ELSEVIER

Systems & Control Letters 34 (1998) 133–140

SYSTEMS & CONTROL LETTERS

The complexity of model classes, and smoothing noisy data[1]

Peter L. Bartlett [a,*], Sanjeev R. Kulkarni [b,2]

[a] *Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, 0200 Australia*

[b] *Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA*

Received 8 January 1997; received in revised form 29 September 1997; accepted 29 September 1997

Abstract

We consider the problem of smoothing a sequence of noisy observations using a fixed class of models. Via a deterministic analysis, we obtain necessary and sufficient conditions on the noise sequence and model class that ensure that a class of natural estimators gives near-optimal smoothing. In the case of i.i.d. random noise, we show that the accuracy of these estimators depends on a measure of complexity of the model class involving covering numbers. Our formulation and results are quite general and are related to a number of problems in learning, prediction, and estimation. As a special case, we consider an application to output smoothing for certain classes of linear and nonlinear systems. The performance of output smoothing is given in terms of natural complexity parameters of the model class, such as bounds on the order of linear systems, the l_1-norm of the impulse response of stable linear systems, or the memory of a Lipschitz nonlinear system satisfying a fading memory condition.

Keywords: System identification; Computational learning theory; Smoothing; Covering numbers

1. Introduction

In this paper, we study the problem of smoothing a set of noisy observations by using a fixed class of models. Specifically, suppose we observe $y_i = f_i + e_i$ for $i = 1, \ldots, n$. Our goal is to obtain an estimate $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_n)$ for $f = (f_1, \ldots, f_n)$, that is close to f in some sense (made precise in the next section). We consider two related formulations. In the first, the true f is assumed to belong to some known class F. In the second, we make no assumptions on f, but restrict attention to estimators $\hat{f}$ that belong to a known class F. In this setting, of course we should be content only in producing an estimate for f that is close to the optimal $g \in F$.

We first consider a deterministic/worst-case setting in which the e_i is a fixed but arbitrary sequence. We obtain deterministic conditions on the noise sequence e_i in terms of the class of models F that are necessary and sufficient to allow smoothing. These conditions have a natural interpretation: the correlation between the noise and certain "model difference" sequences should not be significantly larger than the power of those sequences. This result can be used for stochastic noise models by verifying the deterministic conditions on the sample paths of the noise process. In particular, we treat the case of i.i.d. noise e_i, and show that smoothing is possible if appropriate covering numbers of the model

* Corresponding author.

[1] An earlier version of this paper was presented at the 35th IEEE Conference on Decision and Control, Kobe, 1996.

[2] This work was supported in part by the National Science Foundation under NYI Award IRI-9457645.

0167-6911/98/$19.00
PII S0167-6911(98)00008-5

class grow slowly in terms of n – i.e., if a "richness" constraint is imposed on the class of models. Finally, we consider an application of these results to output prediction of linear and nonlinear systems. In these problems, it is assumed that the underlying system is unknown. An input sequence is applied and noisy outputs are observed. Using knowledge of the inputs we wish to estimate the outputs almost as accurately as the best model in a fixed class. As a special case, this gives near-optimal estimation when the system is known to belong to the model class. For kth-order linear systems, for linear systems of arbitrary order but satisfying a constraint on the l_1 norm of the impulse response, and for nonlinear Lipschitz fading memory systems we obtain explicit bounds on how well we can smooth in terms of the "complexity" parameters of the model classes.

Our formulation is quite general and is related to a number of problems considered in papers on learning, prediction, and estimation. In particular, if the $f_i = f(x_i)$ for some function f and the points $x_1, \ldots, x_n$ are assumed to be known, then the problem considered here is related to work in statistics and computational learning theory (see, e.g., [7] and references therein). Most of the work in these areas that deals with general model classes considers the problem of estimating a target function from noise-corrupted values at a number of randomly and independently chosen points, where the measure of accuracy depends on the probability distribution generating the points. In contrast, we make no assumptions about the process generating the examples, but the conditions we obtain on the target class that are sufficient for the smoothing problem with i.i.d. random noise are similar to corresponding conditions for more standard learning problems. There has been some work that has considered arbitrary x_i for specific estimators for regression or output prediction in a systems context (e.g., see [9, 10] and references therein). However, in addition to arbitrary x_i (or arbitrary inputs), our work also considers deterministic noise, deals with characterizing properties of general model classes that allow smoothing, and considers a different success criterion. In contrast with previous work, we obtain both necessary and sufficient conditions on the noise sequence for general model classes. Our results are also similar in flavor to some work in system identification that uses notions of covering numbers and metric dimension to measure the complexity of identification (e.g., see [13] and references therein), but the specific formulations and results are quite different.

