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

The Complexity of Model Classes and Smoothing Noisy Data

Peter L. Bartlett
Department of Systems Engineering
RSISE, Australian National University
Canberra, 0200 Australia

Sanjeev R. Kulkarni
Department of Electrical Engineering
Princeton University
Princeton, NJ 08544 USA

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.

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 that are necessary and sufficient to allow smoothing in terms of the class of models F. 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 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 k-th 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 computational learning theory (see, for example, [1]). However, most of the recent work in computational learning theory 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. Our formulation is also related to other work on output prediction in a systems

0-7803-3590-2/96 $5.00 © 1996 IEEE

context (e.g., see [2] and references contained therein). However, our success criterion is different and, 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 identification that uses notions of covering numbers and metric dimension to measure the complexity of identification (e.g., see [3] 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 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 define the family of norms over 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 let $e = (e_1, \ldots, e_n) \in \mathbb{R}^n$ 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 $||f_n - f^*||_n^2 \leq \epsilon$.*

For an input sequence $x = (x_1, x_2, \ldots) \in X^{\mathbb{N}}$ and a noise sequence $e = (e_1, e_2, \ldots) \in \mathbb{R}^{\mathbb{N}}$, we say that an estimator smooths 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 $x = (x_1, \ldots, x_n) \in X^n$ be an input sequence and let $e = (e_1, \ldots, e_n) \in \mathbb{R}^n$ 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*

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

For an input sequence $x = (x_1, x_2, \ldots) \in X^{\mathbb{N}}$ and a noise sequence $e = (e_1, e_2, \ldots) \in \mathbb{R}^{\mathbb{N}}$, 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*

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

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 necessary and 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 $\epsilon > 0$, and an $f^* \in F$ such that, for infinitely many values of $n \in \mathbb{N}$, there is an f in F satisfying*

$$\begin{aligned}\epsilon \quad &\le \quad \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - f^*(x_i))^2 \\ &\qquad\le \frac{2}{n}\sum_{i=1}^{n}(f(x_i) - f^*(x_i))e_i. \quad (1)\end{aligned}$$

2. *If some empirical estimator for F fails to optimize over F with respect to x and e, then there is an $\epsilon > 0$ and an m in $\mathbb{R}^X$ (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 $\|f^* - m\|_n = \inf_{g\in F} \|g - m\|_n$,*

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

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

Furthermore, the reverse implications are also true in both cases if the noise satisfies $\limsup_{n\to\infty} \|e\|_n < \infty$ 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 $f_1' - f_2' = \tau(f_1 - f_2)$.

Notice that the condition on F that suffices for the converse results 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 $\epsilon > 0$ and an f^* in F such that, for infinitely many values of n, there is an f in F with $\|f - (f^* + e)\|_n^2 \le \min_{g\in\bar{F}_{|_x}} \|g - (f^* + e)\|_n^2$ and $\|f - f^*\|_n^2 \ge \epsilon$. The second inequality is equivalent to the first inequality in (1), and the first inequality implies $\|f - (f^* + e)\|_n^2 \le \|e\|_n^2$, which is equivalent to the second inequality in (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 (1) is equivalent to

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

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

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

It follows that

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

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

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

and $\|f - f^*\|_n^2 \ge \epsilon$. 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

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

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

$$\begin{aligned}\|g^* - f^*\|_n \quad &\ge \quad \|f^* - y\|_n - \inf_g \|g - y\|_n \\ &= \quad \frac{\|f^* - y\|_n^2 - \inf_{g\in F}\|g - y\|_n^2}{\|f^* - y\|_n + \inf_{g\in F}\|g - y\|_n} \\ &\ge \quad \frac{\beta/2}{2\|e\|_n}.\end{aligned}$$

So some empirical estimator fails.

2. If an empirical estimator fails to optimize over F, then there is an $\epsilon > 0$ 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 $\|f - m\|_n^2 \ge \|f^* - m\|_n^2 + \epsilon$. These inequalities imply

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

and

$$\sum_{i=1}^{n}(f_i - m_i)^2 - \sum_{i=1}^{n}(f_i^* - m_i)^2 \ge n\epsilon.$$

The proof of the converse is similar to the corresponding proof for the first part of the theorem. ■

4 Smoothing with Random Noise

In this section we consider the case of random noise sequences e that are realizations of a uniformly 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 $\epsilon > 0$, 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 $N(F,n,\alpha)$ 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 $|e_i| \leq M$. Suppose that F is a class of real-valued functions defined on X satisfying $|f(x)| \leq B$ 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 $|m(x_i)| \leq B$, 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*

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

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

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

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

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

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

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

Hoeffding's inequality (see for example [4]) shows that this latter probability is no more than

$$\exp(-2\epsilon^2 n/(B^2M^2)),$$

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*

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

The proof is based on that of the main estimation result in [5]. It uses the following consequence of Bernstein's inequality (Lemma 8 in [5]), instead of Hoeffding's inequality.

