SIAM J. Control Optim.
Vol. 43, No. 3, pp. 872–898

LEARNING COMPLEXITY DIMENSIONS FOR A CONTINUOUS-TIME CONTROL SYSTEM*

PIRKKO KUUSELA[†], DANIEL OCONE[†], AND EDUARDO D. SONTAG[†]

Abstract. This paper takes a computational learning theory approach to a problem of linear systems identification. It is assumed that inputs are generated randomly from a known class consisting of linear combinations of k sinusoidals. The output of the system is classified at some single instant of time. The main result establishes that the number of samples needed for identification with small error and high probability, independently from the distribution of inputs, scales polynomially with n, the system dimension, and logarithmically with k.

Key words. linear systems identification, learning theory, VC dimension

AMS subject classifications. 68Q32, 68Q17, 93B30, 93C05

DOI. 10.1137/S0363012901384302

1. Introduction.

Systems identification may be regarded as an instance of the general problem of "learning" an unknown function. Computational learning theory (CLT), which provides a theory for understanding the complexity of a learning problem, can then be used to obtain new insight into identification; conversely, the input/output maps associated to systems theory supply interesting new families of classifiers to consider in CLT. The early paper by Ljung [16] already explored the connection between CLT and identification. Independently, the papers [6, 7] had already developed learning theory complexity results for discrete-time linear systems acting on finite-window data. However, continuous-time linear systems have not been much explored from the CLT viewpoint. The immediate problem is that even for finite-length inputs, the family of maps associated to a continuous-time linear system is not "learnable" in the precise mathematical sense defined in probably approximately correct (PAC) learning theory; for continuous-time systems, the Vapnik–Chervonenkis (VC) dimension, which measures the learning rate, is infinite. This was proved in [23] and, alternatively, can be derived by applying the discrete-time results from [6, 7] at higher and higher sampling rates. However, if we have prior information on the system we wish to learn, or if we are interested in less than the full input/output map, we can ask if the identification problem becomes learnable in the CLT setting, and if so, how many samples are needed. This is the issue motivating the work of this paper.

In practice, it is not necessary or feasible to learn how a linear system classifies every continuous-time input; it usually suffices to know how it acts on a sufficiently rich class. In this paper we consider continuous-time systems acting on linear combinations of k sinusoidals, which constitute a class of inputs used in assessing frequency responses. The parameter k will then measure the richness of the class. To keep things

*Received by the editors January 30, 2001; accepted for publication (in revised form) February 6, 2004; published electronically September 18, 2004.
http://www.siam.org/journals/sicon/43-3/38430.html

[†]Department of Mathematics, Rutgers, The State University of New Jersey, Piscataway, NJ 08854-8019 (pirkko.kuusela@hut.fi, ocone@math.rutgers.edu, sontag@math.rutgers.edu). The first author is now at Networking Laboratory, Helsinki University of Technology, P.O. Box 3000, FIN-02015 HUT, Finland; research financed in part by grants from the Academy of Finland, the Finnish Academy of Sciences and Letters, and the Finnish Cultural Foundation. The third author was supported in part by U.S. Air Force grant F49620-01-1-0063 and NSF grant CCR-0206789.

as simple as possible, we focus our attention on two cases: (a) the map that gives the sign of the output, observed at a single instant of time, namely, zero-one classifiers; and (b) the full output observed at a single instant of time. For the learning theory framework, we have opted for the cleanest setting. In the training stage, inputs are generated randomly from our class of sinusoidals, and our object is to classify further randomly drawn inputs correctly with high probability.

Variations on the above set-up, such as allowing selection of inputs (called active learning) or modeling noise in the observations, can also be formulated. It is important to emphasize that, in this work, we ignore observations at time instants other than the last. This simplifies the problem and makes it easier to formulate our question in the framework of learning theory. However, a full treatment of the identification question in that framework will require further substantial research. One possibility is to see the additional information afforded by data at intermediate instants as "side information" available to the learner. In general, one expects side information to speed learning, so our results here will bound any learning rates that incorporate it. However, quantifying the advantage of side information in a general learning theory framework is a very challenging problem. A preliminary study of how side information can affect learning rates is carried out in [14, 15], in a simple model of learning intervals. One insight of this study is that the side information advantage very much depends on the probability distributions assumed on the training data and the structure of the problem. Because of this, the possibility of distribution free, model-independent side information improvement in learning rates, based on general complexity measures such as VC dimension, is not even clear; in any case, it is an open and difficult problem.

In summary, the system identification problem is posed as a noise-free parametric function identification problem with observations $\{(G_1, S_1), \ldots, (G_n, S_n)\}$, where $G_1, \ldots, G_n$ are independent random variables defining the input as a linear combination of basis input functions $\omega = (\omega_1, \ldots, \omega_k)^T$ and $S_i = \text{sign}(\Phi_\Sigma(G_i\omega; 1))$, where sign $z = 0$ if $z \leq 0$ and sign $z = 1$ if $z > 0$. Here $\Sigma(u; t)$ is the output of the linear system $\Sigma = (A, B, C, x^0)$ at time t when the input is u.

Our results show that the upper bound of samples needed to learn (i.e., identify) Σ in the above setting with a small error and high probability, independently of the distribution of G_i, scales logarithmically with the "bandwidth" k. The sample bound is analogous to the discrete-time case, in which k appeared as the length of the window employed. Also, we provide lower bounds on the number of samples needed. Hence the results can also be seen as unlearnability results where the difficulty arises from the richness of the input signals. Our second class of results concerns systems, with additional assumptions, in which the full output is observed at some time τ.

Our problem setting is selected so that it corresponds to the most standard learning scenario, which serves as the basic problem. In particular, our focus will be on calculating certain complexity dimensions that determine the number of samples needed to learn, given required accuracy and confidence. The CLT framework can accommodate various learning paradigms, and they share a number of common features. In particular, sample complexities are derived from complexity dimensions. Hence, the complexity dimension estimates from the basic problem can be utilized easily also in modified learning settings. We give more pointers to modified problems in the next section.

For papers combining learning theoretic ideas to control theory, see [11, 9] and references there as well as [12] for several results for nonlinear systems in discrete time. The reader is referred to [19] for results that apply to a class of nonlinear continuous-time systems but which are formulated in terms of learning derivatives evaluated at a particular instant (as opposed to time data).

This paper is organized in a top-down fashion. We give definitions and main results in section 2, and in section 3 we state main upper and lower bounds for the complexity dimensions. After that we concentrate on proving the results; central techniques are discussed in section 4, and proofs are in sections 5 and 6. An example of a class with VC dimension k is given in section 7.

2. Definitions and statement of main results.

The simplest learning setting is concept learning, in which there is some known concept class (e.g., "cars") and some target concept (e.g., "a sports car") we wish to learn from a sequence of N randomly chosen observations. Each observation is classified by some "oracle" that knows the target concept. After N classified observations we are required to form an estimate for the unknown target concept so that with high confidence, specified by parameter δ, the misclassification probability for a future unseen sample is smaller than a given level ϵ. The concept class is *learnable* if we can form an estimate that achieves any given confidence level, δ, and misclassification accuracy, ϵ, by taking enough observations. In this case, the number $s(\epsilon, \delta)$ of observations needed to achieve the confidence and misclassification levels is called the *sample complexity.* With this definition, learnability is equivalent to the finiteness of the VC dimension, which describes the "richness" of the concept class; VC dimension together with confidence parameter δ and accuracy parameter ϵ determine the sample complexity $s(\epsilon, \delta)$.

In the following subsections, we provide formal definitions, frame a linear system learning problem, and state some main results.

2.1. VC dimension and fat-shattering dimension.

We begin by defining the key complexity dimension for this work.

DEFINITION 2.1 (Vapnik–Chervonenkis dimension). *The richness of the collection $\mathcal{C}$ can be measured by its Vapnik–Chervonenkis (VC) dimension introduced in* [20]. *A set $S = \{x_1, \dots, x_n\} \subseteq X$ is said to be* shattered *by $\mathcal{C}$ if, for every subset $B \subseteq S$, there exists a set $A \in \mathcal{C}$ such that $S \cap A = B$. The* VC dimension of *$\mathcal{C}$, denoted VC($\mathcal{C}$), equals the largest integer n such that there exists a set of cardinality n that is shattered by $\mathcal{C}$.*

For example, in $\mathbb{R}^k$ the VC dimension of closed half-spaces through the origin is k [22]. Thus, if $\mathrm{VC}(\mathcal{C}) = d$, $\mathcal{C}$ is not rich enough to distinguish all subsets of any $d+1$ element set, but there is some d element set where subsets can be distinguished. Proving exact values of the VC dimension is hard, and typically one looks for upper and lower bounds for the VC dimension, as is also done in this paper.

For our purposes, it is more convenient to work with shattering in terms of dichotomies, i.e., Boolean-valued maps. We identify subsets of D with Boolean functions $\phi : D \to \{0, 1\}$. Similarly, each set $C \in \mathcal{C}$ gives rise to a Boolean function on X, and intersections $C \cap D$ are restrictions of functions to D. In this language, a subset $D \subset X$ is shattered by $\mathcal{F} := \{\phi;\ \phi : X \to \{0, 1\}\}$ if every dichotomy on D is a restriction to D of some $\phi \in \mathcal{F}$.

The VC dimension characterizes learnability of $\{0, 1\}$-valued functions, as formulated in section 2.2. For learning real-valued functions we look for a generalization of the VC dimension with similar properties. One such generalization is the pseudodimension. Unfortunately, pseudodimension does not share the property the VC dimension has; there are learnable function classes with infinite pseudodimension; see [21, p. 206] and [3].

DEFINITION 2.2 (pseudodimension with respect to a loss function). *Given a class of functions $\mathcal{F} : X \to Y$ and a loss function $L : Y \times Y \to [0, r]$, we introduce for each $f \in \mathcal{F}$ the function*

$$A_{f,L} : X \times Y \times \mathbb{R} \to \{0,1\}; \quad (x, y, \rho) \mapsto \operatorname{sign}(L(f(x), y) - \rho), \tag{1}$$

and let $A_{\mathcal{F},L}$ denote all such $A_{f,L}$ with $f \in \mathcal{F}$. The pseudodimension of $\mathcal{F}$ with respect to the loss function L, *$PD[\mathcal{F}, L]$, is defined as*

$$PD[\mathcal{F}, L] := VC(A_{\mathcal{F},L}).$$

Next we define the fat-shattering dimension that corresponds to shattering with fixed "margin" γ. Both the pseudodimension and the fat-shattering dimension can be used to bound certain covering numbers, and in this sense they act like the VC dimension. Moreover, the fat-shattering dimension gives upper and lower bounds for covering numbers of function classes, and the finiteness of the fat-shattering dimension can characterize learnability (see [1] and [2]).

DEFINITION 2.3 (fat-shattering dimension). *Let $\mathcal{F}$ be a set of real-valued functions. We say that a set of points X is* γ-shattered *by $\mathcal{F}$ if there are real numbers r_x indexed by $x \in X$ such that for all binary vectors b_x indexed by X, there is a function $f_b \in \mathcal{F}$ satisfying*

$$\begin{aligned} f_b(x) &\geq r_x + \gamma \quad && \text{if } b_x = 1 \text{ and} \\ f_b(x) &\leq r_x - \gamma \quad && \text{otherwise.} \end{aligned}$$

The fat-shattering dimension $\mathrm{fat}_\gamma(\mathcal{F})$ *is a function from positive real numbers to integers which maps a value γ to the size of the largest fat-shattered set if it is finite or infinity otherwise.*

The shattering dimension when the margin γ equals 0 *is called the* pseudodimension, *and it is denoted by $PD(\mathcal{F})$. Clearly, for all $\gamma > 0$, $\mathrm{fat}_\gamma(\mathcal{F}) \leq PD(\mathcal{F})$.*

2.2. Learning. In this section we discuss the learning setting more formally, beginning with a general introduction to classification problems. In section 2.3.1 we also indicate briefly modified learning settings that can be relevant in control problems.

Assume that a set X, to be called the *input space*, is given together with a collection $\mathcal{C}$ of mappings $X \to \{0,1\}$.[1] Let W be the set of all sequences

$$w = (u_1, \phi(u_1)), \ldots, (u_s, \phi(u_s))$$

over all $s \geq 1$, $(u_1, \ldots, u_s) \in X^s$, and let $\phi \in \mathcal{C}$. An *identifier* is a map $\psi : W \to \mathcal{C}$. The value of ψ on a sequence w above is denoted as ψ_w instead of $\psi(w)$. The "error" of ψ is the probability that ψ will misclassify a future sample. More formally, the *error of ψ with respect to a probability measure P on X, a $\phi \in \mathcal{C}$, and a sequence* $(u_1, \ldots, u_s) \in X^s$, is

$$\mathrm{Err}(P, \phi, u_1, \ldots, u_s) := P\{u \in X;\ \psi_w(u) \neq \phi(u)\}.$$

[1] The set X is assumed to be either countable or a Euclidean space, and the maps in $\mathcal{C}$ are assumed to be measurable. In addition, a mild regularity assumption called "permissibility" is needed so that all sets appearing below are measurable; for further discussion on the topic, see an appendix in [17]. In our context the measurability assumptions are satisfied.

