Learning Dynamical Systems in a Stationary Environment*†

M. C. Campi
Dept. of Electrical Engineering and Automation - University of Brescia
Via Branze 38, 25123 Brescia, Italy

and

P. R. Kumar
Dept. of Electrical and Computer Engineering, and the
Coordinated Science Laboratory
University of Illinois
1308 West Main Street
Urbana, IL 61801,USA

Abstract

We consider the problem of learning the input-output relation of a dynamical system from noisy data. Our method rests on the use of a smooth simultaneous estimator which generalizes the standard empirical estimator. In a stationary environment, our algorithm is shown to select a model which exhibits the Probably Approximately Correct (PAC) property under very mild conditions.

This contribution should be thought of as a first attempt to extend concepts developed in learning theory to the field of system identification where, due to the presence of the system dynamics, the typical i.i.d. assumption on the data made in learning theory is not satisfied.

*Please address all correspondence to the second author.

†The first author would like to acknowledge the financial support of MURST under the 60% project "Adaptive identification, prediction and control" and the 40% project "Model identification, system control and signal processing". The research of the second author has been supported by the U.S. Army Research Office Contract No. DAAH-04-95-1-0090, and the Joint Service Electronics Program under Contract No. N00014-96-1-0129. This work was conducted while M.C. Campi was visiting the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign in Spring 1995.

1 Introduction

Learning theory, as it has developed over the past fifteen years, has been by and large focused on mathematical objects of static structure. For example, in the problem treated by Valiant [1] the goal is to learn a binary valued function $c(x)$ defined on an arbitrary set X, given data which are labeled sample points of the form $(x_t, c(x_t))$, where the x_t's are drawn independently of one another according to an unknown probability distribution P on X. The framework of learning binary valued functions has been thoroughly investigated at least in the two extreme cases in which P is either a fixed known probability (called fixed distribution learning) or it can be any probability on X (called distribution free learning).

Extensions to more general learning settings started to appear in the late 1980's, e.g., [2] -[9]. In particular, in [9] Haussler introduced a learning framework where each datum consists of a point x_t and an outcome y_t belonging to an arbitrary space Y called the outcome space. These data are generated according to an unknown distribution on $X \times Y$. The statistical fluctuation of y for a given x describes the noise and, possibly, extraneous variables affecting the data generation mechanism. This point of view is also adopted in the present contribution.

A common feature of all the above referenced papers is that the points x_t are drawn *independently* of one another. This rules out the possibility of any memory and thus dynamics in the data generation mechanism, and therefore the ability to model the problem of learning dynamic systems, i.e., system identification, within this framework.

In the learning literature, much work has been devoted to the establishment of conditions under which the so called empirical estimator estimates simultaneously over the considered hypothesis class. This property, however, is not necessary for learnability, since one may use estimators other than the empirical estimator. In [5, 6], among other contributions, Buescher and Kumar introduce a new canonical estimator and show that it works even in cases in which the empirical estimator does not.

In this paper our main goal is to study the problem of identifying dynamical systems in the context of a learning theoretical framework. We do so by extending some of the results of [5, 6] from an i.i.d. setting to a stationary environment. Allowing for a stationary setting is important in that it allows one to venture beyond the memoryless case and to enter the world of dynamical systems. Our contribution should be thought of as an attempt to bridge the existing gap between learning theory and the theory of identification of dynamical systems. For other approaches to addressing dependency we refer the reader to [7] and [8].

2 Problem Description

Consider a single input, single output system with input $u_t \in U$ and output $y_t \in Y$, where U and Y are totally bounded subsets of R.

We assume that y_t is conditionally independent of $u_{t+1}^{\infty}, y_{-\infty}^{t-1}$ and y_{t+1}^{∞} given $u_{-\infty}^{t}$

(throughout, $u_r^s := (u_r, u_{r+1}, \ldots, u_s)$ and $y_r^s := (y_r, y_{r+1}, \ldots, y_s)$). This condition serves two purposes. First, it ensures that the system is causal. Second, it also implies that all the correlation in the y process is generated by the input process u.

We assume that the system is in a stationary environment, i.e, that the joint process (u_t, y_t) is strict-sense stationary. Moreover, for convenience, we also assume that it is ergodic (i.e. its invariant σ-algebra is trivial).

Then, we see that the system is completely described by the (unknown) time-invariant conditional distribution of y_t given $u_{-\infty}^t$, henceforth referred to as $P_{y/u}$.

Alternatively, letting $s(u_{-\infty}^t) := \int_Y y P_{y/u}(dy, u_{-\infty}^t)$, the system can also be supposed to be in the form $y_t = s(u_{-\infty}^t) + d_t$, where $d_t := y_t - s(u_{-\infty}^t)$ is conditionally white given the past, i.e., $E[d_t / u_{-\infty}^t, y_{-\infty}^{t-1}] = 0$. The stochastic process d_t can be regarded as an additive disturbance affecting the output y_t.

