ELSEVIER
Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE @ DIRECT®
Journal of Computer and System Sciences 68 (2004) 521–545
JOURNAL OF COMPUTER AND SYSTEM SCIENCES
http://www.elsevier.com/locate/jcss

Learning with side information: PAC learning bounds

P. Kuusela[1] and D. Ocone*

Department of Mathematics, Hill Center, Busch Campus, 110 Frelinghuysen Rd., Rutgers University,
Piscataway, NJ 08854-8019, USA

Received 25 October 2002

Abstract

This paper considers a modification of a PAC learning theory problem in which each instance of the training data is supplemented with side information. In this case, a transformation, given by a side-information map, of the training instance is also classified. However, the learning algorithm needs only to classify a new instance, not the instance and its value under the side information map. Side information can improve general learning rates, but not always. This paper shows that side information leads to the improvement of standard PAC learning theory rate bounds, under restrictions on the probable overlap between concepts and their images under the side information map.

Keywords: Uniform convergence of empirical means; Learning theory; Probably Approximately Correct learning; Dependent data

1. Introduction

Probably Approximately Correct (PAC) learning theory studies algorithms that train on randomly generated data sets with the goal of learning to classify new, randomly presented instances with a small probability of error. Standard PAC theory assumes that the training sample is a string of independent and identically distributed (i.i.d.) random variables and that the training data and the instances to be classified both come from the same probability distribution. In this paper we study a variation of the standard problem in which the training data include extra information, called *side information*, which does not need to be classified. The side information learning model can be stated quite generally, but we focus on side information generated by

*Corresponding author.
E-mail addresses: pirkko.kuusela@hut.fi (P. Kuusela), ocone@math.rutgers.edu (D. Ocone).
[1] Current address: Networking Laboratory, Helsinki University of Technology, P.O.Box 3000, FIN-02015 HUT, Finland.

0022-0000/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcss.2003.07.005

classification of the output of a fixed, known transformation acting on the data. Such a problem occurs for example in simple learning theory formulations of classification by linear systems.

In the companion paper [12], we studied improvement of learning rates for the simple, concrete problem of learning an interval on a circle, when the antipodal point of each sample is provided as side information. In this problem, one is able to calculate exact learning rates for particular classification algorithms and compare them to the learning rates without side information. The analysis shows that, in general, one expects exponential improvement of rates, but that the degree of improvement depends delicately on the concept class and the distribution or class of distributions P from which samples are drawn. In some cases, the exponential improvement vanishes. The main factor affecting the improvement rate is the overlap between the concept to be learned and its transformation by the side information map.

In the present paper, we consider a general learning model with a side information map and incorporate side information into the PAC learning-theoretic setting. The objective is to show improvement in the learning rate bounds of classical learning theory when side information is present. We show that the quantitative rate improvements in the special case studied in [12] carry over qualitatively to the PAC framework; there is an improvement in learning rate bounds that depends on the amount of overlap in concepts introduced by the side information mapping. After formulating the problem precisely in Section 2, we present the main results in Sections 3 and 4. Section 3 gives rate bounds of the form $C_1 e^{-c_2 m\varepsilon}$, while Section 4 treats bounds of the form $C_1 e^{-c_2 m\varepsilon^2}$ in the large deviations setting. In these sections we make the simplifying assumption that the side information map preserves the probability measure governing the sampling of training data, as in the example treated in [12]. We show in Section 5 how this can be relaxed.

Similar side information learning problems have not (to the knowledge of the authors) been considered before, but there are some papers that are close to some aspects of the side information problem. In the presence of side information the training phase utilizes non-i.i.d. data. Meir [13] has studied a learning problem with dependent data. Due to the dependency in the data learning is slower than with an i.i.d. sample. In our case the aim is to show that side information speeds learning. For dependent data see also the paper by Campi and Kumar [5]. For related learning problems see also Blum and Mitchell [4] and learning with hints by Abu-Mostafa [1–3] as well as more theoretical work by Ratsaby and Maiorov [15]. As the focus of this paper is the exponential term in the uniform convergence rate we share similar aims with Vayatis [18] and Vayatis and Azencott [20], who has studied distribution dependent Vapnik–Chervonenkis bounds.

2. Preliminary definitions

We shall use the formulation of PAC learning and the notation found in [21]. Let $\mathscr{C}$ be a concept class, a class of subsets of a space X, and let $T \in \mathscr{C}$ be an unknown target concept to be learned. An i.i.d. sample $x_1, \ldots, x_m$ is drawn according to a probability distribution P (known or unknown) on X and an oracle returns $I_T(x_1), \ldots, I_T(x_m)$, where I_T is the characteristic function of T. Based on the multi-sample $\{(x_1, I_T(x_1)), \ldots, (x_m, I_T(x_m))\}$ an algorithm forms an estimate h_m for the unknown T. The classification error of h_m is the probability $d_P(T, h_m) = P(T \Delta h_m)$ that h_m will misclassify a future random sample drawn from P. The PAC learning error is $\operatorname{Err} h_m = P^m\{\bar{x} \in X^m; d_P(T, h_m) > \varepsilon\}$, the probability that h_m misclassifies with probability greater than ε. The quality of h_m is evaluated by the learning rate $\operatorname{lr}(m,\varepsilon) = \sup_{T \in \mathscr{C}} \operatorname{Err} h_m$, and PAC learning theory provides upper bounds on the size of $\operatorname{Err}(h,m)$, often independent of P, thus allowing one to calculate the training sample size needed to achieve an error rate ε with a specified confidence level.

We introduce next the formal model of learning with side information. Assume that there is a known mapping $s : X \to X$ and that the oracle classifies both x *and* $s(x)$. We refer to the sample value $s(x)$ as the side information. The problem again is to derive bounds on the training sample size needed to achieve a given error rate. The learner can now use the observation $\{(x_1, I_T(x_1), I_T(s(x_1))), \dots, (x_m, I_T(x_m), I_T(s(x_m)))\}$ to form an estimate $\tilde{h}_m$ for the target T, but only needs to classify correctly a future unseen x-sample (not $s(x)$). Clearly, when side information is available, the algorithm chooses from a smaller and more probably accurate set of concepts, and hence it should operate more efficiently. The question is, how much better and how does the improvement of efficiency depend on the particular learning problem?

An example of such a situation arises naturally in PAC learning formulations of linear systems identification. Here one is interested in training a linear system to map a sequence of inputs to an output at a final time. However the training data may contain the outputs at all intermediate times; these supply additional information, but the map from the input sequence to the full output sequence need not be classified.

Let $A = \tilde{h}_m \Delta T \in \mathscr{C} \Delta T$, the symmetric difference of the concept T to be learned and the estimate $\tilde{h}_m$. The analysis of an example in [12] demonstrated that there are two important factors for improvement of the learning rate when side information is present:

(i) the overlap of A and $s^{-1}(A)$
(ii) uniform upper and lower bounds on $\frac{P(s^{-1}(A))}{P(A)}$ for $A \in \mathscr{A}$, where $\mathscr{A} = \mathscr{C} \Delta T$.

The results in this paper state improvement theorems for general learning spaces and algorithms under assumptions concerning (i) and (ii). Our methods follow PAC learning methodology, as treated in Vidyasagar. The learning rates for consistent algorithms are connected to the uniform convergence of empirical probabilities by the following inequality:

$$P^m\{\bar{x} \in X^m; \exists A \text{ consistent, } d_P(A) > \varepsilon\}$$
$$\leqslant P^m \left\{ \bar{x} \in X^m; \quad \sup_{\substack{A \in \mathscr{A} \\ \sum \mathbf{1}_A(x_i) = 0}} \left| \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_A(x_i) - P(A) \right| > \varepsilon \right\}.$$

Similarly, for learning problems in which zero empirical errors (i.e., consistency) cannot be assumed, the error of the best classifier is bounded by the uniform convergence of empirical probabilities and we study (see Section 4)

$$P^m \left\{ \bar{x} \in X^m; \sup_{A \in \mathscr{A}} \left| \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_A(x_i) - P(A) \right| > \varepsilon/2 \right\}.$$

In general, in concept learning bounds of the form

$$P^m\left\{\bar{x}\in X^m;\ \sup_{A\in\mathscr{A};\sum\mathbf{1}_A(x_i)=0}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(x_i)-P(A)\right|\geqslant\varepsilon\right\}\leqslant C_1e^{-c_2m\varepsilon}, \tag{2.1}$$

or

$$P^m\left\{\bar{x}\in X^m;\ \sup_{A\in\mathscr{A}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(x_i)-P(A)\right|\geqslant\varepsilon\right\}\leqslant C_1e^{-c_2m\varepsilon^2}, \tag{2.2}$$

the exponential term comes from a permutation argument or Hoeffding's inequality, and the constant C_1 is usually a combinatorial term describing the richness of the class $\mathscr{A}$.

Note: Throughout the paper, $|y|$ means the absolute value of y, if y is a real number, but the cardinality of y, if y is a set. It should always be apparent which interpretation is meant.

3. Bounds of type $C_1e^{-c_2m\varepsilon}$

In this section we study convergence bounds for consistent algorithms, i.e., those that agree with the observed data. The proof follows the technique by Vapnik and Chervonenkis [17]. The best convergence rate without side information was achieved by Shawe-Taylor et al. [16]:

$$\Pr\left\{\sup_{A\in\mathscr{A};\sum\mathbf{1}_A(X_i)=0}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(X_i)-P(A)\right|>\varepsilon\right\}\leqslant 2\left(\frac{em}{2d}\right)^d\varepsilon^d e^{2\sqrt{2d}}e^{-m\varepsilon},$$

where $m\geqslant 4d/\varepsilon$ and $d=\mathrm{VC}(\mathscr{A})$, the Vapnik–Chervonenkis dimension of the class $\mathscr{A}$ describing the richness of the class. However, our reference point will be the bound

$$\Pr\left\{\sup_{A\in\mathscr{A};\sum\mathbf{1}_A(X_i)=0}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(X_i)-P(A)\right|>\varepsilon\right\}\leqslant 2S(\mathscr{A},m^2)e^{-m\varepsilon},$$

where

$$S(\mathscr{A},m)=\max_{x_1,\dots,x_m}|\{\mathbf{1}_A(x_1),\dots,\mathbf{1}_A(x_m);A\in\mathscr{A}\}|$$

with the same exponential term as the above. This bound allows a more uniform treatment of the two types of convergence bounds analyzed.

Let $X_i\in X$ and let $s:X\to X$ be measure preserving. For a given $A\in\mathscr{A}$ we form

$$Z_i^A=\mathbf{1}_A(x)+\mathbf{1}_A(s(x))=\begin{cases}0, & \text{with probability } p_0,\\ 1, & \text{with probability } p_1,\\ 2, & \text{with probability } p_2.\end{cases}$$

The following random quantity is essential to the bounds:

$$\tilde{N}_{\mathscr{A}}(X_1,\dots,X_{m^2})=|\{(Z_1^A,\dots,Z_{m^2}^A);A\in\mathscr{A}\}|.$$

The first result shows when side information can effectively double the sample size in the exponential term, improving the best exponent $e^{-m\varepsilon}$ to $e^{-2m\varepsilon}$.