The class $\mathcal{C}$ is said to be PAC *learnable* if there is some identifier ψ with the following property: For each *accuracy parameter* $\epsilon > 0$ and *confidence parameter* $\delta > 0$ there is some s so that, for every probability P and every $\phi \in \mathcal{C}$,

$$P^s\{(u_1, \ldots, u_s) \in X^s;\ \mathrm{Err}(P, \phi, u_1, \ldots, u_s) > \epsilon\} < \delta,$$

where P^s is the s-fold product of P. In the learnable case, the function $s(\epsilon, \delta)$ which provides the smallest s achieving for any positive ϵ and δ is called the *sample complexity.* It can be proved that learnability is equivalent to the finiteness of the *VC dimension* $\mathrm{VC}(\mathcal{C})$ of the class $\mathcal{C}$. Moreover, for learning algorithms that classify the observed samples correctly, the sample complexity is bounded by [18]

$$s(\epsilon, \delta) \leq \max\left\{\frac{1}{\epsilon(1-\sqrt{\epsilon})}\left(2\mathrm{VC}(\mathcal{C})\ln\left(\frac{6}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right), \frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right)\right\}.$$

In addition, there is a similar lower bound for the sample complexity.

Classification may be viewed as a problem of identifying systems with binary outputs. More generally, we introduce a problem of identification for systems having bounded outputs ($[0,1]$-valued, for technical reasons) via an L^1-error, following [4, 5] (for similar statements with L^2-error see [2]). Denote by $\mathcal{F}$ a class of mappings from X to $[0,1]$.

By definition, an *identifier* is a mapping from $\cup_{s\in\mathbb{N}}(X \times [0,1])^s$ to $[0,1]^X$. Such a map takes as data a sequence of labeled samples and produces a hypothesis. If h is a $[0,1]$-valued function defined on X and P is a probability measure over $X \times [0,1]$, we define the *error of* h *with respect to* P as

$$\mathrm{Er}_{\mathrm{P}}(h) := \int_{X\times[0,1]} |h(u) - y|\, dP(u, y).$$

For $\epsilon > 0$ and $\delta > 0$ we say that an identifier ψ (ϵ, δ)*-learns in the agnostic sense with respect to* $\mathcal{F}$ *from* s *examples* if, for all distributions P on $X \times [0,1]$,

$$P^s\{w;\ \mathrm{Er}_{\mathrm{P}}(\psi_w) \geq \inf_{f\in\mathcal{F}} \mathrm{Er}_{\mathrm{P}}(f) + \epsilon\} < \delta.$$

Similarly, for $\epsilon > 0$ the function class $\mathcal{F}$ is said to be ϵ*-agnostically learnable* if there is a function $s_0 : (0,1) \to \mathbb{N}$ such that, for all $0 < \delta < 1$, there is an identifier ψ which (ϵ, δ)-learns in the above sense with s_0 samples. In addition, if the identifier always chooses a hypothesis from $\mathcal{F}$, we say that $\mathcal{F}$ is *properly* ϵ*-agnostically learnable.*

For learning $[0,1]$-valued functions, a sample complexity result may be stated in terms of the fat-shattering dimension. For $\epsilon > 0$ and $\delta > 0$ there is an identifier ψ that properly (ϵ, δ)-learns in the agnostic sense with respect to $\mathcal{F}$ from [4, 5]

$$\frac{4}{\alpha^2}\left(\frac{6d}{\ln 2}\ln\frac{7}{\alpha}\left(\frac{336e}{\alpha^3\ln 2}\ln\frac{7}{\alpha}\right) + \ln\frac{8}{\delta}\right) = O\left(\frac{1}{\alpha^2}\left(d\log^2\frac{1}{\alpha} + \log\frac{1}{\delta}\right)\right)$$

samples, where $0 < \alpha < \epsilon/4$ is chosen so that $d = \mathrm{fat}_{\epsilon/4-\alpha}(\mathcal{F})$ is finite. The quantity $\mathrm{fat}_\gamma(\mathcal{F})$ is called the *fat-shattering dimension* of the class $\mathcal{F}$, and it measures the richness of the class $\mathcal{F}$ with scale γ.

The sample complexity results show us that the difficulty of system identification in the learning theoretic setting can be analyzed by studying various complexity dimensions, and deriving bounds on the complexity dimension is the main focus of this paper.

2.3. Linear systems. In the context of learning we discuss continuous-time linear control systems

$$\dot{x} = Ax + Bu, \qquad x(0) = x^0, \qquad y = Cx, \tag{2}$$

where A, B, and C are $n \times n$, $n \times m$, and $p \times n$ real matrices, and the time interval is [0,1]. We study sign-observations (see [13] for related work in control theory)

$$\text{sign } y(1) = (\text{sign } y_1(1), \dots, \text{sign } y_p(1))^T ,$$

where sign $z = 0$ if $z \leq 0$, sign $z = 1$ if $z > 0$, and ^{T} stands for the transpose. For scalar observations this is a classification problem; each output is classified as either 0 or 1. The value of the final time plays no role in the results and is taken to be 1 for notational convenience.

Unlike the VC dimension associated to discrete-time linear systems [6, 12], the VC dimension of the classification problem for continuous-time control systems is unbounded [23], even when $n = 1$, and the identification problem is not learnable in the sense discussed earlier. Therefore, we restrict the class of admissible controls in order to achieve a bound for the VC dimension. We consider controls $u = (u_1, \dots, u_m)$ such that $u = G\omega$, where G is an $m \times k$ matrix that parameterizes the control. The set of basis input functions $\Omega = \{\omega_1, \dots, \omega_k\}$ is fixed. The bounds for the VC dimension or other complexity dimensions will depend on the properties of the set Ω.

For scalar inputs (i.e., $m = 1$) the VC dimension associated to the mapping from inputs G to scalar sign-observations is bounded by k, which in fact can be very large in applications.[2] However, by considering band-limited controls a better bound can be achieved. In this work we consider the set of basis input functions

$$\begin{aligned} \Omega = \Big\{\omega_1, \dots, \omega_k;\ & \omega_1, \dots, \omega_k \text{ linearly independent and} \\ & \omega_j = t^{\ell_j} e^{\alpha_j t} \sin(\beta_j t) \text{ or } \omega_j = t^{\ell_j} e^{\alpha_j t} \cos(\beta_j t) \\ & \text{with } \ell_j \in \mathbb{N}, \alpha_j, \beta_j \in \mathbb{R}, j = 1, \dots, k\Big\}, \end{aligned} \tag{3}$$

and let

$$\ell_{\max} = \max\{\ell_1, \dots, \ell_k\}. \tag{4}$$

The results in this paper hold with straightforward modifications if the basis input functions $\omega_j, j = 1, \dots, k$, are, for example, linear combinations of functions of the above form.

Definition 2.4 (sign system concept class, $\mathcal{C}_{m,p}$). *Order the set of basis input functions Ω and denote $\omega = (\omega_1, \dots, \omega_k)^T$. Let*

$$X_\Omega = \{G\omega : [0,1] \to \mathbb{R}^m;\ G \in \mathbb{R}^{mk}\},$$

and for each linear system $\Sigma = (A, B, C, x^0)$ of dimension n define the mapping $\Phi_\Sigma : X_\Omega \to \mathbb{R}^p$ by $\Phi_\Sigma(G\omega) = y(1)$, where $y(1)$ is the solution of Σ with control $u = G\omega$. Similarly, we define the mapping for sign-observations as

$$S_\Sigma : X_\Omega \to \{0,1\}^p, \quad G\omega \mapsto \text{sign}(\Phi_\Sigma(G\omega)).$$

[2] This bound is tight; we give an example of a function class Ω for which the associated VC dimension is indeed k (see section 7).

We call the class of above mappings the sign system concept class*, $\mathcal{C}_{m,p} = \{S_\Sigma;\ \Sigma$ linear system of dimension $n\}$.*

We formulate two theorems about bounding sample complexities as main results. They are immediate corollaries of learning complexity bounds proved in this paper.

THEOREM 2.5 (sample complexity for concept learning). *For sign systems concept class $\mathcal{C}_{m,1}$ with scalar observations, i.e., $p = 1$, the sample complexity $s(\epsilon, \delta)$ for identifiers which agree with the observed sample can be bounded as*

$$s(\epsilon,\delta) \le \max\left\{ \frac{1}{\epsilon(1-\sqrt{\epsilon})}\left(2\,VC(\mathcal{C}_{m,1})\ln\left(\frac{6}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right), \frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right)\right\},$$

where

$$VC(\mathcal{C}_{m,1}) \le 2(2mn^2+4n+1)\log_2[8e(8mn^2k(n+\ell_{\max})+1)(2nk+2(1+2k)^n)]$$

and $\ell_{\max}$ is given by (4).

Sketch of proof. The VC dimension bound is based on the observation that, due to the structure of the input, Ω_Σ can be written as a function of the parameters of Σ which is piecewise rational. (The parameterization is derived from the eigenstructure of matrix A.) The complexity upper bound utilizes the Goldberg–Jerrum bound. (Matching lower bounds are based on the notion of dual VC dimension and axis shattering.) □

In terms of n (the dimension of the state space) and k (the bandwidth), the upper bound for the VC dimension is of the form $O(n^3 \log_2(nk))$. The next section states also a corresponding VC dimension lower bound, in terms of the bandwidth, of the form $O(\log(k))$, and, together with a lower bound for the sample complexity, this provides an estimate for the number of samples needed in learning. In particular, in a typical setting of fairly small system dimension n and large bandwidth k, the $\log k$ bound is a clear improvement over the linear bound given by elementary analysis.

For learning $[0,1]$-valued functions the role of the VC dimension is replaced by other complexity dimensions such as the pseudodimension or the fat-shattering dimension that give upper bounds on sample complexities in the corresponding learning paradigms. In the setting of learning $[0,1]$-valued functions we consider the time interval $[0,\tau]$, with $\tau > 1$ in order to show the impact of the final time on the sample complexity.

For the system

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx, \end{aligned} \tag{5}$$

we can write $Ce^{At}B = [\gamma_1, \dots, \gamma_m]$, where each γ_i is a linear combination of n functions $\xi_1, \dots, \xi_n$. Each ξ_i is of the form $t^\ell e^{at}\sin(bt)$ or $t^\ell e^{at}\cos(bt)$ with $\ell \in \{0, \dots, n-1\}$ and $a+ib$ an eigenvalue of A. Assume that A has a fixed Jordan block structure, and let a_k+ib_k be an eigenvalue of A. We take $\alpha_{11}, \dots, \alpha_{nm}, a_1, b_1, \dots, a_r, b_r$ to be the system parameters, where $\gamma_i(t) = \sum_{j=1}^n \alpha_{ij}\xi_j(t)$ for $i = 1, \dots, m$ and $a_1, b_1, \dots, a_r, b_r$ are n eigenvalue parameters. For example, the eigenvalue parameters for a real 4×4 matrix A with eigenvalues $a_1 \pm b_1 i$, a_2, and a_3 would be a_1, b_1, a_2, and a_3. Similarly, the eigenvalue parameters with purely complex eigenvalues $a_1 \pm b_1 i$ and $a_2 \pm b_2 i$ would be a_1, b_1, a_2, and b_2, whereas real eigenvalue parameters would be listed as a_1, a_2, a_3, and a_4.

Let $U \subset \mathbb{R}^{mk}$ be a bounded set. Define a mapping $F : \lambda \times U \to \mathbb{R}$ such that $F(\lambda, u) = y(\tau)$, where $\tau \geq 1$ and $y(\tau)$ is a solution of (5) with system parameters λ and initial condition $x(0) = 0$.

Definition 2.6. *Assume that the system $\dot{x} = Ax + Bu$, $y = Cx$, $x(0) = 0$ can be parameterized by $\lambda \in \mathbb{R}^{n(m+1)}$ as above and $\|\lambda\|_\infty = \max_{1 \leq i \leq n(m+1)} \lambda_i < 1$. Let $F(\lambda, u) = y(\tau)$ be the solution of* (5) *with system parameters λ and control $u = (u_1, \dots, u_m) \in U = \{u = (u_1, \dots, u_m);\ \int_0^\tau |u_i(t)|dt \leq M,\ i = 1, \dots, m\}$. Denote $B_\infty^k(c) := \{x \in \mathbb{R}^k;\ \|x\|_\infty < c\}$ and define*

$$\mathcal{F}_B = \{F(\lambda, \cdot) \colon U \to \mathbb{R};\ \lambda \in B_\infty^{n(m+1)}(1)\}.$$

A suitable learning notion for the above function class $\mathcal{F}_B$ is the proper agnostic learning, defined formally in section 2.2. The related sample complexity result is as follows.

Theorem 2.7 (sample complexity for proper agnostic learning). *Let $1/4 > \kappa > 0$ be an arbitrary real number; then the class $\mathcal{F}_B$, given in Definition* 2.6*, is properly agnostically learnable from*

$$O\left(\frac{1}{\epsilon^2}\left(\mathrm{fat}_{(1/4-\kappa)\epsilon}(\mathcal{F}_B)\log^2\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$$

samples, where

$$\mathrm{fat}_{(1/4-\kappa)\epsilon}(\mathcal{F}_B) \leq (m+1)n\log_2\left\lfloor\frac{n^2 m \tau^n e^\tau k M}{(1/4-\kappa)\epsilon}\right\rfloor,$$

and M is a constant satisfying

$$\int_0^\tau |u_i(\tau - t)|dt \leq kM$$

for all $i = 1, \dots, m$. In the above, $\lfloor x \rfloor$ stands for the integer part of x. If the inputs are of the form $u = G\omega$, then also

$$\mathrm{fat}_{(1/4-\kappa)\epsilon}(\mathcal{F}_B) \leq 2(m+4)n\log_2(8e(nmk4(n+\ell_{\max})+1)(2nk+2(2k+1)^n)),$$

where $\ell_{\max}$ is as given by (3) *and* (4).