Once the probability distribution of $\{u_t\}_{t=-\infty}^{+\infty}$ has been specified, $P_{y/u}$ completely defines a probability measure P in the space $U^\infty \times Y^\infty$ of doubly infinite sequences $(u_{-\infty}^{+\infty}, y_{-\infty}^{+\infty})$. We assume that P is unknown but it belongs to some prespecified set $\mathcal{P}$.

The learning problem considered here is to determine a suitable q-dimensional approximation of the function s, where q is an integer that will be regarded as predefined and fixed throughout. The approximation function h is selected from a hypothesis set $\mathcal{H}$ of functions from U^q to Y.

Given an estimate (or model) h, one can measure its accuracy by the error criterion given below.

Definition 1 *(Error between P and h).*

The error between P and h is defined as,

$$err(P,h) := \lim_{t\to\infty} \frac{1}{t-q+1} \sum_{i=q}^{t} \left(y_i - h(u_{i-q+1}^i)\right)^2 = E_P\left[\left(y_t - h(u_{t-q+1}^t)\right)^2\right]. \qquad \square$$

Note that the last equality in Definition 1 is a consequence of the ergodicity assumption.

Clearly, $err(P,h)$ is the expected error we make by using model h to predict y_t; hence it is sometimes referred to as the *generalization error*.

Definition 2 *(Optimal error).*

The optimal error is the minimum error over $h \in \mathcal{H}$,

$opt(P,\mathcal{H}) := \inf_{h\in\mathcal{H}} err(P,h).$ $\square$

Suppose that at time t data (u_1, y_1), ..., (u_t, y_t) have been collected. From this data, one wishes to construct an estimate h. This is done by an *algorithm*, which is an indexed family of maps $a_t : (U \times Y)^t \to \mathcal{H}$; where t is the index.

Our goal is to construct an algorithm which approaches the optimal error asymptotically as follows:

Definition 3 *(Nonuniformly learning algorithm)*

An estimator a_t is said to learn (possibly) nonuniformly over $(\mathcal{P}, \mathcal{H})$ if for all $\varepsilon > 0$,

$\lim_{t\to\infty} P\left\{err(P, a_t(u_1^t, y_1^t)) - opt(P,\mathcal{H}) > \epsilon\right\} = 0, \forall P \in \mathcal{P}.$ $\square$

The reason for the usage of the qualifier "nonuniform" is that, as compared to other notions of learnability, in Definition 3 the convergence is not required to take place uniformly in $P \in \mathcal{P}$.

Such an algorithm may or may not exist, and so we introduce the following definition.

Definition 4 *(Nonuniform learnability)*

$(\mathcal{P}, \mathcal{H})$ is (nonuniformly) learnable if there exists an algorithm a_t which learns nonuniformly. □

Remark 1

In the definition of $err(P,h)$, the second equality follows from the assumption that process (u_t, y_t) is ergodic. In the nonergodic case, we would instead have

$$err(P,h) \;:=\; \lim_{t\to\infty} \frac{1}{t-q+1} \sum_{i=q}^{t} \left(y_i - h(u_{i-q+1}^{i})\right)^2 \;=\; E_P\left[\left(y_t - h(u_{t-q+1}^{t})\right)^2 / \mathcal{J}\right],$$

where $\mathcal{J}$ is the invariant σ-algebra of process (u_t, y_t). As a consequence, $err(P,h)$ is itself a random variable.

When (u_t, y_t) is not ergodic, one should be particularly careful in the definition of optimal error for the reason explained below. Since an algorithm selects an hypothesis on the basis of the realization at hand, in the nonergodic case, it is natural to compare in Definition 3 the quantity $err(P, a_t(u_1^t, y_1^t))$ with the optimal error obtained by minimizing $err(P,h)$, when h is a random element measurable with respect to the invariant σ-algebra $\mathcal{J}$. Correspondingly , the optimal value in Definition 2 should be defined as $opt(P,\mathcal{H}) = \inf err(P,h)$, where inf is taken over the set of $\mathcal{J}$-measurable elements $h \in \mathcal{H}$.

The optimal error obtained by allowing h to be random is in general strictly lower than minimizing over deterministic hypotheses. A simple example suffices to make this point clear. Suppose that $U = Y = \{0,1\}$ and that $y_t = u_t, \forall t$. Moreover, assume that with probability 0.5 the input sequence is $0,0,0,\ldots$ and, with probability 0.5, it is $1,1,1,\ldots$. Finally, let $q = 1$ and consider $\mathcal{H}$ = set of constant functions from U to Y (clearly, $\mathcal{H}$ has only two elements, $h_0 \equiv 0$, $h_1 \equiv 1$). Clearly, if h is constrained to be deterministic, selecting h_0 results in a nonzero error when $u_t = 1,1,\ldots$ and, similarly, selecting h_1 leads to a nonzero error when $u_t = 0,0,\ldots$. On the other hand, if h is allowed to be stochastic, we can take h_0 in the invariant set where $y_t = u_t = 0, \forall t$ and h_1 in its complement. The corresponding error is obviously zero.

In connection with the discussion above, we also note that, given a certain dynamical system, it is possible that a certain hypothesis class is too restricted when the system operates in an ergodic environment and yet it can be rich enough to describe one by one the input-output relations corresponding to different invariant sets in a nonergodic environment. Therefore, we see that learning in a nonergodic framework may be easier than learning in an ergodic setting.