2. Smoothing problems

Suppose we observe $y_i = f_i^* + e_i$ for $i = 1, \ldots, n$ where $f^* = (f_1^*, \ldots, f_n^*)$ is an underlying real sequence we wish to estimate and e_i represents measurement noise. For convenience, it is useful to assume we see a sequence of input points $x_1, \ldots, x_n$ chosen from a set X, and $f_i^* = f^*(x_i)$ for some unknown target function $f^*(\cdot)$. The aim is to estimate the target function at the points, in the sense that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - f^*(x_i))^2$$

is small, where f^* is the target function and $\hat{y}$ is the estimate. We also write this error as $\|\hat{y} - f^*\|_n^2$. That is, we identify the function f^* with the sequence $f^*(x_1), f^*(x_2), \ldots$, and we consider the family of norms over initial segments of real sequences,

$$\|y\|_n = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right)^{1/2}.$$

An estimator can be viewed as a mapping from $X^n \times \mathbb{R}^n$ to $\mathbb{R}^n$.

We will fix a class F of functions defined on the input set X, and consider two smoothing problems. In the first of the two smoothing problems, we want an estimator for which, for all target functions f^* chosen from F, the error of the estimate goes to zero as $n \to \infty$.

Definition 1. Let $x = (x_1, \ldots, x_n) \in X^n$ be an input sequence and $e = (e_1, \ldots, e_n) \in \mathbb{R}^n$ be a noise sequence. We say that an estimator ε-smooths F with respect to x and e if it satisfies the following condition. For all $f^* \in F$, when the estimator sees the sequences x and y, where $y_i = f^*(x_i) + e_i$, it produces an estimate f_n satisfying $\|f_n - f^*\|_n^2 \leqslant \varepsilon$.

For an input sequence $x = (x_1, x_2, \ldots) \in X^\infty$ and a noise sequence $e = (e_1, e_2, \ldots) \in \mathbb{R}^\infty$, we say that an estimator smooth F with respect to x and e if, for all $f^* \in F$, the estimate satisfies

$$\limsup_{n\to\infty} \|f_n - f^*\|_n = 0.$$

In the second problem, we allow modelling error. In this case, for any target function from X to $\mathbb{R}$ the estimate must have error that approaches that of the best approximation in F to the target function. In fact, we restrict the target functions to those for which the

approximation error is bounded (otherwise aiming for a near-optimal approximation seems pointless).

Definition 2. Let $x=(x_1,\dots,x_n)\in X^n$ be an input sequence and $e=(e_1,\dots,e_n)\in\mathbb{R}^n$ be a noise sequence. We say that an estimator ε-optimizes over F with respect to x and e if it satisfies the following condition. For all target functions $m:X\to\mathbb{R}$, when the estimator sees the sequences x and y with $y_i=m(x_i)+e_i$, it produces an estimate f_n satisfying

$$\|f_n-m\|_n^2\leqslant\inf_{g\in F}\|g-m\|_n^2+\varepsilon.$$

For an input sequence $x=(x_1,x_2,\dots)\in X^\infty$ and a noise sequence $e=(e_1,e_2,\dots)\in\mathbb{R}^\infty$, we say that an estimator optimizes over F with respect to x and e if, for all $m:X\to\mathbb{R}$ satisfying

$$\limsup_{n\to\infty}\inf_{g\in F}\|g-m\|_n<\infty,$$

the estimates $\{f_n\}$ satisfy

$$\limsup_{n\to\infty}\left(\|f_n-m\|_n-\inf_{g\in F}\|g-m\|_n\right)=0.$$

We will concentrate on empirical estimators.

Definition 3. For a class F of real-valued functions defined on a set X, an empirical estimator for F is a mapping from $X^n\times\mathbb{R}^n$ to $\mathbb{R}^n$. It returns a sequence f_n in $\bar{F}_{|x}$, the closure (with respect to $\|\cdot\|_n$) of

$$F_{|x}=F_{|x_1,\dots,x_n}=\{(f(x_1),\dots,f(x_n)):\ f\in F\},$$

that satisfies $\|f_n-y\|_n=\inf_{f\in F}\|f-y\|_n$.