Sketch of proof. The proof is developed around a Lipschitz bound on Ω_Σ as a function of unknown parameters, which gives the upper term in the fat-shattering bound. The lower term in the bound is in turn a pseudodimension bound that can be derived from an associated VC dimension bound. □

2.3.1. Modified learning settings. In this paper we have opted for the standard setting in CLT in which inputs are random and observations are noiseless. However, CLT can accommodate various modified learning settings that typically share common features with the standard setting.

A paradigm in which a learner can select the samples to be classified is called active learning. The set of PAC learnable concept classes is not enlarged by active learning but, in general, fewer training samples are needed (concept classes that are "dense in themselves" make an exception) [8].

For dealing with the case of noisy observations the reader is referred to [21]. For example, it is shown that learning a concept class with a "noisy oracle" (that makes a mistake with probability $\beta < 1/2$) to accuracy ϵ is the same as learning the same concept class to an accuracy $\epsilon/(1-2\beta)$ with a perfect oracle.

Finally, we have considered the case in which only the final state output is observed; i.e., observation is done only on a single instant of time. However, typically in control applications observations are made at multiple time instances. If one wishes to learn the mapping from inputs to final state outputs, but one can also see intermediate observations, can one learn faster by utilizing this additional information? This has motivated further research by the authors on "learning with side information" [14, 15]. The main problem is that in the control problem considered the samples become dependent, which poses a challenge in the theory of learning.

3. Complexity dimensions; main upper and lower bounds.

3.1. Bounds. We begin by stating bounds in the easiest learning setting, i.e., classifying the final state observations as either 0 or 1.

THEOREM 3.1 (VC dimension upper bound, $p = 1$). *The VC dimension of the sign system concept class $\mathcal{C}_{m,1}$ with scalar observations can be bounded as*

$$VC(\mathcal{C}_{m,1}) \leq 2(2mn^2 + 4n + 1)\log_2[8e(8mn^2k(n + \ell_{\max}) + 1)(2nk + 2(1 + 2k)^n)],$$

where $\ell_{\max}$ is given by (4).

In terms of n (the dimension of the state space) and k (the bandwidth) the upper bound is of the form $O(n^3 \log_2(nk))$.

All VC dimension upper bounds are based on the fact that input basis functions satisfy a certain rationality condition. Remark 5.3 indicates how the bound is formed when the input functions satisfy the more abstract rationality condition. In that case the degrees of the polynomials and the number of polynomial evaluations are different. However, in terms of n and k, the bound is of the same form. VC dimension or pseudodimension bounds stated in this paper can be modified for the rationality condition in the same way.

The lower bound for the VC dimension is in terms of n and k. It holds for linearly independent, continuous basis input functions, and, compared to upper bounds, no particular form of the functions is needed. The bound is obtained by imposing a specific structure on control systems, and a lower bound for a restricted class of control systems provides a lower bound for more general classes.

THEOREM 3.2 (VC dimension lower bound, $m = 1$, $p = 1$).

$$VC(\mathcal{C}_{1,1}) \geq \max\left\{m' \left\lfloor \log_2 \left\lfloor \frac{k}{m'} \right\rfloor \right\rfloor, m'\right\},$$

where $m' = \min\{n, k\}$.

In terms of k the upper and the lower bound match up to a constant. For n and k the lower bound is typically of the form $O(n \log_2(k/n))$. Note that if the system dimension n is small compared to the bandwidth k, the VC dimension upper and lower bounds in Theorems 3.1 and 3.2 become tighter, both being of the form $c \log_2 k$ (with different values of the constant c).

Extending the upper bounds to the case of vector-valued observations can be done in various ways based on the result obtained for scalar observations. For example, we may consider the p-dimensional output as bits representing a number in $\{0, \dots, 2^p - 1\}$ and introduce a loss function for each $f \in \mathcal{C}_{m,p}$ as $L_{0\text{-}1,f}(z, a) = L_{0\text{-}1}(f(z), a) = 1$, when $f(z) \neq a$, and 0 otherwise. We define the VC dimension of the p-dimensional observation as the VC dimension of the above class of loss functions. Modifying the argument used with scalar observations leads to a bound of the following form.

THEOREM 3.3 (VC dimension upper bound).

$$\begin{aligned} VC(\mathcal{C}_{m,p}) &\leq 2(2pmn^2 + 4n + p) \\ &\quad \times \log_2\Big[8e\big(8mn^2k(n+\ell_{\max})+1\big)\big(2^p - 1 + 2p(2k+1)^n + 2nk\big)\Big], \end{aligned}$$

where $\ell_{\max}$ *is given by* (4).

Next we state the main result concerning learnability of the actual input/output mapping, i.e., learning without taking the sign of the final state observation.

DEFINITION 3.4 (control system concept class, $\mathcal{G}_{p,L}$). *Let* $\widetilde{\mathcal{F}} = \{\Phi_\Sigma : X_\Omega \to \mathbb{R}^p;\ \Sigma$ *linear system of dimension* $n\}$, *and define the* control system concept class *as* $\mathcal{G}_{p,L} = A_{\widetilde{\mathcal{F}},L}$, *where* $A_{\widetilde{\mathcal{F}},L}$ *is given by* (1).

Methods for calculating upper bounds for the VC dimension readily extend to the case of pseudodimension with respect to loss that preserves the rationality structure of the output. A typical example is illustrated by the loss function,

$$L(z_1, z_2) = (z_1 - z_2)^2/(1 + (z_1 - z_2)^2), \tag{6}$$

and the following result.

THEOREM 3.5 (pseudodimension upper bound, $p = 1$).

$$PD(\mathcal{G}_{1,L}) \leq 2\left(2mn^2 + 4n + 1\right)\log_2[16e(8mn^2k(n+\ell_{\max})+1)(2nk + 2(2k+1)^n)],$$

where the loss function L is given by (6) *and* $\ell_{\max}$ *is given by* (4).

This differs from the corresponding VC dimension bound only by the maximum degree of the polynomials, which is doubled. Extending this pseudodimension bound for p-dimensional observations can be done naturally by modifying the loss function. Lower bounds for the VC dimension are lower bounds for the pseudodimension as such.

The next results summarize upper bounds for the fat-shattering dimension. We begin by illustrating how the fat-shattering dimension can be bounded for Lipschitz functions in certain cases.

THEOREM 3.6 (fat-shattering bound). *Let* $F(\lambda, u) : \mathbb{R}^k \times U \to \mathbb{R}$ *be such that* $F(\cdot, u)$ *is Lipschitz with constant L, i.e.,* $|F(\lambda_1, u) - F(\lambda_2, u)| \leq L\|\lambda_1 - \lambda_2\|$ *for all* $u \in U$. *For any subset* $B \subseteq \mathbb{R}^k$, *consider the following class of functions:*

$$\mathcal{F}_B = \{F(\lambda, \cdot) : U \to \mathbb{R};\ \lambda \in B\}.$$

Then

$$\begin{aligned} \mathrm{fat}_\gamma(\mathcal{F}_{B^k_\infty(C)}) &\leq k\log_2\left\lfloor\frac{CL}{\gamma}\right\rfloor, \\ \mathrm{fat}_\gamma(\mathcal{F}_{\bar{B}^k_\infty(C)}) &\leq k\log_2\left(1 + \left\lfloor\frac{CL}{\gamma}\right\rfloor\right), \\ \mathrm{fat}_\gamma(\mathcal{F}_{\bar{B}_{2,1}}) &\leq k\log_2\left(C + \frac{L}{\gamma}\right), \end{aligned}$$

where

$$\begin{aligned} B_\infty^k(C) &= \{x \in \mathbb{R}^k;\ \|x\|_\infty = \max_{1\le i\le k} |x_i| < C\}, \\ \bar{B}_\infty^k(C) &= \{x \in \mathbb{R}^k;\ \|x\|_\infty = \max_{1\le i\le k} |x_i| \le C\}, \\ \bar{B}_2^k(C) &= \left\{x \in \mathbb{R}^k;\ \|x\|_2 = \sqrt{\sum\nolimits_{i=1}^k x_i^2} \le C\right\}. \end{aligned}$$

THEOREM 3.7 (fat-shattering bound for a control system). *Assume that the system $\dot{x} = Ax + Bu$, $y = Cx$, $x(0) = 0$, can be parameterized by $\lambda \in \mathbb{R}^{n(m+1)}$ as in Definition* 2.6 *and $\|\lambda\|_\infty < 1$. Let $F(\lambda, u) = y(\tau)$ be the solution with system parameters λ and control $u = (u_1, \dots, u_m) \in U = \{u = (u_1, \dots, u_m);\ \int_0^\tau |u_i(t)|dt \le M,\ i = 1, \dots, m\}$. Define*

$$\mathcal{F}_B = \{F(\lambda, \cdot)\colon U \to \mathbb{R};\ \lambda \in B_\infty^k(1)\}.$$

Then

$$\mathrm{fat}_\gamma(\mathcal{F}_B) \le n(m+1)\log_2 \left\lfloor \frac{n^2 m \tau^n e^\tau M}{\gamma} \right\rfloor.$$

4. Techniques for proving VC dimension results. Our main results are based on the fact that the basis input functions satisfy a certain rationality condition. In this section we first formulate this rationality condition, and then we summarize existing results that are used in proving upper and lower bounds for the complexity dimensions.

We recall briefly the control system setting. We study systems

$$\begin{aligned} &\dot{x} = Ax + Bu, \quad x(0) = x^0, \quad y = Cx, \\ &u = G\omega, \quad \text{and} \quad \omega_j \in \Omega \text{ for } j = 1, \dots, k, \end{aligned}$$

with basis input functions

$$\begin{aligned} \Omega = \Big\{\omega_1, \dots, \omega_k;\ &\omega_1, \dots, \omega_k \text{ linearly independent and} \\ &\omega_j = t^{\ell_j} e^{\alpha_j t} \sin(\beta_j t) \text{ or } \omega_j = t^{\ell_j} e^{\alpha_j t} \cos(\beta_j t) \\ &\text{with } \ell_j \in \mathbb{N}, \alpha_j, \beta_j \in \mathbb{R}, j = 1, \dots, k\Big\}, \end{aligned}$$

such that $\ell_{\max} = \max\{\ell_1, \dots, \ell_k\}$.

DEFINITION 4.1 (rationality condition (RAT)). *Let n be a positive integer. We say that a bounded function $\omega : [0,1] \to \mathbb{R}$ satisfies the rationality condition relative to the class of n-dimensional systems if there exist h polynomial functions $f_1, \dots, f_h : \mathbb{R}^4 \to \mathbb{R}$ and $2\gamma n$ rational functions $r_{ij\tilde{\ell}}$, $i \in \{1,2\}$, $j \in \{1, \dots, \gamma\}$, and $\tilde{\ell} \in \{1, \dots, n\}$, with no poles on subsets $S_{ij\tilde{\ell}}$ of $\mathbb{R}^4$, such that the following properties hold:*

1. *For each $i \in \{1,2\}$, $\tilde{\ell} \in \{1, \dots, n\}$, $\mathbb{R}^4$ is a disjoint union of $S_{i1\tilde{\ell}}, \dots, S_{i\gamma\tilde{\ell}}$.*
2. *Each $S_{ij\tilde{\ell}}$ can be defined in terms of a Boolean expression involving $[f_1 = 0]$, $\dots$, $[f_h = 0]$, where we say that for functions $f_1, \dots, f_h : \mathbb{R}^4 \to \mathbb{R}$, $[f_i = 0]$ has value 1 if $f_i(x_1, x_2, x_3, x_4) = 0$ and 0 otherwise.*

3. *Let* $r_{i\tilde{\ell}} : \mathbb{R}^4 \to \mathbb{R}$*,* $i \in \{1, 2\}$*,* $\tilde{\ell} \in \{1, \dots, n\}$*, be defined as*

$$r_{i\tilde{\ell}}(v) = \begin{cases} r_{i1\tilde{\ell}}(v) & \textit{if } v \in S_{i1\tilde{\ell}}, \\ \vdots & \vdots \\ r_{i\gamma\tilde{\ell}}(v) & \textit{if } v \in S_{i\gamma\tilde{\ell}}; \end{cases}$$

then for each $a, b \in \mathbb{R}$*, and for all* $\tilde{\ell} \in \{1, \dots, n\}$*,*

$$\int_0^1 t^{\tilde{\ell}-1} e^{at} \cos(bt)\omega(t)dt = r_{1\tilde{\ell}}(a, b, e^a \cos b, e^a \sin b),$$
$$\int_0^1 t^{\tilde{\ell}-1} e^{at} \sin(bt)\omega(t)dt = r_{2\tilde{\ell}}(a, b, e^a \cos b, e^a \sin b).$$

We denote by $d_{\max}$ the maximum degree of any polynomial (i.e., $f_1, \dots, f_h$, numerators and denominators of $r_{ij\tilde{\ell}}$'s) appearing in the rationality condition.