In the present contribution, our attention is focused on the ergodic setting only for notational convenience, no particular difficulty being involved in the extension of our results to the nonergodic case. □

The above learnability definition is amenable to the following obvious interpretation. Set $f(u_{t-q+1}^t) := E_P\left[y_t/u_{t-q+1}^t\right]$ and define $\xi_t := y_t - f(u_{t-q+1}^t)$. Then, the system can be written as $y_t = f(u_{t-q+1}^t) + \xi_t$. If $f \in \mathcal{H}$, learnability implies the existence of an algorithm a_t such that the L^2-norm of the difference between $a_t(u_1^t, y_1^t)$ and f tends to zero in probability. Precisely, for all $\varepsilon > 0$,

$$P\left\{E_P\left[\left(a_t\left(u_1^t, y_1^t\right) - f\right)^2\right] > \varepsilon\right\} \to 0,\ t \to \infty.$$

When $f \notin \mathcal{H}$, $a_t(u_1^t, y_1^t)$ provides an approximation of f whose L^2-error tends in probability to the best one in the class $\mathcal{H}$.

A key point for the design of a successful learning algorithm is the possibility of estimating certain expected values from data sequences. In a stationary environment, this calls for some mixing condition on the involved processes.

Assumption 1

For any positive bounded function $\xi : U^q \times Y \to R$, it holds that

$$P\{|\frac{1}{t-n-q+1}\sum_{i=n+q}^{t}\left(\xi\left(u_{i-q+1}^i, y_i\right) - E_P\left[\xi\left(u_{t-q+1}^t, y_t\right)\right]\right)| > \varepsilon / u_{-\infty}^n, y_{-\infty}^n\}$$

$$\leq \zeta\left(\varepsilon, t-n\right),\ \ \forall P \in \mathcal{P}, \tag{1}$$

where $\zeta(\cdot,\cdot)$ is a real function such that $\zeta\left(\varepsilon, t-n\right) \to 0$, as $(t-n) \to \infty$. □

Remark 2

Some condition on the tail of the probability distribution of process $\xi\left(u_{i-q+1}^i, y_i\right)$ is needed to cope with large deviation problems. The constraint assumed in Assumption 1 that ξ is deterministically bounded is particularly strong, but it is met in our context. Clearly, even though not explicitly indicated, function ζ will depend on the bound on function ξ. □

Remark 3

Suitable expressions for $\zeta(\varepsilon, t-n)$ in Assumption 1 can be derived under standard ψ-mixing conditions (see e.g. [10]).

We recall that the ψ-coefficient of two σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ is defined as

$$\psi(\mathcal{F}_1, \mathcal{F}_2) := \sup_{A \in \mathcal{F}_1, B \in \mathcal{F}_2, P(A) \neq 0, P(B) \neq 0} \left|\frac{P(A \cap B)}{P(A)P(B)} - 1\right|,$$

and that a process α_t is ψ-mixing if $\sum_{n=0}^{\infty} \psi_n < \infty$, where $\psi_n := \sup_t \psi\left(\sigma\left(\alpha_{-\infty}^t\right), \sigma\left(\alpha_{t+n}^{\infty}\right)\right)$ ($\sigma(\alpha_i^j)$ is the σ-algebra generated by $\alpha_i, \alpha_{i+1}, \ldots, \alpha_j$). The sum $\sum_{n=0}^{\infty} \psi_n$ is called the ψ-dependence index of process α_t.

Now, suppose that process (u_t, y_t) is ψ-mixing and that c is an upper bound for the corresponding ψ-dependence index, $\forall P \in \mathcal{P}$. Letting $\eta_t := \xi\left(u_{t-q+1}^t, y_t\right) -$

$E_P\left[\xi\left(u_{t-q+1}^t, y_t\right)\right]$ we want to show that with $\lambda =$ upper bound for ξ,

$$P\{|\frac{1}{t-n-q+1}\sum_{i=n+q}^{t}\eta_i| > \varepsilon/u_{-\infty}^n, y_{-\infty}^n\} \leq \frac{2\lambda^2(c+q-1)}{\varepsilon^2(t-n-q+1)}, \tag{2}$$

that is,

$$\zeta(\varepsilon, t-n) = 2\lambda^2(c+q-1)/\varepsilon^2(t-n-q+1)$$

in this case. To prove (2), recall first that, [11], if β_1 and β_2 are two summable random variables measurable with respect to $\mathcal{F}_1$ and $\mathcal{F}_2$ respectively, it holds that

$$|E_P[\beta_1\beta_2] - E_P[\beta_1]E_P[\beta_2]| \leq \psi(\mathcal{F}_1, \mathcal{F}_2)E_P[|\beta_1|]E_P[|\beta_2|]. \tag{3}$$

