PERGAMON

Automatica 37 (2001) 1515–1528

automatica

www.elsevier.com/locate/automatica

Randomized algorithms for robust controller synthesis using statistical learning theory☆

M. Vidyasagar

Tata Consultancy Services, Coromandel House, 1-2-10 Sardar Patel Road, Hyderabad 500 003, India

Received 16 July 1999; revised 27 February 2001; received in final form 6 April 2001

Abstract

By now it is known that several problems in the robustness analysis and synthesis of control systems are NP-complete or NP-hard. These negative results force us to modify our notion of "solving" a given problem. An approach that is recently gaining popularity is that of using *randomized* algorithms, which can be used to solve a problem *approximately*, *most of the time*. We begin with the premise that many problems in robustness analysis and synthesis can be formulated as the minimization of an objective function with respect to the controller parameters. It is argued that, in order to assess the performance of a controller as the plant varies over a prespecified family, it is better to use the *average* performance of the controller as the objective function to be minimized, rather than its *worst-case* performance, as the worst-case objective function usually leads to rather conservative designs. Then it is shown that a property from statistical learning theory known as uniform convergence of empirical means (UCEM) plays an important role in allowing us to construct efficient randomized algorithms for a wide variety of controller synthesis problems. In particular, whenever the UCEM property holds, there exists an efficient (i.e., polynomial-time) randomized algorithm. Using very recent results in statistical learning theory, it is shown that the UCEM property holds in *any* problem in which the satisfaction of a performance constraint can be expressed in terms of a finite number of polynomial inequalities. In particular, several problems such as robust stabilization and weighted H_2/H_∞-norm minimization are amenable to the randomized approach.

Keywords: Robust control; Randomized algorithms; Statistical learning theory; VC dimension; Polynomial inequalities

1. Introduction

During recent years it has been shown that several problems in the robustness analysis and synthesis of control systems are either NP-complete or NP-hard. The notions of computational complexity, and in particular NP-completeness and NP-hardness, are not widely known amongst control theorists, and the reader is referred to Garey and Johnson (1979) for a dated but still timely exposition, and to Papadimitrou (1994) for a more up to date treatment. Of the large number of results proved in recent years on the NP-hardness of problems arising in robustness analysis and synthesis, only a few examples are quoted here to give a flavour of this body of work. Many of the problems listed below make use of the idea of "interval matrices", which are defined first. Given an integer n, suppose $\mathbf{y} \in [-1,1]^{2n^2}$ is of the form

$$\mathbf{y} := (\alpha_{ij}, \beta_{ij}), \quad i,j = 1, \ldots, n: \ -1 \leqslant \alpha_{ij} \leqslant \beta_{ij} \leqslant 1,$$

$$\alpha_{ij}, \beta_{ij} \in \mathbb{Q} \ \forall i,j.$$

Here $\mathbb{Q}$ denotes the set of rational numbers. The corresponding set $\mathbf{A}_\mathbf{y}$ consisting of all matrices A with $a_{ij} \in [\alpha_{ij}, \beta_{ij}] \forall i,j$ is known as the *interval matrix* corresponding to the parameter vector $\mathbf{y}$. The symbol $\mathbf{A}_{s,\mathbf{y}}$ consisting of all symmetric matrices in $\mathbf{A}_\mathbf{y}$ is called the *symmetric interval matrix* corresponding to the parameter vector $\mathbf{y}$. With this notation, it is known that each of the following problems is NP-hard (see the survey paper by Blondel and Tsitsiklis (2000) and earlier papers by Nemirovskii (1993) and Poljak and Rohn (1993)): Given a vector $\mathbf{y} \in Y$, determine whether every matrix in the set $\mathbf{A}_\mathbf{y}$ is stable, in the sense that all of its eigenvalues have negative real parts (robust stability), or is nonsingular (robust nonsingularity), or has norm less than one

☆This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Shinji Hara under the direction of Editor Roberto Tempo. This work was carried out while the author was working with the Centre for Artificial Intelligence and Robotics, Bangalore, India.

E-mail address: sagar@atc.tcs.co.in (M. Vidyasagar).

0005-1098/01/$ - see front matter
PII: S0005-1098(01)00122-4

(robust norm boundedness). In the case of symmetric matrices, determining whether every symmetric matrix in $\mathbf{A}_{s,y}$ is positive semidefinite (robust positive definiteness) is also NP-hard. The following two problems, which do not pertain to interval matrices, are also NP-hard.

1. *Determination of the structured singular value*: Consider a feedback system with an uncertain plant P and a fixed controller C. Suppose the plant has a nominal model P_0 but is subject to structured perturbations around P_0. Under these conditions, it is well-known (see Doyle, 1982) that the controller C stabilizes every such perturbed plant if and only if an associated structured singular value is less than one. Let γ denote this structured singular value, and consider the decision problem of determining whether $\gamma < 1$. It is shown in Braatz, Young, Doyle, and Morari (1994) that answering this question is NP-hard.
2. *Constant output feedback stabilization with constraints*: For each integer n, an instance of the problem consists of $n \times n$ matrices A, B, C and numbers α_{ij}, β_{ij} as above. The problem is to determine whether there exists an $n \times n$ "output feedback" matrix K such that $\alpha_{ij} \leqslant k_{ij} \leqslant \beta_{ij}\ \forall i, j$, and such that $A + BKC$ is a stable matrix. It is shown in Blondel and Tsitsiklis (1997) that this problem is NP-hard.

In the face of these and other negative results, one is forced to make some compromises in the notion of a "solving" a problem. An approach that is recently gaining popularity is the use of *randomized* algorithms, which are not required to work "all" of the time, only "most" of the time. Specifically, the probability that the algorithm fails can be made arbitrarily small (but of course not exactly equal to zero). In return for this compromise, one hopes that the algorithm is *efficient*, i.e., runs in polynomial-time. The idea of using randomization to solve control problems is suggested, among other places, in Ray and Stengel (1991) and Marrison and Stengel (1994). In Khargonekar and Tikku (1996), Tempo, Bai, and Dabbene (1997), randomized algorithms are developed for a general problem of function minimization, and then these algorithms are applied to some specific problems such as: (i) determining whether a given controller stabilizes every plant in a structured perturbation model, (ii) determining whether there exists a controller of a specified order that stabilizes a given fixed plant, and so on.

The objective of the present paper is to demonstrate that some recent advances in statistical learning theory can be used to develop *efficient* randomized algorithms for a wide variety of controller analysis/synthesis problems. Specifically, the starting point is the observation that a wide variety of problems in controller synthesis can be formulated as the minimization of an objective function that quantifies the performance of the controller as the plant varies over a prespecified plant family. It is argued here that, to avoid overly conservative designs, the objective function that measures the performance of a controller should be taken as its *average* performance as the plant varies, rather than its *worst-case* performance; see Section 2 for an elaboration of this argument. For such problems (i.e., those in which the objective function is an expected value), it is shown that whenever a property known as uniform convergence of empirical means (UCEM) holds, there exists an efficient randomized algorithm for an associated function minimization problem. Moreover, it is shown that the UCEM property holds whenever the question of making the objective function smaller than a specified constant can be expressed in terms of a finite number of polynomial inequalities. Some specific problems that fall within this framework include: robust stabilization of a family of plants by a fixed controller belonging to a specified family of controllers, and the minimization of weighted H_2/H_∞-norms. Note that what is developed here is a *broad framework* that can accommodate a wide variety of problems. The few specific problems solved here are meant only to illustrate the power and breadth of this framework. No doubt other researchers would be able to solve many more problems using the same general approach.

Note that the randomized approach described here is successfully applied to the problem of designing a low-order controller to stabilize the longitudinal axis of an unstable aircraft in Patel, Deodhare, and Viswanath (2000).

2. Paradigm of robust controller synthesis problem

Suppose one is given a family of plants $\{G(x),\ x \in X\}$ parametrized by x, and a family of controllers $\{K(y),\ y \in Y\}$ parametrized by y. Thus x represents the vector of all the uncertain elements in the plant description, e.g., the pole and zero locations, or the coefficients of the numerator and denominator polynomials of the plant transfer function, or the A, B, C, D matrices, and the like. The set X represents the set of all possible plant parameter vectors x. It is assumed that x is a finite-dimensional vector, so that X is a subset of some finite-dimensional Euclidean space. Similarly, y represents the vector of all adjustable parameters in the controller description, e.g., various gains, time constants, etc. Again, it is assumed that y is a finite-dimensional vector, so that Y is a subset of some finite-dimensional Euclidean space. The objective is to find a *a single fixed controller* $K(y_0), y_0 \in Y$ that performs reasonably well for almost all plants $G(x)$.

Suppose $\psi(\cdot, \cdot)$ is a given cost function. Thus $\psi(G, K)$ is a measure of the performance of the system when the plant is G and the controller is K. For instance, if the objective is merely to choose a stabilizing controller, then

one could define

$$\psi(G,K) := \begin{cases} 1 & \text{if the pair } (G,K) \text{ is unstable,} \\ 0 & \text{if the pair } (G,K) \text{ is stable.} \end{cases} \tag{2.1}$$

As a second example, in problems of filtering or disturbance rejection, one could choose

$$\psi(G,K) := \begin{cases} 1 & \text{if the pair } (G,K) \\ & \quad \text{is unstable,} \\ J(G,K)/[1+J(G,K)] & \text{if the pair } (G,K) \\ & \quad \text{is stable,} \end{cases} \tag{2.2}$$

where

$$J(G,K) = \|W(I+GK)^{-1}\|_p,$$

where $p = 2$ or ∞, and W is a given weighting matrix.