3. Deterministic conditions

Our first theorem gives conditions on noise sequences and sequences of function differences that are sufficient for the success of empirical estimators. The conditions can be thought of as a requirement that the correlation between the noise and any sequence of non-negligible function differences should not exceed the power of that sequence.

Theorem 4. *Suppose that F is a class of real-valued functions defined on a set X, and x and e are input and noise sequences.*

(1) *If some empirical estimator for F fails to smooth F with respect to x and e, then there is an $\varepsilon>0$, and an $f^*\in F$ such that, for infinitely many values of $n\in N$, there is an f in $\bar{F}_{|x}$ satisfying*

$$\varepsilon\leqslant\frac{1}{n}\sum_{i=1}^n(f(x_i)-f^*(x_i))^2$$

$$\leqslant\frac{2}{n}\sum_{i=1}^n(f(x_i)-f^*(x_i))e_i. \tag{1}$$

Furthermore, the reverse implication is also true if $\limsup_{n\to\infty}\|e\|_n$ is finite and F satisfies the following property: there is a $\rho>0$ such that, for all $f_1,f_2\in F$, there are $f_1',f_2'\in F$ and $\tau\in(\rho,1-\rho)$ such that $f_1'-f_2'=\tau(f_1-f_2)$.

(2) *If some empirical estimator for F fails to optimize over F with respect to x and e, then there is an $\varepsilon>0$ and an m in $\mathbb{R}^X$ (with m satisfying $\limsup_{n\to\infty}\inf_{g\in F}\|g-m\|_n<\infty$) such that, for infinitely many n there is an f in F such that for f^* in $\bar{F}_{|x}$ satisfying $\|f^*-m\|_n=\inf_{g\in F}\|g-m\|_n$,*

$$\varepsilon\leqslant\frac{1}{n}\sum_{i=1}^n(f(x_i)-m(x_i))^2$$

$$-\frac{1}{n}\sum_{i=1}^n(f_i^*-m(x_i))^2$$

$$\leqslant\frac{2}{n}\sum_{i=1}^n(f(x_i)-f_i^*)e_i. \tag{2}$$

Notice that the condition on F that suffices for the converse result in part 1 is trivially satisfied if F is convex and contains the zero function.

Proof. (1) If an empirical estimator fails to smooth with no modelling error, then there is an $\varepsilon>0$ and an f^* in F such that, for infinitely many values of n, there is an f in $\bar{F}_{|x}$ (the closure is with respect to $\|\cdot\|_n$) with $\|f-(f^*+e)\|_n^2=\inf_{g\in F}\|g-(f^*+e)\|_n^2$ and $\|f-f^*\|_n^2\geqslant\varepsilon$. The inequality is equivalent to the first inequality in Eq. (1), and the equality implies $\|f-(f^*+e)\|_n^2\leqslant\|e\|_n^2$, which is equivalent to the second inequality in Eq. (1).

To see that the converse is true under the conditions on e and F, notice that, for any $\tau>0$ the second inequality in Eq. (1) is equivalent to

$$\sum_{i=1}^n(\tau(f_i-f_i^*))^2+(\tau-\tau^2)\sum_{i=1}^n(f_i-f_i^*)^2$$

$$\leqslant2\sum_{i=1}^n\tau(f_i-f_i^*)e_i.$$

For $\rho < \tau < 1-\rho$, this inequality and the first inequality in Eq. (1) imply

$$2\sum_{i=1}^{n} \tau(f_i - f_i^*)e_i \geqslant \sum_{i=1}^{n} (\tau(f_i - f_i^*))^2 + n(\tau - \tau^2)\varepsilon \geqslant \sum_{i=1}^{n} (\tau(f_i - f_i^*))^2 + n\rho^2\varepsilon.$$

It follows that

$$\sum_{i=1}^{n} (\tau(f_i - f_i^*) - e_i)^2 \leqslant \sum_{i=1}^{n} e_i^2 - n\rho^2\varepsilon.$$

Notice also that $\sum_{i=1}^{n}(\tau(f_i - f_i^*))^2 \geqslant n\tau^2\varepsilon \geqslant n\rho^2\varepsilon$. So the condition of the theorem implies (with appropriate relabeling) that there is an $\alpha > 0$, an $\varepsilon > 0$, and an f^* in F such that, for infinitely many values of n, there is an f in F with

$$\|f - (f^* + e)\|_n^2 \leqslant \|f^* - (f^* + e)\|_n^2 - \alpha,$$

and $\|f - f^*\|_n^2 \geqslant \varepsilon$. This implies that

$$\limsup_{n\to\infty} \left(\|f^* - (f^* + e)\|_n^2 - \inf_{g\in F} \|g - y\|_n^2 \right) = \beta > 0.$$