Consider any set $A \in \sigma\left(u_{-\infty}^n, y_{-\infty}^n\right)$ and apply result (3) with $\beta_1 = I_A\eta_i$ and $\beta_2 = \eta_j$ where i and j are such that $i \geq n+q$, $j \geq i+q-1$. Then,

$$\begin{aligned} E_P[I_A\eta_i\eta_j] &\leq \psi_{j-q+1-i}E_P[|I_A\eta_i|]E_P[|\eta_j|] \\ &\leq \psi_{j-q+1-i}\lambda^2 P(A). \end{aligned}$$

From this, one immediately obtains the estimate

$$E_P[\eta_i\eta_j/\sigma\left(u_{-\infty}^n, y_{-\infty}^n\right)] \leq \psi_{j-q+1-i}\lambda^2.$$

We now use this estimate to bound the conditional variance of $\sum_{i=n+q}^t \eta_i/(t-n-q+1)$ as follows

$$\begin{aligned} &E_P[(\frac{1}{t-n-q+1}\sum_{i=n+q}^{t}\eta_i)^2/\sigma\left(u_{-\infty}^n, y_{-\infty}^n\right)] \\ &\quad\leq \frac{2}{(t-n-q+1)^2}\sum_{\{i,j:n+q\leq i,j\leq t;j\geq i\}} E_P[\eta_i\eta_j/\sigma\left(u_{-\infty}^n, y_{-\infty}^n\right)] \\ &\quad\leq \frac{2}{(t-n-q+1)^2}\{\sum_{\{i,j:n+q\leq i,j\leq t;j\geq i+q-1\}}\psi_{j-q+1-i}\lambda^2 \\ &\qquad\qquad + \sum_{\{i,j:n+q\leq i,j\leq t;i\leq j<i+q-1\}}\lambda^2\} \\ &\quad\leq \frac{2\lambda^2(c+q-1)}{t-n-q+1}. \end{aligned}$$

Expression (2) now easily follows by using Chebyshev's inequality.

Tighter expressions for the rate of convergence of $\zeta(\varepsilon, t-n)$ could also be obtained by resorting to inequalities of the Hoeffding type. □

3 Learning by simultaneous estimation

A very natural and common, even though particular, way of learning a hypothesis consists in first estimating the error associated with each hypothesis h in $\mathcal{H}$, and then secondly selecting the hypothesis with the minimal estimated error. For the first step, a widely used error estimate is the so called *empirical error estimate* which is given by

$$e_{emp}\left(u_1^t, y_1^t; h\right) := \frac{1}{t-q+1}\sum_{i=q}^{t}\left(y_i - h\left(u_{i-q+1}^i\right)\right)^2$$

(see e.g. [9]). If the empirical error estimate gets close to $err(P,h)$ *simultaneously over* $\mathcal{H}$ (that is uniformly over all h in $\mathcal{H}$) as $t \to \infty$, then the second step in the above procedure leads to selecting a hypothesis whose generalization error is in fact small. This approach has stimulated a vast literature on the uniform convergence of empirical estimates of the error (e.g. [9, 12, 13, 14]), whose origins are in the pioneering work of Vapnik and Chervonenkis ([15, 16]).

The present contribution generalizes the above procedure in two respects.

i) We allow any *smooth simultaneous estimator for the error*, rather than insisting on using just the empirical error estimate.

Roughly, a simultaneous error estimator is smooth if it provides similar error estimates for hypotheses which almost agree on the sample input at hand ($h\left(u_{i-q+1}^i\right) \simeq h'\left(u_{i-q+1}^i\right)$, $i = q, q+1, \ldots, t$). (This is made precise in Definition 6 below). This smoothness condition is very natural indeed and rules out only pathological situations. It turns out that the empirical error estimator is in fact smooth (and, therefore, simultaneous error estimability through the empirical error estimator implies the existence of a smooth simultaneous error estimator, and thus implies smooth simultaneous error estimability of the model), but there are many cases in which a smooth simultaneous error estimator exists and yet the empirical estimator fails to simultaneously estimate (see Remark 4 below for an example).

We introduce a learning scheme which works whenever smooth simultaneous estimation is possible. The basic idea is as follows. First, a suitable *finite empirical cover* for $\mathcal{H}$, i.e. a cover based on the *empirical distance* $\rho_{u_1^t}(h,h') := \frac{1}{t-q+1}\sum_{i=q}^{t}\left|h\left(u_{i-q+1}^i\right) - h'\left(u_{i-q+1}^i\right)\right|$ is constructed. Its main feature is that its size (i.e., the number of elements in the cover) is tailored to the characteristics of the involved processes and to the number of available data points, so that a simultaneously accurate estimate of the generalization error of all the cover elements is possible. A key point is that, by employing a smooth simultaneous error estimator, the estimated generalization error of a cover element is a good example of the *true* generalization error of all the hypotheses in a neighborhood of the cover element (even though it can be a bad example of their *empirical* error!). The estimated hypothesis is then selected by minimizing the estimated error over the empirical cover. (This will be detailed further below).

