal{A}(\epsilon, \mathcal{F}, \mu, P)$  as the set of denumerable partitions of  $\mathcal{F}$  into measurable parts, having strictly positive measure and  $L_2(X, P)$ -diameter at most  $\epsilon$ .*

Recall, by Assumption (a) in Definition 1, that every function in  $\mathcal{F}$  is measurable and bounded on  $X$ . Hence,  $\mathcal{F} \subset L_2(X, P)$  and the quantity  $\mathcal{A}(\epsilon, \mathcal{F}, \mu, P)$  is well-defined.

Observe also that since a partition  $A \in \mathcal{A}(\epsilon, \mathcal{F}, \mu, P)$  is a family of measurable sets, the restriction of  $\mu$  over  $A$ ,  $\mu|_A$  is well-defined. Moreover, by Definition 7,  $\mu|_A$  is a probability distribution<sup>9</sup> on  $A$ .

We are now ready to define  $H_r(\epsilon, \mathcal{F}, \mu)$  and  $I_r(\mathcal{F}, \mu)$ .

**Definition 8.** *For every  $\epsilon > 0$ ,  $r \in [1, \infty]$ , we define the uniform entropy of a touchstone class  $(\mathcal{F}, \mu)$  as*

$$H_r(\epsilon, \mathcal{F}, \mu) = \sup_n \sup_{P_n} \inf_A h_r(\mu|_A),$$

where the supremum is over measures of the form  $P_n = \frac{1}{n} \sum_i \delta_{x_i}$ , and the infimum is over  $\mathcal{A}(\epsilon, \mathcal{F}, \mu, P_n)$ .

The corresponding uniform entropy integral is

$$(9) \quad I_r(\mathcal{F}, \mu) = \int_0^\infty \sqrt{H_r(\epsilon, \mathcal{F}, \mu)} d\epsilon.$$

The following theorem collect the relevant properties of the quantities introduced in previous definition.

**Theorem 2.** *The following properties of the uniform entropy hold.*

- (a)  $H_r(\epsilon, \mathcal{F}, \mu)$  is non-increasing with respect to  $\epsilon$ ;
- (b)  $H_r(\epsilon, \mathcal{F}, \mu)$  is non-decreasing with respect to  $r$ ;
- (c)  $H(2\epsilon, \mathcal{F}) \leq H_\infty(2\epsilon, \mathcal{F}, \mu) \leq H(\epsilon, \mathcal{F})$ .

Moreover  $I_r(\mathcal{F}, \mu)$  is non-decreasing in  $r$ , and

$$I(\mathcal{F}) \leq I_\infty(\mathcal{F}, \mu) \leq 2I(\mathcal{F}).$$

Finally we illustrate the results of this subsection through our two examples.

**Example 1 (cont.)** From Theorem 2 and the already known result  $I(\mathcal{F}) \leq C'\sqrt{k}$ , we conclude that for every  $\mu$  fulfilling the assumptions, and  $r \in [1, \infty]$ , it holds  $I_r(\mathcal{F}, \mu) \leq 2C'\sqrt{k}$ .

**Example 2 (cont.)** From Definition 6 it follows (by the monotonicity property of  $(-\log q)^{\frac{r}{2}}$  w.r.t.  $q$ ) for arbitrary  $P$  and  $\mu$ , that

$$\hat{A} = \operatorname{argmax}_{A \in \mathcal{A}(\epsilon, \mathcal{F}, \mu, P)} h_r(\mu|_A) = \{\{b_i\} : t \in \mathbb{N}\}.$$

<sup>9</sup>Recall that the probability measure  $\mu$  is, by definition, a function over the  $\sigma$ -field  $\Sigma$  of  $\mathcal{F}$ , fulfilling  $\mu(\mathcal{F}) = 1$  and, for all  $a$  and  $b$  in  $\Sigma$  with  $a \cap b = \emptyset$ , the equality  $\mu(a \cup b) = \mu(a) + \mu(b)$  holds. Therefore if the denumerable partition  $A$  in the text is  $\{a_1, a_2, \dots\}$ , we get  $\sum_i \mu|_A(a_i) = 1$ .