Consider an infinite subsequence for which

$$\|f^* - y\|_n^2 \geqslant \inf_{g\in F} \|g - y\|_n^2 + \beta/2.$$

For each n, choose a sequence g^* from $\bar{F}_{|x}$ with $\|g^* - y\|_n = \inf_{g\in F} \|g - y\|_n$. Then the triangle inequality implies

$$\|g^* - f^*\|_n \geqslant \|f^* - y\|_n - \inf_g \|g - y\|_n = \frac{\|f^* - y\|_n^2 - \inf_{g\in F} \|g - y\|_n^2}{\|f^* - y\|_n + \inf_{g\in F} \|g - y\|_n} \geqslant \frac{\beta/2}{2\|e\|_n}.$$

So some empirical estimator fails.

(2) If an empirical estimator fails to optimize over F, then there is an $\varepsilon > 0$ and a function m such that, for infinitely many n, if $f^* \in \bar{F}_{|x}$ satisfies $\|f^* - m\|_n = \inf_{g\in F} \|g - m\|_n$, there is an f in F with $\|f - (m + e)\|_n^2 = \min_{g\in \bar{F}_{|x}} \|g - (m + e)\|_n^2$, and $\|f - m\|_n^2 \geqslant \|f^* - m\|_n^2 + \varepsilon$. These inequalities imply

$$\sum_{i=1}^{n} (f_i - m_i)^2 - \sum_{i=1}^{n} (f_i^* - m_i)^2 \leqslant 2\sum_{i=1}^{n} (f_i - f_i^*)e_i,$$

and

$$\sum_{i=1}^{n} (f_i - m_i)^2 - \sum_{i=1}^{n} (f_i^* - m_i)^2 \geqslant n\varepsilon. \qquad \square$$

4. Smoothing with random noise

In this section we consider the case of random noise sequences e that are realizations of a bounded i.i.d. stochastic process. We show that in this case empirical estimators can smooth and optimize a uniformly bounded function class F if F has a slowly growing *covering number*. If (Y, d) is a metric space, $S \subset Y$, and $\varepsilon > 0$, the ε-covering number of Y is the size of the smallest subset T of Y for which every point in S is within ε of some point in T. For a function class F and $\alpha > 0$, let $\mathcal{N}(F, n, \alpha)$ denote the maximum over x in X^n of the α-covering number of $F_{|x} \subseteq \mathbb{R}^n$ with respect to the metric $d(a, b) = \|a - b\|_n$.

Theorem 5. *Suppose that $e_1, \ldots, e_n$ are independent zero-mean random variables satisfying $|e_i| \leqslant M$. Suppose that F is a class of real-valued functions defined on X satisfying $|f(x)| \leqslant B$ for all $x \in X$* and all *$f \in F$. Then for any input sequence $x \in X^n$, and any $m : X \to \mathbb{R}$ satisfying $|m(x_i)| \leqslant B$, the probability of a noise sequence e for which some empirical estimator fails to ε-optimize over F with respect to x and e is no more than*

$$\mathcal{N}(F, n, \varepsilon/(4M)) \exp\left(-\frac{2\varepsilon^2 n}{M^2 B^2}\right).$$

Clearly, since the second factor in the probability bound is exponentially small in n, a sufficient condition to force the probability of failure to go to zero is that the growth of the covering number be slower than exponential in n. For i.i.d. noise, it is possible to show that a slowly growing covering number is also necessary for vanishing failure probability.

Proof. From Theorem 4, if some empirical estimator fails for a given ε and m, some f in F has

$$\frac{1}{n}\sum (f(x_i) - f_i^*)e_i \geqslant \varepsilon/2,$$

where f^* minimizes $\|f^* - m\|_n$. In that case, choose the $\hat{f}$ in an $\varepsilon/(4M)$-cover of $F_{|x}$ that satisfies $\|\hat{f} - f\|_n \leq \varepsilon/(4M)$. This $\hat{f}$ satisfies

$$\begin{aligned}\frac{1}{n}\sum(\hat{f}_i - f_i^*)e_i &= \frac{1}{n}\sum(\hat{f}_i - f_i)e_i \\ &\quad + \frac{1}{n}\sum(f_i - f_i^*)e_i \\ &\geq \varepsilon/2 - \|\hat{f} - f\|_n\|e\|_n \\ &\geq \varepsilon/4.\end{aligned}$$

The probability that the estimator fails is no more than the size of the cover times the probability that some fixed $\hat{f}$ will satisfy

$$\frac{1}{n}\sum(\hat{f}_i - f_i^*)e_i \geq \varepsilon/4.$$

Hoeffding's inequality (see for example [12]) shows that this latter probability is no more than $\exp(-2\varepsilon^2 n/(B^2M^2))$, which gives the desired result. □

In fact, for convex function classes F we can improve the rate of convergence in this result.