In problems of robust stabilization and robust performance, the cost function should reflect the performance of a *fixed* controller for a *variety* of plants. Since $G = G(x)$ and $K = K(y)$, let us define

$$g(x,y) := \psi[G(x), K(y)].$$

Note that g maps $X \times Y$ into [0,1]. The aim is to define an objective function of y alone that quantifies the performance of the controller $K(y)$, so that by minimizing this objective function with respect to y one could find an "optimal" controller.

As a first attempt, one could choose

$$h(y) := \sup_{x \in X} g(x,y) = \sup_{x \in X} \psi[G(x), K(y)]. \tag{2.3}$$

Thus $h(y)$ measures the *worst-case* performance of a controller $K(y)$ as the plant varies over $\{G(x),\ x \in X\}$. For instance, if one chooses $\psi(\cdot,\cdot)$ as in (2.1), then $h(y) = 0$ if and only if the controller $K(y)$ stabilizes *every single* plant in $\{G(x),\ x \in X\}$. If $K(y)$ fails to stabilize even a single plant, then $h(y) = 1$. Thus minimizing the present choice of $h(\cdot)$ corresponds to solving the robust (or simultaneous) stabilization problem. Similarly, if $\psi(G,K)$ is chosen as in (2.2), then minimizing the associated $h(\cdot)$ corresponds to achieving the best possible *guaranteed* performance with robust stabilization.

It is widely believed that methods such as H_∞-norm minimization for achieving robust stabilization, and μ-synthesis for achieving guaranteed performance and robust stabilization, lead to overly conservative designs. Much of the conservatism of the designs can be attributed to the worst-case nature of the associated cost function. To illustrate this point, consider the situation depicted in Fig. 1, where the function $g(\cdot, y)$ is plotted for two different values of y. From this stylized diagram, it can be seen that if the objective function $h(y)$ is defined as in (2.3), then the controller $K(y_1)$ would be preferred to the controller $K(y_2)$, even though the latter controller seems to outperform the former for "most" plants $G(x)$. Thus, if a worst-case objective function of the type (2.3) is used, then the control system designer is forced to focus all the effort and attention to bring down the value of $g(x,y)$ *for the value of x at which the supremum in* (2.3) *occurs*, which could be rather an unrepresentative plant.

It seems much more reasonable to settle for controllers that work satisfactorily "most of the time". One way to capture this intuitive idea in a mathematical framework is to introduce a probability measure P_X on the set X, that reflects one's prior belief on the way that the "true" plant $G(x)$ is distributed in the set of possible plants $\{G(x),\ x \in X\}$. For instance, in a problem of robust stabilization, G_0 can be a nominal, or most likely, plant model, and the probability measure P_X can be "peaked" around G_0. The more confident one is about the nominal plant model G_0, the more sharply peaked the probability measure P_X can be. Once the probability measure P_X is chosen, the objective function to be minimized can be defined as

$$f(y) := E_{P_X}[g(x,y)] = E_{P_X}[\psi(G(x), K(y))]. \tag{2.4}$$

Thus $f(y)$ is the *expected or average* performance of the controller $K(y)$ when the plant is distributed according to the probability measure P_X. For instance, if P_X is chosen as the uniform measure in Fig. 1, then the controller $K(y_2)$ would be preferred to $K(y_1)$ since the area under the curve of $g(x,y_2)$ is less than that of $g(x,y_1)$. Thus the expected value type of objective function captures the intuitive idea that a controller can occasionally be permitted to perform poorly for some plant conditions, provided these plant conditions are not too likely to occur.

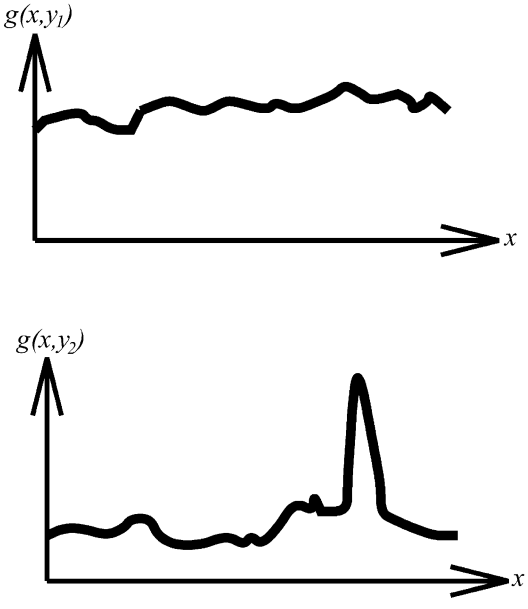

Fig. 1. Conservatism of the worst-case objective function.

While the worst-case objective function defined in (2.3) is easy to understand and to interpret, the interpretation of the expected-value type of objective function defined in (2.4) needs a little elaboration. Suppose $\psi(\cdot,\cdot)$ is defined as in (2.1). Then $f(y)$ is the measure (or "volume" with respect to the measure P_X) of the subset of $\{G(x),\ x\in X\}$ that *fails* to be stabilized by the controller $K(y)$. Alternatively, one can assert with confidence $1-f(y)$ that the controller $K(y)$ stabilizes a plant $G(x)$ selected at random from $\{G(x),\ x\in X\}$ according to the probability measure P_X.

3. Various types of "near" minima

In the previous section, an abstract formulation of the robust controller synthesis problem was given, which ultimately involves finding the minimum (and a minimizer) of a function $f: Y\to[0,1]$, where Y is the set of controller parameter vectors. There are many problems, such as those mentioned in Section 1, in which finding the *exact* minimum value f^* of $f(\cdot)$ is NP-hard. More precisely, given a number f_0, it is NP-hard to determine whether or not $f_0\geqslant f^*$. In such cases, one has to be content with "nearly" minimizing $f(\cdot)$. The objective of this section is to introduce three different definitions of "near minima", and to discuss their significance.

Definition 1. Suppose $f: Y\to\Re$ and that $\varepsilon>0$ is a given number. A number $f_0\in\Re$ is said to be a *Type 1 near minimum of* $f(\cdot)$ *to accuracy* ε, or an *approximate near minimum of* $f(\cdot)$ *to accuracy* ε, if

$$\inf_{y\in Y} f(y)-\varepsilon\leqslant f_0\leqslant \inf_{y\in Y} f(y)+\varepsilon \tag{3.1}$$

or equivalently

$$\left| f_0-\inf_{y\in Y} f(y)\right|\leqslant\varepsilon.$$

An approximate near minimum perhaps corresponds most closely to what we normaly think of as a "near" minimum. Unfortunately, in some robust stability problems, it is NP-hard to compute even an *approximation* to f^*; see Coxson and DeMarco (1994) for a result along this direction. In other words, there exist problems in which it is NP-hard to determine not only the exact minimum value f^*, but even a Type 1 (or approximate) near minimum. Thus it is necessary to look for other notions of a near minimum. One such notion is provided in the next definition.

Definition 2. Suppose $f: Y\to\Re$, that P_Y is a given probability measure on Y, and that $\alpha>0$ is a given number. A number $f_0\in\Re$ is said to be a *Type 2 near minimum of* $f(\cdot)$ *to level* α, or a *probable near minimum of* $f(\cdot)$ *to level* α, if $f_0\geqslant f^*$, and in addition

$$P_Y\{y\in Y: f(y)<f_0\}\leqslant\alpha.$$

The notion of a probable near minimum can be interpreted as follows: f_0 is a probable near minimum of $f(\cdot)$ to level α if there is an "exceptional set" S with $P_Y(S)\leqslant\alpha$ such that

$$\inf_{y\in Y} f(y)\leqslant f_0\leqslant \inf_{y\in Y\setminus S} f(y). \tag{3.2}$$

In other words, f_0 is bracketed by the infimum of $f(\cdot)$ over all of Y, and the infimum of $f(\cdot)$ over "nearly" all of Y.

It turns out that there exist "efficient" algorithms for finding a Type 2 (or probable) near minimum of *any* function $f(\cdot)$. Examples of such algorithms are presented in Khargonekar and Tikku (1996), Tempo et al. (1997), for the case where the underlying probability distribution function is continuous. This assumption is removed in Vidyasagar (1997a, Lemma 11.1, p. 357). However, these algorithms require that the function value $f(y)$ be computable *exactly* for any given $y\in Y$. To cater to situations in which this is not possible, we introduce yet one more type of near minimum.