Remark 4.2. First, entries of e^{At} are functions of the form $t^s e^{at} \cos(bt)$ and $t^s e^{at} \sin(bt)$. Solving (2) involves convolutions of e^{At} and the basis input functions ω_j, and we require those to be rational functions.

Example 4.3. Let $\omega(t) = \sin(ct)$, with nonzero c. Then

$$\int_0^1 e^{at} \sin(bt)\omega(t)dt = \frac{1}{2}\int_0^1 e^{at} \cos((b-c)t)dt - \frac{1}{2}\int_0^1 e^{at} \cos((b+c)t)dt.$$

After integration this can be split into cases with no poles yielding

$$\int_0^1 e^{at} \sin(bt)\omega(t)dt = \begin{cases} \frac{p(a,b,e^a \cos b, e^a \sin b)}{(a^2+(b+c)^2)(a^2+(b-c)^2)} & \text{if } f_1 \neq 0, f_2 \neq 0, \\ \frac{b - \sin b \cos b}{2b} & \text{if } f_1 = 0, f_2 \neq 0, \\ \frac{\sin b \cos b - b}{2b} & \text{if } f_1 \neq 0, f_2 = 0, \end{cases}$$

where $f_1(a, b, e^a \sin b, e^a \cos b) = a^2 + (b+c)^2$, $f_2(a, b, e^a \sin b, e^a \cos b) = a^2 + (b-c)^2$, and $p(a, b, e^a \cos b, e^a \sin b)$ stands for the polynomial

$$\begin{aligned} &-4abc + 4abce^a \cos b \cos c - 2a^2ce^a \cos c \sin b + 2b^2ce^a \cos c \sin b \\ &\quad -2c^3e^a \cos c \sin b - 2a^2be^a \cos b \sin c - 2b^3e^a \cos b \sin c + 2bc^2e^a \cos b \sin c \\ &\qquad +2a^3e^a \sin b \sin c + 2ab^2e^a \sin b \sin c + 2ac^2e^a \sin b \sin c. \end{aligned}$$

LEMMA 4.4. *Each* $\omega_j \in \Omega$ *given by* (3) *satisfies the rationality condition. Further, the maximum degree of polynomials in (RAT) is at most* $4(n + \ell_{\max})$*, where* $\ell_{\max}$ *is given by* (4)*.*

Review of VC dimension techniques. In the context of control theory it is sometimes easier to work with the dual VC dimension. Assume that a function $F : \Lambda \times X \to \{0, 1\}$ is given. This induces two function classes

$$\mathcal{F} := \{F(\lambda, \cdot) : X \to \{0, 1\};\ \lambda \in \Lambda\}$$

and

$$\mathcal{F}^* := \{F(\cdot, x) : \Lambda \to \{0, 1\};\ x \in X\}.$$

The complexity dimension $\mathrm{VC}(\mathcal{F}^*)$ is called the *dual VC dimension* of $\mathcal{F}$, and it is related to $\mathrm{VC}(\mathcal{F})$ as follows [21]:

$$\mathrm{VC}(\mathcal{F}) \geq \lfloor \log_2 \mathrm{VC}(\mathcal{F}^*) \rfloor, \tag{7}$$

where $\lfloor x \rfloor$ is the integer part of x.

A sharper estimate can be obtained if Λ can be written as a product $\Lambda_1 \times \cdots \times \Lambda_n$. The following construction and result are due to DasGupta and Sontag [6]. We study in particular those dichotomies that are defined on "rectangular" subsets of Λ. Let $L = L_1 \times \cdots \times L_n$ be a subset of Λ such that for each i, $L_i \subset \Lambda_i$ is nonempty. Given any index $1 \leq \kappa \leq n$, a *κ-axis dichotomy* on L is any function $\delta : L \to \{0, 1\}$ which depends only on the κth coordinate; i.e., there is some function $\phi : L_\kappa \to \{0, 1\}$ so that $\delta(\lambda_1, \ldots, \lambda_n) = \phi(\lambda_\kappa)$ for all $(\lambda_1, \ldots, \lambda_n) \in L$. We say that a mapping is an axis dichotomy if it is a κ-axis dichotomy for some κ. A rectangular set L is said to be *axis-shattered* by $\mathcal{F}^*$ if every axis dichotomy is a restriction to L of some function of the form $F(\cdot, x) : \Lambda \to \{0, 1\}$ for some $x \in X$.

Theorem 4.5 (axis-shattering bound [6]). *If $L = L_1 \times \cdots \times L_n \subset \Lambda$ can be axis-shattered and each L_i has cardinality $r_i > 0$, then $\mathrm{VC}(\mathcal{F}) \geq \lfloor \log_2(r_1) \rfloor + \cdots + \lfloor \log_2(r_n) \rfloor$.*

Upper bounds for VC dimensions of concept classes that are obtained by evaluating polynomial equalities and inequalities can be obtained in terms of the number and degrees of the polynomials.

Theorem 4.6 (Goldberg–Jerrum bound [10]). *Given a function $F : \Lambda \times X \to \{0, 1\}$ and the associated concept class $\mathcal{F} := \{F(\lambda, \cdot) : X \to \{0, 1\};\ \lambda \in \Lambda\}$, suppose that $\Lambda = \mathbb{R}^\ell$ and $X = \mathbb{R}^k$. Let F be defined in terms of a Boolean formula involving at most s polynomial equalities and inequalities in $\ell + k$ variables, each polynomial being of degree at most d in λ for all $x \in \mathbb{R}^k$. Then, $\mathrm{VC}(\mathcal{F}) \leq 2\ell \log_2(8eds)$.*

The Goldberg–Jerrum bound is based on a result showing that the number of sign assignments $\{-1, 0, 1\}$ to polynomials cannot grow too quickly.

Theorem 4.7 (see [10]). *Suppose that $f_1, \ldots, f_m$ are polynomials of degree at most d in $n \leq m$ variables. Then the number of distinct vectors*

$$[\mathrm{sign}\ f_1(x), \ldots, \mathrm{sign}\ f_m(x)] \in \{-1, 0, 1\}^m$$

that can be generated by varying x over $\mathbb{R}^n$ is at most $((8edm)/n)^n$.

5. Proofs of VC dimension bounds.

5.1. An upper bound for the VC dimension with scalar observations.

We begin this section by proving Lemma 4.4 stating that the input basis functions satisfy the rationality condition (RAT) and bounding the degrees of polynomials appearing in (RAT). As a proposition we formalize how control systems can be parameterized. After that, as a lemma, we develop an upper bound for the VC dimension induced by the control system (2) with its initial state fixed to be zero. Theorem 3.1 with an arbitrary initial condition is then a simple modification of the argument.

Proof of Lemma 4.4. If $\omega(t) = t^\ell e^{\alpha t} \sin(\beta t)$ or $\omega(t) = t^\ell e^{\alpha t} \cos(\beta t)$ with $\ell \leq \ell_{\max}$, then in place of

$$\int_0^1 t^{\tilde{\ell}} e^{at} \sin(bt)\omega(t)dt \quad \text{or} \quad \int_0^1 t^{\tilde{\ell}} e^{at} \cos(bt)\omega(t)dt, \tag{8}$$

by combining exponents and using sum formulae for sin and cos (see Example 4.3), it is enough to study terms of the form $\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \sin(\tilde{b}t)dt$ or $\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \cos(\tilde{b}t)dt$, where $\tilde{k} \in \{0, \dots, n + \ell_{\max} - 1\}$. In fact, each expression in (8) is one of the following types:

$$\frac{1}{2}\left(\int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \cos\left((b-\beta)t\right) dt - \int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \cos\left((b+\beta)t\right) dt\right),$$
$$\frac{1}{2}\left(\int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \cos\left((b+\beta)t\right) dt + \int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \cos\left((b-\beta)t\right) dt\right),$$
$$\frac{1}{2}\left(\int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \sin\left((b-\beta)t\right) dt - \int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \sin\left((b+\beta)t\right) dt\right),$$
$$\frac{1}{2}\left(\int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \sin\left((b+\beta)t\right) dt + \int_0^1 t^{\tilde{\ell}} e^{\tilde{a}t} \sin\left((b-\beta)t\right) dt\right),$$

where $\tilde{a} = a + \alpha$. Because for $\tilde{k} > 0$

$$\begin{aligned}\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \sin(\tilde{b}t)dt = {} & \frac{e^{\tilde{a}}}{\tilde{a}^2 + \tilde{b}^2}(\tilde{a}\sin\tilde{b} - \tilde{b}\cos\tilde{b}) \\ & - \frac{\tilde{k}}{\tilde{a}^2 + \tilde{b}^2}\int_0^1 t^{\tilde{k}-1} e^{\tilde{a}t}(\tilde{a}\sin(\tilde{b}t) - \tilde{b}\cos(\tilde{b}t))dt\end{aligned} \tag{9}$$

and

$$\int_0^1 e^{\tilde{a}t} \sin(\tilde{b}t)dt = \frac{e^{\tilde{a}}(\tilde{a}\sin\tilde{b} - \tilde{b}\cos\tilde{b}) + \tilde{b}}{\tilde{a}^2 + \tilde{b}^2} \tag{10}$$

and similar formulae for $\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \cos(\tilde{b}t)dt$ hold, we see by induction that the numerator of $\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \sin(\tilde{b}t)dt$ is a polynomial of $\tilde{a}, \tilde{b}, e^{\tilde{a}}\cos\tilde{b}$, and $e^{\tilde{a}}\sin\tilde{b}$. By using sum formulae for sin and cos, the previous expression is in turn a polynomial of $a, b, e^a \cos b$, and $e^a \sin b$ because $\tilde{a} = a + \alpha$ and $\tilde{b} = b \pm \beta$ for some fixed α and β. By similar arguments, the denominator is a polynomial of a and b. Note that, for example, e^α equals a constant times e^a, so this process does not change the degrees of the polynomials.

Further, observe that the denominator of $\int_0^1 t^{\tilde{\ell}} e^{at} \sin(bt)\omega(t)dt$ consists of at most two products of variables a and b of the form $((a+\alpha)^2 + (b \pm \beta)^2)^{\tilde{\ell}+\ell+1}$, and similarly with the $\cos(bt)$ term. Let us index the basis input functions $\omega_1, \dots, \omega_k$ so that ω_κ has parameters α_κ and β_κ. Hence the functions f_i in (RAT), defining the subsets without poles, can be taken as

$$\{(a+\alpha_\kappa)^2 + (b-\beta_\kappa)^2, (a+\alpha_\kappa)^2 + (b+\beta_\kappa)^2\ ;\ \kappa = 1, \dots, k\}.$$

Furthermore, the sets $S_{ij\tilde{\ell}}$ are as simple as

$$\cup_{\kappa=1}^k \{\{(x_1, x_2, x_3, x_4);\ x_1 = -\alpha_\kappa, x_2 = -\beta_\kappa\} \cup \{(x_1, x_2, x_3, x_4);\ x_1 = -\alpha_\kappa, x_2 = \beta_\kappa\}\}$$

and

$$\begin{aligned}\mathbb{R}^4 \setminus \cup_{\kappa=1}^k \{ & \{(x_1, x_2, x_3, x_4);\ x_1 = -\alpha_\kappa, x_2 = -\beta_\kappa\} \\ & \cup \{(x_1, x_2, x_3, x_4);\ x_1 = -\alpha_\kappa, x_2 = \beta_\kappa\}\}.\end{aligned}$$

We turn to estimating the maximum degree of polynomials appearing in (RAT). We already saw that functions f_i are polynomials of degree 2. Equations (9) and (10) show that the degree of numerator is not higher than the one of denominator. We claim that

$$\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \begin{cases} \sin(\tilde{b}t) \\ \cos(\tilde{b}t) \end{cases} dt = \frac{P(2(\tilde{k}+1))}{(\tilde{a}^2+\tilde{b}^2)^{\tilde{k}+1}} \quad \text{for } \tilde{k} = 0, 1, \dots,$$

where $P(2(\tilde{k}+1))$ stands for some polynomial in $\tilde{a}$, $\tilde{b}$, $e^{\tilde{a}}\sin(\tilde{b})$, and $e^{\tilde{a}}\cos(\tilde{b})$ of degree $2(\tilde{k}+1)$. Clearly, the claim is true for $\tilde{k}=0$ by (10), and the inductive argument follows from (9). Assuming the claim is true for $\tilde{k}-1$, we get

$$\begin{aligned}
&\int_0^1 t^{\tilde{k}} e^{\tilde{a}t} \sin(\tilde{b}t) dt \\
&\quad = \frac{P(2)}{\tilde{a}^2+\tilde{b}^2} - \frac{\tilde{k}\tilde{a}}{\tilde{a}^2+\tilde{b}^2}\int_0^1 t^{\tilde{k}-1} e^{\tilde{a}t}\sin(\tilde{b}t)dt + \frac{\tilde{k}\tilde{b}}{\tilde{a}^2+\tilde{b}^2}\int_0^1 t^{\tilde{k}-1}e^{\tilde{a}t}\cos(\tilde{b}t)dt \\
&\quad = \frac{P(2)}{\tilde{a}^2+\tilde{b}^2} - \frac{P(1)}{(\tilde{a}^2+\tilde{b}^2)}\frac{P(2\tilde{k})}{(\tilde{a}^2+\tilde{b}^2)^{\tilde{k}}} + \frac{P(1)}{(\tilde{a}^2+\tilde{b}^2)}\frac{P(2\tilde{k})}{(\tilde{a}^2+\tilde{b}^2)^{\tilde{k}}} \\
&\quad = \frac{P(2)(\tilde{a}^2+\tilde{b}^2)^{\tilde{k}} - 2P(2\tilde{k}+1)}{(\tilde{a}^2+\tilde{b}^2)^{\tilde{k}+1}} = \frac{P(2(\tilde{k}+1))}{(\tilde{a}^2+\tilde{b}^2)^{\tilde{k}+1}},
\end{aligned}$$

and similarly for the $\cos(\tilde{b}t)$ term, concluding the proof of the claim. As a corollary of the claim,

$$\int_0^1 t^{\tilde{\ell}} e^{at}\sin(bt)\omega(t)dt = \frac{P(2(\tilde{k}+1))}{(\tilde{a}^2+(b+\beta)^2)^{\tilde{k}+1}} + \frac{P(2(\tilde{k}+1))}{(\tilde{a}^2+(b-\beta)^2)^{\tilde{k}+1}},$$

where $\tilde{k} = \ell + \tilde{\ell}$ and $\tilde{a} = a + \alpha$.

