Proceedings of the 35th
Conference on Decision and Control
Kobe, Japan • December 1996

Learning Dynamical Systems in a Stationary Environment

M.C. Campi† and P.R. Kumar‡
†Dept. of Electrical Engineering and Automation - University of Brescia
Via Branze 38, 25123 Brescia, Italy
‡Dept. of Electrical and Computer Engineering, and the
Coordinated Science Laboratory
University of Illinois
1308 West Main Street
Urbana, IL 61801,USA

Abstract

The recently emerging field of learning theory, pioneered by Vapnik and Chervonenkis [1, 2], has by and large been focused on the problem of learning static relations. As an initial attempt to extend this approach to system identification, we examine the problem of learning input–output relations in a stationary environment.

1 Introduction

The recently emerging field of learning theory [1, 2, 3, 4, 5, 6, 7, 8, 9], pioneered by Vapnik and Chervonenkis, has by and large been focused on the problem of learning static relations. For example in the work of Valiant [3], the goal is to learn a binary valued function $c(x)$ defined on an arbitrary set X, based on labeled sample points $(x_t, c(x_t))$, where the x_t's are drawn independently and identically according to an unknown probability distribution P on X. This treatment has been extended (see Haussler [4]) to situations involving noisy observations.

However, in these treatments, the data points x_t are assumed to be independent, thus disallowing any memory and hence dynamics in the evolution of the x_t-process. Thus the theory has by and large not been applicable to the problem of system identification.

In this paper, we study the problem of learning dynamical systems by casting them in a stationary rather than an i.i.d. framework. We also capitalize on some recent results of Buescher and Kumar [8, 9] which allow the use of a new canonical estimator rather than just the empirical estimator.

2 Problem Description

We consider a single input, single output (SISO) 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:
(i) y_t is conditionally independent of $u_{t+1}^{\infty}, y_{-\infty}^{t-1}$ and y_{t+1}^{∞} given $u_{-\infty}^{t}$ (throughout, $u_r^s := (u_r, u_{r+1}, \ldots, u_s)$). This ensures not only that the system is causal, but also that all the correlation in the y process is generated by the input process u.
(ii) The system is in a stationary environment, i.e, the joint process (u_t, y_t) is strict-sense stationary.
(iii) For convenience, we also assume that it is ergodic (i.e. its invariant σ-algebra is trivial).

Under these conditions, the dependence of y on u is captured by the (unknown) time-invariant conditional distribution $P_{y/u}$ of y_t given $u_{-\infty}^{t}$. Along with the specification of the probability distribution of $\{u_t\}_{t=-\infty}^{+\infty}$, $P_{y/u}$ 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 $\mathcal{P}$.

For an alternative viewpoint, define the conditional mean $s(u_{-\infty}^{t}) := \int_Y y P_{y/u}(dy, u_{-\infty}^{t})$, Then the system can be written as $y_t = s(u_{-\infty}^{t}) + d_t$, where $d_t := y_t - s(u_{-\infty}^{t})$, the additive noise, is conditionally white given the past.

We consider the learning problem of determining an q-dimensional approximation of the function s, where q is a fixed integer. The approximation function h is to be selected from a hypothesis set $\mathcal{H}$ of functions from U^q to Y.

The accuracy of such an estimate will be measured by

0-7803-3590-2/96 $5.00 © 1996 IEEE

the error criterion given below:

Definition 1 *(Error between P and h)*

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

Clearly, $err(P,h)$, also called the *generalization error* is the expected error one makes by using h to predict y_t.

Definition 2 *(Optimal error)*

The optimal error *is the minimum error over $h \in \mathcal{H}$:* $opt(P,\mathcal{H}) := \inf_{h\in\mathcal{H}} err(P,h)$. $\square$

Once the data $(u_1,y_1),\ldots,(u_t,y_t)$ have been collected, one uses an *algorithm*, which is an indexed family of maps $a_t : (U\times Y)^t \to \mathcal{H}$ to construct an estimate h.

Definition 3 *(Nonuniformly learning algorithm)*
We say that an algorithm a_t learns (possibly) nonuniformly over $(\mathcal{P},\mathcal{H})$ if :

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

$\forall P \in \mathcal{P}$. $\square$

The reason for the usage of the qualifier "nonuniform" is that in contrast to a more stringent notion of learnability, the convergence is not required to take place uniformly in $P \in \mathcal{P}$.

Our goal is to construct such an algorithm.

If such an algorithm exists (it may not), we say that $(\mathcal{P},\mathcal{H})$ is (nonuniformly) *learnable*.

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. The optimal error obtained by allowing h also to be a $\mathcal{J}$-measurable random variable is in general strictly lower than minimizing over deterministic hypotheses.

$\square$

Central to the design of a successful learning algorithm is the need to estimate certain expected values from data sequences. In the stationary environment assumed here, this can be met by appropriate mixing condition of the following sort:

Assumption 1
For any positive bounded function $\xi : U^q \times Y \to R$, it holds that

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

where $\zeta(\varepsilon,t-n) \to 0$, as $(t-n)\to\infty$.