Definition 3. Suppose $f: Y\to\Re$, that P_Y is a given probability measure on Y, and that $\varepsilon,\alpha>0$ are given numbers. A number $f_0\in\Re$ is said to be a *Type 3 near minimum of* $f(\cdot)$ *to accuracy* ε *and level* α, or a *probably approximate near minimum of* $f(\cdot)$ *to accuracy* ε *and level* α, if $f_0\geqslant f^*-\varepsilon$, and in addition

$$P_Y\{y\in Y: f(y)<f_0-\varepsilon\}\leqslant\alpha.$$

Another way of saying this is that there exists an "exceptional set" $S\subseteq Y$ with $P_Y(S)\leqslant\alpha$ such that

$$\inf_{y\in Y} f(y)-\varepsilon\leqslant f_0\leqslant \inf_{y\in Y\setminus S} f(y)+\varepsilon. \tag{3.3}$$

A comparison of (3.1), (3.2) and (3.3) brings out clearly the relationships between the various types of near minima.

4. A general approach to randomized algorithms

In this section, a general approach is outlined that could be used to develop randomized algorithms for minimizing an objective function of type (2.4). Subsequent sections contain a study of some specific situations in which this general approach could be profitably applied.

4.1. The UCEM property

Let us return to the specific problem of minimizing the type of objective function introduced in (2.4),

namely

$$f(y) = E_{P_x}[g(x, y)].$$

In general, evaluating an expected value *exactly* is not an easy task, since an expected value is just an integral with respect to some measure. However, it is possible to *approximate* an expected value to arbitrarily small error, as follows: Let us introduce the notation

$$g_y(x) := g(x, y), \quad \forall x \in X, \ \forall y \in Y.$$

Thus for each $y \in Y$, the function $g_y(\cdot)$ maps X into [0,1]. Now define the associated family of functions $\mathscr{G} := \{g_y(\cdot), y \in Y\}$. Suppose now that $\mathbf{x} := [x_1 \ldots x_m]^t \in X^m$ is a collection of i.i.d. samples in X, generated according to the probability measure P_X. For each function $g_y(\cdot) \in \mathscr{G}$, one can define its *empirical mean* based on the multisample $\mathbf{x}$ as

$$\hat{f}(y) = \hat{E}(g_y; \mathbf{x}) := \frac{1}{m} \sum_{j=1}^{m} g_y(x_j)$$

$$= \frac{1}{m} \sum_{j=1}^{m} g(x_j, y), \quad y \in Y. \tag{4.1}$$

In other words, the expected value of $f(y)$ is approximated by the *average* performance of the controller $K(y)$ on the randomly generated plants $G(x_1), \ldots, G(x_m)$. Now let

$$q(m, \varepsilon; \mathscr{G}) := P_X^m \left\{ \mathbf{x} \in X^m : \sup_{g_y \in \mathscr{G}} |\hat{E}(g_y; \mathbf{x}) - E_{P_x}(g_y)| > \varepsilon \right\}. \tag{4.2}$$

Observe that an equivalent way of writing $q(m,\varepsilon;\mathscr{G})$ is as follows:

$$q(m, \varepsilon; \mathscr{G}) := P_X^m \left\{ \mathbf{x} \in X^m : \sup_{y \in Y} |\hat{E}(g_y; \mathbf{x}) - f(y)| > \varepsilon \right\}.$$

Thus after m i.i.d. plants have been generated and an empirical mean $\hat{E}(g_y; \mathbf{x})$ has been computed for each function, it can be said with confidence $1 - q(m, \varepsilon; \mathscr{G})$ that *every single* empirical mean is within ε of the corresponding true value $f(y) = E_{P_x}(g_y)$. The family $\mathscr{G}$ is said to *have the UCEM property* if $q(m, \varepsilon; \mathscr{G}) \to 0$ as $m \to \infty$ for each $\varepsilon > 0$.

4.2. An approach to finding approximate near minima with high confidence

Suppose the family $\mathscr{G}$ does indeed have the UCEM property. Let $\varepsilon, \delta \in (0,1)$ be specified accuracy and confidence parameters, respectively. Choose m large enough that $q(m, \varepsilon; \mathscr{G}) < \delta$. Then it can be said with confidence $1 - \delta$ that

$$|f(y) - \hat{f}(y)| \leqslant \varepsilon, \quad \forall y \in Y.$$

In other words, the function $\hat{f}(\cdot)$ is a *uniformly close* approximation to the original objective function $f(\cdot)$. Hence it readily follows that an *exact* minimizer of $\hat{f}(\cdot)$ is also an *approximate* near minimizer of $f(\cdot)$ to accuracy ε. Thus we can minimize the function $\hat{f}(\cdot)$ using our favourite method, and in this way we would be able to generate an approximate near minimizer of $f(\cdot)$

Note that the above algorithm is an example of a *randomized* algorithm, in the sense that there is a nonzero probability (namely, $q(m, \varepsilon; \mathscr{G})$) that the algorithm may fail to produce an approximate near minimum of $f(\cdot)$. By increasing the integer m of x-samples used in computing the empirical mean $\hat{E}(g_y; \mathbf{x})$, this failure probability can be made *arbitrarily small*, but it can never be made exactly equal to zero.

Now we tackle a different problem. How may we go about finding a *probable* near minimum of $\hat{f}(\cdot)$ using randomized methods? By the above discussion, such a *probable* near minimum of the approximate objective function $\hat{f}(\cdot)$ would automatically be a *probably approximate* near minimum of the original objective function $f(\cdot)$. We now recall a method introduced in Tempo et al. (1997) and Khargonekar and Tikku (1996). In these references, the validity of this method is established under the assumption that the underlying probability distribution function is continuous; this assumption is removed in Vidyasagar (1997a, Lemma 11.1, p. 357).

Suppose P_Y is a probability measure on Y, and that $\alpha, \delta \in (0,1)$ are specified level and confidence parameters, respectively. Choose an integer n such that

$$(1 - \alpha)^n \leqslant \delta, \text{ or equivalently } n \geqslant \frac{\lg(1/\delta)}{\lg[1/(1 - \alpha)]}. \tag{4.3}$$

Generate independent identically distributed (i.i.d.) samples $y_1, \ldots, y_n \in Y$ distributed according to P_Y. Define

$$\bar{h} := \min_{1 \leqslant i \leqslant n} h(y_i).$$

Then it can be said with confidence at least $1 - \delta$ that $\bar{h}$ is a probable near minimum of $h(\cdot)$ to level α. This is because

$$P_Y^n\{\mathbf{y} \in Y^n : P_Y\{y \in Y : h(y) < \bar{h}(\mathbf{y})\} > \alpha\} \leqslant (1 - \alpha)^n.$$

See e.g. Vidyasagar (1997a, Lemma 11.1, p. 357). Note that a similar bound is proved in Khargonekar and Tikku (1996) with the additional assumption that the distribution function of the probability measure P_Y is continuous.

Finally, a word about the "sample complexity" estimate given in (4.3) for the number n of i.i.d. samples that need to be generated to apply the algorithm. A perusal of the proof given above or in Vidyasagar (1997a) shows that the estimate for n is *the best possible*, because one can always envisage a situation whereby

$P_Y\{y \in Y: h(y) < \bar{h}(\mathbf{y})\}$ exactly equals α, in which case all the inequalities above become equalities.

4.3. An algorithm for finding probably approximate near minima

The two ideas in the preceding subsection can be combined to produce two distinct randomized algorithms for finding a *probably approximate* (or Type 3) near minimum of an objective function $f(\cdot)$ of the form (2.4). The first algorithm is "universal", while the second algorithm is applicable only to situations where an associated family of functions has the UCEM property. The sample complexity estimates for the first "universal" algorithm are the best possible, whereas there is considerable scope for improving the sample complexity estimates of the second algorithm.

Suppose real parameters $\varepsilon,\alpha,\delta > 0$ are given; the objective is to develop a randomized algorithm that constructs a probably approximate (Type 3) near minimum of

$$f(y) := E_{P_x}[g(x,y)]$$

to accuracy ε and level α, with confidence $1-\delta$. In other words, the probability that the randomized algorithms fails to find a probably approximate near minimum to accuracy ε and level α must be at most δ.

Algorithm 1. *Choose integers*

$$n \geqslant \frac{\lg(2/\delta)}{\lg[1/(1-\alpha)]} \quad \text{and} \quad m \geqslant \frac{1}{2\varepsilon^2}\ln\frac{4n}{\delta}. \tag{4.4}$$

Generate i.i.d. samples $y_1,\ldots,y_n \in Y$ *according to* P_Y *and* $x_1,\ldots,x_m \in X$ *according to* P_X. *Define*

$$\hat{f}_i := \frac{1}{m}\sum_{j=1}^{m} g(x_j,y_i),\ i=1,\ldots,n \quad \text{and} \quad \hat{f}_0 := \min_{1\leqslant i\leqslant n} \hat{f}_i.$$

Then with confidence $1-\delta$, *it can be said that* $\hat{f}_0$ *is a probably approximate* (*Type* 3) *near minimum of* $f(\cdot)$ *to accuracy* ε *and level* α.