Theorem 6. *Suppose that F is a convex class of real-valued functions defined on X satisfying $|f(x)| \leq B$ for all $x \in X$ and all $f \in F$. Let $e \in \mathbb{R}^n$ be a realization of an i.i.d. stochastic process satisfying $|e_i| \leq M$ and $Ee_i = 0$. Then for any input sequence $x \in X^n$, and any $m : X \to \mathbb{R}$ satisfying $|m(x_i)| \leq B$, the probability of a noise sequence e for which some empirical estimator fails to ε-optimize over F with respect to $x \in X^n$ and e is no more than*

$$\mathcal{N}(F, n, \varepsilon/(28B + 12M))\exp\left(\frac{-\varepsilon n}{54(B+M)^2}\right).$$

The proof is based on that of the main estimation result in [11]. It uses the following consequence of Bernstein's inequality ([11], (Lemma 8)), instead of Hoeffding's inequality.

Lemma 7 (Lee et al. [11]). *For i.i.d. random variables $V_1, \ldots, V_n$ satisfying $|V_i| \leq K$, $EV_i \geq 0$, and $EV_i^2 < K_2 EV_i$ for $i = 1, \ldots, n$ with $K_2 \geq 1$, we have*

$$\Pr\left(\frac{E(1/n\sum_{i=1}^n V_i) - 1/n\sum_{i=1}^n V_i}{v + E(1/n\sum_{i=1}^n V_i)} \geq \alpha\right) \leq \exp\left(\frac{-3\alpha^2 vn}{2(K_1 + K_2)}\right).$$

Proof of Theorem 6 . If some f in F minimizes $\|f - y\|_n$ and $\|f - m\|_n^2 \geq \|f^* - m\|_n^2 + \varepsilon$, then that f also satisfies

$$E(\|f - y\|_n^2 - \|f^* - y\|_n^2) \geq \varepsilon$$

and minimizes $\|f - y\|_n^2 - \|f^* - y\|_n^2$. Since this latter quantity is zero for $f = f^*$, it follows that, for any $\alpha > 0$, this f has

$$\begin{aligned}&E(\|f - y\|_n^2 - \|f^* - y\|_n^2) \\ &\quad \geq \varepsilon + \alpha(\|f - y\|_n^2 - \|f^* - y\|_n^2).\end{aligned}$$

Set $\alpha = 2$ and choose an $\hat{f}$ in an ε_0-cover of $F_{|x}$ such that $\|f - \hat{f}\|_n \leq \varepsilon_0$ (ε_0 will be chosen later). Then for any y it is easy to show that

$$\begin{aligned}&\|f - y\|_n^2 - 2(2B + M)\varepsilon_0 \\ &\quad \leq \|\hat{f} - y\|_n^2 \leq \|f - y\|^2 + (4B + 2M + \varepsilon_0)\varepsilon_0.\end{aligned}$$

If we set $\varepsilon_0 = \varepsilon/(28B + 12M)$ then, provided $\varepsilon \leq B^2$, we have that

$$\begin{aligned}&E(\|\hat{f} - y\|_n^2 - \|f^* - y\|_n^2) \\ &\quad \geq \varepsilon/2 + 2(\|\hat{f} - y\|_n^2 - \|f^* - y\|_n^2).\end{aligned}$$

So the probability that an empirical estimator does not ε-optimize over F is no more than the size of an ε_0-cover of $F_{|x}$ times the probability that some fixed $\hat{f}$ satisfies this inequality.