Hence the maximum degree of denominators of expressions in (8) is $2(\tilde{k}+1) + 2(\tilde{k}+1) = 4(\tilde{k}+1)$ with $\tilde{k} \in \{0, \dots, \ell_{\max}+n-1\}$. Thus the maximum degree of polynomials appearing in the (RAT) is $4(n+\ell_{\max})$. □

The next proposition indicates how control systems are parameterized and later the concept or function classes associated to control systems are obtained by varying the parameter vector.

PROPOSITION 5.1. *Denote the basis input functions by* $\omega = (\omega_1, \dots, \omega_k)^T$, *assume that each* ω_i, $i = 1, \dots, k$, *satisfies the rationality condition (RAT), and let* $\Lambda = \mathbb{R}^{2pn^2m} \times \mathbb{R}^{4n} \times \mathbb{R}^p$. *Then there exists a mapping* $H : \Lambda \times \mathbb{R}^{mk} \to \mathbb{R}^p$ *(depending on* ω*) such that for each* $\Sigma = (A, B, C, x^0)$ *there exists a* $\lambda \in \Lambda$ *satisfying*

$$\Phi_\Sigma(G\omega) = H(\lambda, G) \quad \forall\ G \in \mathbb{R}^{mk}.$$

Proof. Given a system $\Sigma = (A, B, C, x^0)$,

$$\Phi_\Sigma(u) = y(1) = Ce^A x^0 + C\int_0^1 e^{A(1-t)}Bu(t)dt.$$

By an argument based on the real Jordan form of e^{At}, the entries of $e^{A(1-t)}$ are linear combinations of functions of the form $t^{\tilde{\ell}}e^{at}\cos(bt)$ and $t^{\tilde{\ell}}e^{at}\sin(bt)$, where $\tilde{\ell} \in \{0, \dots, n-1\}$ and $a+ib$ is an eigenvalue of A. Hence we define the $2n$ functions $\xi_j(a,b,t) = t^{j-1}e^{at}\cos(bt)$, $\xi_{n+j}(a,b,t) = t^{j-1}e^{at}\sin(bt)$ for $j = 1, \dots, n$.

By the rationality condition (RAT), for all $\ell = 1, \dots, 2n$,

$$\int_0^1 \xi_\ell(a,b,t)\omega_j(t)dt = \frac{\hat{P}_{\ell j}(a,b,e^a\cos b, e^a \sin b)}{\hat{Q}_{\ell j}(a,b,e^a\cos b, e^a \sin b)}$$

for all $a, b \in \mathbb{R}$ and where $\hat{P}_{\ell j}$ and $\hat{Q}_{\ell j}$ are piecewise polynomial expressions.

Let $H(\mathbf{A}, \mathbf{X}, h, G) = (H_1, \dots, H_p)^T$, where for $1 \le \kappa \le p$

$$H_\kappa(\mathbf{A}, \mathbf{X}, h, G) = \sum_{i=1}^m \sum_{r=1}^n \sum_{\ell=1}^{2n} \alpha_{ir\ell\kappa} \sum_{j=1}^k g_{ij} \frac{\hat{P}_{\ell j}(x_{r1}, x_{r2}, x_{r3}, x_{r4})}{\hat{Q}_{\ell j}(x_{r1}, x_{r2}, x_{r3}, x_{r4})} + h_\kappa$$

and

$$\mathbf{A} = (\alpha_{ir\ell\kappa})_{\substack{i=1,\dots,m \\ r=1,\dots,n \\ \ell=1,\dots,2n \\ \kappa=1,\dots,p}}, \qquad \mathbf{X} = (x_{r\eta})_{\substack{r=1,\dots,n, \\ \eta=1,\dots,4}}$$

$$h = (h_1, \dots, h_p)^T, \qquad G = (g_{ij})_{\substack{i=1,\dots,m. \\ j=1,\dots,k}}$$

Next, we relate Φ_Σ and H and we write

$$Ce^{A(1-t)}B = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & & \vdots \\ \gamma_{p1} & \cdots & \gamma_{pm} \end{bmatrix}.$$

We list the eigenvalues of A as $a_r + ib_r$ for $r = 1, \dots, n$ and let $\xi_{r\ell}(t) = \xi_\ell(a_r, b_r, t)$ for $r = 1, \dots, n$ and $\ell = 1, \dots, 2n$. Then there exists some $(\alpha_{ir\ell\kappa})$ such that

$$\gamma_{\kappa i}(t) = \sum_{r=1}^n \sum_{\ell=1}^{2n} \alpha_{ir\ell\kappa} \xi_{r\kappa}(t). \tag{11}$$

Let $\lambda = (\mathbf{A}, \mathbf{X}, h)$, where $\mathbf{A}$ satisfies (11), $\mathbf{X} = (x_{r\eta})$, where $x_{r1} = a_r, x_{r2} = b_r, x_{r3} = e^{a_r}\cos b_r$ and $x_{r4} = e^{a_r}\sin b_r$, and $h = Ce^A x^0$. We claim that

$$H(\lambda, G) = y(1) = \Phi_\Sigma(G\omega) \quad \forall\ G \in \mathbb{R}^{mk}.$$

Note that the κth component of $\Phi_\Sigma(G\omega)$ is given by

$$\begin{aligned}
&\int_0^1 \sum_{i=1}^m \gamma_{\kappa i}(t) u_i(t) dt + h_\kappa \\
&= \sum_{i=1}^m \sum_{r=1}^n \sum_{\ell=1}^{2n} \alpha_{ir\ell\kappa} \int_0^1 \xi_{r\ell}(t) \sum_{j=1}^k g_{ij}\omega_j(t)dt + h_\kappa \\
&= \sum_{i=1}^m \sum_{r=1}^n \sum_{\ell=1}^{2n} \alpha_{ir\ell\kappa} \sum_{j=1}^k g_{ij} \int_0^1 \xi_\ell(a_r, b_r, t)\omega_j(t)dt + h_\kappa \\
&= \sum_{i=1}^m \sum_{r=1}^n \sum_{\ell=1}^{2n} \alpha_{ir\ell\kappa} \sum_{j=1}^k g_{ij} \frac{\hat{P}_{\ell j}(x_{r1}, x_{r2}, x_{r3}, x_{r4})}{\hat{Q}_{\ell j}(x_{r1}, x_{r2}, x_{r3}, x_{r4})} + h_\kappa \\
&= H_\kappa(\mathbf{A}, \mathbf{X}, h, G). \qquad \square
\end{aligned}$$

Next we take $p = 1$ and study the VC dimension of the sign system concept class, $\mathcal{C}_{m,1}$, where each control parameterized by G gives rise to $\operatorname{sign} y(1)$.

LEMMA 5.2. *The sign system concept class $\mathcal{C}_{m,1}$ with initial condition $x(0) = 0$ satisfies*

$$VC(\mathcal{C}_{m,1}) \leq 2\left(2mn^2 + 4n\right) \log_2[8e(8mn^2k(n + \ell_{\max}) + 1)(2nk + 2(1 + 2k)^n)],$$

where $\ell_{\max}$ is given by (4).

Proof. By Proposition 5.1, $y(1) = H(\lambda, G)$, where $\lambda \in \mathbb{R}^{2mn^2} \times \mathbb{R}^{4n}$ are considered as parameters. In fact, $y(1) = \frac{P}{Q}$, where P and Q denote piecewise polynomial functions. As in the statement of Goldberg–Jerrum bounds, we have a function $F : \Lambda \times \mathbb{R}^{mk} \to \{0, 1\}$ defined by $F(\lambda, G) = \operatorname{sign} H(\lambda, G)$. The concept class associated to the system identification problem is $\mathcal{F} := \{F(\lambda, \cdot) : \mathbb{R}^{mk} \to \{0, 1\};\ \lambda \in \Lambda\}$, where $\Lambda = \mathbb{R}^{2mn^2+4n}$. Before applying the Goldberg–Jerrum bound, we need to determine the possible degrees of P and Q with respect to the parameters.

The rationality condition implies that

$$\max_{\substack{i \leq \ell \leq n \\ 1 \leq j \leq k}} \left\{\deg(\hat{P}_{\ell j}), \deg(\hat{Q}_{\ell j})\right\} \leq d_{\max}.$$

Then

$$\frac{\widetilde{P}_{i\ell}}{\widetilde{Q}_{i\ell}} = \sum_{j=1}^{k} g_{ij} \frac{\hat{P}_{\ell j}}{\hat{Q}_{\ell j}},$$

so $\deg(\widetilde{Q}_{i\ell}) \leq kd_{\max}$ and $\deg(\widetilde{P}_{i\ell}) \leq kd_{\max}$. Note here that we are calculating the degree with respect to the system parameters, and the inputs g_{ij} do not contribute. By continuing in a similar fashion and combining r-summation to the ℓ-summation in Proposition 5.1, we write $P_i/Q_i = \sum_{\ell=1}^{2n^2} \alpha_{i\ell} \widetilde{P}_{i\ell}/\widetilde{Q}_{i\ell}$ to conclude that $\deg(Q_i) \leq 2n^2 \deg(\widetilde{Q}_{i\ell}) = 2n^2kd_{\max}$ and $\deg(P_i) \leq 2n^2kd_{\max} + 1$. Finally, $P/Q = \sum_{i=1}^{m} P_i/Q_i$ with $\deg(Q) \leq m2n^2kd_{\max}$ and $\deg(P) \leq m2n^2kd_{\max} + 1$.

Recall that with $p = 1$ and initial condition $x(0) = 0$, using the notation of Proposition 5.1,

$$y(1) = \sum_{i=1}^{m} \sum_{r=1}^{n} \sum_{\ell=1}^{2n} \alpha_{ir\ell} \sum_{j=1}^{k} g_{ij} \int_0^1 \xi_\ell(x_{r1}, x_{r2}, t)\omega_j(t)dt.$$

The proof of Lemma 4.4 indicates that the denominator of $\int_0^1 \xi_\ell(x_{r1}, x_{r2}, t)\omega_j(t)dt$ equals

$$((x_{r1} + \alpha_j)^2 + (x_{r2} + \beta_j)^2)^{z_{\ell j}}((x_{r1} + \alpha_j)^2 + (x_{r2} - \beta_j)^2)^{z_{\ell j}},$$

where α_j, β_j are fixed parameters of the basis input function ω_j and $z_{\ell j} \in \mathbb{N}$.

By carrying out the summations we get $y(1) = P/Q$, where Q consists of powers of polynomials f_{ij1}, f_{ij2} with

$$f_{ij1}(\mathbf{A}, \mathbf{X}, G) = (x_{i1} + \alpha_j)^2 + (x_{i2} + \beta_j)^2,$$
$$f_{ij2}(\mathbf{A}, \mathbf{X}, G) = (x_{i1} + \alpha_j)^2 + (x_{i2} - \beta_j)^2,$$

and $i = 1, \ldots, n$, $j = 1, \ldots, k$.

Our final step before applying the Goldberg–Jerrum bound is finding out the number of polynomial inequalities s needed in the Boolean formula and evaluating the sign of the final state output. This is done by studying the number of different P/Q expressions without poles.

An upper bound for different P/Q expressions without poles can be obtained by applying Theorem 4.7 to $2nk$ polynomials f_{ij1}, f_{ij2}, $i = 1, \ldots, n$ and $j = 1, \ldots, k$, and viewing those as polynomials of $2n$ variables and each polynomial having degree 2. This gives the upper bound $(16ek)^{2n}$.

However, a more specific bound can be obtained in this problem. Note that varying x_{i1} and x_{i2} we can make at most one of the $2k$ polynomials f_{ij1}, f_{ij2}, $j = 1, \ldots, k$, to be zero. For example, γ zeros among f_{ij1}, f_{ij2}, $i = 1, \ldots, n$ and $j = 1, \ldots, k$, can be obtained in $(2k)^\gamma \binom{n}{\gamma}$ ways, and the number of possible sign assignments is obtained by summing over γ yielding

$$\sum_{\gamma=0}^{n} (2k)^\gamma \binom{n}{\gamma} = (1 + 2k)^n.$$

Thus the number of P/Q expressions without poles is $(1 + 2k)^n$, which gives rise to $2(1 + 2k)^n$ polynomials.