Theorem 3.1. *If* $\mathbf{1}_A(X_i)+\mathbf{1}_A(s(X_i))\in\{0,1\}$ *for all i and* $s: X\to X$ *is measure preserving then for* $m\geqslant 2$

$$\Pr\left\{\sup_{\substack{A\in\mathscr{A}\\ \sum_{i=1}^{m} Z_i^A=0}}\left|\frac{1}{2m}\sum_{i=1}^{m} Z_i^A - P(A)\right|>\varepsilon\right\}$$
$$\leqslant 2e^{2(1+\varepsilon)}e^{-2m\varepsilon}E[\tilde{N}_{\mathscr{A}}(X_1,\dots,X_{m^2})]\leqslant 2S(\mathscr{A},2m^2)e^{2(1+\varepsilon)}e^{-2m\varepsilon}.$$

The main term in the next result is also an improved exponential convergence rate but there is also a fast decaying correction term present.

Each $A\in\mathscr{A}$ gives rise to $\bar{p}_A=(p_0,p_1,p_2)$ such that $Z^A(x)=\mathbf{1}_A(x)+\mathbf{1}_A(s(x))=j$ with probability p_j for $j=0,1,2$ and let $\mathbb{P}_{\mathscr{A}}=\{\bar{p}_A; A\in\mathscr{A}\}$.

Theorem 3.2. *Assume that* $s: X\to X$ *is measure preserving. If* $\mathbb{P}_{\mathscr{A}}=\{(p_0,p_1,p_2); p_0+p_1+p_2=1, p_2\leqslant\gamma\}$ *and* $\gamma<\varepsilon$ *then*

$$\Pr\left\{\sup_{\substack{A\in\mathscr{A}\\ \sum_{i=1}^{m} Z_i^A=0}}\left|\frac{1}{2m}\sum_{i=1}^{m} Z_i^A - P(A)\right|>\varepsilon\right\}$$
$$\leqslant 2E[\tilde{N}_{\mathscr{A}}(X_1,\dots,X_{m^2})](e^{-mf(m^2\rho)}+C(d,m^2)e^{1+\varepsilon}e^{-m\varepsilon})$$
$$\leqslant 2S(\mathscr{A},2m^2)e^{-mf(m^2\rho)}+C(d,m^2)e^{1+\varepsilon}e^{-m\varepsilon},$$

where

$$f(k)=\frac{k}{m^2}+\frac{2((m\varepsilon-1)(m-1)-k)}{m^2-k},$$
$$C(d,m)=4e^{4d^2+4d}S(\mathscr{A},2m^2)e^{-2d^2m},$$

$m\geqslant 2$, $\rho=\gamma+d$ *and* $m^2\rho\leqslant(m\varepsilon-1)(m-1)$.

We emphasize that the sample size in the "correction term" $C(d,m^2)=4e^{4d^2+4d}S(\mathscr{A},2m^4)e^{-2d^2m^2}$ is m^2 and hence that term becomes negligible very quickly and the exponential term $e^{-mf(m^2\rho)}$ dominates the bound.

One cannot derive exponential improvement when p_2 is of size ε or larger, as we shall explain later. Indeed, an example in [12] shows that exponential improvement may not occur. In this example, we want to learn a target interval (a,b) in $(0,1)$ when the side information map is $s(x)=(x+1/2)\bmod 1$. When the length of the target interval $T=(a,b)$ is $b-a=1/2+\varepsilon/2$ and the learning algorithm is the smallest interval containing positive samples, then for each estimate $\tilde{h}_m$ the p_2 corresponding to $\tilde{h}_m\Delta T$ satisfies $p_2\geqslant\varepsilon$. However, the exact learning rate is asymptotically of the form $e^{-m\varepsilon}$.

We begin by proving a fundamental lemma setting up the problem, and then separately we prove the main theorems.

The bounds on a training sample of size m are obtained by analyzing a larger sample $(X_1, \dots, X_{m^2})$ of length m^2. Informally, one thinks of the additional $m^2 - m$ training samples beyond m as a "ghost" sample. Throughout, we denote $T = X_1, \dots, X_m$ and $V = X_{m+1}, \dots, X_{m+m'}$, where $m \geqslant 2$ and $m' = m^2 - m$. Now $Z_1^A, \dots, Z_{m^2}^A$ are i.i.d. random variables and let

$$\hat{P}_T(A) = \frac{1}{2m} \sum\nolimits_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) = \frac{1}{2m} \sum_{i=1}^{m} Z_i^A$$

and

$$\hat{P}_V(A) = \frac{1}{2m'} \sum\nolimits_{i=1}^{m'} \mathbf{1}_A(X_{m+i}) + \mathbf{1}_A(s(X_{m+i})) = \frac{1}{2m'} \sum_{i=1}^{m'} Z_{m+i}^A.$$

The probability laws of $(T, V) \in X^{m^2}$, $T \in X^m$ and $V \in X^{m'}$ will be denoted by P, P_T and P_V, respectively.

Lemma 3.3. *For* $m\varepsilon > 1$

$$\Pr\left\{ \sup_{\substack{A \in \mathscr{A} \\ \hat{P}_T(A)=0}} |\hat{P}_T(A) - P(A)| \geqslant \varepsilon \right\} \leqslant 2 \Pr\left\{ \sup_{\substack{A \in \mathscr{A} \\ \hat{P}_T(A)=0}} |\hat{P}_T(A) - \hat{P}_V(A)| \geqslant (1-\alpha)\varepsilon \right\}$$

$$\leqslant 2E\left[\tilde{N}_{\mathscr{A}}(X_1, \dots, X_{m^2}) e^{\frac{-k^A}{m}} e^{\frac{-ml^A}{m^2-k^A}} \mathbf{1}_D \right] \leqslant 2\tilde{S}(\mathscr{A}, m^2) E\left[e^{\frac{-k^A}{m}} e^{\frac{-ml^A}{m^2-k^A}} \mathbf{1}_D \right],$$

where

$$l^A = \text{number of } 1\text{'s in } z_1^A, \dots, z_{m^2}^A,$$
$$k^A = \text{number of } 2\text{'s in } z_1^A, \dots, z_{m^2}^A,$$
$$D = \{m^2 - m \geqslant \ell^A + k^A > 2(m^2 - m)(1-\alpha)\varepsilon - k^A\},$$
$$\alpha = 1/(m\varepsilon),$$
$$\tilde{N}_{\mathscr{A}}(X_1, \dots, X_{m^2}) = |\{(Z_1^A, \dots, Z_{m^2}^A); A \in \mathscr{A}\}|$$

and

$$\tilde{S}(\mathscr{A}, m) = \max_{(x_1, \dots, x_m)} |\{(\mathbf{1}_A(x_1) + \mathbf{1}_A(s(x_1)), \dots, \mathbf{1}_A(x_m) + \mathbf{1}_A(s(x_m))); A \in \mathscr{A}\}|.$$

Proof. Let $\rho_A = |\hat{P}_V(A) - \hat{P}_T(A)|$ and denote $\rho = \sup_{A \in \mathscr{A}; \hat{P}_T(A)=0} \rho_A$ and $\sigma = \sup_{A \in \mathscr{A}; \hat{P}_T(A)=0}| P(A) - \hat{P}_T(A)|$. Throughout the proof we assume that

$$\sup_{A \in \mathscr{A}} \left| \frac{1}{2m} \sum\nolimits_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) - P(A) \right|$$
$$\sup_{A \in \mathscr{A}} \frac{1}{2m} \sum\nolimits_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i))$$
$$\sup_{A \in \mathscr{A}} |\hat{P}_T(A) - \hat{P}_V(A)|$$

are measurable and $\sigma > \varepsilon$ implies that there exists a random $A^* \in \mathscr{A}$ such that $|P(A^*) - \hat{P}_T(A^*)| > \varepsilon$. This happens if the class is *permissible*, for details see the appendix in [14], which in turn relies on [6] for proofs.

Step 1: In this step we link the uniform convergence of empirical probabilities in the m sample to the difference in empirical probabilities between the original sample and the ghost sample of size m'. This link is the following result taken from Devroye [8].

Proposition 3.4. *For all* $0 < \alpha < 1$ *and* $\varepsilon > 0$,

$$\Pr\{\rho > (1-\alpha)\varepsilon\} \geqslant \left(1 - \frac{1}{4\alpha^2\varepsilon^2 m'}\right)\Pr\{\sigma > \varepsilon\}.$$

Proof. The proof is in Devroye [8]. We recall the essentials. The event $\{\sigma > \varepsilon\}$ implies the existence of $A^* \in \mathscr{A}$ such that $|P(A^*) - \hat{P}_T(A^*)| > \varepsilon$ and on $\{\sigma > \varepsilon\}$, the following holds: $\{|\hat{P}_V(A^*) - P(A^*)| \leqslant \alpha\varepsilon\} \subseteq \{\rho_{A^*} > (1-\alpha)\varepsilon\} \subseteq \{\rho > (1-\alpha)\varepsilon\}$.

Thus following Devroye

$$\begin{aligned}\Pr\{\rho > (1-\alpha)\varepsilon\} &\geqslant \int_{X^{m^2}} \mathbf{1}_{\{\rho > (1-\alpha)\varepsilon\}} dP = \int_{X^m}\int_{X^{m'}} \mathbf{1}_{\{\rho > (1-\alpha)\varepsilon\}} dP_V dP_T \\ &\geqslant \int_{X^m} \mathbf{1}_{\{\sigma > \varepsilon\}} \int_{X^{m'}} \mathbf{1}_{\{\rho > (1-\alpha)\varepsilon\}} dP_V dP_T \\ &\geqslant P_T\{\sigma > \varepsilon\} \inf_{A \in \mathscr{A}} P_V\{|\hat{P}_V(A) - P(A)| \leqslant \alpha\varepsilon\}\left(1 - \tfrac{1}{4\alpha^2\varepsilon^2 m'}\right).\end{aligned}$$

Remark 3.5.

$$\begin{aligned}P_V\{|\hat{P}_V(A) - P(A)| > \alpha\varepsilon\} &= P_V\left\{\left|\frac{1}{2m'}\sum_{i=1}^{m'} Z^A_{m+i} - P(A)\right| > \alpha\varepsilon\right\} \\ &= P_V\left\{\left|\frac{1}{m'}\sum_{i=1}^{m'} Z^A_{m+i} - 2P(A)\right| > 2\alpha\varepsilon\right\} \leqslant \frac{\operatorname{Var} Z^A_{m+i}}{m'(2\alpha\varepsilon)^2} \leqslant \frac{1}{4\alpha^2\varepsilon^2 m'},\end{aligned}$$

where the first inequality follows from Chebychev's inequality and the last from the fact that $\operatorname{Var}(Z) \leqslant 1$.