Now, consider Lemma 7 with $V_i = (\hat{f}_i - y_i)^2 - (f_i^* - y_i)^2$. Clearly, $K_1 = (2B + M)$ and it is easy to show that we can choose $K_2 = 16(B + M)^2$ if the closure of $F_{|x}$ is convex (see [11], (Lemma 14)). Substituting $\alpha = \frac{1}{2}$ and $v = \varepsilon/2$ into Lemma 7 shows that the probability that an empirical estimator does not ε-optimize over F is no more than

$$\begin{aligned}&\mathcal{N}\left(F, n, \frac{\varepsilon}{28B + 12M}\right) \\ &\quad \times \Pr\left(E\left(\frac{1}{n}\sum_{i=1}^n V_i\right) \geq \frac{2}{n}\sum_{i=1}^n V_i + \varepsilon/2\right) \\ &\quad \leq \mathcal{N}\left(F, n, \frac{\varepsilon}{28B + 12M}\right)\exp\left(\frac{-\varepsilon n}{54(B+M)^2}\right).\end{aligned}$$

□

Clearly, corresponding results for smoothing (with no modeling errors) follow as special cases of Theorems 5 and 6. In fact, we can always obtain the improved rate of convergence (approximately $1/n$ rather than $1/\sqrt{n}$) in this case, even if F is not convex. The proof is essentially identical to that of Theorem 6 (except we use the fact that m is in F to provide the bound on K_2).

5. Systems applications

We consider discrete-time systems $f:\mathbb{R}^\infty \to \mathbb{R}^\infty$ where $\mathbb{R}^\infty = \{(u_1,u_2,\ldots): u_i \in \mathbb{R},\ i \geqslant 0\}$. We assume these systems are causal and time-invariant, so we will write $f(u_1,\ldots,u_n)$ for the initial length n subsequence of $f(u)$. As above, an empirical estimator sees a sequence $y_1,\ldots,y_n$ where $y_i = m_i + e_i$, and chooses an f in the model class that minimizes $\|f(u)-y\|_n$. Clearly, this includes smoothing as the special case in which $m = f(u)$ for some f in the model class.

Theorem 8. *Suppose that $e \in \mathbb{R}^n$ is a realization of an i.i.d. stochastic process satisfying $|e_i| \leqslant M$ and $Ee_i = 0$. Let F_k be a subset of the set of kth order linear systems. Let $u_1,\ldots,u_n$ be a real sequence satisfying $|f(u_1,\ldots,u_i)| \leqslant B$ for all $i \leqslant n$ and all f in F_k. Let the real numbers $m_1,\ldots,m_n$ satisfy $|m_i| \leqslant B$ for all i. Then with probability $1-\delta$ over the noise sequence e, the output f of an empirical estimator satisfies*

$$\|f-m\|_n^2 \leqslant \inf_{g\in F_k} \|g-m\|_n^2 + \frac{c}{n}\left((B+M)^2 k \log^2 n + \log\frac{1}{\delta}\right),$$

where c is a universal constant.

Proof. Represent f using parameters $y_{-n+1},\ldots,y_0$, $a_1,\ldots,a_k$, $b_0,\ldots,b_{k-1}$ so that $f(u_1,\ldots,u_n) = y_n$ and

$$y_i = \sum_{j=1}^{k} a_j y_{i-j} + \sum_{j=0}^{k-1} b_j u_{i-j}$$

for $i = 1,\ldots,n$. Then f is a polynomial of degree no more than n in its $3k$ real parameters. Results of Goldberg and Jerrum [6] and Pollard [12] imply that the covering number of this class satisfies $\mathcal{N}(F,n,\varepsilon) \leqslant (B/\varepsilon)^{ck\log n}$ (Dasgupta and Sontag [5] used a similar argument in their study of all-pole scalar systems with a thresholded output). Applying Theorem 6 gives the desired result. □

Of course, Theorem 8 immediately gives a similar result for ε-smoothing with respect to the class of linear systems of bounded order. In fact, we need not restrict the order of the linear systems. The following theorem shows that for stable systems, the l_1-norm of the impulse response provides an alternative measure of complexity.

Theorem 9. *Suppose F_S is the class of causal time-invariant linear systems with impulse response coefficients satisfying $\sum_{i=0}^{\infty} |h_i| \leqslant S$. Suppose that $e \in \mathbb{R}^n$ is a realization of an i.i.d. stochastic process satisfying $|e_i| \leqslant M$ and $Ee_i = 0$, and that $u_1,\ldots,u_n$ and $m_1,\ldots,m_n$ are real sequences satisfying $|u_i| \leqslant B$ and $|m_i| \leqslant B$ for all i. Then with probability $1-\delta$ over the noise sequence, the output f of an empirical estimator satisfies*

$$\|f-m\|_n^2 \leqslant \inf_{g\in F_S} \|g-m\|_n^2 + 48(BS+M)^2\left(\frac{\log(2n)}{n}\right)^{1/3} + \frac{108(BS+M)^2}{n}\log\frac{1}{\delta}.$$

The proof uses ideas from [11, 2]. It needs the following approximation result (which Barron in [1] attributes to Maurey).