Note that in order to write $\operatorname{sign} y(1)$ as a Boolean formula evaluating polynomial inequalities and equalities one also has to include the $2nk$ polynomials f_{ij1}, f_{ij2}, $i = 1, \ldots, n$, $j = 1, \ldots, k$. Values of these polynomials determine which P/Q expression is the valid one to determine $\operatorname{sign} y(1)$. The Boolean formula for $\operatorname{sign} y(1)$ can be given as a truth table involving polynomial inequalities of $2nk$ f_{ij1}, f_{ij2} expressions and $2(1 + 2k)^n$ different P and Q expressions.

Using Lemma 4.4 for bound on $d_{\max}$, we apply the Goldberg–Jerrum bound with $s = 2nk + 2(2k + 1)^n$, $d = m2n^2k4(n + \ell_{\max}) + 1$, and $\ell = 2mn^2 + 4n$. □

A simple example of a piecewise polynomial function P/Q together with the decision table for the final output is provided in the appendix.

Remark 5.3. The VC dimension bound is modified for the more abstract rationality conditions as follows. Evaluating the sign of the output involves the evaluation of $2(8ed_{\max}2n^2kh/4n)^{4n} + 2n^2kh$ polynomials; $2n^2kh$ evaluations are needed to find an appropriate piece, and by Theorem 4.7 the maximum number of possible expressions of the type P/Q is bounded by $(8ed_{\max}2n^2kh/4n)^{4n}$. Applying the Goldberg–Jerrum bound with $s = 2(8ed_{\max}2n^2kh/4n)^{4n}+2n^2kh$, $d = m2n^2kd_{\max}+1$, and $\ell = 2mn^2+4n$ gives the result.

Proof of Theorem 3.1*, the VC dimension upper bound, $p = 1$.* By using the previous notation, $y = Ce^Ax^0 + C\int_0^1 e^{A(1-t)}Bu(t)dt$. Let $\widetilde{x} = Ce^Ax^0$. Then $y = \widetilde{x} + P/Q = (\widetilde{x}Q + P)/Q = \widetilde{P}/Q$. This has $2mn^2 + 4n + 1$ parameters and $\deg(\widetilde{P}) \leq m2n^2kd_{\max} + 1$. □

5.2. Lower bounds for the VC dimension. The lower bounds for the VC dimension are developed for a single-input single-output system with initial state zero. The control is

$$u = \sum_{j=1}^{k} g_j\omega_j : [0, 1] \to \mathbb{R}.$$

We derive lower bounds by fixing the structure of A, B, and C and using the dual VC dimension and axis shattering following the ideas of DasGupta and Sontag [6].

Lemmas 5.5, 5.6, and 5.7 given in this section together prove Theorem 3.2. These lower bounds are very general; we just assume that the input functions are continuous and linearly independent; thus no particular structure of input functions is required as in the upper bounds.

To make the next proof cleaner we formulate a part of it as a separate proposition. (The proposition is a standard fact and we omit the proof.)

PROPOSITION 5.4. *Let $\omega_j : [0,1] \to \mathbb{R}$, $j = 1, \ldots, k$, be continuous and linearly independent. Then the functions*

$$h_j(\lambda) = \int_0^1 e^{\lambda t} \omega_j(t) dt, \quad j = 1, \ldots, k,$$

are linearly independent.

LEMMA 5.5 (lower bound 1). *The sign system concept class $\mathcal{C}_{1,1}$ with scalar inputs and scalar outputs satisfies*

$$VC(\mathcal{C}_{1,1}) \geq m' \left\lfloor \log_2 \left\lfloor \frac{k}{m'} \right\rfloor \right\rfloor,$$

where $m' = \min\{n, k\}$.

Proof. Let $\omega_j(t)$, $j = 1, \ldots, k$, be continuous and linearly independent. Let A have n distinct real eigenvalues $-\lambda_1, \ldots, -\lambda_n$, and take B and C so that

$$Ce^{A(1-t)}B = \sum_{i=1}^{m'} e^{\lambda_i t},$$

where $m' = \min\{n, k\}$. Then the final output of the system is

$$y(1) = \int_0^1 Ce^{A(1-t)}B \sum_{j=1}^{k} g_j \omega_j(t) dt = \sum_{i=1}^{m'} \sum_{j=1}^{k} g_j \int_0^1 e^{\lambda_i t} \omega_j(t) dt.$$

Define $h_j(\lambda) = \int_0^1 e^{\lambda t} \omega_j(t) dt$. By Proposition 5.4 the h_j's are linearly independent and we can find $\lambda_1, \ldots, \lambda_k$ such that the matrix

$$\begin{bmatrix} h_1(\lambda_1) & \cdots & h_k(\lambda_1) \\ \vdots & & \vdots \\ h_1(\lambda_k) & \cdots & h_k(\lambda_k) \end{bmatrix}$$

has rank k.

The control system with sign-observations gives the mapping $F : \mathbb{R}^{m'} \times \mathbb{R}^k \to \{0,1\}$ by

$$(\lambda_1, \ldots, \lambda_{m'}, g_1, \ldots, g_k) \mapsto \operatorname{sign} \left[\sum_{i=1}^{m'} \sum_{j=1}^{k} g_j h_j(\lambda_i) \right].$$

We show that the mapping from parameters $\lambda_1, \ldots, \lambda_{m'}$ to $\{0,1\}$ can be axis-shattered. Let $L = \{\lambda_1, \ldots, \lambda_k\}$ be so that $[h_j(\lambda_i)]_{i,j}$ has rank k. Denote by $L_1, \ldots, L_{m'}$ disjoint subsets of L such that $|L_i| = \lfloor k/m' \rfloor$, and let $M = L \setminus \{\bigcup_{i=1}^{m'} L_i\}$. Next we want to interpolate in the points of L.

Fix s, $1 \le s \le m'$, and let $\phi : L_s \to \{0,1\}$ be any dichotomy. Next find $g_1, \dots, g_k$ such that

$$\begin{aligned} &\sum_{j=1}^{k} g_j h_j(\lambda_s) = \phi(\lambda_s) \qquad \forall\, \lambda_s \in L_s, \\ &\sum_{j=1}^{k} g_j h_j(\lambda) = 0 \qquad\qquad \forall\, \lambda \in (L \cup M) \setminus L_s. \end{aligned} \tag{12}$$

Let $g_1^*, \dots g_k^*$ satisfy (12). (A unique solution exists because $[h_j(\lambda_i)]$ has rank k.) Then

$$F[\lambda_1, \dots, \lambda_{m'}, g_1^*, \dots, g_k^*] = \operatorname{sign}\left[\sum_{i=1}^{m'} \sum_{j=1}^{k} g_j^* h_j(\lambda_i)\right] = \phi(\lambda),$$

when $\lambda \in L_s$ and for all $(\lambda_1, \dots, \lambda_{m'}) \in L_1 \times \cdots \times L_{m'}$.

Let $\widetilde{\mathcal{F}} = \{F(\lambda_1, \dots, \lambda_{m'}, \cdot) : \mathbb{R}^k \to \{0,1\};\ (\lambda_1, \dots, \lambda_{m'}) \in \mathbb{R}^{m'}\}$. By the axis-shattering bound given in Theorem 4.5,

$$\mathrm{VC}(\widetilde{\mathcal{F}}) \ge m' \left\lfloor \log_2 \left\lfloor \frac{k}{m'} \right\rfloor \right\rfloor,$$

and thus $\mathrm{VC}(\mathcal{C}_{1,1}) \ge \mathrm{VC}(\widetilde{\mathcal{F}})$, where $\mathcal{C}_{1,1}$ is the control system concept class with $p = m = 1$. □

LEMMA 5.6 (lower bound 2). *If $k \le n$, then*

$$VC(\mathcal{C}_{1,1}) \ge k.$$

Proof. We make a small modification of the above argument. Assume that $k \le n$, and let A have n real eigenvalues $\lambda_1, \dots, \lambda_n$. Next we take B and C so that $Ce^{A(1-t)}B = \sum_{i=1}^{n} e^{\lambda_i t}\beta_i$, where $(\beta_1, \dots, \beta_n, \lambda_1, \dots, \lambda_n)$ are considered as system parameters.

We study the mapping

$$\begin{aligned} (\beta_1, \dots, \beta_n, \lambda_1, \dots, \lambda_n, g_1, \dots, g_k) &\mapsto \operatorname{sign}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} g_j h_j(\lambda_i)\beta_i\right] \\ = \operatorname{sign}\left[\sum_{j=1}^{k} g_j \underbrace{\sum_{i=1}^{n} h_j(\lambda_i)\beta_i}_{\gamma_j}\right] &= \operatorname{sign}\left[\sum_{j=1}^{k} g_j \gamma_j\right]. \end{aligned}$$

Given $(\gamma_1, \dots, \gamma_k)$, by linear independence of $h_1, \dots, h_k$, we can find $\lambda_1, \dots, \lambda_n, \beta_1, \dots, \beta_n$ such that $\sum_{i=1}^{n} h_j(\lambda_i)\beta_i = \gamma_j$, $j = 1, \dots, k$. But $(\gamma_1, \dots, \gamma_k)$ can be viewed as a normal vector for a hyperplane through the origin in $\mathbb{R}^k$, and the concept class associated to the mapping $(g_1, \dots, g_k) \mapsto \operatorname{sign}[\sum_{j=1}^{k} g_j\gamma_j]$ as $(\gamma_1, \dots, \gamma_k)$ varies has VC dimension k. Hence $\mathrm{VC}(\mathcal{C}_{1,1}) \ge k$. □

LEMMA 5.7 (lower bound 3). *If $n \le k$, then*

$$VC(\mathcal{C}_{1,1}) \ge n.$$

Proof. Our construction for the control system is as in the previous proof, but now we assume that $n \leq k$, and we study

$$
\begin{aligned}
(\beta_1, \dots, \beta_n, \lambda_1, \dots, \lambda_n, g_1, \dots, g_k) &\mapsto \operatorname{sign}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} g_j h_j(\lambda_i)\beta_i\right] \\
= \operatorname{sign}\left[\sum_{i=1}^{n}\underbrace{\sum_{j=1}^{k} g_j h_j(\lambda_i)}_{\tilde{g}_i}\beta_i\right] &= \operatorname{sign}\left[\sum_{i=1}^{n}\tilde{g}_i\beta_i\right],
\end{aligned}
$$

and again by linear independence and the above hyperplane argument (now via first transforming $(g_1, \dots, g_k)$) we can conclude that the above mapping has VC dimension n. Thus $\mathrm{VC}(\mathcal{C}_{1,1}) \geq n$. □

5.3. VC dimension upper bounds for *p*-dimensional outputs. We begin by proving Theorem 3.3.

Proof of the VC dimension upper bound. We develop an upper bound based on the bound for a scalar sign-observation. We have seen that under the rationality assumption (RAT) the scalar output is a piecewise rational expression P/Q. In general, the control system maps G to $(\operatorname{sign}(P_1/Q_1), \dots, \operatorname{sign}(P_p/Q_p))^T$, which is understood as a binary representation of a number in $\{0, 1, \dots, 2^p - 1\}$. Let $f : \mathbb{R}^{mk} \to \{0, \dots, 2^p - 1\}$ be the mapping given by the control system, and denote the class of all such mappings by $\mathcal{F}$. For each $f \in \mathcal{F}$ introduce a loss function $L_{0\text{-}1,f}(z, a) = L_{0\text{-}1}(f(z), a) = 1$, when $f(z) \neq a$, and 0 otherwise. Define the class $L_{0\text{-}1,\mathcal{F}} = \{L_{0\text{-}1,f};\ f \in \mathcal{F}\}$.

In order to calculate the value of the output, after determining an appropriate piece, one needs to know the truth values of the expressions $P_1 > 0, Q_1 > 0, \dots, P_p > 0$, and $Q_p > 0$, where P's and Q's are polynomials on inputs and parameters of the control system. To evaluate the value of the loss function $L_{0\text{-}1,f}(z, a)$, one needs the truth values of $y = 0, y = 1, \dots, y = 2^p - 2$.

In the general case one needs $2nk + 2p(2k + 1)^n + 2^p - 1$ truth values. As this procedure evaluates only polynomials, we can use the Goldberg–Jerrum bound again. The maximum degree of the polynomials is $m2n^2k4(n + \ell_{\max}) + 1$, and the total number of parameters is $2pn^2m + 4n + p$, where the last term comes from the initial condition. □

6. A fat-shattering bound. We begin this section by proving Theorems 3.6 and 3.7. As a corollary of Theorem 3.7 we prove the fat-shattering bound appearing in Theorem 2.7 bounding the sample complexity for proper agnostic learning.