The proof of the claim in Algorithm 1 is easy. Once the i.i.d. samples $y_1,\ldots,y_n$ are generated where n satisfies (4.4), one can define

$$\bar{f} := \min_{1\leqslant i\leqslant n} f(y_i).$$

Then it follows from Lemma 11.1 of Vidyasagar (1997a) that, with confidence $1-\delta/2$ (*not* $1-\delta$ – compare (4.4) with (4.3)), the number $\bar{f}$ is a probable near minimum of $f(\cdot)$ to level α. Now consider the *finite* family of functions $\mathscr{A} := \{g(\cdot,y_i),\ i=1,\ldots,n\}$, and note that $2ne^{-2m\varepsilon^2} \leqslant \delta/2$ in view of Hoeffding's inequality. Hence it follows that with confidence $1-\delta/2$, we have

$$|f(y_i) - \hat{f}_i| \leqslant \varepsilon, \quad \text{for } i=1,\ldots,n.$$

In particular, it follows that

$$|\hat{f}_0 - \bar{f}| \leqslant \varepsilon. \tag{4.5}$$

Combining the two statements shows that, with confidence $1-\delta$, $\hat{f}_0$ is a probably approximate (Type 3) near minimum of $f(\cdot)$ to accuracy ε and level α.

While Algorithm 1 is "universal" in the sense that it requires no assumptions about the nature of the function $g(\cdot,\cdot)$, it has the drawback that the number m of x-samples is dependent on n, the number of y samples. It is now shown that, if the associated family of functions $\mathscr{G}$ defined previously has the UCEM property, then it is possible to make the integer m *independent* of the level parameter α.

To state the algorithm, recall the notation

$$q(m,\varepsilon;\mathscr{G}) := P_X^m\left\{\mathbf{x}\in X^m: \sup_{g_y\in\mathscr{G}} |\hat{E}(g_y;\mathbf{x}) - E_{P_x}(g_y)| > \varepsilon\right\}.$$

Suppose the family $\mathscr{G}$ has the UCEM property, i.e., that $q(m,\varepsilon;\mathscr{G}) \to 0$ as $m \to \infty$ for each $\varepsilon > 0$.

Algorithm 2. *Select integers* n,m *such that*

$$n \geqslant \frac{\lg(2/\delta)}{\lg[1/(1-\alpha)]} \quad \text{and} \quad q(m,\varepsilon;\mathscr{G}) \leqslant \delta/2. \tag{4.6}$$

Generate i.i.d. samples $y_1,\ldots,y_n \in Y$ *according to* P_Y *and* $x_1,\ldots,x_m \in X$ *according to* P_X. *Define*

$$\hat{f}_i := \hat{f}(y_i) = \frac{1}{m}\sum_{j=1}^{m} f(x_j,y_i),\ i=1,\ldots,n \quad \text{and}$$

$$\hat{f}_0 := \min_{1\leqslant i\leqslant n} \hat{f}_i.$$

Then with confidence $1-\delta$, *it can be said that* $\hat{f}_0$ *is a probably approximate* (*Type* 3) *near minimum of* $f(\cdot)$ *to accuracy* ε *and level* α.

It can be seen by comparing (4.4) and (4.6) that the only difference between Algorithms 1 and 2 is in the number m of x-samples. The key point to note is that m is *independent* of the integer n, which in turn depends on the level parameter α. We shall return to this point later.

Finally, a general philosophical remark about the nature of the "optimal" controller that is generated using either of the above randomized algorithms. Note that the objective function $f(\cdot)$ that is being minimized is the expected performance of a controller under an assumed probability distribution on the family of plants. Now it can certainly happen that a particular controller $K(y)$ has a very low value of $f(y)$ for one probability measure P_X, and a very high value for another choice of P_X. However, since the "optimal" controller is chosen by minimizing the empirical mean $\hat{f}(y)$ *corresponding to the given set of randomly generated plants*, which in turn correspond to the chosen probability measure P_X, the optimal

controller generated by the algorithms used here will always look good under the chosen P_X.

5. Some sufficient conditions for the UCEM property

In this section, brief review is given of some known sufficient conditions for a family of functions to have the UCEM property. These sufficient conditions are based on two powerful notions known as the Vapnik–Chervonenkis (VC-) dimension, and the Pollard (P-) dimension, respectively. For a thorough and contemporary treatment of the UCEM property and related issues, see Vidyasagar (1997a).

5.1. Definitions of the VC- and P-dimension

We begin with a definition of the VC-dimension, which is applicable to *binary-valued* functions. Note that it is more common to define the VC-dimension for collections of sets; see e.g., Vidyasagar (1997a, p. 69). However, it is easy to see that there is a one-to-one correspondence between collections of subsets of X, and families of *binary-valued* functions defined on X. Specifically, suppose $A \subseteq X$; then its indicator function $I_A(\cdot)$ maps X into $\{0,1\}$. Conversely, suppose $a(\cdot)$ maps X into $\{0,1\}$; then its "support" is a subset of X.

Definition 4. Suppose $\mathscr{A}$ is a family of measurable functions mapping X into $\{0,1\}$. A set $S = \{x_1, \ldots, x_n\}$ is said to be *shattered* by $\mathscr{A}$ if each of the 2^n functions mapping S into $\{0,1\}$ is the restriction to S of some function in $\mathscr{A}$. Equivalently, S is shattered by $\mathscr{A}$ if, for every subset $A \subseteq S$, there exists a corresponding function $a_A(\cdot) \in \mathscr{A}$ such that $a_A(x_i) = 1$ if $x_i \in A$ and $a_A(x_i) = 0$ if $x_i \notin A$. The *Vapnik–Chervonenkis* (*VC-*) *dimension* of $\mathscr{A}$, denoted by *VC-dim*($\mathscr{A}$), is the largest integer n such that there exists a set of cardinality n that is shattered by $\mathscr{A}$.

See Vidyasagar (1997a), Section 4.1 for several examples of the explicit computation of the VC-dimension.

The VC-dimension is defined for families of *binary-valued* functions. The corresponding notion for families of [0,1]-valued functions is referred to by various authors as the Pollard dimension or the pseudo-dimension; it is referred to here by the neutral symbol P-dimension.

Definition 5. Suppose $\mathscr{A}$ is a family of measurable functions mapping X into $[0,1]$. A set $S = \{x_1, \ldots, x_n\}$ is said to be *P-shattered* by $\mathscr{A}$ if there exists a vector $\mathbf{c} \in [0,1]^n$ such that, for every binary vector $\mathbf{e} \in \{0,1\}^n$, there exists a corresponding function $a_{\mathbf{e}}(\cdot) \in \mathscr{A}$ such that

$$a_{\mathbf{e}}(x_i) \geqslant c_i \quad \text{if } e_i = 1 \quad \text{and} \quad a_{\mathbf{e}}(x_i) < c_i \quad \text{if } e_i = 0.$$

The *P-dimension* of $\mathscr{A}$, denoted by *P-dim*($\mathscr{A}$), is the largest integer n such that there exists a set of cardinality n that is P-shattered by $\mathscr{A}$.

It is easy to see that if every function in $\mathscr{A}$ is binary-valued, then VC-dim($\mathscr{A}$) = P-dim($\mathscr{A}$). More generally, the two dimensions are related as follows: Let $\eta : \Re \to \{0,1\}$ denote the *Heaviside* function (sometimes referred to also as the "step" function). Given a family of functions $\mathscr{A}$ mapping X into $[0,1]$, define an associated family of functions $\bar{\mathscr{A}}$ mapping $X \times [0,1]$ into $\{0,1\}$ as follows: For each $a(\cdot) \in \mathscr{A}, x \in X, c \in [0,1]$, let $\bar{a}(x,c) = \eta[a(x) - c]$. Now let a vary over $\mathscr{A}$; the corresponding family of functions $\bar{a}$ is the collection $\bar{\mathscr{A}}$.

Lemma 1. *With all symbols defined as above, we have*

$$P\text{-}dim(\mathscr{A}) = VC\text{-}dim(\bar{\mathscr{A}}).$$

The above observation is explicitly stated and proved in Macintyre and Sontag (1993); see also Vidyasagar (1997a, Lemma 10.1, p. 300).

5.2. Finiteness of the VC- and P-dimensions implies the UCEM property

The significance of the VC- and P-dimension arises from the fact that the finiteness of these dimensions is sufficient for a family of functions to have the UCEM property. Specifically, we have the following results:

Theorem 1. *Suppose $\mathscr{A}$ is a family of measurable functions X mapping into $\{0,1\}$, and that $VC\text{-}dim(\mathscr{A}) \leqslant d < \infty$. Then $\mathscr{A}$ has the UCEM property, whatever be the probability measure P_X. Moreover*

$$q(m, \varepsilon; \mathscr{A}) \leqslant 4\left(\frac{2em}{d}\right)^d \exp(-m\varepsilon^2/8), \quad \forall m, \varepsilon. \tag{5.1}$$

Moreover, $q(m, \varepsilon; \mathscr{A}) \leqslant \delta$ provided

$$m \geqslant \max\left\{\frac{16}{\varepsilon^2}\ln\frac{4}{\delta}, \frac{32d}{\varepsilon^2}\ln\frac{32e}{\varepsilon^2}\right\}. \tag{5.2}$$

This seminal theorem is proved in Vapnik and Chervonenkis (1971); see also Vapnik (1982). A complete proof "from scratch" is also given in Vidyasagar (1997a, see Theorem 7.2, p. 198). Since the above result holds for *every* probability measure P_X on X, the UCEM property in this case is said to be "distribution-free". The bound proved in (5.1) has since been superceded by other bounds that are more difficult to prove. These bounds differ from (5.1) in that the exponent $-m\varepsilon^2/8$ is replaced by $-m\varepsilon^2$ in Parrondo and van den Broeck (1993), and by the "best possible" exponent $-2m\varepsilon^2$, which is the same as the Hoeffding bound (1963), Hoeffding (1963) in Fine (1999).