Step 2: *Symmetrization by permuting*. The distribution of $\sup_{A;\hat{P}_T(A)=0} |\hat{P}_T(A) - \hat{P}_V(A)|$ is the same as the distribution of

$$\beta(\pi) := \sup_{A;\sum_{i=1}^m Z^A_{\pi(i)}=0} \left|\frac{1}{2m}\sum_{i=1}^{m} Z^A_{\pi(i)} - \frac{1}{2m'}\sum_{i=1}^{m'} Z^A_{\pi(m+i)}\right|,$$

where π is a permutation of the indices from 1 to m^2.

There are $m^2!$ possible permutations $\pi_1, \dots, \pi_{m^2!}$. Thus

$$\Pr\left\{\sup_{A;\sum_{i=1}^{m}(\pi Z_i)^A=0} |\hat{P}_T(A)-\hat{P}_V(A)|>(1-\alpha)\varepsilon\right\} = E\left[\frac{1}{m^2!}\sum_{j=1}^{m^2!}\mathbf{1}_{\{\beta(\pi_j)>(1-\alpha)\varepsilon\}}\right].$$

The last expression is equivalent to an expectation of $\mathbf{1}_{\{\beta(\pi_j)>(1-\alpha)\varepsilon\}}$ where first a random sample $X_1, \dots, X_{m^2}$ is drawn and then, independently, a permutation π is drawn from the uniform distribution on permutations of m^2 letters.

Step 3: *Conditioning.* In this step, the average over the permutations is analyzed, the sample $X_1, \dots, X_{m^2}$ being fixed. To emphasize that the sample is fixed, it will be denoted by lower case letters. Fix $x_1, \dots, x_m, x_{m+1}, \dots, x_{m+m'}$ and let $\bar{\mathscr{A}} \subset \mathscr{A}$ be a collection of sets such that any two sets in $\bar{\mathscr{A}}$ give rise to different vectors $(z_1^A, \dots, z_m^A, z_{m+1}^A, \dots, z_{m+m'}^A)$. Now we can take the supremum over $\bar{\mathscr{A}}$ instead of over $\mathscr{A}$. To ease the notation for a moment let $P(m, j) = 1/(2m)\sum_{i=1}^{m}(\pi_j z_i)^A$ and $P(m', j) = 1/(2m')\sum_{i=1}^{m'}(\pi_j z_{m+i})^A$. Then we can bound

$$\begin{aligned}
&\frac{1}{m^2!}\sum_{j=1}^{m^2!}\sup_{A;P(m,j)=0}\mathbf{1}_{\{|P(m,j)-P(m',j)|>(1-\alpha)\varepsilon\}}\\
&\quad=\frac{1}{m^2!}\sum\nolimits_{j=1}^{m^2!}\sup_{A\in\bar{\mathscr{A}};P(m,j)=0}\mathbf{1}_{\{P(m',j)>(1-\alpha)\varepsilon\}}\\
&\quad\leqslant\frac{1}{m^2!}\sum\nolimits_{j=1}^{m^2!}\sum\nolimits_{A\in\bar{\mathscr{A}};P(m,j)=0}\mathbf{1}_{\{P(m',j)>(1-\alpha)\varepsilon\}}\\
&\quad=\frac{1}{m^2!}\sum_{j=1}^{m^2!}\sum_{A\in\bar{\mathscr{A}}}\mathbf{1}_{\{P(m,j)=0\}}\mathbf{1}_{\{P(m',j)>(1-\alpha)\varepsilon\}}\\
&\quad=\sum_{A\in\bar{\mathscr{A}}}\frac{1}{m^2!}\sum_{j=1}^{m^2!}\mathbf{1}_{\{P(m,j)=0\}}\mathbf{1}_{\{P(m',j)>(1-\alpha)\varepsilon\}}.
\end{aligned}$$

Step 4: *Counting.* Fix a set $A \in \bar{\mathscr{A}}$ and observe that

$$\frac{1}{m^2!}\sum_{j=1}^{m^2!}\mathbf{1}_{\{P(m,j)=0\}}\mathbf{1}_{\{P(m',j)>(1-\alpha)\varepsilon\}}$$

is the fraction of permutations satisfying

$$\mathbf{1}_{\{P(m,j)=0\}}\mathbf{1}_{\{P(m',j)>(1-\alpha)\varepsilon\}} = 1. \tag{3.1}$$

Let l^A and k^A denote the number of 1's and 2's, respectively, in the sequence $z_1^A, \dots, z_m^A, z_{m+1}^A, \dots, z_{m+m'}^A$, where $m + m' = m^2$. We call a permutation admissible if it satisfies (3.1). There are no admissible permutations if $l^A + k^A \geqslant m'$ or $(l^A + 2k^A)/(2m') \leqslant (1-\alpha)\varepsilon$. Hence we require that $m' \geqslant l^A + k^A > 2m'(1-\alpha)\varepsilon - k^A$.

The fraction of permutations satisfying the above condition is bounded by

$$\begin{aligned}\frac{\binom{m'}{k^A}\binom{m'-k^A}{l^A}}{\binom{m+m'}{k^A}\binom{m+m'-k^A}{l^A}} &\leqslant \left(\frac{m'}{m+m'}\right)^{k^A}\left(\frac{m'-k^A}{m+m'-k^A}\right)^{l^A}\\ &= \left(1-\frac{m}{m+m'}\right)^{k^A}\left(1-\frac{m}{m+m'-k^A}\right)^{l^A}\\ &\leqslant \exp\left(-\frac{m}{m+m'}k^A\right)\exp\left(-\frac{m}{m+m'-k^A}l^A\right)\\ &= \exp\left(\frac{-k^A}{m}\right)\exp\left(\frac{-ml^A}{m^2-k^A}\right),\end{aligned}$$

where in the last line we have substituted $m'=m^2-m$.

Now we can collect all the steps together. Note that the cardinality of $\bar{\mathscr{A}}$ is given by $\tilde{N}_{\mathscr{A}}(X_1,\dots,X_{m^2})$ which is in turn bounded by $\tilde{S}(\mathscr{A},m^2)$, see Proposition 3.6 below. We get

$$\begin{aligned}\Pr\left\{\sup_{\substack{A\in\mathscr{A}\\ \hat{P}_T(A)=0}}|\hat{P}_T(A)-P(A)|>\varepsilon\right\} &\leqslant 2\Pr\left\{\sup_{\substack{A\in\mathscr{A}\\ \hat{P}_T(A)=0}}|\hat{P}_T(A)-\hat{P}_V(A)|>(1-\alpha)\varepsilon\right\}\\ &\leqslant 2E\left[\tilde{N}_{\mathscr{A}}(X_1,\dots,X_{m^2})e^{\frac{-k^A}{m}}e^{\frac{-ml^A}{m^2-k^A}}\mathbf{1}_D\right]\\ &\leqslant 2\tilde{S}(\mathscr{A},m^2)E\left[e^{\frac{-k^A}{m}}e^{\frac{-ml^A}{m^2-k^A}}\mathbf{1}_D\right], \qquad (3.2)\end{aligned}$$

where $D=\{m^2-m\geqslant l^A+k^A>2(m^2-m)(1-\alpha)\varepsilon-k^A\}$. □

Proposition 3.6.

$$\tilde{N}_{\mathscr{A}}(X_1,\dots,X_m)\leqslant\tilde{S}(\mathscr{A},m)\leqslant S(\mathscr{A},2m),$$

where

$$\begin{aligned}f_A(x) &= \mathbf{1}_A(x)+\mathbf{1}_A(s(x)),\\ \tilde{S}(\mathscr{A},m) &= \max_{(x_1,\dots,x_m)}|\{(f_A(x_1),\dots,f_A(x_m));A\in\mathscr{A}\}|,\\ S(\mathscr{A},m) &= \max_{(x_1,\dots,x_m)}|\{(\mathbf{1}_A(x_1),\dots,\mathbf{1}_A(x_m));A\in\mathscr{A}\}|.\end{aligned}$$

Proof. Fix $(x_1,\dots,x_m)$. Each $A\in\mathscr{A}$ gives rise to

$$(\mathbf{1}_A(x_1),\mathbf{1}_A(s(x_1)),\dots,\mathbf{1}_A(x_m),\mathbf{1}_A(s(x_m)))$$

and if

$$(f_A(x_1),\dots,f_A(x_m))\neq(f_{A'}(x_1),\dots,f_{A'}(x_m))$$

for some $A, A' \in \mathscr{A}$, then also

$$\begin{aligned}&(\mathbf{1}_A(x_1), \mathbf{1}_A(s(x_1)), \dots, \mathbf{1}_A(x_m), \mathbf{1}_A(s(x_m)))\\ &\quad \neq (\mathbf{1}_{A'}(x_1), \mathbf{1}_{A'}(s(x_1)), \dots, \mathbf{1}_{A'}(x_m), \mathbf{1}_{A'}(s(x_m))).\end{aligned}$$

Thus

$$\begin{aligned}&|\{(f_A(x_1), \dots, f_A(x_m)); A \in \mathscr{A}\}|\\ &\quad \leqslant |\{(\mathbf{1}_A(x_1), \mathbf{1}_A(s(x_1)), \dots, \mathbf{1}_A(x_m), \mathbf{1}_A(s(x_m))); A \in \mathscr{A}\}|\\ &\quad \leqslant \max_{(x_1, x_2, \dots, x_{2m-1}, x_{2m})} |\{(\mathbf{1}_A(x_1), \mathbf{1}_A(x_2), \dots, \mathbf{1}_A(x_{2m-1}), \mathbf{1}_A(x_{2m})); A \in \mathscr{A}\}|\\ &\quad = S(\mathscr{A}, 2m).\end{aligned}$$

Hence $\tilde{S}(\mathscr{A}, m) \leqslant S(\mathscr{A}, 2m)$. □

Proof (of Theorem 3.1). If $p_2 = 0$ then $k^A = 0$ and to obtain admissible permutations in the counting step of Lemma 3.3, $l^A > 2m'(1-\alpha)\varepsilon$. By substituting $\alpha = 1/(m\varepsilon)$ and $m' = m^2 - m$ to (3.2) and simplifying we obtain the result. □

The remainder of this section gives the proof of Theorem 3.2, which indicates improvement when p_2 is small. It turns out that we need to apply various uniform convergence results at different "levels".

Proof (of Theorem 3.2). Let l^A and k^A denote the number of 1's and 2's in the sequence $z_1^A, \dots, z_m^A, z_{m+1}^A, \dots, z_{m+m'}^A$. By taking $m' = m^2 - m$ and $\alpha = 1/(m\varepsilon)$ the condition for admissible permutations in Step 4 of the proof of Lemma 3.3 translates to condition $l^A + 2k^A > 2(m\varepsilon - 1)(m-1)$.