ii) Learning is performed over a *nested family of hypothesis classes.*

We consider the case in which $\mathcal{H}$ has the substructure $\mathcal{H} = \bigcup_k \mathcal{H}^k$, $\mathcal{H}^k \subseteq \mathcal{H}^{k+1}$, and we try to learn over $\mathcal{H}$ by learning as time goes on over progressively increasing classes $\mathcal{H}^k$. Using such nested classes helps avoid overfitting problems by preventing the selection of hypotheses that agree too well with the noisy data. A crucial technical point is that the empirical cover for $\mathcal{H}$ is constructed in such a way that it always contains empirical covers for $\mathcal{H}^k$, $\forall k$, formed solely by elements of $\mathcal{H}^k$. In this way, the *true* generalization error of each hypothesis in $\mathcal{H}^k$ is close to the *empirical* error of an element of the cover if each pair $(\mathcal{P}, \mathcal{H}^k)$ is smoothly simultaneously estimable. Ultimately, this permits us to prove learnability under the mild condition that each single pair $(\mathcal{P}, \mathcal{H}^k)$ is smoothly simultaneously estimable. □

We now define precisely the notions of simultaneous nonuniform error estimation and smoothness.

Definition 5 *(Simultaneous nonuniform error estimability)*

$(\mathcal{P}, \mathcal{H})$ is simultaneously (nonuniformly) error estimable if there exists an error estimator $\{e_t\}$ (i.e. an indexed family of maps from $(U \times Y)^t \times \mathcal{H}$ to R) such that for all $\varepsilon > 0$,

$$\lim_{t\to\infty} P\{\sup_{h\in\mathcal{H}} \left| err\,(P,h) - e_t\left(u_1^t, y_1^t; h\right)\right| > \varepsilon\} = 0, \forall P \in \mathcal{P}. \qquad \square$$

In Definition 5 and throughout where $\forall P \in \mathcal{P}$ follows a statement of convergence, it is intended that the convergence rate may depend upon $P \in \mathcal{P}$.

Definition 6 *(Smooth estimators)*

The error estimator e_t is smooth if $\forall \vartheta > 0$, $\exists \sigma_t(\vartheta) > 0$ ($\sigma_t(\vartheta)$ may depend on $P \in \mathcal{P}$) such that

$$\lim_{t\to\infty} P\{\sup_{h,h'\in\mathcal{H}\ s.t.\ \rho_{u_1^t}(h,h')<\sigma_t(\vartheta)} \left|e_t\left(u_1^t, y_1^t; h\right) - e_t\left(u_1^t, y_1^t; h'\right)\right| > \vartheta\} = 0, \forall P \in \mathcal{P}. \qquad \square$$

Remark 4

It is easily seen that the empirical error estimator is smooth. On the other hand, smooth simultaneous error estimability is more general than insisting that the empirical error estimator be simultaneously good, as shown by the following simple, and somehow pathological, example.

Let $U = [0,1]$ and $Y = \{0,1\}$. Pairs (u_t, y_t) are drawn independently and such that u_t is uniformly distributed in $[0,1]$ and $y_t = 1$, $\forall u_t$. Finally, take $\mathcal{H}$ as the collection of all the indicator functions I_A, where A is a finite subset of $[0,1]$. Given any data sample (u_1^t, y_1^t) there always exists an hypothesis $\bar{h} \in \mathcal{H}$ that agrees with data, that is $e_{emp}\,(u_1^t, y_1^t; \bar{h}) = 0$. On the other hand, $err(P,h) = 1$, $\forall h$. Therefore $P\left\{\sup_{h\in\mathcal{H}} \left|err\,(P,h) - e_{emp}\,(u_1^t, y_1^t; h)\right| > \varepsilon\right\} = 1$, whenever $\varepsilon < 1$, and the empirical estimator fails to simultaneously estimate the error well for all h in $\mathcal{H}$. On the other hand, $e_t\,(u_1^t, y_1^t; h) = 1$, $\forall\,(u_1^t, y_1^t)$, $\forall h$, is clearly a smooth simultaneous error estimator in the present situation. □

Our goal is to construct an algorithm able to learn $(\mathcal{P}, \mathcal{H})$ whenever a smooth simultaneous error estimator exists for each pair $(\mathcal{P}, \mathcal{H}^k), k = 1, 2, \ldots$. With this objective in mind, we start by introducing the notion of an *empirical cover*.

Given an input sample u_1^n, the associated empirical distance $\rho_{u_1^n}$ is a pseudo-metric on $\mathcal{H}$. A set $\mathcal{H}_{n,u} \subset \mathcal{H}$ is an *empirical ε-cover* based on u_1^n if, for each $h \in \mathcal{H}$, there exists $\bar{h} \in \mathcal{H}_{n,u}$ such that $\rho_{u_1^n}(h, \bar{h}) < \varepsilon$. Associated with an empirical ε-cover, there is a mapping $m_{n,u} : \mathcal{H} \to \mathcal{H}_{n,u}$ such that $\rho_{u_1^n}(h, m_{n,u}(h)) < \varepsilon$, $\forall h \in \mathcal{H}$. Consider now the situation in which $\mathcal{H}$ is given as the union of nested classes: $\mathcal{H} = \bigcup_k \mathcal{H}^k$, $\mathcal{H}^k \subseteq \mathcal{H}^{k+1}$. In this case, an important concept is that of simple empirical ε-cover: an empirical ε-cover is *simple* if the associated mapping $m_{n,u}$ is such that $m_{n,u}(h) \in \mathcal{H}^k$, for any $h \in \mathcal{H}^k$, $\forall k$. Roughly, "simple" in this context means that $m_{n,u}(h)$ is not allowed to be too complex with respect to h. This will play a crucial role in our learning procedure.