Lemma 10 (Maurey). *Let μ be a probability measure on X and let F be a class of real-valued functions defined on X satisfying $|f(x)| \leqslant 1$. Then for any sequence $w_1,\ldots,w_n$ of positive real numbers that satisfy $\sum_i w_i = 1$ and any sequence $f_1,\ldots,f_n$ of functions from F, there are functions $\hat{f}_1,\ldots,\hat{f}_k$ in F for which*

$$\int \left(\sum_{i=1}^{n} w_i f_i(x) - \frac{1}{k}\sum_{i=1}^{k} \hat{f}_i(x)\right)^2 \mathrm{d}\mu(x) \leqslant \frac{1}{k}.$$

Proof of Theorem 9. Fix an input sequence u. Then for any $f \in F_S$, we have

$$f(u_1,\ldots,u_i) = \sum_{j=0}^{i-1} h_j u_{i-j}.$$

Let μ be the uniform distribution on $\{(u_1,u_2,\ldots,u_i): i = 1,\ldots,n\}$. By Lemma 10, there is some subsequence $j_1,\ldots,j_k$ of indices in $\{1,\ldots,i\}$ and a sequence $\alpha_1,\ldots,\alpha_k$ from $\{-1,1\}$ such that

$$\left\| f(u_1,\ldots,u_i) - \frac{S}{k}\sum_{l=1}^{k} \alpha_l u_{i-j_l} \right\|^2 \leqslant \frac{B^2S^2}{k}.$$

It follows that there is a $BS/\sqrt{k}$-cover of $F_{S_{1/}}$ of size no more than $\binom{2n}{k} \leqslant (2n)^k$. Now, by rescaling F, e, and ε, we can assume that $BS + M = 1$. Theorem 6 shows that if

$$\frac{\varepsilon n}{54} \geqslant \frac{28^2}{\varepsilon^2}\log(2n) + \log\frac{1}{\delta},$$

then with probability $1-\delta$ any empirical estimator will ε-optimize over F. For this it suffices if

$$\varepsilon \geq 48\left(\frac{\log(2n)}{n}\right)^{1/3} + \frac{108}{n}\log\frac{1}{\delta}.$$

Rescaling F, e, and ε gives the desired result.

It is possible to extend this result to give an estimator that optimizes over the class of all stable linear systems, by allowing the bound on the impulse response coefficients to increase slowly as n increases (roughly as $\sqrt{n}$, ignoring log factors). □

The next theorem considers output smoothing for nonlinear systems that satisfy a Lipschitz constraint. For a real-valued function f defined on a metric space, define $\|f\|_L$ as

$$\inf\{K>0: |f(x)-f(y)| \leq K\|x-y\| \text{ for all } x, y\},$$

and let $\|f\|_{BL} = \|f\|_L + \|f\|_\infty$. We shall consider $u_i \in [-1,1]$ and $x_i = (u_1,\ldots,u_i)$, and define $\|\cdot\|_L$ with respect to the Euclidean distance on $[-1,1]^i$.

Theorem 11. *Suppose F is a class of bounded Lipschitz systems (that is, there is an L such that for all f in F, $\|f_{|[-1,1]^n}\|_{BL} \leq L$, where $f_{|[-1,1]^n}$ is the restriction of f to $[-1,1]^n$), and F satisfies the following fading memory condition: there is a sequence $\phi_i \to 0$ such that, for all n, all $(u_1,\ldots,u_n) \in [-1,1]^n$, and all f in F,*

$$|f(u_1,\ldots,u_n) - f(u_{n-i+1},\ldots,u_n)| \leq \phi_i.$$

We define $\phi^{-1}(\alpha) = \min\{i: \phi_j \leq \alpha \text{ for all } j \geq i\}$.

Suppose that $e \in \mathbb{R}^n$ is a realization of an i.i.d. stochastic process satisfying $|e_i| \leq M$ and $Ee_i = 0$. Suppose that the real sequences $u_1,\ldots,u_n$ and $m_1,\ldots,m_n$ satisfy $|u_i| \leq 1$ and $|m_i| \leq L$ for all i. Then the probability of a noise sequence e for which the output f of an empirical estimator has $\|f-m\|_n^2 > \inf_{g\in F}\|g-m\|_n^2 + \varepsilon$ is no more than

$$\exp\left(\left(\frac{c_1 L(B+M)}{\varepsilon}\right)^{\phi^{-1}(c_2\varepsilon/(B+M))} - \frac{\varepsilon n}{c_3(B+M)^3}\right)$$

for universal constants c_1, c_2, and c_3.

