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 $\mathbf{x} = (x_1, \ldots, x_n) \in X^n$ 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 [1]

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

where $\mathcal{P}(X)$ is the set of all probability measures on X. Function classes fulfilling condition (1) are called (universal) 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 13, 2009.

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

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

[1]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 μ 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 characteristics 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 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 $\mathbb{R}^k$. We denote by $\mathcal{M}(X)$ the space of (signed) bounded measures over X and by $\mathcal{P}(X)$ the subset of probability measures. Given $M \in \mathcal{M}(X)$ and a bounded measurable function $f : X \to \mathbb{R}$, we define pairing as

$$Mf = \int_X f(x) dM_+(x) - \int_X f(x) dM_-(x), \tag{2}$$

where $M = M_+ - M_-$ is Hahn decomposition of M as sum of two positive bounded measures. The above pairing suggests that any class of functions $\mathcal{F}$ defines a metrics on $\mathcal{P}(X)$ by means of

$$d(P, P') = \sup_{f \in \mathcal{F}} |Pf - P'f|. \tag{3}$$

For example, if $\mathcal{F}$ is the unit ball in the Banach space $C_0(X)$ of continuous functions on X vanishing at infinity, then d is the distance induced by the total variation, see, for example, [2, Appendix C.18]. Another classical example 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).

According to the definition (3) two probability measures are ϵ-close to each other whenever for every $f \in \mathcal{F}$ they have ϵ-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, we would like two probability measures to be ϵ-close even if they do not have ϵ-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 μ, and considering, for some $r \geq 1$, the pseudo-distance

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

The distance d given by (3) will be recovered as the limit of d_r when r goes to infinity.

Inspired by [12] we name *touchstone class* a class of functions $\mathcal{F}$ inducing a metric over $\mathcal{P}(X)$ through equation (4). The definition below formalizes the notion of touchstone class.

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 μ, 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*[2]

$$\mu \left(A_f \cap B(f, \delta) \right) > 0 \qquad \forall \delta > 0$$

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

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

The *measurability* is always relative to the σ-algebra induced by metric of $\mathcal{F}$. In most applications (see for example Examples 1 and 2 below) the metric space structure over $\mathcal{F}$ is naturally induced by a suitable space of parameters $\mathcal{T}$ through a parametrization function $t \mapsto f_t$. Assumption (a) ensures that every function in $\mathcal{F}$ is measurable and bounded on X, so that the following definition makes sense (see Proposition 1 below).

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

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

[2]Here $B(f, \delta)$ is the open ball in $\mathcal{F}$, with center f and radius δ.

The pseudo-metric introduced in equations (3) and (4) can be expressed in the form

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

As shown by next proposition, Assumption (b) ensures that

$$\lim_{r\to\infty} d_r(P,P') = d_\infty(P,P') = d(P,P'). \tag{5}$$

Clearly this assumption implies that the support of μ is $\mathcal{F}$, but this condition is not sufficient. Two examples of sufficient conditions are that the map $f \mapsto f(x)$ is continuous for all $x \in X$, or that $\mathcal{F}$ is discrete (in the definition, let $A_f = \{\mathcal{F}\}$ or $A_f = \{f\}$, respectively). 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).

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. For every $f \in \mathcal{F}$, Mf is well defined since f is measurable and bounded. Given $r \geq 1$, Fubini theorem implies that the map $f \mapsto Mf$ is r-integrable with respect to μ, so that $\|M\|_{\mu,r}$ is finite.

Part (1) follows from the finiteness of μ 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

$$|Mf' - Mf| \leq \epsilon \qquad \forall f' \in A_f \cap B(f,\delta), \tag{6}$$

and, hence,

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

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 (6) 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

$$\lim_{i\to\infty} Mf_i' = Mf,$$

which is a contradiction. □

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 $f_t(x) = \mathbf{1}\{x_i \le t_i, \ \forall i \in \{1, \dots, n\}\}$, 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 μ be an arbitrary probability measure on the metric space $\mathcal{F}$, satisfying the condition supp $\mu = \mathcal{F}$. In this example the evaluation functionals $f \mapsto f(x)$ are not continuous at $t = x$, nevertheless Assumption (b) in Definition 1 may be fulfilled thank to the upper semi-continuity property of the functions in $\mathcal{F}$. In fact, it easy to verify that a suitable choice for the sets A_f is

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

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[3] $x = \sum_i b_i(x) 2^{-i}$.