We now show that a simple empirical ε-cover for $\mathcal{H}$ of finite cardinality always exists. The proof (borrowed from [6]) is constructive and, therefore, provides a way to build such a cover. Note first that, $Y \subset R$ being totally bounded, for each $\varepsilon > 0$ there exists an $\varepsilon/2$-cover $Y_{\varepsilon/2}$ of Y of finite cardinality $N(\varepsilon/2)$. There are $N(\varepsilon/2)^{n-q+1}$ possible mappings M_j from $i \in \{q, q+1, \ldots, n\}$ to cover elements $\left(M_j(i) \in Y_{\varepsilon/2},\ i = q, q+1, \ldots, n\right)$. The simple empirical ε-cover for $\mathcal{H}$ is recursively constructed as follows. Initially, let $\mathcal{C}_0 = 0$. For $j = 1$ to $N(\varepsilon/2)^{n-q+1}$, check whether there exist hypotheses $h \in \mathcal{H}$ such that

$$\left| h\left(u_{i-q+1}^i\right) - M_j(i) \right| < \varepsilon/2,\ i \in \{q, q+1, \ldots, n\}. \tag{4}$$

If any, pick an hypothesis h_j satisfying (4) which is as simple as possible ($h_j \in \mathcal{H}^j$ and there exists no $h \in \mathcal{H}^r$, $r < j$, such that (4) is satisfied) and set $\mathcal{C}_j = \mathcal{C}_{j-1} \cup \{h_j\}$. It is easy to verify that, for each $h \in \mathcal{H}^k$, there exists an element $h' \in \mathcal{C}_{N(\varepsilon/2)^{n-q+1}} \cap \mathcal{H}^k$ such that $\rho_{u_1^n}(h, h') < \varepsilon$. Set $\mathcal{H}_{n,u} := \mathcal{C}_{N(\varepsilon/2)^{n-q+1}}$ and $m_{n,u}(h) = h'$.

The learning algorithm

Fix three sequences of real numbers $\tau_n \downarrow 0$, $\mu_n \downarrow 0$, $\nu_n \downarrow 0$.

At time t do the following:

Let n be the largest integer such that

$$N\left(\frac{\tau_n}{2(n-q+1)}\right)^{n-q+1} \zeta\left(\mu_n, t-n\right) < \nu_n.$$

Construct a simple empirical $\left(\frac{\tau_n}{n-q+1}\right)$-cover $\mathcal{H}_{n,u}$ based on the pseudo-metric $\rho_{u_1^n}$ with cardinality less than or equal to $N(\frac{\tau_n}{2(n-q+1)})^{n-q+1}$ and denote by $m_{n,u}$ the corresponding mapping from $\mathcal{H}$ to $\mathcal{H}_{n,u}$.

Select $\bar{a}_t(u_1^t, y_1^t) = arg \min_{h \in \mathcal{H}_{n,u}} e_{emp}(u_{n+1}^t, y_{n+1}^t; h)$. □

The following theorem highlights the effectiveness of the proposed algorithm.

Theorem 1 *If $(\mathcal{P}, \mathcal{H}^k)$ is smoothly simultaneously nonuniformly error estimable for every k, then $(\mathcal{P}, \mathcal{H})$ is nonuniformly learnable through algorithm $\bar{a}_t$.*

Remark 5
Theorem 1 only guarantees that algorithm $\bar{a}_t$ *nonuniformly* learns $(\mathcal{P}, \mathcal{H})$. On the other hand, the assumption that pairs $(\mathcal{P}, \mathcal{H}^k)$ are smoothly simultaneously error estimable is very mild indeed. As a matter of fact, it is not hard to find examples in which $\mathcal{H}$ is a complex class (even with infinite VC-dimension) and yet a nested family $\mathcal{H}^k$ exists such that $(\mathcal{P}, \mathcal{H}^k)$ are smoothly simultaneously nonuniformly error estimable.

We also note that, by a slight modification of the proof of Theorem 1, it is possible to show that *uniform* (i.e. with convergence taking place uniformly in P in Definition 4) learnability holds for $(\mathcal{P}, \mathcal{H})$ provided that $(\mathcal{P}, \mathcal{H})$ is itself smoothly simultaneously uniformly (i.e. with convergence uniform in P in Definitions 5 and 6) error estimable. Details of this extension are left to the reader. □

Proof
Considering that a random hypothesis $h \in \mathcal{H}_{n,u}$ depends on data up to time n, we have

$$\begin{aligned} P\{|e_{emp}\left(u_{n+1}^t, y_{n+1}^t; h\right) &- err\,(P,h)| > \mu_n\} \\ &= E_P[P\{|\frac{1}{t-n-q+1}\sum_{i=n+q}^{t}((y_i - h(u_{i-q+1}^i))^2 \\ &\quad -E_P[(y_t - h(u_{t-q+1}^t))^2])| > \mu_n / u_{-\infty}^n, y_{-\infty}^n\})] \\ &\leq \zeta\,(\mu_n, t-n)\,,\ \forall h \in \mathcal{H}_{n,u},\ \forall P \in \mathcal{P}, \end{aligned}$$

where the last inequality follows from Assumption 1.