Proof. For any $\alpha>0$, let

$$F_\alpha = \{f': f'(u_1,\ldots,u_n) = f(u_{n-k+1},\ldots,u_n), \text{ for some } f \text{ in } F\},$$

where $k = \phi^{-1}(\alpha)$. Clearly, F_α forms an α-cover of F under the infinity norm, and F_α is a subset of bounded Lipschitz functions defined on $[-1,1]^k$. Theorem IV of [8] shows that

$$\log_2 \mathcal{N}(F_\alpha, n, \beta) \leq \left(\frac{cL}{\beta}\right)^k$$

for all β and some universal constant c. By the triangle inequality, this implies $\log_2 \mathcal{N}(F, n, \alpha+\beta) \leq (cL/\beta)^k$. Substituting into Theorem 6 with $\alpha = \beta = \varepsilon/(56B + 24M)$ gives the result. □

As for the linear case, if the Lipschitz constant, infinity norm bound, or fading memory property $\phi^{-1}(\cdot)$ of f^* are not known in advance, we can consider classes defined using estimates of these quantities, and the estimates can be increased as n increases.

6. Final remarks

We conjecture that the deterministic conditions given in Section 3 are equivalent, under some suitable richness conditions on F, and that empirical estimators will smooth F with respect to noise and input sequences if and only if they will optimize over F with respect to the same noise and input sequences.

We have used similar techniques [4] to construct a universally consistent estimator for arbitrary bounded measurable functions that has error converging to zero in $L_2(d\mu)$ (μ is a probability distribution), provided the data asymptotically approximates the distribution in the sense that the Lèvy–Prohorov metric between the empirical distribution and the probability distribution converges to zero at a known rate. (This mild assumption includes i.i.d. data as a special case.) We have also used these techniques in a signal restoration problem [3], obtaining generalizations (to noisy data and more arbitrary function classes) of the Nyquist sampling theorem.

Acknowledgements

Thanks are due to Mostefa Golea for helpful discussions, particularly related to the pseudodimension bounds in the proof of Theorem 8.

References

[1] A.R. Barron, Universal approximation bounds for superposition of a sigmoidal function, IEEE Trans. Inform. Theory 39 (1993) 930–945.

[2] P.L. Bartlett, Pattern classification in neural networks: the size of the weights is more important than the size of the network, IEEE Trans. Inform. Theory, March 1998, to appear.

[3] P.L. Bartlett, S.R. Kulkarni, Signal restoration for general function classes, Tech. Report, 1998, in preparation.

[4] P.L. Bartlett, S.R. Kulkarni, A universally consistent estimator with deterministic data conditions, Tech. Report, 1998, in preparation.

[5] B. Dasgupta, E.D. Sontag, Sample complexity for learning recurrent perceptron mappings, IEEE Trans. Inform. Theory 42 (1996) 1479–1487.

[6] P.W. Goldberg, M.R. Jerrum, Bounding the Vapnik–Chervonenkis dimension of concept classes parametrized by real numbers, Mach. Learning 18 (2/3) (1995) 131–148.

[7] D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, Inform. Comput. 100 (1992) 78–150.

[8] A.N. Kolmogorov, V.M. Tihomirov, ε-entropy and ε-capacity of sets in functional spaces, AMS Translations Ser. 2, 17 (1961) 277–364.

[9] S.R. Kulkarni, S.E. Posner, Rates of convergence of nearest neighbor estimation under arbitrary sampling, IEEE Trans. Inform. Theory 39 (1995) 1028–1039.

[10] S.R. Kulkarni, S.E. Posner, Universal prediction of nonlinear systems, in: Proc. 34th Conf. on Decision and Control, 1995, pp. 4024–4029.

[11] W.S. Lee, P.L. Bartlett, R.C. Williamson, Efficient agnostic learning of neural networks with bounded fan-in, IEEE Trans. Inform. Theory 42 (1996) 2118–2132.

[12] D. Pollard, Convergence of Stochastic Processes, Springer, Berlin, 1984.

[13] G. Zames, J.G. Owen, A note on metric dimension and feedback in discrete time, IEEE Trans. Automat. Control 38 (4) (1993) 664–667.