Therefore, by Definition 8, for  $\epsilon \in (0, 1)$  we get

$$H_r(\epsilon, \mathcal{F}, \mu) \leq h_r(\mu|_{\hat{A}}),$$

and for  $\epsilon \geq 1$ ,  $H_r(\epsilon, \mathcal{F}, \mu) = 0$ . From the estimate above we see that the uniform entropy integral  $I_r(\mathcal{F}, \mu)$  is upper bounded by  $h_r(\mu|_{\hat{A}})$ .

Computing the function  $h_r$  for an arbitrary probability distribution over  $\mathbb{N}$  is not an easy task. However, assuming that  $\mu(\{b_t\}) = O(t^{-\eta})$  for some  $\eta > 1$ , it is straightforward to show that  $h_r(\mu|_{\hat{A}})$  is finite for every  $r \in [1, \infty)$ .

**4.2. Upper bounds on  $d_r(P, P_n)$ .** In this subsection, we extend the results of Section 3, from the analysis of  $d(P, P_n)$  to that of  $d_r(P, P_n)$  for arbitrary  $r \in [1, \infty)$ .

We already observed (see equation (6)) that the pseudo-metric  $d$  can be seen as the limit of the pseudo-metric  $d_r$  as  $r$  goes to  $\infty$ . The next definition introduces the quantity  $R_{r,n}(\mathcal{F}, \mu)$  which, as  $d_r$  does with  $d$ , generalizes the Rademacher averages  $R_n(\mathcal{F})$  introduced in Definition 3.

**Definition 9.** For every  $r \in [1, \infty]$ , the empirical Rademacher averages  $R_{r,n}(\mathcal{F}, \mu)$  of the touchstone class  $(\mathcal{F}, \mu)$ , relative to the samples  $\mathbf{x} = (x_1, \dots, x_n)$  are defined by

$$R_{r,n}(\mathcal{F}, \mu) = \mathbb{E}_\sigma \left\| \frac{1}{n} \sum_i \sigma_i \delta_{x_i} \right\|_{\mu, r}$$

where  $\sigma = (\sigma_1, \dots, \sigma_n)$  is an  $n$ -tuple of Rademacher variables, and  $\delta_x$  is the Dirac delta measure at  $x$ .

The relation between  $R_{r,n}(\mathcal{F}, \mu)$  and  $R_n(\mathcal{F})$  is clarified by observing that

$$R_n(\mathcal{F}) = \mathbb{E}_\sigma \left\| \frac{1}{n} \sum_i \sigma_i \delta_{x_i} \right\|_{\mathcal{F}}.$$

Therefore, from Proposition 1 we conclude that  $R_{r,n}(\mathcal{F}, \mu)$  is increasing as a function of  $r$ , and

$$(10) \quad \lim_{r \rightarrow \infty} R_{r,n}(\mathcal{F}, \mu) = R_{\infty,n}(\mathcal{F}, \mu) = R_n(\mathcal{F}).$$

We also note that the Symmetrization Lemma stated in Proposition 2 may be naturally extended to the  $L_r$  setting.

**Proposition 4.** Let  $P$  be in  $\mathcal{P}(X)$  and  $\mathbf{x} = (x_1, \dots, x_n)$  be i.i.d. samples drawn from  $P$ . For every  $\delta \in (0, 1)$  and  $r \in [1, \infty]$ , with probability at least  $1 - \delta$ , it holds

$$d_r(P, P_n) \leq 2\mathbb{E}_{\mathbf{x}} R_{r,n}(\mathcal{F}, \mu) + \sqrt{\frac{\log \frac{1}{\delta}}{n}}.$$

More interestingly, the chaining technique used to derive Theorem 1 can still be applied in the  $L_r$  setting. This is possible by exploiting the properties of the uniform entropies  $H_r(\epsilon, \mathcal{F}, \mu)$  which have been shown in the previous subsection.

**Theorem 3.** Let  $P$  be in  $\mathcal{P}(X)$  and  $\mathbf{x} = (x_1, \dots, x_n)$  be i.i.d. samples drawn from  $P$ . For every  $\delta \in (0, \frac{1}{2})$ , with probability at least  $1 - \delta$ , it holds, for  $r \in [1, \infty]$ , the inequality

$$(11) \quad d_r(P, P_n) \leq \frac{C}{\sqrt{n}} \left( I_r(\mathcal{F}, \mu) + \sqrt{\log \frac{1}{\delta}} \right),$$

where  $C$  is a universal constant and the uniform entropy integral  $I_r(\mathcal{F}, \mu)$  is defined in equation (9).