Then the upper bound for the fraction of admissible permutations becomes

$$e^{\frac{-k^A}{m}} e^{\frac{-ml^A}{m^2 - k^A}} \leqslant e^{\frac{-k^A}{m}} e^{\frac{-2m}{m^2 - k^A}((m\varepsilon - 1)(m-1) - k^A)} = e^{-mf(k^A)},$$

where

$$f(k) = \frac{k}{m^2} + \frac{2((m\varepsilon - 1)(m-1) - k)}{m^2 - k}$$

for $0 \leqslant k \leqslant (m\varepsilon - 1)(m-1)$. $f(k) > 0$ and f is a decreasing function of k. Note that $f(0)$ gives the bound of the previous theorem.

To proceed we need to solve a technical difficulty. Let $\hat{p}_1 = \ell^A/m^2$, $\hat{p}_2 = k^A/m^2$ and $\hat{p}_0 = 1 - \hat{p}_1 - \hat{p}_2$. Unfortunately, the $(\hat{p}_0, \hat{p}_1, \hat{p}_2)$ obtained above is not guaranteed to be in $\mathbb{P}_{\mathscr{A}}$ although (p_0, p_1, p_2) (or A) that generated the sequence $z_1, \dots, z_{m^2}$ is in $\mathbb{P}_{\mathscr{A}}$. However, note that $\hat{p}_2 \to p_2$ in probability and we get a bound:

Proposition 3.7

$$\Pr\left\{ \sup_{A \in \mathscr{A}} \hat{p}_2 - p_2 > d \right\} \leqslant 4e^{4d^2 + 4d} S(\mathscr{A}, 2m^4) e^{-2d^2 m^2}.$$

Proof. We use the bound by Devroye [16]:

$$\Pr\left\{\sup_{\hat{A}\in\mathscr{B}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\hat{A}}(X_i)-P(\hat{A})\right|>\varepsilon\right\}\leqslant 4e^{4\varepsilon^2+4\varepsilon}S(\mathscr{B},n^2)e^{-2\varepsilon^2n^2}.$$

Now $\hat{p}_2=1/m^2\sum_{i=1}^{m^2}\mathbf{1}_{\{Z_i^A=2\}}$ and define $W_i^A=W^A(X_i)=1$, if $Z_i^A=2$ and 0 otherwise. This give rise to a new collection of sets $\mathscr{A}$ (namely sets $\hat{A}=A\cap s^{-1}(A)$ as A ranges through $\mathscr{A}$). For any $(x_1,\dots,x_{m^4})$ the number of vectors in $\{(W^A(x_1),\dots,W^A(x_{m^4}));A\in\mathscr{A}\}$ is less than the number of vectors in $\{(Z^A(x_1),\dots,Z^A(x_{m^4}));A\in\mathscr{A}\}\leqslant S(\mathscr{A},2m^4)$, where there the last bound follows from Proposition 3.6. Hence we can apply Devroye's result with $\mathscr{B}=\hat{\mathscr{A}}$, $n=m^2$ and $S(\hat{\mathscr{A}},n^2)\leqslant S(\mathscr{A},2m^4)$. □

To solve the technical difficulty the idea is to relax $\mathbb{P}_{\mathscr{A}}$ to $\mathbb{P}_{\mathscr{A},d}=\{(p_0,p_1,p_2);p_0+p_1+p_2=1,p_2\leqslant\gamma+d\}$ and because $\Pr\{\sup_{A\in\mathscr{A}}\hat{p}_2\leqslant\gamma+d\}\geqslant 1-4e^{4d^2+4d}S(\mathscr{A},2m^4)e^{-2d^2m^2}$ this occurs with high probability. In the remaining part we have no control over $\hat{p}_2$ and we apply the existing bounding technique for general $\{0,1,2\}$-valued random variable.

Therefore, if $p_2\leqslant\gamma$, then for $d>0$ such that $\rho=\gamma+d<\varepsilon$, Proposition 3.7 gives that $\Pr\{\sup_{A\in\mathscr{A}}k^A/m^2-\gamma>d\}\leqslant 4e^{4d^2+4d}S(\mathscr{A},2m^4)e^{-2d^2m^2}$.

Thus

$$E\left[e^{\frac{-k^A}{m}}e^{\frac{-ml^A}{m^2-k^A}}\right]\leqslant e^{-mf(m^2\rho)}+C(d,m^2)e^{1+\varepsilon}e^{-m\varepsilon}.$$

This with Lemma 3.3 implies the result. □

4. Bounds of the form $C_1e^{-c_2\varepsilon^2}$

In this section we study the bound

$$P^m\left\{\bar{x}\in X^m;\sup_{A\in\mathscr{A}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(x_i)-P(A)\right|\geqslant\varepsilon\right\}\leqslant C_1e^{-c_2m\varepsilon^2}$$

in the large deviations setting and we indicate how side information can improve the exponential term. The main results are stated and proved in Section 4.2.

4.1. Preliminaries

Our derivation of convergence bounds in the presence of side information will use the upper bound from Cramér's Large Deviation Theorem in conjunction with a bound on uniform convergence. The Cramér upper bound states that if Y is a random variable and F is a closed set, then

$$\Pr(F)\leqslant 2\exp\left\{-\inf_{y\in F}\sup_{\lambda\in\mathbb{R}}(\lambda y-\ln E[e^{\lambda Y}])\right\}.$$

See the proof of Cramér's large deviation theorem in any text on large deviations, for example [7]. If $Y_1, \ldots, Y_n$ are i.i.d., mean zero, and $P(a \leqslant Y \leqslant b) = 1$, then an application of the Cramér bound, together with simple bounding, yields Hoeffding's inequality:

$$P\left((1/n) \sum_{1}^{n} Y_i \geqslant \varepsilon \right) \leqslant e^{-2n\varepsilon^2/(b-a)^2}. \tag{4.1}$$

The following lemma of Hoeffding will also be useful.

Lemma 4.1 (Hoeffding's lemma [9]). *Let $W_1, \ldots, W_m$ denote a random sample without replacement and let $Y_1, \ldots, Y_m$ be a sample with replacement (and hence i.i.d.) from a finite population of real values.*

If f is continuous and convex then

$$E\left[f\left(\sum_{i=1}^{m} W_i \right) \right] \leqslant E\left[f\left(\sum_{i=1}^{m} Y_i \right) \right].$$

Recall the learning setup. The random variables $X_1, \ldots, X_m$ are i.i.d. with probability law P, the transformation $s : X \to X$ is a measure preserving, and $Z_i^A = \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i))$. Let $\bar{p} = (p_0, p_1, p_2)$, where $p_j = P(Z_i^A = j)$, represent the probability distribution of Z_i^A and let

$$\Lambda_{\bar{p}}^*(z) = \sup_{\lambda \in \mathbb{R}} (\lambda z - \ln E[e^{\lambda Z^A}]) \tag{4.2}$$

denote the corresponding *rate function*. Then, letting $S_m = Z_1^A + \cdots + Z_m^A$, one finds from applying the Cramér theorem upper bound to $Y = S_m$ and using independence that

$$\begin{aligned} \Pr\left\{ \left| \frac{1}{2m} S_m - P(A) \right| \geqslant \varepsilon \right\} &= \Pr\left\{ \frac{1}{m} S_m \in (-\infty, 2P(A) - 2\varepsilon] \cup [2P(A) + 2\varepsilon, \infty) \right\} \\ &\leqslant 2 \exp\left(-m \inf_{z \in F_\varepsilon(P(A))} \Lambda_{\bar{p}}^*(z) \right), \end{aligned} \tag{4.3}$$

where $F_\varepsilon(P(A)) = (-\infty, 2P(A) - 2\varepsilon] \cup [2P(A) + 2\varepsilon, \infty)$. Note here that $2mP(A) = E[S_m]$.

The ultimate goal is to obtain a bound similar to (4.3), uniformly over the concept class, in order to obtain an upper bound on the learning rate. This section presents a preliminary bound, from which the learning rate bounds will follow easily. The derivation of this preliminary bound follows standard procedure; the particular choices of parameter values used here are similar to those of Devroye [8]. There are two tricks. First, extend the training sample by adding $m' = m^2 - m$ "ghost" samples to get a random sample $X_1, \ldots, X_{m^2}$ of m^2 observations. Second, reorder these m^2 observations according to an independent random permutation. If the permutations are chosen from the uniform distribution, the final distribution of the re-ordered sample will be the same as that of the un-permuted sample. The bound is obtained by conditioning on the original sample and then applying the Hoeffding lemma and the Cramér bound to its random permutations.

The following notation will be used for the extended sample of length m^2. First, $\mathbb{Z}_j^A = \mathbf{1}_A(X_j) + \mathbf{1}_A(s(X_j))$, for $1 \leqslant j \leqslant m^2$, as usual. Then, we use $\hat{p}(A) = (\hat{p}_0(A), \hat{p}_1(A), \hat{p}_2(A))$ to denote the empirical distribution of $(Z_1^A, \ldots, Z_{m^2}^A)$; that is

$$\hat{p}_j(A) = \frac{|\{s; 1 \leqslant s \leqslant m^2, Z_s^A = j\}|}{m^2}, \quad j = 0, 1, 2.$$

Let $\hat{P}(A) = (1/m^2)\sum_{i=1}^{m^2} Z_i^A = \hat{p}_1 + 2\hat{p}_2$ denote the corresponding empirical mean.

Next, consider two sets A and B in $\mathscr{A}$. Given a sample $(X_1, \ldots, X_{m^2})$, say that A and B are equivalent if

$$A \cap \{X_1, s(X_1), \ldots, X_{m^2}, s(X_{m^2})\} = B \cap \{X_1, s(X_1), \ldots, X_{m^2}, s(X_{m^2})\},$$

and choose a finite subclass $\hat{\mathscr{A}}$ of $\mathscr{A}$ such that

(i) $A, B \in \hat{\mathscr{A}} \Rightarrow A$ and B are not equivalent;
(ii) for every $A \in \mathscr{A}$ there exists $B \in \hat{\mathscr{A}}$ that is equivalent to A.

Note that $\hat{\mathscr{A}}$ is random; we should really write $\hat{\mathscr{A}}(X_1, \ldots, X_{m^2})$, but we suppress the dependence on the sample for notational simplicity.

Lemma 4.2. *Let $s : X \to X$ be measure preserving. Then for $m \geqslant 2$,*

$$\Pr\left\{\sup_{A \in \mathscr{A}} \left|\frac{1}{2m}\sum_{i=1}^{m} Z_i^A - P(A)\right| \geqslant \varepsilon\right\} \leqslant 4E\left[\sum_{A \in \hat{A}} e^{-m \inf_{z \in F_{\varepsilon'}(\hat{P}(A))} \Lambda^*_{\hat{p}(A)}}\right] \tag{4.4}$$

where $\varepsilon' = (m\varepsilon - 1)(m-1)/m^2$.