Theorem 2. *Suppose $\mathscr{A}$ is a family of measurable functions X mapping into* [0,1], *and that* $2 \leqslant P\text{-}dim(\mathscr{A}) \leqslant d < \infty$. *Then $\mathscr{A}$ has the UCEM property, whatever be the probability measure P_X. Moreover, for all $\varepsilon < 1$, we have*

$$q(m,\varepsilon;\mathscr{A}) \leqslant 8\left(\frac{16e}{\varepsilon}\ln\frac{16e}{\varepsilon}\right)^d \exp(-m\varepsilon^2/32), \ \forall m, \quad \forall \varepsilon. \tag{5.3}$$

Thus $q(m,\varepsilon;\mathscr{A}) \leqslant \delta$ provided

$$m \geqslant \frac{32}{\varepsilon^2}\left[\ln\frac{8}{\delta} + d\left(\ln\frac{16e}{\varepsilon} + \ln\ln\frac{16e}{\varepsilon}\right)\right]. \tag{5.4}$$

The above theorem is stated and proved in this form in Vidyasagar (1997a, Theorem 7.1, p. 196), and is a slight refinement of some earlier results of Haussler (1992). In particular, the exponent in Vidyasagar (1997) has been computed more carefully than in previous works, and the condition $d \geqslant 2$ (which is a very trivial restriction) has been added.

5.3. Upper bounds for the VC-dimension

Theorems 1 and 2 bring out the importance of the VC- and P-dimension in the context of the UCEM property. Moreover, Lemma 1 shows that the P-dimension of a family of [0,1]-valued functions is just the VC-dimension of an associated family of binary-valued functions. Because of these theorems, several researchers have explored the VC-dimension of various families of binary-valued functions. Over the years an impressive array of upper bounds have been derived for the VC-dimension of a wide variety of function families. Note that upper bounds are enough to apply Theorems 1 and 2. Many of these bounds are collected and rederived in Vidyasagar (1997a, Chapter 10). However, there is one result that is particularly appropriate for the class of problems studied here; it is presented next.

A little terminology is first introduced towards this end. Suppose $X \subseteq \Re^k$, $Y \subseteq \Re^l$ for some integers k, l respectively. Now suppose $\tau_1(x,y), \ldots, \tau_t(x,y)$ are polynomials in x, y, and suppose that the degree *with respect to* y of each polynomial $\tau_i(x,y)$ is no larger than r (the degree with respect to x does not matter). For each $x \in X, y \in Y$, each polynomial inequality "$\tau_i(x,y) > 0$" evaluates to either "true" or "false". Now suppose $\phi(x,y)$ is a *Boolean formula* obtained from the expressions "$\tau_i(x,y) > 0$" using the standard logical connectives $\neg$ (not), $\vee$ (or), $\wedge$ (and) and $\Rightarrow$ (implies).[1] Let 1 correspond to "true", 0 to "false", and define, for each $y \in Y$,

$$A_y := \{x \in X : \phi(x,y) = 1\}, \quad \text{and} \quad \mathscr{A} := \{A_y : y \in Y\}.$$

Then $\mathscr{A}$ is a collection of subsets of X. The objective is to obtain an upper bound for the VC-dimension of $\mathscr{A}$.

The following theorem is a refinement of a result from Karpinski and Macintyre (1995), (1997), and is proved in this precise form in Vidyasagar (1997a, Corollary 10.2, p. 330).

Theorem 3. *With all symbols as above, we have*

$$VC\text{-}dim(\mathscr{A}) \leqslant 2l\lg(4ert). \tag{5.5}$$

6. Robust stabilization

In this section, the approach of the preceding section is applied to the problem of robustly stabilizing a given family of plants $\{G(x),\ x \in X\}$ using a single fixed controller selected from the family $\{K(y),\ y \in Y\}$. As before, define

$$g_y(x) := g(x,y) = \begin{cases} 1 & \text{if the pair } (G(x), K(y)) \text{ is unstable,} \\ & \quad \text{and} \\ 0 & \text{if the pair } (G(x), K(y)) \text{ is stable.} \end{cases}$$

Now it is shown that, under some reasonable assumptions, the family of binary-valued functions $\mathscr{G} := \{g_y(\cdot), y \in Y\}$ has the UCEM property.

Suppose the plants $G(x,s)$ and controllers $K(y,s)$ are of the following form:

$$G(x,s) = \frac{n_G(x,s)}{d_G(x,s)}, \quad \forall x \in X,$$

$$K(y,s) = \frac{n_K(y,s)}{d_K(y,s)}, \quad \forall y \in Y,$$

where n_G, d_G are *polynomials* in x, s, and the degree in s of d_G is at most α_s, while the degree in s of n_g is at most $\alpha_s - 1$ (that is, each plant $G(x,s)$ is strictly proper). The assumptions about $K(y,s)$ are entirely analogous. It is assumed that is a *proper* rational function of s, with McMillan degree β_s. Also, it is assumed that $n_K(y,s)$, $d_K(y,s)$ are polynomials in y of degree no larger than β_y. Finally, it is assumed that $X \subseteq \Re^k$, $Y \subseteq \Re^l$ for suitable integers k, l.

Theorem 4. *Under the above assumptions, the family of binary-valued functions $\mathscr{G} := \{g_y : y \in Y\}$ has finite VC-dimension. In particular*

$$VC\text{-}dim(\mathscr{G}) \leqslant 2l\lg[4e(\alpha_s + \beta_s)^2\beta_y]. \tag{6.1}$$

Remarks. It is important to note that the quantity $\alpha_s + \beta_s$, which is the order of the closed-loop system, appears *inside* the $\lg(\cdot)$ function. Thus the bounds given above will be rather small even for high-order control systems. However, the integer l, representing the number

[1] Note that such a formula is *quantifier-free* in the sense that it does not contain the modifiers $\exists$ (there exists) and $\forall$ (for all).

of "degrees of freedom" in the controller parameter y, appears linearly in the upper bound.

Proof. The proof consists of writing down the conditions for the closed-loop system to be stable as a Boolean formula entailing several polynomial inequalities in x and y, and then appealing to Theorem 3. For a fixed $x \in X, y \in Y$, the pair $[G(x,s), K(y,s)]$ is stable if and only if two conditions hold: (i) the closed-loop transfer function is proper, and (ii) the characteristic polynomial of the closed-loop system is Hurwitz. Since every plant $G(x,s)$ is *strictly* proper and every controller $K(y,s)$ is *proper*, the first condition is automatically satisfied and we are left with only the second condition. Next, the closed-loop characteristic polynomial equals

$$\theta(x,y,s) := n_G(x,s)\, n_K(y,s) + d_G(x,s)\, d_K(y,s). \tag{6.2}$$

This is a polynomial of degree $\alpha_s + \beta_s$ in s, where each coefficient of θ is a polynomial in x, y; moreover, the degree of each coefficient with respect to y is no larger than β_y. The stability of the polynomial θ can be tested by forming its Hurwitz determinants as in Gantmacher (1959, pp. 190ff). Let $\delta_c := \alpha_s + \beta_s$ denote the degree of θ with respect to s, and write

$$\theta(x,y,s) = \sum_{i=0}^{\delta_c} a_{\delta_c - i}(x,y)\, s^i.$$

Let H_i denote the ith Hurwitz determinant of θ; then H_i is of the form

$$H_i(x,y) = \begin{vmatrix} a_1 & a_3 & a_5 & \dots & a_{2i-1} \\ a_0 & a_2 & a_4 & \dots & a_{2i-2} \\ 0 & a_1 & a_3 & \dots & a_{2i-3} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & a_i \end{vmatrix},$$

where $a_i(x,y)$ is written as a_i in the interests of clarity. Note that a_i is taken as zero if $i > \delta_c$. Now θ is a Hurwitz polynomial if and only if $H_i(x,y) > 0$ for $i = 1, \dots, \delta_c$. Hence the pair $[G(x,s), K(y,s)]$ is stable, and $\psi(x,y) = 0$, if and only if

$$[H_1(x,y) > 0] \wedge \dots \wedge [H_{\delta_c}(x,y) > 0]. \tag{6.3}$$

This is a Boolean formula. Moreover, $\psi(x,y) = 1$ is just the *negation* of the above Boolean formula, and is thus another Boolean formula with just the same atomic polynomial inequalities, with the connectives changed according to De Morgan's formula.