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

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

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

In the next sections, we fix a touchstone class $\mathcal{F}$ and a probability measure $P \in \mathcal{P}(X)$. For any sample $\mathbf{x} = (x_1, \dots, x_n)$ drawn i.i.d from P, we denoted by $P_n = \frac{1}{n} \sum_i \delta_{x_i}$ the corresponding empirical measure and we study the convergence of P_n to P with respect to d_r, $1 \le r \le \infty$. As discussed above, if $r = +\infty$, $d_\infty = d$, so that the results are well known, see for instance [5, 13, 6]. The convergence is possible if and only if $\mathcal{F}$ is a Glivenko-Cantelli class and an explicit non-asymptotic upper bound on $d(P, P_n)$ can be given in terms of the Koltchinskii-Pollard entropy integral $I(\mathcal{F})$. We review these results in the following section, as a preliminary step toward the generalization presented in Section 4.

3. Uniform entropy condition and Glivenko-Cantelli property

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

[3]For rational x, the expansion is not unique. In this case ties are broken by choosing the unique finite expansion.

Definition 3. *The empirical Rademacher averages of a touchstone class $\mathcal{F}$, relative to the samples $\mathbf{x} = (x_1, \ldots, x_n)$ are defined by* [4]

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

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a n-tuple of Rademacher variables[5].

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

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

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

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 $[-1, 1]$. □

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

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*

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

and the covering number of $\mathcal{F}$ as [6]

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

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 $I(\mathcal{F})$.

[4]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 $\frac{1}{2} R_n(\mathcal{F}) \leq \bar{R}_n(\mathcal{F}) \leq R_n(\mathcal{F})$.

[5]The Rademacher variables $(\sigma_1, \ldots, \sigma_n)$ are $\{-1, 1\}$-valued and independent, with $\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = \frac{1}{2}$.

[6]We denote by $|C|$ the cardinality of the set C.

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

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

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 ϵ. The inequality follows from Proposition 2 and [13, Corollary 2.2.8]. □

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 $P \in \mathcal{P}(X)$.

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 $p : I \to [0,1]$ be a probability distribution over a denumerable set* [7] *I. For every $r \in [1,\infty]$, we define the quantity*

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

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

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

and for $r = \infty$ we have

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

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 $I \times J$*

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

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

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 ϵ.*

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 μ over A, $\mu_{|A}$ is well-defined. Moreover, by Definition 7, $\mu_{|A}$ is a probability distribution[8] on A.

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

---

[7] A set is denumerable if and only if it is finite or countably infinite.

[8] Recall that the probability measure μ is, by definition, a function over the σ-field Σ of $\mathcal{F}$, fulfilling $\mu(\mathcal{F}) = 1$ and, for all a and b in Σ 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 $\sum_i \mu_{|A}(a_i) = 1$.

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

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

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 ϵ;*
(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 μ 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 μ, that

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

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 (5)) that the pseudo-metric d can be seen as the limit of the pseudo-metric d_r as r goes to infinity. 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 sample $\mathbf{x} = (x_1, \ldots, x_n)$ 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 $\sigma = (\sigma_1, \ldots, \sigma_n)$ is an n-tuple of Rademacher variables, and δ_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

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

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 $\mathbf{x} = (x_1, \ldots, x_n)$ 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*

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

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 $\mathbf{x} = (x_1, \ldots, x_n)$ 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*

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

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 (5) 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 μ 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 $\mathbf{x} = (x_1, \ldots, x_n)$ 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*

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

for some universal constant C.

The important point here is that the estimate (12) does not accounts 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 by City University of Hong Kong grant No.7200111(MA), 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), $\|\sqrt{-\log q}\|^2_{L_r(I,p)}$, 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

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

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

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 $\epsilon' \le \epsilon$.

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 $|A| \ge |C|$ associating every element $a \in A$ with a ball in C having radius 2ϵ and center in a.

The right inequality follows by constructing from every $C \in \mathcal{C}(\epsilon, \mathcal{F}, P_n)$, a $A \in A(2\epsilon, \mathcal{F}, \mu, P_n)$ with $|A| \le |C|$. 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

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

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

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

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

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 $x_1, \ldots, x_n$, 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

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

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. □

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 sample $\mathbf{x}'$ independent of $\mathbf{x}$, and the measure $P'_n = \frac{1}{n}\sum_i \delta_{x'_i}$,

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

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

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

□

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 $\mathbf{x} = (x_1, \ldots, x_n)$ 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 independent random variables $\{\sigma_i f(x_i)\}_i$, we get that with probability not less than $1 - \delta q(f)$ it holds

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

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. μ, we get that with probability not less than $1-\delta$ it holds

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

The lemma follows from

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

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 $\mathbf{x} = (x_1, \ldots, x_n) \in X^n$ be an arbitrary sample. For every $r \in [1, \infty]$ it holds*

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

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

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

and operators $C_j : \mathcal{F} \to \mathcal{F}$ fulfilling

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

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

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

Observe that Δ_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

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

endowed with the probability measures μ_j given by

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

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

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

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

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

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 $N \to \infty$. □

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.

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

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.

□

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. Tscherwonenkis, *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 Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong, Hong Kong
E-mail address: caponnet@cityu.edu.hk

Ernesto De Vito, DSA, 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