Theorem 3 generalizes Theorem 1 since by equation (6) and Theorem 2, for  $r = \infty$  equation (11) becomes equation (7).

The advantage of the new result is that for some touchstone classes, the uniform entropy integral  $I_r(\mathcal{F}, \mu)$  may be finite for arbitrarily large  $r$  while  $I(\mathcal{F})$  is infinite. Under these circumstances Theorem 3 gives quantitative probabilistic bounds for the defects  $|Pf - P_n f|$  while the standard uniform analysis in Theorem 1 is ineffective. This is the case for Example 2 when a suitably fast decaying probability measure  $\mu$  is chosen.

We conclude this section with an important remark about the presented result. We want to stress that the content of Theorem 3 resides in the non-asymptotic character of equation (11) and in the explicit evaluation of  $I_r(\mathcal{F}, \mu)$ . In fact, an asymptotic result analogous to Corollary 1 for the  $L_r$  setting can be directly obtained exploiting the uniform boundedness of  $\mathcal{F}$ .

**Proposition 5.** *Let  $P$  be in  $\mathcal{P}(X)$  and  $\mathbf{x} = (x_1, \dots, x_n)$  be i.i.d. samples drawn from  $P$ . For every  $\delta \in (0, 1)$  and  $r \in [1, \infty]$ , with probability at least  $1 - \delta$ , it holds*

$$(12) \quad d_r(P, P_n) \leq \frac{C}{\sqrt{n}} \left( \sqrt{r} + \sqrt{\log \frac{1}{\delta}} \right),$$

for some universal constant  $C$ .

The important point here is that the estimate (12) does not account for any specific structure of  $\mathcal{F}$ . For instance, when  $r \gg n$  equation (12) gives no information on  $d_r(P, P_n)$ , while equation (11) may give a tight bound, for specific classes of functions with small uniform entropy integral  $I_r(\mathcal{F}, \mu)$ .

*Acknowledgments.* This paper was supported by the FIRB project RBIN04PARL and the EU Integrated Project Health-e-Child IST-2004-027749. The first author was also partially supported by the NSF grant 0325113, and the third author by EPSRC Grant GR/T18707/01 and by the IST Programme of the European Community, under the PASCAL Network of Excellence IST-2002-506778.

#### APPENDIX A. PROOFS OF RESULTS FROM SUBSECTION 4.1

*Proof of Proposition 3.* Property (a) follows by noting that the argument of the infimum in equation (8),  $\|\sqrt{-\log \bar{q}}\|_{L_r(I, p)}^2$ , is non-decreasing in  $r$  by Hölder's inequality.

To prove (b) we let  $N = |\{i : p(i) \neq 0\}|$  and note that

$$\begin{aligned} h_\infty(p) &= \inf_q \sup \{-\log q(i) : i \in I, p(i) \neq 0\} \\ &= -\log \left\{ \sup_q \inf \{q(i) : i \in I, p(i) \neq 0\} \right\}. \end{aligned}$$

The quantity inside the logarithm cannot be greater than  $\frac{1}{N}$  because this would imply the existence of  $\bar{q}$  with  $\bar{q}(i) > \frac{1}{N}$  for every  $i$  such that  $p(i) \neq 0$ , which violates the normalization constraint on  $\bar{q}$ . To prove the claim we note that the infimum is achieved for  $q(i) = \frac{1}{N}$  for  $i$  such that  $p(i) \neq 0$  and  $q(i) = 0$  otherwise.

Property (c) follows from well-known properties of KL-divergence, see, for example [3, Chapter 2].