To apply Theorem 3, it is necessary to count the number of atomic polynomial inequalities, and their maximum degree with respect to y. First, the number of atomic inequalities is $t = \delta_c = \alpha_s + \beta_s$. Next, let us examine the degree of each of the atomic polynomial inequalities with respect to y. Each Hurwitz determinant $H_i(x,y)$ is the determinant of an $i \times i$ matrix, each of whose entries is a polynomial in y of degree β_y or less. Therefore the degree of H_i with respect to y is at most $\beta_y i$. Since the number of Hurwitz determinants is $\delta_c = \alpha_s + \beta_s$, we can take $r = (\alpha_s + \beta_s)\beta_y$. Finally, applying Theorem 3 and in particular the bound (5.5) leads to the estimate

$$\text{VC-dim}(\mathscr{G}) \leqslant 2l \lg(4ert) \leqslant 2l \lg[4e(\alpha_s + \beta_s)^2 \beta_y].$$

This concludes the proof. □

The reader might wonder why, in the above proof, the Hurwitz test for stability is used instead of the more widely known Routh test (which is almost always mislabelled as the "Routh–Hurwitz test"). The reason is economy in estimating the integer r. We leave it to the reader to show that, if the Hurwitz determinants are replaced by the elements in the first column of the Routh array, the estimate for r now jumps to $(\alpha_s + \beta_s)(\alpha_s + \beta_s + 1)\beta_y/2$, resulting in a roughly 50% increase in the estimate of the VC-dimension.

7. Weighted H_∞- or H_2-norm minimization

By using arguments entirely analogous to those in Section 6, it is possible to estimate the P-dimension of the family $\mathscr{G}$ in the case where the objective is to minimize the weighted H_∞-norm of the weighted H_2-norm of the closed-loop transfer function.

7.1. Weighted H_∞-norm minimization

We begin with the problem of minimizing the weighted H_∞-norm of the closed-loop transfer function. Let $W(s)$ be a given weighting function, and define the performance measure to be

$$\psi(G,K) := \begin{cases} 1 & \text{if the pair } (G,K) \\ & \text{is unstable,} \\ J(G,K)/[1 + J(G,K)] & \text{if the pair } (G,K) \\ & \text{is stable,} \end{cases}$$

where

$$J(G,K) = \|W(1 + GK)^{-1}\|_p, \tag{7.1}$$

where $p = \infty$ and $\|\cdot\|_\infty$ denotes the H_∞-norm. Let $G = G(x,s)$, $K = K(y,s)$ be as in Section 6, wherein $G = n_G/d_G$, $K = n_K/d_K$, and the coefficients of n_G, d_G are polynomials in x, and the coefficients of n_K, d_K are polynomials in y. As before, let β_y denote the maximum degree of any coefficient of n_K, d_K with respect to y. For convenience, let $\delta_c := \alpha_s + \beta_s$ denote the McMillan degree of the closed-loop transfer function, and let δ_w denote the McMillan degree of the *weighted* closed-loop transfer function $W(1 + GK)^{-1}$. As before, let $g(x,y) := \psi[G(x), K(y)]$. For each $y \in Y$ define the

corresponding function $g_y : X \to [0,1]$ by $g_y(x) := g(x,y)$, and define $\mathscr{G} := \{g_y : y \in Y\}$.

Theorem 5. *With all symbols as above, we have*

$$P\text{-}dim(\mathscr{G}) \leqslant 2l \lg[8e\beta_y\delta_w(2\delta_c + 2\delta_w + 1)]. \tag{7.2}$$

Remarks. As in the case of the bounds in (6.1) given in Theorem 4, the quantities δ_c which is the order of the closed-loop system and δ_w which is the order of the weighted closed-loop transfer function, both appear *inside* the $\lg(\cdot)$ function. Thus the bounds given above will be rather small even for high-order control systems. However, the integer l, representing the number of "degrees of freedom" in the parameter y, appears linearly in the upper bound.

Proof. Let $G = N_G/d_G, K = n_K/d_K, W = n_W/d_W$ denote *polynomial* coprime factorizations with respect to s of G, K, W, respectively. Then the closed-loop system is stable if and only the characteristic polynomial θ defined in (6.2) is Hurwitz. Moreover, the weighted sensitivity function can be expressed as

$$\frac{W}{1+KG} = \frac{n_W n_G n_K}{d_W(n_G n_K + d_G d_K)} =: \frac{n_M(x,y,s)}{d_M(x,y,s)}.$$

By assumption, each element of n_K, d_K is a polynomial in y of degree at most β_y. Hence it follows that each element of d_M is also a polynomial in y of degree at most β_y. Moreover, the degree of d_M with respect to s is δ_w.

For each $y \in Y$, define the associated function $\bar{g}_y : X \times [0,1] \to \{0,1\}$ by $\bar{g}_y(x,c) = \eta[g_y(x) - c]$, where $\eta(\cdot)$ is the Heaviside function. Let $\bar{\mathscr{G}} := \{\bar{g}_y : y \in Y\}$. Then it follows from Lemma 1 that P-dim$(\mathscr{G})$ = VC-dim$(\bar{\mathscr{G}})$. Hence we concentrate on estimating VC-dim$(\bar{\mathscr{G}})$.

For this purpose, we write "$\eta[g_y(x) - c] = 0$" as a Boolean formula involving several polynomial inequalities in x, y, c and then apply Theorem 3. Note that $\eta[g_y(x) - c] = 0 \Leftrightarrow g(x,y) - c < 0$. Let us write J as a shorthand for $J[G(x), K(y)]$. Then[2]

$$g(x,y) - c < 0 \Leftrightarrow [(c = 1) \wedge (J < \infty)] \vee [(c < 1) \wedge (J < c/(1-c))].$$

Now $J < \infty$ is just closed-loop stability. The results of Section 6 imply that this condition can be written in terms of δ_c polynomial inequalities, each of which has degree at most $\beta_y\delta_c$ with respect to y. Next, the condition $J < c/(1-c)$ says that the weighted transfer function $F := W(1+GK)^{-1}$ has H_∞-norm less than $b := c/(1-c)$. This is the case if and only if

(G,K) is stable, $b > F(\infty)$ and

$b^2 - F^*(jw)F(jw) > 0 \ \forall\omega$.

[2] Strictly speaking, we should write $\neg(c < 1)$ instead of $c = 1$.

Again, the first requirement involves δ_c polynomial inequalities of degree at most $\beta_y\delta_c$ in y. Next, since $F(s)$ is a *stable* rational function, we can express the condition "$b^2 - F^*(jw)F(jw) > 0 \ \forall\omega$" equivalently as a *polynomial* inequality in ω of the form "$p(\omega) > 0 \ \forall\omega$", where the coefficients of $p(\cdot)$ are rational functions of x, y, c. Now $p(\omega)$ is an *even* polynomial in ω of degree $2\delta_w$. Since $b > F(\infty)$, the polynomial $p(\cdot)$ is positive for sufficiently large ω. Hence the condition $p(\omega) > 0 \ \forall\omega$ is equivalent to the requirement that $p(\cdot)$ does not have any positive real zeros. By Jury (1977, Theorem 5.3, p. 154), this requirement can be expressed as follows: (i) Both the highest and smallest coefficient of $p(\cdot)$ are positive, and (ii) the number of sign changes in the Routh table for $p(\cdot)$ is $\delta_w - 1$. Now it is known (see, e.g., Gantmacher (1959, p. 198)) that there is a sign change in the first column of the ith row in the Routh table if and only if the ith Hurwitz determinant is negative. Thus Theorem 5.3 of Jury (1977) requires that, out of the $2\delta_w$ Hurwitz determinants of $p(\cdot)$, exactly $\delta_w - 1$ are negative. It is easy to see that this requirement can be written as a (very messy) Boolean formula involving the $2\delta_w$ atomic inequalities $H_1(p) > 0$ through $H_{2\delta_w}(p) > 0$, where $H_i(p)$ denotes the ith Hurwitz determinant of $p(\cdot)$. As in the case of the Hurwitz determinants for closed-loop stability, each of these determinants $H_i(p)$ is a polynomial in x, y, c, and the maximum degree of any determinant with respect to y is $2\beta_y\delta_w$. Since $\delta_w \geqslant \delta_c$, we can now apply Theorem 3 with

t = No. of inequalities $= 2\delta_c + 2\delta_w + 1$,

r = Max. degree w.r.t. $y = 2\beta_y\delta_w$.

This completes the proof. □

7.2. Weighted H_2-norm minimization

In this subsection, an upper bound is derived for the P-dimension of the family $\mathscr{G}$ in the case where the objective is to minimize the weighted H_2-norm of the closed-loop transfer function. Let all symbols be as in Section 7.1, except that we set $p = 2$ in (7.1).