As a consequence,

Corollary 4.3. *Let $s : X \to X$ be measure preserving. Then for $m\varepsilon > 1$,*

$$\begin{aligned}
&\Pr\left\{\sup_{A \in \mathscr{A}} \left|\frac{1}{2m}\sum_{i=1}^{m} Z_i^A - P(A)\right| \geqslant \varepsilon\right\} \\
&\quad \leqslant 4E\left[\tilde{N}_{\mathscr{A}}(X_1, \ldots, X_{m^2}) \exp\left(-m \inf_{A \in \mathscr{A}} \inf_{z \in F_{\varepsilon'}} \tilde{\Lambda}^*_{\bar{p}}(z)\right)\right] \\
&\quad \leqslant 4S(\mathscr{A}, 2m^2) E\left[\exp\left(-m \inf_{A \in \mathscr{A}} \inf_{z \in F_{\varepsilon'}} \tilde{\Lambda}^*_{\bar{p}}(z)\right)\right],
\end{aligned}$$

where

$$\begin{aligned}
&\varepsilon' = (m\varepsilon - 1)(m-1)/m^2, \\
&\tilde{N}_{\mathscr{A}}(X_1, \ldots, X_{m^2}) = |\{(Z_1^A, \ldots, Z_{m^2}^A); A \in \mathscr{A}\}|, \\
&S(\mathscr{A}, m) = \max_{x_1, \ldots, x_m} |\{\mathbf{1}_A(x_1), \ldots, \mathbf{1}_A(x_m); A \in \mathscr{A}\}|,
\end{aligned}$$

and $\tilde{\Lambda}^*_{\bar{p}}$ is the rate function (4.2).

In the statement of Lemma 4.2, it is formally assumed that the integrand is a measurable random variable. This is not actually necessary for passing to the bound in Corollary 4.3 or, later, to the bounds in theorems, because these all arise by bounding inequality (4.8) below.

Proof. *Step* 1: Divide the sample of size m^2 into two parts T and V, where $T = X_1, \dots, X_m$ and $V = X_{m+1}, \dots, X_{m+m'}$, where $m' = m^2 - m$. Define the empirical probabilities associated to each part:

$$\hat{P}_T(A) := \frac{1}{2m} \sum_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) = \frac{1}{2m} \sum_{i=1}^{m} Z_i^A$$

and

$$\hat{P}_V(A) := \frac{1}{2m'} \sum\nolimits_{i=1}^{m'} \mathbf{1}_A(X_{m+i}) + \mathbf{1}_A(s(X_{m+i})) = \frac{1}{2m'} \sum_{i=1}^{m'} Z_{m+i}^A.$$

The event of interest, that is the event whose probability we are bounding in Lemma 4.2, is $\sup_{A \in \mathscr{A}} |P(A) - \hat{P}_T(A)| \geqslant \varepsilon$. As in Step 1 of Lemma 3.3, one can show that for all $0 < \alpha < 1$ and $\varepsilon > 0$,

$$\Pr\{\rho > (1-\alpha)\varepsilon\} \geqslant \left(1 - \frac{1}{4\alpha^2\varepsilon^2 m'}\right) \Pr\left\{\sup_{A \in \mathscr{A}} |P(A) - \hat{P}_T(A)| > \varepsilon\right\}, \tag{4.5}$$

where $\rho = \sup_{A \in \mathscr{A}} |\hat{P}_V(A) - \hat{P}_T(A)|$. We are now concerned with bounding the probability of the left-hand side. For convenience of notation, take $\rho_A := |\hat{P}_V(A) - \hat{P}_T(A)|$.

Step 2: *Permutations.* Let π denote a permutation of the indices from 1 to m^2, and let $\{\pi_i; 1 \leqslant i \leqslant m^2!\}$ be a list of all such permutation. Let (T_π, V_π) denote the sample obtained from (T, V) by permuting the indices using π (T, V), and introduce $\rho_A(\pi) = |\hat{P}_{V_\pi}(A) - \hat{P}_{T_\pi}(A)|$ and all other notations similarly. Consider now the sample obtained first by choosing $X_1, \dots, X_{m^2}$ i.i.d.. according to P and then permuting them by a drawing π at random from the uniform measure on the permutation group on m^2 letters. The result does not change the final distribution of the sample. Thus

$$\Pr\{\rho > (1-\alpha)\varepsilon\} = \Pr\left\{\frac{1}{m^2!} \sum\nolimits_{i=1}^{m^2!} \mathbf{1}_{\{\rho(\pi_i) > (1-\alpha)\varepsilon\}}\right\}. \tag{4.6}$$

Now bound the fraction of permutations such that $\rho(\pi) > (1-\alpha)\varepsilon$ as follows:

$$\begin{aligned}
\frac{1}{m^2!} \sum_{i=1}^{m^2!} \mathbf{1}_{\{\rho(\pi_i) > (1-\alpha)\varepsilon\}} &= \frac{1}{m^2!} \sum_{i=1}^{m^2!} \sup_{A \in \mathscr{A}} \mathbf{1}_{\{\rho_A(\pi_i) > (1-\alpha)\varepsilon\}} \\
&\leqslant \frac{1}{m^2!} \sum_{i=1}^{m^2!} \sum_{A \in \mathscr{A}(T,V)} \mathbf{1}_{\{\rho_A(\pi_i i) > (1-\alpha)\varepsilon\}} \\
&= \sum_{A \in \mathscr{A}(T,V)} \frac{1}{m^2!} \sum\nolimits_{i=1}^{m^2!} \mathbf{1}_{\{\rho_A(\pi_i) > (1-\alpha)\varepsilon\}},
\end{aligned}$$

Step 3: Condition on $X_1, \dots, X_{m^2}$ and hence consider the corresponding sequence $Z_1^A, \dots, Z_{m^2}^A$ as fixed and given. For notational convenience we drop the superscript A for a moment. Observe that the full empirical probability of A,

$$\hat{P}_{(T_\pi, V_\pi)}(A) = \frac{1}{2(m+m')}\left(\sum\nolimits_{j=1}^{m} Z_{\pi(j)} + \sum_{j=1}^{m'} Z_{\pi(m+j)}\right)$$

is invariant with respect to the permutation π and is equal to what we have defined as $\hat{P}(A)$. Thus, $\hat{P}_{V_\pi}(A) = \frac{1}{2m'}\left(2m^2\hat{P}(A) - \sum_{j=1}^{m} Z_{\pi(j)}\right)$. As a result, we easily find that

$$\{\rho_A(\pi) > (1-\alpha)\varepsilon\} = \left\{\left|\frac{1}{2m}\sum_{j=1}^{m} Z_{\pi(j)} - \hat{P}(A)\right| > (1-\alpha)\varepsilon\frac{m'}{m^2}\right\}. \tag{4.7}$$

Let us denote $\varepsilon' = (1-\alpha)\varepsilon m'/m^2$. If we choose the permutation π uniformly from the set of all permutations of the letters $1, \dots, m^2$, $Z_{\pi(1)}, \dots, Z_{\pi(m)}$ will be distributed as a random sample with replacement of the m^2 values $Z_1, \dots, Z_{m^2}$ according to the probability distribution $\hat{p}_A$. By applying the Cramér bound (4.3) and then bounding the exponent rate function using Hoeffding's lemma (4.7),

$$\begin{aligned}\frac{1}{m^2!}\sum_{i=1}^{m^2!} \mathbf{1}_{\{\rho_A(\pi_i) > (1-\alpha)\varepsilon\}} &\leqslant 2\exp\left\{-\inf_{y\in F_\varepsilon(\hat{P}(A))}\ \sup_{\lambda\in\mathbb{R}}\ \left(\lambda x - \ln E\left[e^{\lambda\left(\sum_{i=1}^{m} W_i\right)}\right]\right)\right\}\\ &\leqslant 2\exp\left\{-m \inf_{y\in F_\varepsilon(\hat{P}(A))} \Lambda^*_{\hat{p}(A)}(y)\right\}.\end{aligned} \tag{4.8}$$

Step 4: The proof of Lemma 4.2 now follows by applying (4.7) in step 2 and (4.5) and taking $\alpha = (m\varepsilon)^{-1}$. □

Remark. Application of Hoeffding's inequality in (4.7) implies

$$\frac{1}{m^2!}\sum\nolimits_{i=1}^{m^2!} \mathbf{1}_{\{\rho_A(\pi_i) > (1-\alpha)\varepsilon\}} \leqslant 2\exp\{-2m\varepsilon'^2\} \leqslant 2e^{-2m\varepsilon^2+4\varepsilon+4\varepsilon^2} \quad \text{when } \alpha = \frac{1}{m\varepsilon}. \tag{4.9}$$

This gives a general bound on the convergence rate, but does not capture any improvement due to the expected magnitude of the overlap probability. Lemma 4.1 includes an explicit minimization over the rate functions to obtain an improved bound.

4.2. Bounding the uniform convergence, exponential term

So far we have just assumed that the side information mapping $s(x)$ is measure preserving. However, more has to be assumed for improved rates. In this case we illustrate cases in which we can derive an improved bound.

4.2.1. Case 1: "No overlap"

Assume that $\bar{p}_A = (p, 1-p, 0)$ for each $A \in \mathscr{A}$. This situation was studied (see [12]) in the interval learning problem when the target $T \subset (0, 1/2)$ and $s(x) = (x + 1/2) \bmod 1$.

Theorem 4.4. *If $\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) \in \{0, 1\}$ for all i and $s : X \to X$ is measure preserving then for $m \geqslant 2$*

$$\begin{aligned}
&\Pr\left\{\sup_{A \in \mathscr{A}} \left|\frac{1}{2m}\sum\nolimits_{i=1}^m \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) - P(A)\right| \geqslant \varepsilon\right\} \\
&\quad \leqslant 4e^{16\varepsilon^2+16\varepsilon} E[\tilde{N}_{\mathscr{A}}(X_1, \dots, X_{m^2})] e^{-8m\varepsilon^2} \\
&\quad \leqslant 4e^{16\varepsilon^2+16\varepsilon} S(\mathscr{A}, 2m^2) e^{-8m\varepsilon^2}.
\end{aligned}$$

Proof. In this case, $\hat{p}_2 = 0$ always and for $0 < p_1 < 1$

$$\Lambda^*_{\hat{p}}(x) = \Gamma_{\hat{p}_1}(x) := \begin{cases} x \ln \dfrac{x}{\hat{p}_1} + (1-x) \ln \dfrac{1-x}{1-\hat{p}_1} & \text{if } x \in [0, 1], \\ \infty & \text{otherwise.} \end{cases}$$

When $p_0 = 0$ or $p_0 = 1$, this rate function is infinite. Recall that $F_{\varepsilon'} = (-\infty, \hat{p}_1 - 2\varepsilon'] \cup [\hat{p}_1 + 2\varepsilon', \infty)$. By convexity,

$$\inf_{x \in F_{\varepsilon'}} \Gamma^*_{\hat{p}_1}(x) = \min\{\Gamma_{\hat{p}_1}(\hat{p}_1 + 2\varepsilon'), \Gamma_{\hat{p}_1}(\hat{p}_1 - 2\varepsilon')\}.$$

It is well known from Chernoff's bound that $\Gamma_{\hat{p}_1}(\hat{p}_1 + 2\varepsilon') \geqslant 8\varepsilon'^2$ and similarly for $\Gamma^*_p(p - 2\varepsilon')$. Thus the exponential term in the bound of Corollary 4.3 is less than $e^{-8m\varepsilon'^2} \leqslant e^{16\varepsilon^2+16\varepsilon} e^{-8m\varepsilon^2}$. □

In comparison to the best exponent by Devroye [8]

$$\Pr\left\{\sup_{A \in \mathscr{A}} \left|\frac{1}{m}\sum\nolimits_{i=1}^m \mathbf{1}_A(X_i) - P(A)\right| \geqslant \varepsilon\right\} \leqslant 4e^{4\varepsilon+4\varepsilon^2} S(\mathscr{A}, m^2) e^{-2m\varepsilon^2} \tag{4.10}$$

we see that we have been able to improve the exponent by the factor of 4. Although the combinatorial term is worse, that term grows only polynomially in the sample size and the improvement in the exponential term gives a better bound unless m is small.