Finally, property (d) follows by observing that for every  $\epsilon > 0$ , there exist probability distributions  $q_1$  and  $q_2$  over  $I$  and  $J$  respectively, such that the following chain of inequalities holds

$$\begin{aligned}
h_r(p) &= \inf_q \left\| \sqrt{-\log q} \right\|_{L_r(I \times J, p)}^2 \leq \left\| \sqrt{-\log(q_1 q_2)} \right\|_{L_r(I \times J, p)}^2 \\
&= \left\| \sqrt{-\log q_1 - \log q_2} \right\|_{L_r(I \times J, p)}^2 \leq \left\| \sqrt{-\log q_1} + \sqrt{-\log q_2} \right\|_{L_r(I \times J, p)}^2 \\
&\leq \left( \left\| \sqrt{-\log q_1} \right\|_{L_r(I \times J, p)} + \left\| \sqrt{-\log q_2} \right\|_{L_r(I \times J, p)} \right)^2 \\
&\leq 2 \left( \left\| \sqrt{-\log q_1} \right\|_{L_r(I, p_1)}^2 + \left\| \sqrt{-\log q_2} \right\|_{L_r(J, p_2)}^2 \right) \\
&\leq 2(h_r(p_1) + h_r(p_2) + \epsilon),
\end{aligned}$$

where the third inequality follows from Minkowsky's inequality for  $L_r(I \times J, p)$  norm.  $\square$

*Proof of Theorem 2.* Property (a) follows from Definition 7 which implies that  $\mathcal{A}(\epsilon', \mathcal{F}, \mu, P_n) \subset \mathcal{A}(\epsilon, \mathcal{F}, \mu, P_n)$  whenever  $\epsilon' \leq \epsilon$ .

Property (b) follows directly from property (a) in Proposition 3.

To prepare for the proof of property (c), we observe, by Definitions 4 and 5, that

$$H(\epsilon, \mathcal{F}) = \log \sup_n \sup_{P_n} \inf\{|C| : C \in \mathcal{C}(\epsilon, \mathcal{F}, P_n)\}$$

and, by Definition 7, Definition 8 and property (b) in Proposition 3, that

$$H_\infty(\epsilon, \mathcal{F}, \mu) = \log \sup_n \sup_{P_n} \inf\{|A| : A \in \mathcal{A}(\epsilon, \mathcal{F}, \mu, P_n)\}.$$

Now, the left inequality follows by noting that for any  $A \in \mathcal{A}(2\epsilon, \mathcal{F}, \mu, P_n)$  we can build a  $C \in \mathcal{C}(2\epsilon, \mathcal{F}, P_n)$  with  $|A| \geq |C|$  associating every element  $a \in A$  with a ball in  $C$  having radius  $2\epsilon$  and center in  $a$ .

The right inequality follows by constructing from every  $C \in \mathcal{C}(\epsilon, \mathcal{F}, P_n)$ , a  $A \in \mathcal{A}(2\epsilon, \mathcal{F}, \mu, P_n)$  with  $|A| \leq |C|$ . The case  $|C| = \infty$  is trivial, hence let us assume that  $|C|$  is finite.

First we observe that by definition the elements of  $C$  have the form

$$c_k = \{f \in \mathcal{F} : \frac{1}{n} \sum_i |f(x_i) - f_k(x_i)|^2 < \epsilon^2\} \quad f_k \in \mathcal{F}, \quad k = 1, \dots, |C|$$

and without loss of generality we assume that  $\|f_k - f_h\|_{L_2(X, P_n)} > 0$  for  $k \neq h$ .

Let us consider the partition  $A = \{a_1, \dots, a_{|C|}\}$  defined by

$$a_k = \{f \in \mathcal{F} : \forall h < k, \Delta_{k,h}(f) < 0 \wedge \forall h > k, \Delta_{k,h}(f) \leq 0\},$$

where  $\Delta_{k,h}(f) = \frac{1}{n} \sum_i |f(x_i) - f_k(x_i)|^2 - \frac{1}{n} \sum_i |f(x_i) - f_h(x_i)|^2$ .

By Assumption (a) in Definition 1, the maps

$$f \mapsto \frac{1}{n} \sum_i |f(x_i) - f_k(x_i)|^2$$

are measurable, and hence subsets  $a_k$  of  $\mathcal{F}$  are measurable. Moreover, by Assumption (b) in Definition 1 applied to  $x_1, \dots, x_n$ , for every  $a_k$ , there exists  $\delta_k > 0$  such that  $B(f_k, \delta_k) \subset a_k$ , so that  $\mu(a_k) > 0$ .