Since $|\mathcal{H}_{n,u}| \leq N(\frac{\tau_n}{2(n-q+1)})^{n-q+1}$, this equation immediately gives

$$P\{\max_{h \in \mathcal{H}_{n,u}} |e_{emp}(u_{n+1}^t, y_{n+1}^t; h) - err(P,h)| > \mu_n\} < \nu_n, \forall P \in \mathcal{P}. \tag{5}$$

Next, we show that $err\,(P,h)$ is close to $err\,(P,h')$ with probability approaching 1 as $t \to \infty$ when h and h' range over $\mathcal{H}^k$ and are close to each other with respect to the empirical distance, namely, for all $\varepsilon > 0$,

$$\lim_{t\uparrow\infty, \tau\downarrow 0} P\{\sup_{h,h' \in \mathcal{H}^k\ s.t.\ \rho_{u_1^t}(h,h') < \frac{\tau}{t-q+1}} |err\,(P,h)\ - err\,(P,h')| > \varepsilon\} = 0,\ \forall P \in \mathcal{P},\ \forall k. \tag{6}$$

(One should note that, condition $\rho_{u_1^t}(h,h') < \tau/(t-q+1)$ being equivalent to $\sum_{i=q}^{t} \left|h\left(u_{i-q+1}^i\right) - h'\left(u_{i-q+1}^i\right)\right| < \tau$, this limit will be the same, no matter how $t \uparrow \infty$ and $\tau \downarrow 0$).

To prove (6), note first that, $(\mathcal{P}, \mathcal{H}^k)$ being smoothly simultaneously error estimable, there exists an error estimator e_t such that the following two conditions

i) $P\{\sup_{h \in \mathcal{H}^k} \left|err\,(P,h) - e_t\left(u_1^t, y_1^t; h\right)\right| > \frac{\varepsilon}{3}\} \leq \frac{\delta}{2}$,

ii) $P\{\sup_{h,h' \in \mathcal{H}^k\ s.t.\ \rho_{u_1^t}(h,h') < \sigma_t\left(\frac{\varepsilon}{3}\right)} \left|e_t\left(u_1^t, y_1^t; h\right)\ - e_t\left(u_1^t, y_1^t; h'\right)\right| > \frac{\varepsilon}{3}\} \leq \frac{\delta}{2}$.

hold true simultaneously for arbitrary $\delta > 0$ and $\varepsilon > 0$ and for t greater than a suitable $\bar{t}(\delta, \varepsilon, P)$.

On the other hand, by the expansion

$$\begin{aligned} |err\,(P,h) - err\,(P,h')\,| \;\leq\; & |err\,(P,h) - e_t\left(u_1^t, y_1^t; h\right)| \\ & + |e_t\left(u_1^t, y_1^t; h\right) - e_t\left(u_1^t, y_1^t; h'\right)| \\ & + |e_t\left(u_1^t, y_1^t; h'\right) - err\,(P,h')\,|, \end{aligned}$$

it is easily seen that the set $\{\sup_{h,h' \in \mathcal{H}^k \ s.t.\ \rho_{u_1^t}(h,h') < \sigma_t\left(\frac{\varepsilon}{3}\right)} |err\,(P,h) - err\,(P,h')\,| > \varepsilon\}$ is included in

$$\{\sup_{h \in \mathcal{H}^k} |err\,(P,h) - e_t\left(u_1^t, y_1^t; h\right)| > \frac{\varepsilon}{3}\}$$

$$\cup\{\sup_{h,h' \in \mathcal{H}^k \ s.t.\ \rho_{u_1^t}(h,h') < \sigma_t\left(\frac{\varepsilon}{3}\right)} |e_t\left(u_1^t, y_1^t; h\right) - e_t\left(u_1^t, y_1^t; h'\right)| > \frac{\varepsilon}{3}\}.$$

Thus, from *i)* and *ii)*, the conclusion can be drawn that

$$P\{\sup_{h,h' \in \mathcal{H}^k \ s.t.\ \rho_{u_1^t}(h,h') < \sigma_t\left(\frac{\varepsilon}{3}\right)} |err\,(P,h) - err\,(P,h')| > \varepsilon\} \leq \delta, \text{ for } t > \bar{t}(\delta, \varepsilon, P)\,.$$

Setting $\tau = (t - q + 1)\,\sigma_t\left(\frac{\varepsilon}{3}\right)$, result (6) readily follows.

The thesis will now be proven by the joint use of equations (5) and (6).