4.2.2. Case 2: "Small overlap"

We assume that the overlap of A and $s^{-1}(A)$ is small and we consider

$$\mathbb{P}_{\mathscr{A}} = \left\{(p_0, p_1, p_2); p_i \geqslant 0, \sum p_i = 1, p_2 \leqslant \beta\right\}.$$

The rate function for $p_i > 0$, $p_0 + p_1 + p_2 = 1$ and $0 \leqslant x \leqslant 2$ is given by

$$\begin{aligned}\Lambda^*_{\bar{p}}(x) &= \Lambda^*_{p_1,p_2}(x)\\ &= \inf_{q_0,q_1,q_2}\left\{ q_0 \ln\frac{q_0}{p_0} + q_1 \ln\frac{q_1}{p_1} + q_2 \ln\frac{q_2}{p_2}; q_1 + 2q_2 = x, q_0 + q_1 + q_2 = 1\right\}\\ &= \inf_{\substack{\max(0,x-1)\leqslant q\\ \leqslant \min(1,x/2)}}\left\{(1-x+q)\ln\frac{1-x+q}{1-p_1-p_2} + (x-2q)\ln\frac{x-2q}{p_1} + q\ln\frac{q}{p_2}\right\}.\end{aligned}$$

We begin by showing that, as in the proof of Theorem 3.2, $\mathbb{P}_{\mathscr{A}}$ has to be relaxed to allow larger values for p_2 and the convergence bound can be split into two parts: one occurring with high probability in which we have control over p_2 and one with small probability, where we apply existing bounding techniques.

Theorem 4.5. *If $s : X \to X$ is measure preserving and $\mathbb{P}_{\mathscr{A}} \subset \{(p_0, p_1, p_2); p_0 + p_1 + p_2 = 1, p_2 \leqslant \beta\}$ then*

$$\begin{aligned}&\Pr\left\{\sup_{A\in\mathscr{A}}\left|\frac{1}{2m}\sum_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) - P(A)\right| > \varepsilon\right\}\\ &\leqslant 4E[\tilde{N}_{\mathscr{A}}(X_1, \dots, X_{m^2})e^{-m \inf_{p\in\mathbb{P}_{\mathscr{A}}} \inf_{x\in F_{\varepsilon'}} \Lambda^*_{\bar{p}}(x)}]\\ &\leqslant 4S(\mathscr{A}, 2m^2)e^{-m \inf_{p\in\mathbb{P}_{\mathscr{A},d}} \inf_{x\in F_{\varepsilon'}} \Lambda^*_{\bar{p}}(x)} + C(d, m^2)e^{-2m\varepsilon^2+4\varepsilon+4\varepsilon^2},\end{aligned}$$

where $m \geqslant 2$, $\mathbb{P}_{\mathscr{A},d} = \{(p_0, p_1, p_2); p_0 + p_1 + p_2 = 1, p_2 \leqslant \beta + d\}$ and $C(d, m) = 4e^{4d^2+4d}$ $S(\mathscr{A}, 2m^2)e^{-2d^2m}$.

Again we note that the sample size in the "correction term" $C(d, m^2) = 4e^{4d^2+4d}$ $S(\mathscr{A}, 2m^4)e^{-2d^2m^2}$ is m^2 and hence that term becomes negligible very quickly and the exponential term from the rate function dominates.

Proof. We relax $\mathbb{P}_{\mathscr{A}}$ to $\mathbb{P}_{\mathscr{A},d} = \{(p_0, p_1, p_2); p_0 + p_1 + p_2 = 1, p_2 \leqslant \beta + d\}$. By the uniform bound in Devroye, $\Pr\{\sup_{A\in\mathscr{A}} \hat{p}_2 \geqslant \beta + d\} \leqslant 4e^{4d^2+4d}S(\mathscr{A}, 2m^4)e^{-2d^2m^2}$. We employ the method of the proof of Lemma 4.2. In case of the rare event $\{\sup_{A\in\mathscr{A}} \hat{p}_2 \geqslant \beta + d\}$ we apply the general bound (4.8) from Hoeffding's inequality in bounding the average over the permutations. Then on the event $\{\sup_{A\in\mathscr{A}} \hat{p}_2 \leqslant \beta + d\}$ we use the bound of (4.7). Putting these two bounds together gives the bound of Theorem 4.5. □

Next we illustrate that we can get the exponential term to be as close to $e^{-8m\varepsilon^2}$ as we wish by taking the bound on p_2 to be small enough:

Theorem 4.6. *For fixed $\varepsilon > 0$, given $\lambda > 0$ there exists a ρ^0 such that*

$$\min\{\Lambda^*_{\bar{p}}(p_1 + 2p_2 + 2\varepsilon), \Lambda^*_{\bar{p}}(p_1 + 2p_2 - 2\varepsilon)\} \geqslant (8 - \lambda)\varepsilon^2$$

whenever $0 < p_2 \leqslant \rho^0$, $p_1 \geqslant 0$ and $p_1 + p_2 \leqslant 1$.

Corollary 4.7. *Using Theorem* 4.5 *if $m > M$ such that $\varepsilon' = (m\varepsilon - 1)(m-1)/m^2 > \varepsilon_M$ and $\beta + d < \rho^0$ then*

$$\Pr\left\{ \sup_{A \in \mathscr{A}} \left| \frac{1}{2m} \sum_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) - P(A) \right| > \varepsilon \right\}$$
$$\leqslant 4S(\mathscr{A}, 2m^2) e^{-(8-\lambda)m\varepsilon_M^2} + C(d, m^2) e^{-2m\varepsilon^2 + 4\varepsilon + 4\varepsilon^2},$$

where the "correction term" $C(d, m^2) = 4e^{4d^2+4d} S(\mathscr{A}, 2m^4) e^{-2d^2m^2}$.

Proof. The proof can be found in Appendix A. □

Remark 4.8. In general, finding

$$\min_{\substack{0 \leqslant p_2 \leqslant \rho \\ 0 \leqslant p_1 \leqslant 1 \\ \overline{0 \leqslant p_1 + p_2 \leqslant 1}}} \min\{\Lambda^*_{\bar{p}}(p_1 + 2p_2 + 2\varepsilon'), \Lambda^*_{\bar{p}}(p_1 + 2p_2 - 2\varepsilon')\} \tag{4.11}$$

analytically is very difficult. However, numerical experiments suggest that if we have a bound $p_2 \leqslant \rho \in [0.032, 1/4]$ then $R(\rho, \varepsilon)\varepsilon^2$, where

$$R(\rho, \varepsilon) = \frac{8}{1 + 8\rho} - \frac{8}{6} \frac{192\rho\varepsilon}{(1 + 8\rho)^3},$$

may serve as a lower bound for (4.11).

4.2.3. Case 3: "Complete covering"

The case of "complete covering" in which $\bar{p}_A = (0, 1-p, p)$ for each $A \in \mathscr{A}$ parallels the case $p_2 = 0$. Note, however, that the problem is not symmetric due to the relation $p_1 + 2p_2 = x$ in the relative entropy. Methods and ideas in the proofs correspond to ones in the case $p_2 = 0$ or p_2 small. We begin by showing an improvement in the rate function:

Theorem 4.9. *If $\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) \in \{1, 2\}$ for all i and $s : X \to X$ is measure preserving then for $m \geqslant 2$*

$$\Pr\left\{ \sup_{A \in \mathscr{A}} \left| \frac{1}{2m} \sum_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) - P(A) \right| \geqslant \varepsilon \right\}$$
$$\leqslant 4e^{16\varepsilon^2 + 16\varepsilon} E[\tilde{N}_{\mathscr{A}}(X_1, \ldots, X_{m^2})] e^{-8m\varepsilon^2}$$
$$\leqslant 4e^{16\varepsilon^2 + 16\varepsilon} S(\mathscr{A}, 2m^2) e^{-8m\varepsilon^2}.$$

Proof. In this case the $\hat{p} = (\hat{p}_0, \hat{p}_1, \hat{p}_2)$ from the Step 3 in the proof of Lemma 4.2 satisfies $\hat{p} \in \mathbb{P}_{\mathscr{A}} = \{\bar{p}_A = (p_0, p_1, p_2); A \in \mathscr{A}\}$ and

$$E[e^{-m \inf_{x \in F_{\varepsilon'}} \tilde{\Lambda}^*_{\hat{p}}(x)}] \leqslant e^{-m \inf_{\bar{p} \in \mathbb{P}_{\mathscr{A}}} \inf_{x \in F_{\varepsilon'}} \Lambda^*_{\bar{p}}(x)},$$

where

$$\Lambda_{\bar{p}}^*(x) = \Lambda_p^*(x) = \begin{cases} (2-x)\ln\dfrac{2-x}{p} + (x-1)\ln\dfrac{x-1}{1-p} & \text{if } x \in [1,2], \\ \infty & \text{otherwise.} \end{cases}$$

Because

$$\Lambda_p^*(x) = \frac{1}{-2+3x-x^2} \geqslant 4$$

we conclude that

$$\inf_{x \in F_{\varepsilon'}} \Lambda_p^*(x) = \min\{\Lambda_p^*(p+2\varepsilon'), \Lambda_p^*(p-2\varepsilon')\} \geqslant 8\varepsilon'^2.$$

As $e^{-8m\varepsilon'^2} \leqslant e^{16\varepsilon^2+16\varepsilon}e^{-8m\varepsilon^2}$, the result follows from Lemma 4.2. □

Also it is possible to show improvement when p_0 is small as was done in Theorem 4.5.

5. Relaxing the measure preservability assumption

We illustrate how the measure preservability assumption on the side information mapping $s(x): X \to X$ can be relaxed. It turns out that in this case the analysis for both kinds of exponential bounds can be repeated with a smaller value of ε.

Assume that there exists a probability measure Q on X such that the side information mapping $s: X \to X$ is measure preserving with respect to Q. Assume further that P is absolutely continuous with respect to Q so that the Radon–Nikodym derivative $dP/dQ = f$ satisfies, for some constants c_1, c_2, $0 < c_1 \leqslant f \leqslant c_2$ for all $x \in X$. Thus Q is also absolutely continuous with respect to P and $dQ/dP = 1/f$.