Finally observe that, for every  $f, f' \in a_k$  it holds

$$\|f - f'\|_{L_2(X, P_n)} \leq \|f - f_k\|_{L_2(X, P_n)} + \|f' - f_k\|_{L_2(X, P_n)} \leq 2\epsilon,$$

which proves that  $A \subset A(2\epsilon, \mathcal{F}, \mu, P_n)$ .

The second part of the theorem follows straightforwardly from equation (9) and the properties (b) and (c) already proved.  $\square$

#### APPENDIX B. PROOFS OF RESULTS FROM SUBSECTION 4.2

*Proof of Proposition 4.* The first step is to use a symmetrization technique introducing the ghost samples  $\mathbf{x}'$  independent of  $\mathbf{x}$ , and the measure  $P'_n = \frac{1}{n} \sum_i \delta_{x'_i}$ ,

$$\begin{aligned} \mathbb{E}_{\mathbf{x}} d_r(P, P_n) &= \mathbb{E}_{\mathbf{x}} \|P - P_n\|_{\mu, r} = \mathbb{E}_{\mathbf{x}} \|\mathbb{E}_{\mathbf{x}'} P'_n - P_n\|_{\mu, r} \\ \text{[Minkowski's + Jensen's ineq.]} &\leq \mathbb{E}_{\mathbf{x}, \mathbf{x}'} \left\| \frac{1}{n} \sum_i (\delta_{x'_i} - \delta_{x_i}) \right\|_{\mu, r} \\ \text{[symmetry of } \delta_{x'_i} - \delta_{x_i}] &= \mathbb{E}_{\mathbf{x}, \mathbf{x}', \sigma} \left\| \frac{1}{n} \sum_i \sigma_i (\delta_{x'_i} - \delta_{x_i}) \right\|_{\mu, r} \\ \text{[Minkowski's ineq.]} &\leq 2 \mathbb{E}_{\mathbf{x}, \sigma} \left\| \frac{1}{n} \sum_i \sigma_i \delta_{x_i} \right\|_{\mu, r} = 2 \mathbb{E}_{\mathbf{x}} R_{r, n}(\mathcal{F}, \mu). \end{aligned}$$

The proposition follows from the estimate above applying McDiarmid's inequality (see, for example, [4, Theorem 9.2]) to the random variable  $d_r(P, P_n)$  and observing that since, for every  $x \in X$ ,  $f(x) \in [-1, 1]$ , whenever  $\mathbf{x}'$  is obtained from  $\mathbf{x}$  replacing  $x_i$  with  $x'_i$ , it holds

$$\begin{aligned} |d_r(P, P_n) - d_r(P, P'_n)| &= \left| \|P - P_n\|_{\mu, r} - \|P - P'_n\|_{\mu, r} \right| \\ \text{[Minkowski's ineq.]} &\leq \|P_n - P'_n\|_{\mu, r} \\ &= \frac{1}{n} \left( \int_{\mathcal{F}} |f(x_i) - f(x'_i)|^r d\mu(f) \right)^{\frac{1}{r}} \leq \frac{2}{n}. \end{aligned}$$

$\square$

The proof of Theorem 3 is based on the following two lemmas.

**Lemma 1.** *Let  $(\mathcal{F}, \mu)$  be a denumerable touchstone class. If for a given  $\mathbf{x} = (x_1, \dots, x_n)$  the inequality  $\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i f^2(x_i) \leq R^2$  is fulfilled, then for every  $r \in [1, \infty]$  it holds*

$$R_{r, n}(\mathcal{F}, \mu) \leq \sqrt{\frac{2R^2}{n}} \left( \sqrt{h_r(\mu|_{\mathcal{F}})} + 2 \right).$$

*Proof.* Let us fix an arbitrary probability distribution  $q$  over  $\mathcal{F}$ . For every  $f \in \mathcal{F}$  and  $\delta \in (0, 1)$ , by Hoeffding's inequality applied to the random variable  $\frac{1}{n} \sum_i \sigma_i f(x_i)$ , we get that with probability not less than  $1 - \delta q(f)$  it holds

$$(13) \quad \left| \frac{1}{n} \sum_i \sigma_i f(x_i) \right|^2 \leq \frac{2R^2}{n} \left( \log \frac{1}{q(f)} + \log \frac{2}{\delta} \right).$$

Since  $\sum_{f \in \mathcal{F}} q(f) = 1$ , with probability not less than  $1 - \delta$ , the inequality above holds uniformly over  $\mathcal{F}$ .