Fix $\varepsilon > 0$. Select $\bar{h} \in \mathcal{H}$ such that $err\,(P,\bar{h}) \leq opt\,(P,\mathcal{H}) + \varepsilon$ and let $\bar{k} = \min\left\{k : \bar{h} \in \mathcal{H}^k\right\}$. Consider the expansion

$$\begin{aligned} & err\left(P, \bar{a}_t\left(u_1^t, y_1^t\right)\right) - opt\,(P,\mathcal{H}) \\ = \;& err\left(P, \bar{a}_t\left(u_1^t, y_1^t\right)\right) - e_{emp}\left(u_{n+1}^t, y_{n+1}^t; \bar{a}_t\left(u_1^t, y_1^t\right)\right) \\ & + e_{emp}\left(u_{n+1}^t, y_{n+1}^t; \bar{a}_t\left(u_1^t, y_1^t\right)\right) - e_{emp}\left(u_{n+1}^t, y_{n+1}^t; m_{n,u}(\bar{h})\right) \\ & + e_{emp}\left(u_{n+1}^t, y_{n+1}^t; m_{n,u}\left(\bar{h}\right)\right) - err\,(P, m_{n,u}\left(\bar{h}\right)) \\ & + err\,(P, m_{n,u}\left(\bar{h}\right)) - err\,(P,\bar{h}) \\ & + err\,(P,\bar{h}) - opt\,(P,\mathcal{H})\,. \end{aligned}$$

The probability that the first term is greater than μ_n tends to zero as $t \to \infty$ (see equation (5)). The second term is less than or equal to zero by the very definition of algorithm $\bar{a}_t$, whereas the third term can be bounded similarly to the first one. As for the fourth term, note that both $\bar{h}$ and $m_{n,u}(\bar{h})$ belong to $\mathcal{H}^k$ and they are at most $\frac{\tau_n}{n-q+1}$ apart with respect to the empirical distance $\rho_{u_1^n}$. Therefore, equation (6) implies that the probability that the fourth term is greater than ε approaches zero as $t \to \infty$. Finally, the last term is no greater than ε by construction. In conclusion, $P\left\{err\,(P, \bar{a}_t\,(u_1^t, y_1^t)) - opt\,(P,\mathcal{H}) > 2\mu_n + 2\varepsilon\right\}$ approaches zero as $t \to \infty$ for any ε. This completes the proof. □

4 Concluding remarks

Even though the fields of system identification and learning theory have many points in common, so far there have been only sporadic contacts between them. The reason for this can probably be found in the different technical assumptions adopted in these two fields.

In this paper, we have proposed a method to learn dynamical relations in a stationary framework. This study is a first attempt to bridge the existing gap between learning theory and system identification. An important aspect completely neglected in the present contribution is the computational effort required by the learning techniques.

References

[1] L.G. Valiant, "A theory of the learnable", *Comm. ACM*, vol.27, 1984, pp. 1134-1142.

[2] D. Angluin and P.D. Laird, "Learning from noisy samples", *Machine Learning*, vol.2, 1988, pp. 343-370.

[3] B.K. Natarajan, "On learning sets and functions", *Machine Learning*, vol.4, 1989, pp. 67-97.

[4] B.K. Natarajan, "Probably approximate learning of sets and functions", *SIAM Journal on Computing*, vol.20, 1991, pp. 328-351.

[5] K.L. Buescher and P.R. Kumar, "Learning by canonical smooth estimation, part I: simultaneous estimation", *IEEE Trans. on Automatic Control, AC-41*, 1996, pp. 545-556.

[6] K.L. Buescher and P.R. Kumar, "Learning by canonical smooth estimation, part II: learning and choice of model complexity", *IEEE Trans. on Automatic Control, AC-41*, 1996, pp. 557-569.

[7] D. Aldous and U. Vazirani, "A Markovian extension of Valiant's learning model", *Information and Computation*, vol. 117, pp. 181–186, March 1995.

[8] P. L. Bartlett, P. Fischer and K-U. Hoffgen, "Exploiting random walks for learning," *COLT94*, 1994.

[9] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications", *Information and Computations*, vol.100, 1992, pp. 78-150.

[10] M. Iosifescu and R. Theodorescu, *Random processes and learning*, New York, Springer Verlag, 1969.

[11] W. Phillip, "The central limit problem for mixing sequences of random variables",*Z. Wahrscheinlichkeitstheorie und verw. Geb.*, vol.12, 1969, pp. 155-171.

[12] D. Pollard, *Convergence of stochastic processes*, New York, Springer Verlag, 1984.

[13] R.M. Dudley, *A course on empirical processes*, in Lecture Notes in Mathematics, vol.1097, Springer Verlag, 1984.

[14] B.K. Natarajan, "Probably approximate learning over classes of distributions", *SIAM Journal on Computing*, vol.21, 1992, pp. 438-449.

[15] V.N. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies to their probabilities", *Theory of Prob. and its Appl.*, vol.16, 1971, pp. 264-280.

[16] V.N. Vapnik and A. Chervonenkis, "Necessary and sufficient conditions for the uniform convergence of means to their expectations", *Theory of Prob. and its Appl.*, vol.26, 1981, pp. 532-553.