SOME RESULTS IN STATISTICAL LEARNING THEORY WITH RELEVANCE TO NONLINEAR SYSTEM IDENTIFICATION

Robert C. Williamson

Department of Engineering, Australian National University, Canberra, 0200, Australia

Abstract: Statistical Learning Theory comprises a collection of techniques that have been developed in order to theoretically analyse the performance of neural network and other "learning" algorithms. In this paper, a number of recent results in statistical learning theory are summarised in the context of nonlinear system identification. A top-down approach to the problem is taken, leading to the statement of a number of characterisation results. Specific topics covered include empirical risk minimisation, various types of Glivenko-Cantelli classes, scale-sensitive dimensions, degree of approximation in a Hilbert space setting, the importance of convexity of function classes in the agnostic learning model, and the development of new inductive principles.

Keywords: learning system, learning algorithm, identification, statistical analysis, models

1. INDUCTION OF MODELS

> *The players . . . threw these abstract formulas at one another displaying the sequences and possibilities of their science.*
> — Herman Hesse: The Glass Bead Game

It is well to consider the philosophical foundations of anything one does; that is particularly true when one is trying to solve technologically a philosophically impossible problem. System identification is the induction of models from observed events (data). In general it is impossible in the sense that you can never *know* you have identified the truth: "for all is but a woven web of guesses." You can not even assert with high probability that you are close to the truth Popper [1981, 1980]. The (implicit) philosophical stance of an engineer identifying a dynamical system might be described as a game. The game is this: *If* certain assumptions I am willing to make (with no epistemological justification) are true, *then* on the basis of data available, how reliable a conclusion can I draw?

It is important to realize that in solving a system identification problem, one is not necessarily seeking to discover the "true" model; one is after a model that attains a certain level of performance. One could say that of the two goals of a scientific theory (predictive performance and explanatory power, the former is valued rather more). The postulation of a "true" model is sometimes useful for theoretical analysis, but as we shall show, an insightful analysis can be performed without the assumption. Furthermore we shall show how one can theoretically determine the number of data points required to attain a pre-specified level of performance. Nevertheless, one is still attempting to reason inductively, and cognisance needs to be taken of the dangers inherent in doing so.

Black-box identification [Sjöberg et al., 1995, Juditsky et al., 1995] is an approach taken when the engineer neither can hypothesise a physical model nor has a specific parametric model imposed upon him (perhaps from extrinsic considerations). However a model structure which "belongs to families that are known to have good flexibility and have been successful in the past" [Sjöberg et al., 1995, p. 1692] may be sought. This is, of course, inductivism by another name, and appeal to such an infinite regress does not really help [see Popper, 1980, page 29].

[1] Thanks to Erik Weyer for helpful comments on a draft. This work was supported by the Australian Research Council.

The extent to which past success will help is indeed uncertain; however all is not lost. One can derive very useful insight from an analytical consideration of the game mentioned above. One could say that the goal is to find how black is the box, and what determines its blackness? The game might be called "Assumption Engineering."

One can play the game in many ways. The best games follow an intrinsic logic rather than an extrinsic one. Thus we might care to determine what are the essential limitations of particular approaches to system identification. For example, to what extent is the *parametrisation* chosen for a class of models an essential feature, and to what extent is it a choice of convenience?

In this paper I will review some recent work in the field of statistical learning theory that relates to these questions. Specifically, I will explain what is now known about the characterisation of problem difficulty in a number of formal frameworks (or specific game rules to stretch our metaphor). I will show by example the extent to which the exact rules can significantly affect the conclusions one can infer. I have attempted to follow the *logical* rather than *chronological* structure of the field in order to argue where to go next. I have included technical definitions of the key ideas discussed in order to make the paper reasonably self-contained, but the reader should not be blinded by those technicalities: above anything else, it is the *flavour* of the results and there logical structure that I wish to communicate. There is a longer version of this paper with material on a number of other related topics.

2. REGRESSION AND NONLINEAR BLACK-BOX STRUCTURES

We will adopt the general setting and some of the notation in Sjöberg et al. [1995], Juditsky et al. [1995] to which the reader is referred for a rather fuller introduction. Suppose we have a black-box dynamical system with input $u(t)$ and output $y(t)$. We observe

$$u^t := [u(1), \ldots, u(t)]$$
$$y^t := [y(1), \ldots, y(t)]$$

and would like to infer a relationship between past observations $[u^{t-1}, y^{t-1}]$ and future outputs $y(t)$. A general model for this is

$$y(t) = g(u^{t-1}, y^{t-1}) + v(t).$$

With this setup, the problem becomes: Choose $g \in \mathcal{G}$, a class of candidate models. The quality of a model is often measured by means of a criterion such as