Taking the  $\frac{r}{2}$ -th power of (13) and averaging over  $\mathcal{F}$  w.r.t.  $\mu$ , we get that with probability not less than  $1 - \delta$  it holds

$$\begin{aligned} \left\| \frac{1}{n} \sum_i \sigma_i \delta_{x_i} \right\|_{\mu, r}^2 &\leq \frac{2R^2}{n} \left\| \sqrt{\log \frac{1}{q} + \log \frac{2}{\delta}} \right\|_{L_r(\mathcal{F}, \mu)}^2 \\ &\leq \frac{2R^2}{n} \left\| \sqrt{\log \frac{1}{q}} + \sqrt{\log \frac{2}{\delta}} \right\|_{L_r(\mathcal{F}, \mu)}^2 \\ &\leq \frac{2R^2}{n} \left( \left\| \sqrt{\log \frac{1}{q}} \right\|_{L_r(\mathcal{F}, \mu)} + \sqrt{\log \frac{2}{\delta}} \right)^2 \end{aligned}$$

The lemma follows from

$$\begin{aligned} \mathbb{E}_\sigma \left\| \frac{1}{n} \sum_i \sigma_i \delta_{x_i} \right\|_{\mu, r} &= \int_0^\infty \mathbb{P}_\sigma \left[ \left\| \frac{1}{n} \sum_i \sigma_i \delta_{x_i} \right\|_{\mu, r} > t \right] dt \\ &\leq \sqrt{\frac{2R^2}{n}} \left( \left\| \sqrt{\log \frac{1}{q}} \right\|_{L_r(\mathcal{F}, \mu)} + 2 \int_0^\infty e^{-\frac{mt^2}{2R^2}} dt \right) \\ &\leq \sqrt{\frac{2R^2}{n}} \left( \sqrt{\left\| \log \frac{1}{q} \right\|_{L_{\frac{r}{2}}(\mathcal{F}, \mu)}} + 2 \right), \end{aligned}$$

by taking the infimum of the last term, relative to  $q$  over the class of probability distributions on  $\mathcal{F}$ . □

**Lemma 2.** *Let  $(\mathcal{F}, \mu)$  be a touchstone class and  $\mathbf{x} = (x_1, \dots, x_n) \in X^n$  be arbitrary samples. For every  $r \in [1, \infty]$  it holds*

$$R_{r,n}(\mathcal{F}, \mu) \leq \frac{48}{\sqrt{n}} \left( I_r(\mathcal{F}, \mu) + \frac{1}{2} \right)$$

*Proof.* For every  $j \in \mathbb{N}$ , choose arbitrary partitions

$$A_j \in \mathcal{A}(2^{-j}, \mathcal{F}, \mu, P_n),$$

and functions  $C_j : \mathcal{F} \rightarrow \mathcal{F}$  fulfilling

$$\forall a \in A_j \quad \forall f, f' \in a \quad C_j(f) = C_j(f') \in a.$$

Moreover for  $j > 0$  define the functions  $\Delta_j : \mathcal{F} \rightarrow L_2(X, P_n)$  by

$$\forall f \in \mathcal{F} \quad \Delta_j(f) = C_j(f) - C_{j-1}(f).$$

Observe that  $\Delta_j$  is piecewise constant on the partition  $A_j \cap A_{j-1}$  composed of intersections between elements of  $A_j$  and  $A_{j-1}$ .

We define the denumerable classes of functions

$$\mathcal{F}_j = \text{Im } \Delta_j,$$

endowed with the probability measures  $\mu_j$  given by

$$\forall \hat{f} \in \mathcal{F}_j \quad \mu_j(\{\hat{f}\}) = \mu(\Delta_j^{-1}(\hat{f})).$$

Observe that for all  $\hat{f} \in \mathcal{F}_j$ , for some  $f \in \mathcal{F}$  it holds

$$\begin{aligned}
 (14) \quad \frac{1}{n} \sum_i \hat{f}^2(x_i) &= \|\Delta_j(f)\|_{L_2(X, P_n)} \\
 &\leq \|C_j(f) - f\|_{L_2(X, P_n)} + \|f - C_{j-1}(f)\|_{L_2(X, P_n)} \\
 &\leq 2^{-j} + 2^{-j+1} = 3 \cdot 2^{-j}.
 \end{aligned}$$