Lemma 7 ([5]) *For i.i.d. random variables $V_1, \ldots, V_n$ satisfying $|V_i| \leq K$, $EV_i \geq 0$, and $EV_i^2 < K_2EV_i$ for $i = 1, \ldots, n$ with $K_2 \geq 1$, we have*

$$\Pr\left(\frac{E\left(\frac{1}{n}\sum_i V_i\right) - \frac{1}{n}\sum_i V_i}{\nu + E\left(\frac{1}{n}\sum_i V_i\right)} \geq \alpha\right) \leq \exp\left(\frac{-3\alpha^2\nu n}{2(K_1+K_2)}\right).$$

Proof: (of Theorem 6) If some f in F has $||f-y||_n$ minimized and $||f-m||_n^2 \geq ||f^*-m||_n^2 + \epsilon$, then that f also satisfies

$$E\left(||f-y||_n^2 - ||f^*-y||_n^2\right) \geq \epsilon$$

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

$$E\left(||f-y||_n^2 - ||f^*-y||_n^2\right) \geq \epsilon + \alpha\left(||f-y||_n^2 - ||f^*-y||_n^2\right).$$

Set $\alpha = 2$ and choose an $\hat{f}$ in an ϵ_0-cover of $F_{|_x}$ such that $||f - \hat{f}||_n \leq \epsilon_0$ (ϵ_0 will be chosen later). Then for any y it is easy to show that

$$||f-y||_n^2 - 2(2B+M)\epsilon_0 \leq ||\hat{f}-y||_n^2 \leq ||f-y||^2 + (4B+2M+\epsilon_0)\epsilon_0.$$

If we set $\epsilon_0 = \epsilon/(28B+12M)$ then, provided $\epsilon \leq B^2$, we have that

$$E\left(||\hat{f}-y||_n^2 - ||f^*-y||_n^2\right) \geq \epsilon/2 + 2\left(||\hat{f}-y||_n^2 - ||f^*-y||_n^2\right).$$

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 Lemma 14 in [5]). Substituting $\alpha = 1/2$ and $\nu = \epsilon/2$ into Lemma 7 shows that the probability that an empirical estimator does not ϵ-optimize over F is no more than

$$\begin{aligned} & N\left(F, n, \frac{\epsilon}{28B + 12M}\right) \times \\ & \qquad \Pr\left(E\left(\frac{1}{n}\sum_i V_i\right) \geq \frac{2}{n}\sum_i V_i + \epsilon/2\right) \\ \leq \; & N\left(F, n, \frac{\epsilon}{28B + 12M}\right) \exp\left(\frac{-\epsilon n}{54(B + M)^2}\right). \end{aligned}$$

■

Clearly, corresponding results for smoothing (with no modelling 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 $\mathbb{R}^\infty = \{(u_1, u_2, \ldots) : u_i \in \mathbb{R},\ i \geq 0\}$. 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 $|e_i| \leq M$ and $Ee_i = 0$. Let F_k be a subset of the set of k-th order linear systems. Let $u_1, \ldots, u_n$ be a real sequence satisfying $|f(u_1, \ldots, u_i)| \leq B$ for all $i \leq n$ and all f in F_k. Let the real numbers $m_1, \ldots, m_n$ satisfy $|m_i| \leq B$ for all i. Then with probability $1 - \delta$ over the noise sequence e, the output f of an empirical estimator satisfies*

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

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 [4] imply that the covering number of this class satisfies $N(F, n, \epsilon) \leq (B/\epsilon)^{ck \log n}$ (Dasgupta and Sontag [7] 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 $\sum_{i=0}^{\infty} |h_i| \leq S$. Suppose that $e \in \mathbb{R}^n$ is a realization of an i.i.d. stochastic process satisfying $|e_i| \leq M$ and $Ee_i = 0$, and that $u_1, \ldots, u_n$ and $m_1, \ldots, m_n$ are real sequences satisfying $|u_i| \leq B$ and $|m_i| \leq B$ for all i. Then with probability $1 - \delta$ over the noise sequence, the output f of an empirical estimator satisfies*

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

The proof uses ideas from [5] and [8]. It needs the following approximation result (which Barron in [9] 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 $|f(x)| \leq 1$. 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*

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

Proof: **(of Theorem 9)** Fix an input sequence u. Let μ be the uniform distribution on $\{(u_1, u_2, \ldots, u_i) :$

$i = 1, \ldots, n\}$. Then for any $f \in F_S$, we have

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

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

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

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

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

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

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

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

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

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 with probability $1 - \delta$ over the noise sequence e, the output f of an empirical estimator satisfies

$$\|f - m\|_n^2 \leq \inf_{g \in F} \|g - m\|_n^2 + \frac{c(L + M)^2}{n} \left(\phi^{-1}((L + M)/n) + \log(1/\delta) \right).$$

Proof Sketch: 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 $[-1, 1]^k$. Standard bounds on covering numbers with respect to the infinity norm for this class imply 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.

Acknowledgements

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

References

[1] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[2] S.R. Kulkarni and S.E. Posner. Universal prediction of nonlinear systems. In *Proc. IEEE CDC*, pages 4024–4029, 1995.

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

[4] D. Pollard. *Convergence of Stochastic Processes.* Springer-Verlag, 1984.

[5] W.S. Lee, P.L. Bartlett, and R.C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, (to appear), 1996.

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

[7] B. Dasgupta and E.D. Sontag. Sample complexity for learning recurrent perceptron mappings. Technical Report 95-17, DIMACS, Rutgers University, 1995.

[8] P.L. Bartlett. Pattern classification in neural networks: the size of the weights is more important than the size of the network. Technical report, Australian National University, April 1996.

[9] A. R. Barron. Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory*, 39:930–945, 1993.