Proof of Theorem 3.6. For the first part of the proof we use a generic set B for the parameters. Assume that we can γ-shatter a set of inputs $\{u_1, \dots, u_d\}$ and there exists $\{r_1, \dots, r_d\}$ such that, for each assignment $b \in \{0, 1\}^d$, there exists a $\lambda \in B$ such that

$$
\begin{aligned}
F(\lambda, u_i) &\geq r_i + \gamma \quad \text{if } b_i = 1, \text{ and} \\
F(\lambda, u_i) &\leq r_i - \gamma \quad \text{otherwise.}
\end{aligned}
$$

We write $\lambda \sim \mu$ if and only if the parameters λ and μ give the same assignment for all $\{u_1, \dots, u_d\}$. Further, let $\Lambda = \{\lambda_1, \dots, \lambda_{2^d}\}$ be a collection of parameters that shatter $\{u_1, \dots, u_d\}$, and let $\lambda_i, \lambda_j \in \Lambda$. Now $\lambda_i \not\sim \lambda_j$ implies that there exist $u^* \in \{u_1, \dots, u_d\}$ and $r^* \in \{r_1, \dots, r_d\}$ such that $F(\lambda_i, u^*) \geq \gamma + r^*$ and $F(\lambda_j, u^*) \leq$

$\gamma - r^*$, or vice versa. Hence $2\gamma \leq |F(\lambda_i, u^*) - F(\lambda_j, u^*)| \leq L\|\lambda_i - \lambda_j\|$ and so $\|\lambda_i - \lambda_j\| \geq 2\gamma/L$. That is, the set Λ of cardinality 2^d is a $2\gamma/L$-separated set in B. Now the fat-shattering bounds follow by calculating $2\gamma/L$-packing numbers for different sets B.

If $B = B_\infty^k(C)$, the maximum possible cardinality for an ϵ-separated set is $\lfloor 2C/\epsilon \rfloor^k$, and thus

$$2^d \leq \left\lfloor \frac{2C}{2\gamma/L} \right\rfloor^k = \left\lfloor \frac{CL}{\gamma} \right\rfloor^k,$$

and solving for d yields $d \leq k \log_2 \lfloor CL/\gamma \rfloor$.

Similarly, if $B = \bar{B}_\infty^k(C)$, the maximum possible cardinality for an ϵ-separated set is $(1 + \lfloor 2C/\epsilon \rfloor)^k$ and by a similar argument we arrive at the bound $d \leq k \log_2(1 + \lfloor LC/\gamma \rfloor)$.

For $B = \bar{B}_2^k(C)$, let $P(\epsilon)$ be a collection of ϵ-separated sets in $\bar{B}_2^k(C)$, and let $|P(\epsilon)|$ denote its cardinality. As all open balls with radius $\epsilon/2$ with centers at ϵ-separated points have to be disjoint and their union has to be inside a ball of radius $C+\epsilon/2$, we get that $|P(\epsilon)|\alpha(k)(\epsilon/2)^k \leq \alpha(k)(C+\epsilon/2)^k$, where $\alpha(k) = \pi^{k/2}/\Gamma(k/2+1)$ is the volume of a unit ball in $\mathbb{R}^k$. Hence $|P(\epsilon)| \leq (C + 2/\epsilon)^k$ and $2^d \leq (C + L/\gamma)^k$; i.e., $d \leq k \log_2(C + L/\gamma)$. □

Next we prove Theorem 3.7 by applying the Lipschitz bound to a control system.

Proof of Theorem 3.7. Our aim is to compute the Lipschitz constant associated to the control system in Definition 2.6, and then we apply Theorem 3.6.

Denote the system parameters $(\alpha_{11}, \ldots, \alpha_{nm}, a_1, b_1, \ldots, a_r, b_r)$ by λ and assume $\|\lambda\|_\infty < 1$. Let

$$F(\lambda, u) = y(\tau) = \int_0^\tau \sum_{i=1}^m \sum_{\ell=1}^n \alpha_{i\ell}\xi_\ell(t)u_i(\tau - t)dt.$$

Functions $\xi_1(t), \ldots, \xi_n(t)$ are of the form $\xi(t) = t^c e^{at}\sin(bt)$ or $\xi(t) = t^c e^{at}\cos(bt)$, where $a + ib$ is an eigenvalue of A and $c \in \{0, \ldots, n-1\}$. Thus taking a partial derivative with respect to a or b will increase the power of t by one and change the trigonometric functions. Therefore,

$$\begin{aligned}\left|\frac{\partial F(\lambda, u)}{\partial \alpha_{\kappa\rho}}\right| = \left|\frac{\partial y(\tau)}{\partial \alpha_{\kappa\rho}}\right| &= \left|\sum_{i=1}^m \sum_{\ell=1}^n d(i,\ell) \int_0^\tau \xi_\ell(t)u_i(\tau - t)dt\right| \\ &\leq nm \int_0^\tau |\xi_\ell(t)u_i(\tau - t)|\, dt \leq nm\tau^{n-1}e^\tau M,\end{aligned}$$

because $\sup_{t\in[0,\tau]} |\xi_\ell(t)| \leq e^\tau \tau^n$ and $d(i,\ell) = \partial\alpha_{ij}/\partial\alpha_{\kappa\rho} = 1$ if $(i,\ell) = (\kappa,\rho)$ and zero otherwise. Similarly we calculate

$$\begin{aligned}\left|\frac{\partial F(\lambda, u)}{\partial a_\kappa}\right| = \left|\frac{\partial y(\tau)}{\partial a_\kappa}\right| &\leq nm\tau^n e^\tau M \text{ and} \\ \left|\frac{\partial F(\lambda, u)}{\partial b_\kappa}\right| = \left|\frac{\partial y(\tau)}{\partial b_\kappa}\right| &\leq nm\tau^n e^\tau M\end{aligned}$$

as $\sup_{t\in[0,\tau]} |\frac{\partial \xi_\ell(t)}{\partial_\kappa}| \leq e^\tau \tau^n$ and $\sup_{t\in[0,\tau]} |\frac{\partial \xi_\ell(t)}{\partial b_\kappa}| \leq e^\tau \tau^n$.

Now the Lipschitz constant can be taken to be $L = n^2 m e^\tau \tau^n M$ as

$$|F(\lambda, u) - F(\lambda^*, u)| = |\nabla F \cdot (\lambda - \lambda^*)| \leq L\|\lambda - \lambda^*\|_\infty.$$

The number of system parameters is at most $nm+n=(m+1)n$ and we get the level fat-shattering bound by applying Theorem 3.6 with space dimension $n(m+1)$ and $L=n^2me^{\tau}\tau^nM$. □

As a corollary, we combine the above result together with a pseudodimension bound to prove the fat-shattering bound given in Theorem 2.7.

Corollary 6.1 (fat-shattering bound in Theorem 2.7). *Assume that the system* $\dot{x}=Ax+Bu$, $y=Cx$, $x(0)=0$, *can be parameterized by* $\lambda\in\mathbb{R}^{n(m+1)}$ *as in Definition* 2.6 *with* $\|\lambda\|_\infty<1$, *and assume in addition that the control is given by* $u=G\omega$, *where the input basis functions* ω_j *are in* Ω *given by* (3). *We denote the corresponding control system class by*

$$\mathcal{F}_B=\{F(\lambda,\cdot):U\to\mathbb{R};\ \lambda\in B\}.$$

Then

$$\mathrm{fat}_\gamma(\mathcal{F}_B)\le\min\begin{cases}(m+1)n\log_2\left\lfloor\frac{n^2m\tau^ne^{\tau}kM}{\gamma}\right\rfloor,\\ 2(m+4)n\log_2\left(8e\big(nmk4(n+\ell_{\max})+1\big)\big(2nk+2(2k+1)^n\big)\right),\end{cases}$$

where $\ell_{\max}$ *is given by* (3) *and* (4) *and* M *is a constant satisfying*

$$\int_0^\tau|u_i(\tau-t)|dt\le kM$$

for all $i=1,\dots,m$.

Proof. The first part of the bound follows from Theorem 3.7 with kM in place of M.

The remaining part of the bound comes from the pseudodimension bound. First we derive the associated VC dimension bound. As we assumed that A has a fixed Jordan block structure, every entry of $e^{A(1-t)}$ is a linear combination of n functions $\xi_1(t),\dots,\xi_n(t)$. (That is, we do not need to consider all possible functions over different Jordan block structures.) This implies that in the Goldberg–Jerrum argument of section 5.1 we can take $\ell=mn+4n$, $d=nmk4(n+\ell_{\max})+1$, and $s=2nk+2(2k+1)^n$. Moreover, in that section the VC dimension bounds were derived for the time interval $[0,1]$. However, the upper bound depends on the number of system parameters and the degrees of polynomials to be evaluated. Changing the time interval to $[0,\tau]$ means just that we replace the eigenvalue parameters (referring to the proof of Proposition 5.1) $a,b,e^a\cos b,e^a\sin b$ by $a\tau,b\tau,e^{a\tau}\cos b\tau,e^{a\tau}\sin b\tau$.

The above bound is also a bound for the pseudodimension. Observe that for $\mathcal{G}=\{g:X\to\mathbb{R}\}$, the pseudodimension can be defined as $\mathrm{PD}(\mathcal{G})=\mathrm{VC}\{\mathrm{Ind}(x,y)=\mathrm{sign}(g(x)-y);\ g\in\mathcal{G}\}$. Hence we want to study the VC dimension associated to $\mathrm{sign}(y(\tau)-z)=\mathrm{sign}(P/Q-z)=\mathrm{sign}(\hat{P}/Q)$, where $\hat{P}=P-zQ$ has the same degree as P with respect to the parameters. Here z is a new input, but the bound utilizing Goldberg–Jerrum technique does not depend on the dimension of the inputs, and hence the above VC dimension bound is also a bound for the pseudodimension. (Note that here in the scale sensitive setting we do not apply the pseudodimension results of section 5.3 using loss functions, as those rescaled the outputs.) □

7. A class of systems with VC dimension k.

For the control system (2) with scalar control $u(t)=\sum_{i=1}^k g_i\omega_i(t)$ and unrestricted $\omega_1,\dots,\omega_k$, the standard half-space argument gives an upper bound k. This bound is tight. We will give an example of a single-input, single-output one-parameter family of control systems in dimension two that has VC dimension k, when the controls are of the form $u(t) = \sum_{i=1}^{k} g_i\omega_i(t)$ and $\omega_i(1-t) = 1_{[2^{-i},2^{-i}+2^{\alpha}]}$, where $\alpha = -2(k+1)$.

Consider a control system

$$\dot{x_1} = x_2,\\ \dot{x_2} = -\lambda^2 x_1 + u,\\ y = -x_1. \tag{13}$$

For time interval [0,1] and initial condition $(x_1, x_2) = (0, 0)$, the output is given by $y(1) = \int_0^1 \sin(\lambda t)u(1-t)dt$.

LEMMA 7.1. *Controls* $\{\omega_1, \dots, \omega_k\}$ *such that* $\omega_i(1-t) = 1_{[2^{-i},2^{-i}+2^{\alpha}]}$*, where* $\alpha = -2(k+1)$*, are shattered by the control system* (13) *with sign-observations.*

Proof. Let $T = \{2^{-i},\ i = 1, \dots, k\}$ and $J \subseteq T$. Define $\lambda_J = \pi \sum_{i=1}^{k} a_i 2^i$, where $a_i = 1$ if $2^{-i} \notin J$ and $a_i = 0$ otherwise. Now if $t = 2^{-\ell}$, then

$$\lambda_J t = \pi \sum_{i=1}^{k} a_i 2^{i-\ell} = \pi \left(\underbrace{\sum_{i=1}^{\ell-1} a_i 2^{i-\ell}}_{1/2\, c_2} + a_\ell + \underbrace{\sum_{i=\ell+1}^{k} a_i 2^{i-\ell}}_{2c_1} \right),$$

where $c_1 \in \mathbb{N}$ and $0 \le c_2 < 2$.

Hence $\sin(\lambda_J t) = \sin(\pi(1/2\, c_2 + a_\ell))$. Note that if $a_\ell = 0$, then $1/2\, c_2 + a_\ell \in [0, 1 - 2^{-\ell}]$, and if $a_\ell = 1$, then $1/2\, c_2 + a_\ell \in [1, 2 - 2^{-\ell}]$. Thus $\sin(\pi(1/2\, c_2 + a_\ell)) \ge 0$ if $a_\ell = 0$, and $\sin(\pi(1/2\, c_2 + a_\ell)) \le 0$ if $a_\ell = 1$. Therefore, $\sin(\lambda_J t) \ge 0 \ \Leftrightarrow\ a_\ell = 0$ $\sin(\lambda_J t) \ge 0 \ \Leftrightarrow\ t \in J$. Further,

$$\int_{2^{-\ell}}^{2^{-\ell}+2^{\alpha}} \sin(\lambda_J t)dt \ge 0 \ \Leftrightarrow\ a_\ell = 0,$$

where α is taken so that

$$\sum_{j=1}^{k} 2^j 2^{\alpha} \le 2^{-(k+1)}. \tag{14}$$

This ensures that when $\ell \le k$ and $t \in [2^{-\ell}, 2^{-\ell} + 2^{\alpha}]$, $\lambda_J t \in [0, \pi(1 - 2^{-\ell} + \sum_{j=1}^{k} 2^j 2^{\alpha})] \subset [0, \pi)$ if $a_\ell = 0$ or similarly $\lambda_J t \in [\pi, 2\pi)$ if $a_\ell = 1$. In (14) we can take $\alpha = -2(k+1)$ as $\sum_{j=1}^{k} 2^j = 2^{k+1} - 2$.

In this way the integrand in $\int_{2^{-\ell}}^{2^{-\ell}+2^{\alpha}} \sin(\lambda_J t)dt$ is either positive or negative.