Therefore, since  $f = f - C_N(f) + \sum_{j=1}^N \Delta_j(f)$  for every  $N \in \mathbb{N}$ , we get

$$\begin{aligned}
 R_{r,n}(\mathcal{F}, \mu) &= \mathbb{E}_\sigma \left\| \frac{1}{n} \sum_i \sigma_i \delta_{x_i} \right\|_{\mu, r} \\
 &\leq \mathbb{E}_\sigma \left\| \frac{1}{n} \sum_i \sigma_i (\delta_{x_i} - \delta_{x_i} \circ C_N) \right\|_{L_r(\mathcal{F}, \mu)} \\
 &\quad + \sum_{j=1}^N \mathbb{E}_\sigma \left\| \frac{1}{n} \sum_i \sigma_i (\delta_{x_i} \circ C_j - \delta_{x_i} \circ C_{j-1}) \right\|_{L_r(\mathcal{F}, \mu)} \\
 (\text{Cauchy-Schwartz ineq.}) &\leq \sup_{f \in \mathcal{F}} \|f - C_N(f)\|_{L_2(X, P_n)} + 2 \sum_{j=1}^N R_{r,n} \left( \frac{1}{2} \mathcal{F}_j, \mu_j \right) \\
 (\text{Lemma 1, eq. (14)}) &\leq 2^{-N} + \sqrt{\frac{18}{n}} \sum_{j=1}^N 2^{-j+1} \left( \sqrt{h_r(\mu|_{A_j \cap A_{j-1}})} + 2 \right) \\
 (\text{Prop. 3, (d)}) &\leq 2^{-N} + \sqrt{\frac{36}{n}} \sum_{j=1}^N 2^{-j+1} \left( \sqrt{h_r(\mu|_{A_j})} + \sqrt{h_r(\mu|_{A_{j-1}})} + \sqrt{2} \right)
 \end{aligned}$$

Minimizing w.r.t. the partitions  $A_j$ , the inequality above becomes

$$\begin{aligned}
 R_{r,n}(\mathcal{F}, \mu) &= 2^{-N} + \sqrt{\frac{36}{n}} \sum_{j=1}^N 2^{-j+1} \left( \inf_{A_j} \sqrt{h_r(\mu|_{A_j})} + \inf_{A_{j-1}} \sqrt{h_r(\mu|_{A_{j-1}})} + \sqrt{2} \right) \\
 &\leq 2^{-N} + \sqrt{\frac{36}{n}} \sum_{j=1}^N 2^{-j+2} \left( \inf_{A_j} \sqrt{h_r(\mu|_{A_j})} + 1 \right) \\
 &\leq 2^{-N} + \frac{48}{\sqrt{n}} \sum_{j=1}^N (2^{-j} - 2^{-j-1}) \left( \sqrt{H_r(2^{-j}, \mathcal{F}, \mu)} + 1 \right) \\
 (\text{Th. 2, (a)}) &\leq 2^{-N} + \frac{48}{\sqrt{n}} \left( \int_0^\infty \sqrt{H_r(\epsilon, \mathcal{F}, \mu)} d\epsilon + \frac{1}{2} \right).
 \end{aligned}$$

The lemma follows taking the limit  $N \rightarrow \infty$ . □

*Proof of Theorem 3.* The proposition follows from Lemma 2 and Proposition 4 for a suitable value of  $C$  since by assumption  $-\log \delta \geq \log 2$ . □

*Proof of Proposition 5.* The proposition follows recalling Assumption (a) in Definition 1 and that  $|f(x)| \leq 1$ , by the following chain of inequalities.