Then

$$P(s^{-1}(A)) = \int_{s^{-1}(A)} f dQ \leqslant c_2 Q(A) \leqslant \frac{c_2}{c_1} P(A)$$

and similarly

$$P(s^{-1}(A)) = \int_{s^{-1}(A)} f dQ \geqslant c_1 Q(A) \geqslant \frac{c_1}{c_2} P(A).$$

Bounds for consistent algorithms. Assume that there exists a constant $\rho > 0$ such that $P(A) > \varepsilon$ implies $P(s^{-1}(A)) > \rho\varepsilon$, and then

$$\left\{ \sup_{\substack{A \in \mathscr{A} \\ \sum \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) = 0}} \left| \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_A(X_i) + \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_A(s(X_i)) - 2P(A) \right| > 2\varepsilon \right\}$$

is contained in

$$\left\{ \sup_{\substack{A\in\mathscr{A} \\ \sum \mathbf{1}_A(X_i)+\mathbf{1}_A(s(X_i))=0}} \left| \frac{1}{m}\sum\nolimits_{i=1}^m \mathbf{1}_A(X_i) + \frac{1}{m}\sum\nolimits_{i=1}^m \mathbf{1}_A(s(X_i)) - P(A) - P(s^{-1}(A)) \right| > \varepsilon + \rho\varepsilon \right\}.$$

Hence we can study

$$\Pr\left\{ \sup_{\substack{A\in\mathscr{A} \\ \sum Z_i^A=0}} \left| \frac{1}{m}\sum\nolimits_{i=1}^m Z_i^A - (P(A) + P(s^{-1}(A))) \right| > \varepsilon(1+\rho) \right\}$$

in place of

$$\Pr\left\{ \sup_{\substack{A\in\mathscr{A} \\ \sum Z_i^A=0}} \left| \frac{1}{m}\sum\nolimits_{i=1}^m Z_i^A - 2P(A) \right| > 2\varepsilon \right\},$$

and the analysis of $C_1 e^{-c_2 m\varepsilon}$ bounds can be repeated with $\frac{1+\rho}{2}\varepsilon$ in place of ε. This means that the improvement from side information is smaller if $s(x)$ is not measure preserving. For example, Theorem 3.1 translates to:

Theorem 5.1. *If $\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) \in \{0,1\}$ for all i and $s: X \to X$ is such that for some $\rho > 0$, for all $A \in \mathscr{A}$ $P(A) > \varepsilon$ implies $P(s^{-1}(A)) > \rho\varepsilon$ then*

$$\Pr\left\{ \sup_{\substack{A\in\mathscr{A} \\ \sum_{i=1}^m Z_i^A=0}} \left| \frac{1}{2m}\sum_{i=1}^m Z_i^A - P(A) \right| > \varepsilon \right\} \leqslant 2S(\mathscr{A}, 2m^2) e^{2\left(1+\frac{1+\rho}{2}\varepsilon\right)} e^{-m(1+\rho)\varepsilon}.$$

General convergence bounds. For general convergence bounds assume that there is $\rho < 2\varepsilon$ so that $P(A) \leqslant (1+\rho)P(s^{-1}(A))$ and $P(A) \geqslant (1-\rho)P(s^{-1}(A))$ for all $A \in \mathscr{A}$ (we can take $\rho = (c_2 - c_1)/c_1$).

We claim that the event

$$\left\{ \sup_{A\in\mathscr{A}} \left| \frac{1}{m}\sum\nolimits_{i=1}^m \mathbf{1}_A(X_i) + \frac{1}{m}\sum\nolimits_{i=1}^m \mathbf{1}_A(s(X_i)) - 2P(A) \right| > 2\varepsilon \right\}$$

is contained in

$$\left\{ \sup_{A\in\mathscr{A}} \left| \frac{1}{m}\sum\nolimits_{i=1}^m \mathbf{1}_A(X_i) + \frac{1}{m}\sum\nolimits_{i=1}^m \mathbf{1}_A(s(X_i)) - P(A) - P(s^{-1}(A)) \right| > 2(\varepsilon - \rho/2) \right\}.$$

Assume that $|\frac{1}{m}\sum \mathbf{1}_{A^*}(X_i) + \frac{1}{m}\sum \mathbf{1}_{A^*}(s(X_i)) - 2P(A^*)| > 2\varepsilon$. In this case $\frac{1}{m}\sum \mathbf{1}_{A^*}(X_i) + \frac{1}{m}\sum \mathbf{1}_{A^*}(s(X_i)) < 2P(A^*) - 2\varepsilon$ or $\frac{1}{m}\sum \mathbf{1}_{A^*}(X_i) + \frac{1}{m}\sum \mathbf{1}_{A^*}(s(X_i)) > 2P(A^*) + 2\varepsilon$. Using $P(A^*) \leqslant (1+\rho)P(s^{-1}(A^*))$

in the former and $P(A^*)\geqslant(1-\rho)P(s^{-1}(A^*))$ in the latter case, we get

$$\frac{1}{m}\sum\nolimits_{i=1}^{m}\mathbf{1}_{A^*}(X_i)+\frac{1}{m}\sum\nolimits_{i=1}^{m}\mathbf{1}_{A^*}(s(X_i))<P(A^*)+P(s^{-1}(A^*))-2\left(\varepsilon-\frac{\rho}{2}\right)$$

or

$$\frac{1}{m}\sum\nolimits_{i=1}^{m}\mathbf{1}_{A^*}(X_i)+\frac{1}{m}\sum\nolimits_{i=1}^{m}\mathbf{1}_{A^*}(s(X_i))>P(A^*)+P(s^{-1}(A^*))+2\left(\varepsilon-\frac{\rho}{2}\right).$$

Thus

$$\sup_{A\in\mathscr{A}}\left|\frac{1}{m}\sum\nolimits_{i=1}^{m}\mathbf{1}_{A}(X_i)+\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{A}(s(X_i))-P(A)-P(s^{-1}(A))\right|>2\left(\varepsilon-\frac{\rho}{2}\right).$$

That is, we can repeat the analysis of Section 4 with $\tilde{\varepsilon}=\varepsilon-\rho/2$ is place of ε. Then, for example, Theorem 4.4 translates to the following:

Theorem 5.2. *If $\mathbf{1}_A(X_i)+\mathbf{1}_A(s(X_i))\in\{0,1\}$ for all i and $s:X\to X$ is such that for some $\rho<2\varepsilon$, $(1-\rho)P(s^{-1}(A))\leqslant P(A)\leqslant(1+\rho)P(s^{-1}(A))$ for all $A\in\mathscr{A}$, then*

$$\Pr\left\{\sup_{A\in\mathscr{A}}\left|\frac{1}{2m}\sum\nolimits_{i=1}^{m}\mathbf{1}_{A}(X_i)+\mathbf{1}_{A}(s(X_i))-P(A)\right|\geqslant\varepsilon\right\}$$
$$\leqslant 4e^{16(\varepsilon-\rho/2)^2+16(\varepsilon-\rho/2)}S(\mathscr{A},2m^2)e^{-8m(\varepsilon-\rho/2)^2}.$$

6. Open problems and conclusions

In this paper we showed how to incorporate into the learning theoretic complexity bounds improvements due to side information arising from the classification of a function, called the side information map, of each new training sample. The essential approach was to use $\sum\mathbf{1}_A(x_i)+\mathbf{1}_A(s(x_i))$, where s is the side information mapping, to approximate $P(A)$.

The analysis of this approximation then led to conditions under which the exponential terms in the learning bounds are improved and by how much. The degree of improvement depends on the overlap between target concepts and their images under the side information map. The case of no overlap yields the strongest improvement. By comparing Theorem 3.1, in which the rate improves from $-m\varepsilon$ to $-2m\varepsilon$, to Theorem 4.4, in which the rate improves from $-2m\varepsilon^2$ to $-8m\varepsilon^2$, one sees that ε is effectively doubled when there is no overlap. On the other hand, without sufficient control on overlap, no improvement may occur. This is not caused by a deficiency of technique, because exact rates obtained earlier in [12] in the analysis of a concrete model problem of learning an interval on a circle when the antipodal point is classified as side information, show the loss of improvement. If there are several known side information maps the problem becomes considerably harder as the overlap phenomenon will be even more complex, and require more sophisticated overlap assumptions. However, in a simple toy problem[2] with two suitably taken

[2] In an interval problem take, say, $T=(0,1/6), s_1(x)=x+1/3 \bmod 1, s_2(x)=x+2/3 \bmod 1$, and use the smallest interval containing positive original samples or side information translates as an approximation of the target. Apply similar reasoning as in [12] to the intervals $T, s_1^{-1}(T)=s_2(T)$ and $s_2^{-1}(T)=s_1(T)$.

side information maps the best improvement from the standard $\exp(-m\varepsilon)$ bound is $\exp(-3m\varepsilon)$, suggesting that further side information will increase the effective size of ε yet further.

Our analysis considered a restricted class of side information maps. In particular, we assumed that the side information map was measure preserving or almost measure preserving. Inevitably such an assumption translates into an assumption on the underlying measure. It would be interesting to see how distribution dependent results, such as [20] and [19] compare.

In another direction, one can imagine side information of a very different type than that generated by a side information map. For other problems, and even for the problems considered here, one could consider instead refining analysis of the polynomially growing combinatorial term in Learning Theory convergence bounds. The idea would be to use side information to restrict the Vapnik–Chervonenkis (VC) dimension of the class. The dissertation [11] introduced a modified VC-dimension where side information acted as a complexity constraint. However, in examples considered it appeared that side information was not as restrictive as expected.

Incorporating side information into empirical VC-dimension analysis might be more promising. A recent paper [10] introduced improved sample complexities through a stopping time which is based on available data. One may ask if such an approach can also be applied to side information problems. A data dependent approach could also be applied to the problem of empirically determining a bound on the overlap probability p_2 so as to apply the error bounds when sufficient upper bounds on p_2 are not known a priori, because the data will restrict the set of possible concepts. Clearly, the possibility of doing this is tied very closely to the geometry of the underlying space X and the structure of the side information map, and requires a data dependent formulation of the rates problem.

Appendix A. Proof of Theorem 4.6

Proof. First for $p_1 > 0$, $p_2 > 0$, $\min\{\Lambda^*_{p_1,p_2}(p_1 + 2p_2 + 2\varepsilon), \Lambda^*_{p_1,p_2}(p_1 + 2p_2 - 2\varepsilon)\}$ is a continuous function on p_1 and p_2. We begin by studying the rate function as (p_1, p_2) approaches the corner $(0,0)$ or $(1,0)$.