Remark 2

Some such 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. Though the assumption that ξ is deterministically bounded is particularly strong, it is met in our context. Note also that, even though not explicitly indicated, the function ζ will depend on the bound on the 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]). $\square$

3 Learning by simultaneous estimation

A common way of learning a hypothesis is to first estimate the error associated with *each* hypothesis h in $\mathcal{H}$, Subsequently, one then chooses a hypothesis with the minimal estimated error. For the first step, a widely used error estimate is the *empirical error estimate* 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. [4]). 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 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. [4, 12, 13, 14]), whose origins are in the pioneering work of Vapnik and Chervonenkis ([1, 2]).

In this paper, following Buescher and Kumar [8, 9], we generalize 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 $(h(u_{i-q+1}^i) \simeq h'(u_{i-q+1}^i),\ i = q, q+1, \ldots, t)$ (see below). Such a smoothness condition is natural and only rules out pathological situations. It turns out that the empirical error estimator is in fact smooth (and, therefore, our methodology covers it as well). However, there are many cases in which a smooth simultaneous error estimator exists and yet the empirical estimator fails to simultaneously estimate.

Our procedure is as follows. First, a suitable *finite empirical cover* for $\mathcal{H}$, i.e. a cover based on the *empirical distance* $\rho_{u_1^t}(h, h') := \frac{1}{t-q+1}\sum_{i=q}^{t} \left|h(u_{i-q+1}^i) - h'(u_{i-q+1}^i)\right|$ 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.

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

We suppose that $\mathcal{H}$ 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.

We now define precisely the above notions.

Definition 4 *(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
$\forall \varepsilon > 0$, $\lim_{t\to\infty} P\{\sup_{h\in\mathcal{H}} \left|err(P,h) - e_t(u_1^t, y_1^t; h)\right| > \varepsilon\} = 0$,
$\forall P \in \mathcal{P}$. □

Definition 5 *(Smooth estimators)*
The error estimator e_t is smooth if $\forall \vartheta > 0$, $\exists \sigma_t(\vartheta) > 0$ such that
$\lim_{t\to\infty} P\{\sup_{h,h'\in\mathcal{H}\ s.t.\ \rho_{u_1^t}(h,h')<\sigma_t(\vartheta)} \left|e_t(u_1^t, y_1^t; h) - e_t(u_1^t, y_1^t; h')\right| > \vartheta\} = 0$,
$\forall P \in \mathcal{P}$. □

Our goal now is to construct an algorithm able to learn $(\mathcal{P}, \mathcal{H})$ whenever a smooth simultaneous error estimator exists for each pair $(\mathcal{P}, \mathcal{H}^k)$, $k = 1, 2, \ldots$.

Towards this end, we introduce 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, one wishes to choose a *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.

It can be shown, through the constructive procedure below (borrowed from [9]), that a simple empirical ε-cover for $\mathcal{H}$ of finite cardinality always exists. Since $Y \subset R$ is 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_k from $i \in \{q, q+1, \ldots, n\}$ to cover elements $(M_k(i) \in Y_{\varepsilon/2},\ i = q, q+1, \ldots, n)$. The simple empirical ε-cover for $\mathcal{H}$ is recursively constructed as follows. Initially, let $\mathcal{C}_0 = 0$. For $k = 1$ to $N(\varepsilon/2)^{n-q+1}$, check whether there exist hypotheses $h \in \mathcal{H}$ such that

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

If any, pick an hypothesis h_k satisfying (2) which is as simple as possible ($h_k \in \mathcal{H}^k$ and there exists no $h \in \mathcal{H}^j$, $j < k$, such that (2) is satisfied) and set $\mathcal{C}_k = \mathcal{C}_{k-1} \cup \{h_k\}$. 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 $m_{n,u}(h) = h'$.

Now we specify our *learning algorithm*:
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

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

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

Our main result is the following Theorem on the efficacy of this 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 4

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, it is possible to show that *uniform* (i.e. convergence takes place uniformly in P) learnability holds for $(\mathcal{P}, \mathcal{H})$ provided that $(\mathcal{P}, \mathcal{H})$ is itself smoothly simultaneously uniformly error estimable. □

4 Concluding remarks

The reason that the fields of system identification and learning theory have had only sporadic contacts between them is probably that the two fields have adopted different sets of technical assumptions .

Here 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.

We should note that we have completely neglected the issue of computational effort required by the learning techniques.

Acknowledgment

The first author would like to acknowledge the financial support of MURST under the 60% project "Adaptive identification, prediction and control".

The research of the second author has been supported by the U. S.!Army Research Office under 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.

References

[1] 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.

[2] 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.

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

[4] 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.

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

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

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

[8] 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.

[9] 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.

[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.