Theorem 6. *With all symbols as above, we have*

$$P\text{-}dim(\mathscr{G}) \leqslant 2l \lg[8e(\delta_c + 1)(\delta_w + 1)\beta_y].$$

Proof. Given the plant parameter vector x, the controller parameter y, and the real parameter $c \in [0,1]$, we express the inequality "$\eta[g_y(x) - c] = 0$" as a Boolean formula involving polynomial inequalities in x, y, c. As before, we have

$$\eta[g_y(x) - c] = 0 \Leftrightarrow [(c = 1) \wedge (J < \infty)] \vee [(c < 1) \wedge (J < c/(1-c))].$$

Again as before, the condition $J < \infty$ is the closed-loop stability requirement, and as such, it can be expressed in

terms of δ_c polynomial inequalities of degree at most $\delta_c\beta_y$ in y. The only thing that is different in the present instance is the treatment of the condition

$$J < \frac{c}{1-c} \Leftrightarrow J^2 < \left[\frac{c}{1-c}\right]^2.$$

Now it is shown that $J^2[G(x), K(y)]$ is a rational function of x, y, whose degree with respect to y is no larger than $(\delta_w + 1)\beta_y$. For this purpose, let

$$H(x, y, s) := \frac{W(s)}{1 + G(x, s)K(y, s)}$$

denote the weighted closed-loop sensitivity function, and note that $J^2 = \|H\|_2^2$. Next, we use the results in Newton, Gould and Kaiser (1967), Appendix E, to compute J^2. Write

$$H(x, y, s) := \frac{\sum_{i=0}^{\delta_w - 1} c_i s^i}{\sum_{i=0}^{\delta_w} d_i s^i}.$$

This is the same notation as in Newton, et al. (1967), except that δ_w plays the role of the symbol n. Then (cf. Eq. (E.1-7) of Newton et al., 1967) $\|H\|_2^2 = a_{\delta_w - 1}/d_{\delta_w}$, where the vector $A := [a_0\ a_1 \ldots a_{\delta_w - 1}]$ satisfies a linear matrix equation $DA = C$ (cf. Eq. (E.1-19)). Now the solution for $a_{\delta_w - 1}$ can be written using Cramer's rule in the form

$$a_{\delta_w - 1} = \frac{\det[\bar{D}|C]}{\det D},$$

where $\bar{D}$ consists of all but the last column of D. Also (cf. (E.1-22) and (E.1-23)), the entries of D are just the coefficients d_i, whereas the entries of C (cf. (E.1-15)) are sums of products of two of the c_i's. We have already seen in the proof of Theorem 5 that each of the c_i, d_i is a polynomial in y of degree at most β_y. Hence each element of C is a polynomial in y of degree at most $2\beta_y$. Also, $\det D$ is degree at most $\delta_w\beta_y$ whereas $\det[\bar{D}|C]$ has degree at most $(\delta_w + 1)\beta_y$ with respect to y. This shows that $a_{\delta_w - 1}$ is a rational function in y of degree at most $(2\delta_w + 1)\beta_y$, and $\|H\|_2^2$ is a rational function in y of degree at most $2(\delta_w + 1)\beta_y$. Now we can apply Theorem 3 with the following values:

$$t = \delta_c + 1, \quad r = 2(\delta_w + 1)\beta_y.$$

(Here we use the obvious fact that $\delta_c \leqslant \delta_w$.) This leads to the desired conclusion. □

8. Sample complexity considerations

Vapnik–Chervonekis theory of uniform convergence of empirical means to their true values represents a great intellectual achievement in Vapnik–Chervonekis (1981). In as much as the theory provides both *necessary as well as sufficient conditions* for the UCEM property to hold, the only room for improvement in the theory is that of obtaining less conservative estimates for the sample complexity, that is, the number of randomly generated samples that are required to achieve a specified level of accuracy and confidence. In this section, we study the issue of sample complexity for both the general UCEM problem as well as the specific problem studied in this paper of minimizing an expected-value type of objective function. In particular, the sample complexity bounds obtained from VC theory are compared with those obtained from the older Hoeffding's inequality, applied to a finite family of functions. The conclusion is that, in most practical situations, the bounds based on VC theory *as they are at present* lead to much larger numbers of samples than those based on Hoeffding's inequality, applied to a finite family of functions. However, the bounds obtained from Hoeffding's inequality are in some sense the "best possible", whereas there is considerable scope for improvement in the bounds based on VC theory. Hence it is possible that as the latter bounds improve, they may eventually prove to be less conservative than the bounds based on Hoeffding's inequality. Thus it clearly behaves the statistical learning theory research community to continue studying the issue of sample complexity using VC theory.

Let us compare the sample complexity estimates obtained through Hoeffding's inequality and VC theory, in the case where the underlying family of functions is finite. Suppose an accuracy parameter $\varepsilon > 0$ and a confidence parameter $\delta > 0$ are specified. The aim is to estimate an integer $m_0(\varepsilon,\delta)$, referred to as the *sample complexity*, to ensure that

$$P_X^m\left\{\mathbf{x} \in X^m: \max_{1 \leqslant i \leqslant n} |\hat{E}(f_i; \mathbf{x}) - E_{P_X}(f_i)| > \varepsilon\right\} \leqslant \delta,$$
$$\forall m \geqslant m_0(\varepsilon, \delta).$$

Suppose Hoeffding's inequality is used to estimate the left side. Then it is enough to choose m_0 large enough to ensure that $2ne^{-2m_0\varepsilon^2} \leqslant \delta$, or

$$m_{0,\text{Hoeff}}(\varepsilon, \delta) = \frac{\ln(2n/\delta)}{2\varepsilon^2}.$$

It is important to note that the integer n, corresponding to the number of functions, appears *inside* the $\ln(\cdot)$. Now let d denote the VC-dimension of the family of functions $\{f_1, \ldots, f_n\}$. Then it follows from (5.2) that

$$m_{0,\text{VC}}(\varepsilon, \delta) = \max\left\{\frac{16}{\varepsilon^2}\ln\frac{4}{\delta}, \frac{32d}{\varepsilon^2}\ln\frac{32e}{\varepsilon^2}\right\}\Xi$$

To facilitate comparison, let us replace $2n/\delta$ by the larger term $4n/\delta$, so that

$$m_{0,\text{Hoeff}}(\varepsilon, \delta) = \frac{\ln(4n/\delta)}{2\varepsilon^2} = \frac{\ln n + \ln(4/\delta)}{2\varepsilon^2}.$$

Then a simple computation shows that

$$m_{0,\mathrm{VC}}(\varepsilon,\delta) \leqslant m_{0,\mathrm{Hoeff}}(\varepsilon,\delta) \Rightarrow \frac{16\ln(4/\delta)}{\varepsilon^2} \leqslant \frac{\ln n + \ln(4/\delta)}{2\varepsilon^2}$$

$$\Rightarrow \frac{31\ln(4/\delta)}{2\varepsilon^2} \leqslant \frac{\ln n}{2\varepsilon^2} \Rightarrow \ln n \geqslant 31\ln(4/\delta)$$

$$\Rightarrow n \geqslant \left(\frac{4}{\delta}\right)^{31}.$$

It is obvious that, for any reasonable value of δ, the corresponding value of n is astronomically large. Thus, in any practical situation, one is always better off using the Hoeffding bound (4.4) instead of the VC-bound (5.1).

Now let us study the specific problem forming the subject of the present paper, namely, the minimization of an expected-value type of objective function. The above computation can be modified to compare the sample complexities of Algorithms 1 and 2. Suppose an accuracy $\varepsilon > 0$, a confidence $\delta > 0$, and a level $\alpha > 0$ are specified. Then, in both algorithms, we have

$$n = \frac{\lg(2/\delta)}{\lg[1/(1-\alpha)]} \approx \frac{2/\delta}{\alpha}.$$

In Algorithm 1, we have

$$m_{\mathrm{Hoeff}} = \frac{1}{2\varepsilon^2}\ln\frac{4n}{\delta} \approx \frac{\ln(1/\alpha) + \ln(4/\delta) + \ln\ln(2/\delta)}{2\varepsilon^2}.$$

In Algorithm 2, we have, as before

$$m_{\mathrm{VC}} = \max\left\{\frac{16}{\varepsilon^2}\ln\frac{4}{\delta}, \frac{32d}{\varepsilon^2}\ln\frac{32e}{\varepsilon^2}\right\}.$$

If we ignore the $\ln\ln(2/\delta)$ term in m_{Hoeff} as being insignificantly small, then

$$m_{\mathrm{VC}} \leqslant m_{\mathrm{Hoeff}} \Rightarrow \ln(1/\alpha) \geqslant 31\ln(4/\delta) \Rightarrow \alpha \leqslant \left(\frac{\delta}{4}\right)^{31}.$$

Again, for any reasonable values of α and δ, the above inequality will not hold, which implies that in practical situations, one is better off using Algorithm 1 instead of Algorithm 2.

However, this does not mean that all of the VC-dimension estimates derived in the preceding sections are of mere academic interest. The conservatism of the sample complexity estimates in (5.1) and (5.2) is well known in the statistical learning theory research community, and many researchers have attempted to improve these bounds. See Kowalczyk, Ferra, and Szymanski (1995) for an example of such an approach. In Patel et al. (2000) the authors consider the problem of designing a first-order controller for the longitudinal axis of an aircraft. In this instance, it turns out that $d = 118$ using Theorem 3. Setting $\alpha = 0.1$, $\varepsilon = 0.1$, $\delta = 0.01$ leads to $n = 51$. Applying the bound of (5.4) leads to the estimate $m = 2{,}996{,}647$, which is clearly unrealistically large. On the other hand, using the estimate of Kowalczyk et al. (1995) with the "smoothness constant" C equals 1 leads to $m = 620$, which is quite a reasonable number. Note that in the present case, the sample complexity estimate of Algorithm 1 gives $m_{\mathrm{Hoeff}} \approx 500$, which is comparable to the VC-type of bound *provided the bound of* Kowalczyk et al. (1995) *holds*. However, it is not known at present whether this particular bound is applicable to the class of problems studied here, or even if it does, how one goes about estimating the smoothness constant C. Clearly this issue merits further investigation.