First let $(p_1, p_2) \to (0,0)$. Then

$$\Lambda^*_{p_1,p_2}(p_1 + 2p_2 + 2\varepsilon) = \inf_{q_0,q_1,q_2} \left\{ q_0 \ln \frac{q_0}{p_0} + q_1 \ln \frac{q_1}{p_1} + q_2 \ln \frac{q_2}{p_2}; q_1 + 2q_2 = p_1 + 2p_2 + 2\varepsilon, \sum q_i = 1 \right\}$$

and $q_1 + 2q_2 \to 2\varepsilon$ as $(p_1, p_2) \to (0,0)$. There exists $r < r_0 = \frac{\varepsilon}{8} e^{-128/\varepsilon}$ such that $p_1^2 + p_2^2 < (2r)^2$ implies that $q_1 > \varepsilon/2$ or $q_2 > \varepsilon/4$. Then by using $x \ln(x/p) > x \ln x > -1/e$ for $p \in (0,1)$ and the fact that $q \ln(q/p) > 32$ when $q > \varepsilon/4$ and $p < 2r_0$, we have

$$q_0 \ln \frac{q_0}{p_0} + q_1 \ln \frac{q_1}{p_1} + q_2 \ln \frac{q_2}{p_2} > -\frac{2}{e} + 32.$$

Hence $\Lambda^*_{p_1,p_2}(p_1 + 2p_2 + 2\varepsilon) > 16$. Observe that as $(p_1, p_2) \to (0,0)$ we need not study $\Lambda^*_{p_1,p_2}(p_1 + 2p_2 - 2\varepsilon)$ as by definition $\Lambda^*_{p_1,p_2}(x) = \infty$ if $x \notin [0,2]$.

Similarly, as $(p_1, p_2) \to (1, 0)$, $q_1 + 2q_2 = p_1 + 2p_2 + 2\varepsilon \to 1 + 2\varepsilon$. Because $q_1 \leqslant 1$, there exists $r < r_0$ such that $(1 - p_1)^2 + p_2^2 < (2r)^2$ implies that $q_2 > \varepsilon/2$. Then again $q_0 \ln \frac{q_0}{p_0} + q_1 \ln \frac{q_1}{p_1} + q_2 \ln \frac{q_2}{p_2} > -\frac{2}{e} + 32$ and hence $\Lambda^*_{p_1,p_2}(p_1 + 2p_2 + 2\varepsilon) > 16$.

Finally, when $q_1 + 2q_2 = p_1 + p_2 - 2\varepsilon \to 1 - 2\varepsilon$ there exists $r < r_0$ such that $(1 - p_1)^2 + p_2^2 < (2r)^2$ implies that $q_1 < 1 - \varepsilon$. Hence either $q_0 > \varepsilon/2$ or $q_2 > \varepsilon/2$ and by repeating the above argument we conclude that $\Lambda^*_{p_1,p_2}(p_1 + 2p_2 - 2\varepsilon) > 16$.

Now we take neighborhoods $V_{(0,0)} = \{(p_1, p_2); p_1^2 + p_2^2 < (2r)^2, p_1 + p_2 \leqslant 1, p_1 > 0, p_2 > 0\}$ and $V_{(1,0)} = \{(p_1, p_2); (1 - p_1)^2 + p_2^2 < (2r)^2, p_1 + p_2 \leqslant 1, p_1 > 0, p_2 > 0\}$, where r is such that $(p_1, p_2) \in V_{(0,0)}$ implies that $\min\{\Lambda^*_{p_1,p_2}(p_1 + 2p_2 + 2\varepsilon), \Lambda^*_{p_1,p_2}(p_1 + 2p_2 - 2\varepsilon)\} > 16 \geqslant 8\varepsilon^2$, and $V_{(1,0)}$ similarly.

Fix p_1 such that $r \leqslant p_1 \leqslant 1 - r$ and assume that $p_0 > 0$ and $p_2 > 0$. For $x \leqslant 2$, $\Lambda^*_{p_1,p_2}(x) = F(x, \hat{q}(x))$, where

$$F(x, q) = (1 - x + q)\ln \frac{1 - x + q}{1 - p_1 - p_2} + (x - 2q)\ln \frac{x - 2q}{p_1} + q \ln \frac{q}{p_2} \qquad \text{(A.1)}$$

and $\hat{q}(x)$ is the unique solution in $(\max\{0, x - 1\}, x/2)$ of $\partial F/\partial q(x, \hat{q}) = 0$. $\partial F/\partial q(x, \hat{q}) = 0$ implies that

$$\frac{\hat{q}(x)(1 + \hat{q}(x) - x)}{(x - 2\hat{q}(x))^2} = \frac{p_0 p_2}{p_1^2}.$$

Fix $p_1 > 0$ and as $p_2 \to 0$ either $\hat{q}(x) \to 0$ or $1 + \hat{q}(x) - x \to 0$. The solution minimizing the relative entropy is $\hat{q}(x) \to 0$.

We claim that $\hat{q}(x) \to 0$ uniformly in p_1 provided that $p_1 \in [r, 1 - r]$. This can be seen by actually solving for $\hat{q}(x)$ by using the quadratic formula. Thus

$$\hat{q}(x) = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

where $p = p_0 p_2 / p_1^2$, $a = 1 - 4p$, $b = 1 - x + 4xp$ and $c = -px^2$ and the correct sign is taken to give the solution in $(\max\{0, x - 1\}, x/2)$. Now $2a \geqslant r^2$ if $p_2 \leqslant r^2/8$. For $p_1 \in [r, 1 - r]$, $p \leqslant p_2/r^2$ and as $p_2 \to 0$, $p \to 0$ uniformly in p_1 and hence also $|\hat{q}(x)| \to 0$ as $p \to 0$. Hence $\hat{q}(x) \to 0$ uniformly in p_1 as $p_2 \to 0$, provided that $p_1 \in [r, 1 - r]$.

Let $x = p_1 + 2p_2 + 2\varepsilon$. Then $\hat{q}(x) \to 0$, $1 + \hat{q}(x) - x \to 1 - (p_1 + 2\varepsilon)$ and $x - \hat{q}(x) \to p_1 + 2\varepsilon$ as $p_2 \to 0$. Thus

$$\begin{aligned} \Lambda^*_{p_1,p_2}(x) &= (1 - x + \hat{q}(x))\ln \frac{1 - x + \hat{q}(x)}{1 - p_1 - p_2} + (x - 2\hat{q}(x))\ln \frac{x - 2\hat{q}(x)}{p_1} + \hat{q}(x)\ln \frac{\hat{q}(x)}{p_2} \\ &\to (1 - (p_1 + 2\varepsilon))\ln \frac{1 - (p_1 + 2\varepsilon)}{p_0} + (p_1 + 2\varepsilon)\ln \frac{p_1 + 2\varepsilon}{p_1} = \Lambda^*_{p_2=0}(p_1 + 2\varepsilon) \end{aligned}$$

uniformly in p_1 and $\Lambda^*_{p_1,p_2}(p_1 + 2p_2 - 2\varepsilon)$ similarly.

We have previously shown that $\min\{\Lambda^*_{p_2=0}(p_1+2\varepsilon), \Lambda^*_{p_2=0}(p_1-2\varepsilon)\} \geqslant 8\varepsilon^2$ and thus by taking p_2 small enough

$$\min\{\Lambda^*_{p_1,p_2}(p_1+2\varepsilon), \Lambda^*_{p_1,p_2}(p_1-2\varepsilon)\} \geqslant (8-\lambda)\varepsilon^2.$$

If $p_1 = 0$,

$$\begin{aligned}\Lambda^*_{p_1=0}(x) &= \min\left\{(1-q)\ln\frac{1-q}{1-p_2} + q\ln\frac{q}{p_2}; 2q = x\right\}\\ &= \left(1-\frac{x}{2}\right)\ln\frac{1-\frac{x}{2}}{1-p_2} + \frac{x}{2}\ln\frac{\frac{x}{2}}{p_2}\end{aligned}$$

and

$$\Lambda^*_{p_1=0}(2p_2+2\varepsilon) = (1-(p_2+\varepsilon))\ln\frac{1-(p_2+\varepsilon)}{1-p_2} + (p_2+\varepsilon)\ln\frac{p_2+\varepsilon}{p_2} \to \infty$$

as $p_2 \to 0$ establishing the result for $p_1 = 0$. □

References

[1] Y. Abu-Mostafa, Hints and the VC dimension, Neural Comput. 5 (1993) 278–288.

[2] Y. Abu-Mostafa, Learning from hints, J. Complexity 10 (1994) 165–178.

[3] Y. Abu-Mostafa, Hints, Neural Comput. 7 (1995) 639–671.

[4] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of the 11th Annual Conference on Computational Learning Theory, 1998, pp. 92–100.

[5] M. Campi, P. Kumar, Leaning dynamical systems in a stationary environment, Syst. Control Lett. 34 (3) (1998) 125–132.

[6] C. Dellacherie, P.-A. Meyer, Probabilities and Potential, North-Holland, Amsterdam, 1978.

[7] A. Dembo, O. Zeitouni, Large Deviations Techniques and Applications, Applications of Mathematics, Vol. 38, 2nd Edition, Springer, Berlin, 1998.

[8] L. Devroye, Bounds for the uniform deviation of empirical measures, J. Multivariate Anal. 12 (1982) 72–79.

[9] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc. 58 (1963) 13–30.

[10] V. Koltchinskii, C. Abdullah, M. Ariola, P. Dorato, D. Panchenko, Improved sample complexity estimates for statistical learning control of uncertain systems, IEEE Trans. Automat. Control 45 (12) (2000) 2383–2388.

[11] P. Kuusela, Learning theory techniques in control theory, Ph.D. Thesis, Department of Mathematics, Rutgers, the State University of New Jersey, available from www.math.rutgers.edu/kuusela 1999.

[12] P. Kuusela, D. Ocone, Learning with side information: an example, 2001, submitted for publication.

[13] R. Meir, Performance bounds for nonlinear time series prediction, in: Proceedings of the Tenth Annual Conference on Computational Learning Theory, Nashville, Tennessee, 1997.

[14] D. Pollard, Convergence of Stochastic Processes, Springer Series in Statistics, Springer, Berlin, 1984.

[15] J. Ratsaby, V. Maiorov, On the value of partial information for learning from examples, J. Complexity 13 (4) (1997) 509–544.

[16] J. Shawe-Taylor, M. Anthony, N.L. Biggs, Bounding sample size with the Vapnik–Chervonenkis dimension, Discrete Appl. Math. 42 (1993) 65–73.

[17] V. Vapnik, A. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, J. Probab. Appl. 16 (1971) 264–280.

[18] N. Vayatis, Inégalités de Vapnik-Chervonenkis et measures de complexité, Ph.D. Thesis, Ecole Polytechnique (in English), available from www./www.cmla.ens-cachan.fr/Utilisateurs/vayatis/ 2000.

[19] N. Vayatis, The role of critical sets in Vapnik–Chervonenkis theory, in: Proceedings of COLT, Palo Alto, California, 2000, pp. 75–80.

[20] N. Vayatis, R. Azencott, Distribution dependent Vapnik–Chervonenkis bounds, Computational Learning Theory, Lecture Notes in Artificial Intelligence, Vol. 1572, 1999, pp. 230–240.

[21] M. Vidyasagar, A Theory of Learning and Generalization; With Applications to Neural Networks and Control Systems, Springer, Berlin, 1997.