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}} d_r(P, P_n) &= \mathbb{E}_{\mathbf{x}} \|P - P_n\|_{\mu, r} \\
(\text{H\"older's ineq.}) &\leq (\mathbb{E}_{\mathbf{x}} \mathbb{E}_f |Pf - P_n f|^r)^{\frac{1}{r}} \\
(\text{Fubini's Th.}) &= (\mathbb{E}_f \mathbb{E}_{\mathbf{x}} |Pf - P_n f|^r)^{\frac{1}{r}} \\
&= \left( \mathbb{E}_f \int_0^\infty \mathbb{P}_{\mathbf{x}} \left[ \left| \frac{1}{n} \sum_i f(x_i) - \mathbb{E}f \right|^r \geq \epsilon \right] d\epsilon \right)^{\frac{1}{r}} \\
(\text{Hoeffding's ineq.}) &\leq \left( 2 \int_0^\infty \exp\left(-\frac{n\epsilon^2}{2}\right) d\epsilon \right)^{\frac{1}{r}} \\
&= \sqrt{\frac{2}{n}} \left( 2r \int_0^\infty t^r e^{-t^2} dt \right)^{\frac{1}{r}} = \sqrt{\frac{2}{n}} \left( r \Gamma\left(\frac{r-1}{2}\right) \right)^{\frac{1}{r}} \leq C \sqrt{\frac{r}{n}},
\end{aligned}$$

where the last inequality is derived from Stirling's series for the Gamma function.

Hence the proposition is proved by reasoning as in the second part of the proof Proposition 4.  $\square$

#### REFERENCES

- [1] N. Alon and S. Ben-David and N. Cesa-Bianchi and D. Haussler. *Scale-sensitive dimensions, uniform convergence, and learnability*. J. ACM, 44(4):616–631, 1997.
- [2] J. B. Conway. *A Course in Functional Analysis*. 2nd edition, Springer-Verlag, 1994.
- [3] T.M. Cover and J.A.Thomas. *Elements of Information Theory*. Wiley, 1991.
- [4] L. Devroye and L. Györfi and G. Lugosi *A Probabilistic Theory of Pattern Recognition* Springer, New York, 1996.
- [5] R. M. Dudley. *A course on empirical processes*. Lecture notes in mathematics, 1097:2-142,1984.
- [6] R. M. Dudley. *Uniform central limit theorems* Cambridge University Press, Cambridge, MA, 1999.
- [7] A. Kolmogorov. *Sulla determinazione empirica di una legge di distribuzione*. Giorn. Ist. Ital. Attuari, A. IV, n 1, pages 4-11, 1933.
- [8] V. I. Koltchinskii. *On the central limit theorem for the empirical measures*. Theory of Probability and Mathematical Statistics, 24, pages 71-82, 1981.
- [9] D. Pollard. *A central limit theorem for the empirical processes*. Journal of the Australian Mathematical Society, A 33, pages 235-248, 1982.
- [10] S. Mukherjee and V. Vapnik. *Multivariate Density Estimation: An SVM Approach*. CBCL Paper #170/AI Memo #1653, MIT, Cambridge, MA, April 1999.
- [11] L. Schwartz. *Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants)*. J. Analyse Math., 13:115–256, 1964.
- [12] J. Shawe-Taylor and A. Dolia. *A new framework for probability density estimation*. Preprint, 2006.
- [13] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.
- [14] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
- [15] V. N. Vapnik and A. Chervonenkis. *On the uniform convergence of relative frequencies of events to their probabilities*. Theory Probab. Appl. 16, pages 264-280, 1971.
- [16] V. N. Vapnik and A. Chervonenkis. *Theory of pattern recognition [in Russian]*. Nauka, Moscow, 1974. (German translation: W. Wapnik & A. Tschervonenkis, *Theorie der Zeichenerkennung*, Akademie-Verlag, Berlin, 1979).

- [17] V. N. Vapnik and A. Chervonenkis. *Necessary and sufficient conditions for the uniform convergence of empirical means to their expectations*. Theory Probab. Appl. 3, pages 532-553, 1981.

ANDREA CAPONNETTO, DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF CHICAGO, 1100 EAST 58TH STREET, CHICAGO, IL 60637 and DISI, UNIVERSITÀ DI GENOVA, VIA DODECANESO 35, 16146 GENOVA, ITALY

*E-mail address:* `caponnet@uchicago.edu`

ERNESTO DE VITO, DIMA, UNIVERSITÀ DI GENOVA, VIA DODECANESO 35, 16146 GENOVA, ITALY

*E-mail address:* `devito@unimo.it`

MASSIMILIANO PONTIL, DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY COLLEGE LONDON, MALET PLACE, LONDON WC1E 6BT, UK

*E-mail address:* `m.pontil@cs.ucl.ac.uk`