For $S \subseteq \{1, \dots, k\}$, let $J = \{2^{-i}, i \in S\}$. For each ω_i,

$$\int_0^1 \sin(\lambda_J t)\omega_i(1-t)dt = \int_{2^{-i}}^{2^{-i}+2^{\alpha}} \sin(\lambda_J t)dt > 0 \Leftrightarrow\ i \in S;$$

i.e., the set of controls $\{\omega_1, \dots, \omega_k\}$ is shattered by the mapping

$$\omega_i \mapsto \text{sign}\left[\int_0^1 \sin(\lambda_J t)\omega_i(1-t)dt\right]. \qquad \square$$

Appendix. An example of the Goldberg–Jerrum bound.

We begin this appendix with an informal discussion on the Goldberg–Jerrum technique used to prove the VC dimension upper bounds in this paper.

We want to write $y(1) = P/Q$, where P and Q are polynomials. Unfortunately, the value of $\operatorname{sign} y(1)$ cannot be obtained by just evaluating P and Q since Q may have zeros. Therefore, we need to write

$$y(1) = \begin{cases} P_1/Q_1 & \text{if } f_1 \neq 0, \dots, f_\mu \neq 0, \\ \vdots & \\ P_\gamma/Q_\gamma & \text{if } f_1 = 0, \dots, f_\mu \neq 0 \end{cases}$$

so that after evaluating μ polynomials $f_1, \dots, f_\mu$ we can pick a definition P_i/Q_i without poles in a region defined by the μ polynomials. When $y(1)$ is defined in this way, $\operatorname{sign} y(1)$ can be easily expressed by a Boolean formula evaluating $2\gamma + \mu$ polynomial inequalities and equalities.

For simplicity we assume that $p = 1$ and the initial condition $x(0) = 0$. Then using the notation of Proposition 5.1 we write

$$y(1) = \sum_{i=1}^{m} \sum_{r=1}^{n} \sum_{\ell=1}^{2n} \alpha_{ir\ell} \sum_{j=1}^{k} g_{ij} \int_0^1 \xi_\ell(x_{r1}, x_{r2}, t)\omega_j(t)dt,$$

and by the proof of Lemma 4.4

$$\int_0^1 \xi_\ell(x_{r1}, x_{r2}, t)\omega_j(t)dt = \frac{P_{\ell j}}{((x_{r1}+\alpha_j)^2 + (x_{r2}+\beta_j)^2)^{z_{\ell j}} \, ((x_{r1}+\alpha_j)^2 + (x_{r2}-\beta_j)^2)^{z_{\ell j}}},$$

where $P_{\ell j}$ is some polynomial, $z_{\ell j} \in \mathbb{N}$, and $\omega_j(t) = t^{\ell_j} e^{\alpha_j t} \sin(\beta_j t)$ or $\omega_j(t) = t^{\ell_j} e^{\alpha_j t} \cos(\beta_j t)$. Hence the denominator of $\sum_{j=1}^{k} g_{ij} \int_0^1 \xi_\ell(x_{r1}, x_{r2}, t)\omega_j(t)dt$ is

$$\begin{aligned} &\left((x_{r1}+\alpha_1)^2 + (x_{r2}+\beta_1)^2\right)^{z_{\ell 1}} \left((x_{r1}+\alpha_1)^2 + (x_{r2}-\beta_1)^2\right)^{z_{\ell 1}} \\ &\quad \times \cdots \times \left((x_{r1}+\alpha_k)^2 + (x_{r2}+\beta_k)^2\right)^{z_{\ell k}} \left((x_{r1}+\alpha_k)^2 + (x_{r2}-\beta_k)^2\right)^{z_{\ell k}}. \end{aligned}$$

By carrying out all summations $y(1) = P/Q$. The denominator Q consists of the product

$$\begin{aligned} \prod_{r=1}^{n} \Big(&\left((x_{r1}+\alpha_1)^2 + (x_{r2}+\beta_1)^2\right)^* \left((x_{r1}+\alpha_1)^2 + (x_{r2}-\beta_1)^2\right)^* \\ &\times \cdots \times \left((x_{r1}+\alpha_k)^2 + (x_{r2}+\beta_k)^2\right)^* \left((x_{r1}+\alpha_k)^2 + (x_{r2}-\beta_k)^2\right)^* \Big), \end{aligned}$$

where $*$'s stand for some unspecified powers. Hence the zeros of Q are determined by $2nk$ polynomials $f_{ij1} = (x_{i1}+\alpha_j)^2 + (x_{i2}+\beta_j)^2$, $f_{ij2} = (x_{i1}+\alpha_j)^2 + (x_{i2}-\beta_j)^2$, and $i = 1, \dots, n$, $j = 1, \dots, k$. The number of different sign assignments determining γ is calculated as in the proof of Lemma 5.2.

Example. The purpose of the following example is to illustrate the function $y = P/Q$ used in the Goldberg–Jerrum technique together with the sequence of polynomial evaluations involved and a table for the final output depending on the outcomes of the polynomial evaluations.

Take $m = 1$, $n = 2$, $k = 2$, and assume that A has complex eigenvalues $a \pm ib$. Take basis input functions to be $\omega_1(t) = e^t$ and $\omega_2(t) = e^{2t}$. Then $y(1) = \sum_{l=1}^{2} \alpha_l \sum_{j=1}^{2} g_j \int_0^1 \xi_l(t)\omega_j(t)\,dt$, where $\xi_1(t) = e^{at}\sin(bt)$, $\xi_2(t) = e^{at}\cos(bt)$, α_1, α_2, a, b, $e^a \sin b$, and $e^a \cos b$ are system parameters and g_1, g_2 are input parameters.

By using formulae

$$\int_0^1 e^{\tilde{a}t}\sin(\tilde{b}t)\,dt = \frac{e^{\tilde{a}}(\tilde{a}\sin\tilde{b} - \tilde{b}\cos\tilde{b}) + \tilde{b}}{\tilde{a}^2+\tilde{b}^2} \quad \text{and}$$
$$\int_0^1 e^{\tilde{a}t}\cos(\tilde{b}t)\,dt = \frac{e^{\tilde{a}}(\tilde{a}\cos\tilde{b} + \tilde{b}\sin\tilde{b}) - \tilde{a}}{\tilde{a}^2+\tilde{b}^2},$$

we calculate the integrals appearing in the rationality condition, and we call them r_{11}, r_{12}, r_{21}, and r_{22}:

$$\int_0^1 \xi_1(t)\omega_1(t)\,dt = \int_0^1 e^{(a+1)t}\sin(bt)\,dt = \begin{cases} r_{11} & \text{if } (a+1)^2+b^2 \neq 0, \\ 0 & \text{if } (a+1)^2+b^2 = 0, \end{cases}$$
$$\int_0^1 \xi_1(t)\omega_2(t)\,dt = \begin{cases} r_{12} & \text{if } (a+2)^2+b^2 \neq 0, \\ 0 & \text{if } (a+2)^2+b^2 = 0, \end{cases}$$
$$\int_0^1 \xi_2(t)\omega_1(t)\,dt = \begin{cases} r_{21} & \text{if } (a+1)^2+b^2 \neq 0, \\ 1 & \text{if } (a+1)^2+b^2 = 0, \end{cases}$$
$$\int_0^1 \xi_2(t)\omega_2(t)\,dt = \begin{cases} r_{22} & \text{if } (a+2)^2+b^2 \neq 0, \\ 1 & \text{if } (a+2)^2+b^2 = 0. \end{cases}$$

The computation of $\operatorname{sign} y(1) = \operatorname{sign}(\sum_{l=1}^2 \alpha_l \sum_{j=1}^2 g_j \int_0^1 \xi_l(t)\omega_j(t)\,dt)$ is divided into three cases:

- Case $(a+1)^2+b^2 \neq 0$, $(a+2)^2+b^2 \neq 0$:
$$\operatorname{sign} y(1) = \operatorname{sign}(\alpha_1 g_1 r_{11} + \alpha_1 g_2 r_{12} + \alpha_2 g_1 r_{21} + \alpha_2 g_2 r_{22}) = \operatorname{sign}\left(\frac{P_1}{Q_1}\right).$$
- Case $(a+1)^2+b^2 = 0$, $(a+2)^2+b^2 \neq 0$:
$$\operatorname{sign} y(1) = \operatorname{sign}(\alpha_1 g_2 r_{12} + \alpha_2 g_1 + \alpha_2 g_2 r_{22}) = \operatorname{sign}\left(\frac{P_2}{Q_2}\right).$$
- Case $(a+1)^2+b^2 \neq 0$, $(a+2)^2+b^2 = 0$:
$$\operatorname{sign} y(1) = \operatorname{sign}(\alpha_1 g_1 r_{11} + \alpha_2 g_2 r_{21} + \alpha_2 g_2) = \operatorname{sign}\left(\frac{P_3}{Q_3}\right).$$

Thus we have three different expressions of the form $\frac{P}{Q}$.

Next we form the Boolean formula, $F = \operatorname{sign} y(1)$, evaluating polynomials $f_1 = (a+1)^2+b^2 = 0$, $f_2 = (a+2)^2+b^2 = 0$, $P_i > 0$, $Q_i > 0$ for $i \in \{1,2,3\}$. In the following table 1 means true and 0 means false for the above polynomial evaluation ($** = 1$ or 0, i.e., extend the table).

$f_1 = 0$	$f_2 = 0$	$P_1 > 0$	$Q_1 > 0$	$P_2 > 0$	$Q_2 > 0$	$P_3 > 0$	$Q_3 > 0$	F
0	0	1	1	**	**	**	**	1
0	0	1	0	**	**	**	**	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	0	**	**	1	1	**	**	1
1	0	**	**	1	0	**	**	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	1	**	**	**	**	0	0	1

In this case (see the statement of Goldberg–Jerrum bounds), $\Lambda = \{\alpha_1, \alpha_2, a, b, e^a \cos b, e^a \sin b\}$, $X = \{g_1, g_2\}$, $s = 8$, $d = 12$, and $l = 6$.

REFERENCES

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, *Scale-sensitive dimensions, uniform convergence and learnability*, J. ACM, 44 (1997), pp. 615–631.

[2] M. Anthony and B. Bartlett, *Learning in neural networks: Theoretical foundations*, Cambridge University Press, Cambridge, UK, 1999.

[3] P. Bartlett, P. Long, and R. Williamson, *Fat-shattering and learnability of real-valued functions*, in Proceedings of the 7th Annual Conference on Computational Learning Theory, ACM, New York, 1994, p. 299.

[4] P. L. Bartlett and P. M. Long, *More theorems about scale-sensitive dimensions and learning*, in Proceedings of the 8th Annual Conference on Computational Learning Theory, Santa Cruz, CA, 1995, ACM, New York, 1995, pp. 392–401.

[5] P. L. Bartlett and P. M. Long, *Prediction, learning, uniform convergence and scale-sensitive dimensions*, J. Comput. System Sci., 56 (1998), pp. 174–190.

[6] B. DasGupta and E. D. Sontag, *Sample complexity for learning recurrent perceptron mappings*, IEEE Trans. Inform. Theory, 42 (1996), pp. 1479–1487.

[7] B. DasGupta and E. D. Sontag, *Sample complexity for learning recurrent perceptron mappings*, in Advances in Neural Information Processing Systems (NIPS'95), MIT Press, Cambridge, MA, 1996, pp. 204–210.

[8] B. Eisenberg and R. Rivest, *On the sample complexity of pac-learning using random and chosen examples*, in Proceedings of the 3rd Annual Workshop on Computational Learning Theory, Morgan Kaufmann, San Francisco, 1990, pp. 154–162.

[9] C.-N. Fiechter, *PAC adaptive control of linear systems*, in Proceedings of the 10th Annual Conference on Computational Learning Theory, ACM, New York, 1997, pp. 72–80.

[10] P. Goldberg and M. Jerrum, *Bounding the Vapnik-Chervonenkis dimension of concept classes parametrized by real numbers*, Machine Learning, 18 (1995), pp. 131–148.

[11] T. Johansen and E. Weyer, *On convergence proofs in system identification—a general principle using ideas from learning theory*, Systems Control Lett., 34 (1998), pp. 85–92.

[12] P. Koiran and E. D. Sontag, *Vapnik-Chervonenkis dimension of recurrent neural networks*, Discrete Appl. Math., 86 (1998), pp. 63–80.

[13] R. Koplon and E. D. Sontag, *Linear systems with sign-observations*, SIAM J. Control Optim., 31 (1993), pp. 1245–1266.

[14] P. Kuusela and D. Ocone, *Learning with side information: An example*, 2002, submitted.

[15] P. Kuusela and D. Ocone, *Learning with side information: PAC learning bounds*, J. Comput. System Sci., 68 (2004), pp. 521–545.

[16] L. Ljung, *PAC-learning and symptotic system identification theory*, in Proceedings of the 35th IEEE Conference on Decision and Control, Kobe, Japan, 1996, pp. 2303–2307.

[17] D. Pollard, *Convergence of Stochastic Processes*, Springer Ser. Statist., Springer-Verlag, New York, 1984.

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

[19] E. D. Sontag, *A learning result for continuous-time recurrent neural networks*, Systems Control Lett., 34 (1998), pp. 151–158.

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

[21] M. Vidyasagar, *A Theory of Learning and Generalization; with Applications to Neural Networks and Control Systems*, Springer-Verlag, New York, 1997.

[22] R. Wenocur and R. Dudley, *Some special Vapnik–Chervonenkis classes*, Discrete Math., 33 (1981), pp. 313–318.

[23] A. Zador and B. Pearlmutter, *VC dimension of an integrate-and-fire neuron model*, Neural Computation, 8 (1996), pp. 611–624.