9. Conclusions

In this paper, a beginning has been made towards the application of recent results from statistical learning theory, in particular the so-called Vapnik–Chervonenkis (VC)-theory, to the development of randomized algorithms for solving a wide variety of controller analysis and synthesis problems. Specifically, it has been shown that the randomized approach is applicable to problems of robust and simultaneous stabilization, and of weighted H_2/H_∞-norm minimization. Moreover, the results presented here show that a randomized approach is in fact applicable to *any* control problem wherein the desired design objective can be formulated as the satisfaction of a finite number of polynomial inequalities. Using this broad framework, future researchers would undoubtedly be able to extend this approach to many more problems.

In a related paper by Vidyasagar and Blondel (2001), it is shown that there exist efficient randomized algorithms for each of the NP-hard matrix problems mentioned in Section 1. The present paper goes farther by showing that even problems of H_2- and H_∞-optimal control are amenable to a randomized approach.

In order for the randomized approach to be useful in practice, it is imperative that the number of randomly generated plants and controllers be "reasonable". It has been shown that, if one applies the known results from VC-theory *in their original form*, then in fact one is better off using the naive "universal" Algorithm 1 based on the older Hoeffding inequality, rather than Algorithm 2 which is based on VC-theory. However, it has also been shown that, under suitable circumstances (see Section 8), VC-theory can lead to sample complexity estimates that are comparable to those obtained using Hoeffding's inequality. Moreover, whereas the estimates obtained using Hoeffding's inequality are the "best possible", there is still considerable room for improvement in the estimates obtained using VC-theory.

When all is said and done, however, it is perhaps natural to feel a twinge of disappointment about the underlying philosophy of the randomized approach, which can be crudely (and perhaps uncharitably) summarized as "Just try lots of plant-controller pairs and

pick the best controller". One could argue that this is what control system designers have been doing for decades, so what is so new about the randomized approach? What is new here is the *mathematical justification* for using randomization, and in particular, the *sample complexity estimates* that allow the designer to say how many randomly generated plants/controllers are enough. This is the added input that had been missing previously.

All in all, the randomized approach presents an interesting approach to solving problems in controller analysis and/or synthesis, which in their "pure" form are either intractable (NP-hard) or not amenable to other forms of analysis.

Acknowledgements

The author would like to acknowledge Prof. Roberto Tempo for having introduced him to this particular area of research, and Prof. Brian Anderson for drawing his attention to the work of Jury used in Section 7. He also thanks Prof. Michel Gevers for having invited him to give the semi-plenary talk (Vidyasagar, 1997b) which stimulated him to start working seriously on the various problems studied here.

References

Braatz, R., Young, P., Doyle, J., & Morari, M. (1994). Computational complexity of the μ calculation. *IEEE Transactions on Automatic Control*, *39*, 1000–1002.

Blondel, V., & Tsitsiklis, J. N. (1997). NP-hardness of some linear control design problems. *SIAM Journal of Control and Optimization*, *35*(6), 2118–2127.

Blondel, V., & Tsitsiklis, J. N. (2000). A survey of computational complexity results in systems and control. *Automatica*, *45*(9), 1249–1274.

Coxson, G. E., & DeMarco, C. (1994). The computational complexity of approximating the minimal perturbation scaling to achieve instability in an interval matrix. *Mathematics of Control, Signals and Systems*, *7*, 279–291.

Doyle, J. (1982). Analysis of feedback systems with structured uncertainties. *Proceedings of the IEEE*, *129*, 242–250.

Fine, T. L. (1999). *Feedforward neural network methodology*. New York: Springer.

Gantmacher, F.R. (1959). *Matrix theory*, Vol. II, New York: Chelsea.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. New York: W.H. Freeman.

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

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of American Statistics Association*, *58*, 13–30.

Jury, E. I. (1977). *Inners and stability of dynamical systems*. New York: Wiley.

Karpinski, M., & Macintyre, A. J. (1995). Polynomial bounds for VC dimension of sigmoidal neural networks. *Proceedings of the 27th ACM Symposium on Theory of Computing* (pp. 200–208).

Karpinski, M., & Macintyre, A. J. (1997). Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. *Journal of Computer and Systems Sciences*, *54*, 169–176.

Khargonekar, P. P., & Tikku, A. (1996). Randomized algorithms for robust control analysis have polynomial complexity. *Proceedings of the Conference on Decision and Control*, 1996.

Kowalczyk, A., Ferra, H., & Szymanski, J. (1995). Combining statistical physics with VC-bounds on generalisation in learning systems. *Proceedings of the Sixth Australian Conference on Neural Networks* (*ACNN' 95*), Sydney (pp. 41–44).

Macintyre, A. J., & Sontag, E. D. (1993). Finiteness results for sigmoidal neural networks. *Proceedings of the 25th ACM Symposium Theory of Computing* (pp. 325–334).

Marrison, C., & Stengel, R. (1994). The use of random search and genetic algorithms to optimize stochastic robustness functions. *Proceedings of the American Control Conference*, Baltimore, MD (pp. 1484–1489).

Nemirovskii, A. (1993). Several NP-hard problems arising in robust stability analysis. *Mathematics of Control, Signals, and Systems*, *6*(2), 99–105.

Newton Jr., G. C, Gould, L. A., & Kaiser, J. F. (1967). *Analytic design of linear feedback controls*. New York: Wiley.

Papadimitrou, C. (1994). *Computational complexity*. Reading, MA, USA: Addison-Wesley.

Patel, V. V., Deodhare, G. S., & Viswanath, T. (2000). Some applications of randomized algorithms for control system design. *Automatica*, submitted.

Parrondo, J. M., & van den Broeck, C. (1993). Vapnik–Chervonenkis bounds for generalization. *Journal of Physica A*, *26*, 2211–2223.

Poljak, S., & Rohn, J. (1993). Checking robust nonsingularity is NP-hard. *Mathematics of Control, Signals, and Systems*, *6*(1), 1–9.

Ray, L. R., & Stengel, R. F. (1991). Stochastic robustness of linear time-invariant control systems. *IEEE Transactions on Automatic Control*, *36*, 82–87.

Tempo, R., Bai, E. W., & Dabbene, F. (1997). Probabilistic robustness analysis: Explicit bounds for the minimum number of sampling points. *Systems & Control Letters*, *30*, 237–242.

Vapnik, V. N. (1982). *Estimation of dependences based on empirical data*. Berlin: Springer.

Vapnik, V. N., & Chervonenkis, A. Ya. (1971). On the uniform convergence of relative frequencies to their probabilities. *Theory of Probability and its Applications*, *16*(2), 264–280.

Vapnik, V. N., & Chervonenkis, A. Ya. (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and its Applications*, *26*(3), 532–553.

Vidyasagar, M. (1997a). *A theory of learning and generalization: With applications to neural networks and control systems*. London: Springer.

Vidyasagar, M. (1997b). Statistical learning theory and its applications to randomized algorithms for robust controller synthesis, In: G. Basten, & M. Gevers (Eds.), Semi-Plenary Lecture, *European Control Conference*, Brussels, Belgium (pp. 161–189).

Vidyasagar, M., & Blondel, V. (2001). Probabilistic solutions to some NP-hard matrix problems. *Automatica*, to appear.

Mathukumalli Vidyasagar was born in Guntur, Andhra Pradesh, India on 29 September 1947. He received the B.S., M.S., and Ph.D. degrees, all in electrical engineering, from the University of Wisconsin, in 1965, 1967, and 1969, respectively. He has taught at Marquette University, U.S.A. (1969–1970), Concordia University, Canada (1970–1980), and the University of Waterloo, Canada (1980–1989). In 1989 he returned to India as the Director of the Centre for Artificial

Intelligence and Robotics, (under the Defence Research and Development Organisation) in Bangalore. In 2000 he took up his current assignment as Executive Vice President (Advanced Technology) in Tata Consultancy Services, which is India's largest IT firm. At present he is based in the city of Hyderabad. In his current position, his responsibilities are to create an Advanced Technology Centre (ATC) within TCS, to develop futuristictechnologies of relevance to the IT industry. At the present time, the scope of activities of the ATC includes PKI (Public Key Infrastructure), security in e- and m-commerce, advanced cryptography including elliptic curve cryptography, and neural networks.

In the past, he has held visiting positions at several universities including M.I.T., California (Berkeley), California (Los Angeles), C.N.R.S. Toulouse, France, Indian Institute of Science, University of Minnesota, and Tokyo Institute of Technology. He is the author or co-author of seven books and more than one hundred and twenty papers in archival journals. He has received several honours in recognition of his research activities, including the Distinguished Service Citation from his Alma Mater (The University of Wisconsin). In addition, he is a Fellow of Institute of Electrical and Electronics Engineers (IEEE), the Indian Academy of Sciences, the Indian National Science Academy, the Indian National Academy of Engineering, and the Third World Academy of Sciences. His current research interests are control theory, machine learning, and cryptography.