$$\sum_t \|y(t) - g(u^{t-1}, y^{t-1})\|^2 \qquad (1)$$

where the sum is over input-output pairs not used in choosing g. Writing $\phi(t) := \phi(u^{t-1}, y^{t-1})$ for the *regression vector* (some *fixed* function of the data), a more precise statement of the problem becomes: Choose $\phi(\cdot)$, choose $\mathcal{G}$, then pick $g \in \mathcal{G}$ which gives the "best" model according to a criterion such as (1). However this becomes a general problem of statistical regression. Thus to a large extent, the study of nonlinear black-box models for system identification *is* the study of nonlinear regression, perhaps with some added complications due to the dependence structure inherent in $\phi(u^{t-1}, y^{t-1})$.

3. A GENERAL FRAMEWORK FOR REGRESSION PROBLEMS

Each discipline which seized upon the
Game created its own language of formulas,
abbreviations, and possible combinations.

We would like to theoretically understand the intrinsic limitations of the above regression problem. To that end we want to be very careful in proceeding with an analysis that we do not infer fundamental limitations of our analysis *tools*. One recipe to avoid doing this is to aim for *characterisations* of the problem in hand: that is necessary and sufficient conditions to achieve the goals we set ourselves. We can then be reasonably sure that the only place we have ourselves imposed an answer on ourselves is in the *initial* formulation of the problem. This is the approach adopted by Vapnik and we borrow heavily from his books Vapnik [1995, 1982]. A concise presentation of a number of the ideas below can also be found in Vapnik [1993]. We have omitted without further mention when measurability assumptions must be made. (See van der Vaart and Wellner [1996] for a detailed discussion.)

We require a *loss function*. Let $\mathcal{X}$ be the "input" space and $\mathcal{Y}$ the "output" space. Then a *loss function* is a map $L: \mathcal{Y} \times \mathcal{Y}$ which is used to assign a cost of a particular hypothesis $f\colon \mathcal{X} \to \mathcal{Y}$ on a given *data point* $(x, y) \in \mathcal{X} \times \mathcal{Y}$ via $\phi((x, y), f) := L(y, f(x))$. We judge the overall performance of f by the expected value of the loss function (or the *risk*) where the expectation is taken with respect to $P_{\mathcal{X}}$, the distribution of the x data points:

$$R(f) := \mathbf{E}[L(Y, f(X))] = \int \phi(z, f) \mathrm{d}P(z)$$

where we have written $z = (x, y)$, $\phi(z, f) := L(y, f(x))$ and P is the distribution of (X, Y) on $\mathcal{X} \times \mathcal{Y}$. We can again restate our problem: Given $(x_i, y_i)_{i=1}^n$, $y_i = f^*(x_i) + e_i$, (e_i is a "noise" sequence) and a class of functions $\mathcal{F}$, determine

$$\hat{f}^* := \arg\min_{f \in \mathcal{F}} R(f).$$

This is impossible, as we are only given a *sample* of n data points and thus do not know P and hence can not compute the risk.

We thus need to adopt an *inductive principle*, which (we hope!) will lead to estimates somehow "close" to $\hat{f}^*$. Perhaps the most widely used inductive principle is called *Empirical Risk Minimisation* (ERM). There one picks an hypothesis from $\mathcal{F}$ via

$$\hat{f}_n := \arg\min_{f\in\mathcal{F}} R_{\text{emp}}(f)$$

where the *empirical risk* is defined by

$$R_{\text{emp}}(f) := \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i))$$

and is clearly computable from the data provided. The "hope" we mentioned can be expressed formally as requiring that the estimator $\hat{f}_n$ satisfies the two relationships

$$R(\hat{f}_n) \longrightarrow \inf_{f\in\mathcal{F}} R(f)$$
$$R_{\text{emp}}(\hat{f}_n) \longrightarrow \inf_{f\in\mathcal{F}} R(f)$$

in probability as $n \to \infty$. (This is a minimal requirement; we would also like to know how fast the convergence is — more on that below.)

It turns out that a modification of the above requirements is necessary to rule out certain pathologies. One particular modification is given below, although it is not known if it is the "best" one. The situations ruled out include a coded solution whereby a single observation (data point) suffices to identify exactly a function from within some large class — see Bartlett et al. [1996] for an explicit example. As in Vapnik and Chervonenkis [1991], let $\Lambda(c) := \{f \in \mathcal{F}\colon R(f) > c\}$. Say that ERM is *strictly consistent* if for all c such that $\Lambda(c) \neq \emptyset$, $\inf\{R_{\text{emp}}(f)\colon f \in \Lambda(c)\} \longrightarrow \inf\{R(f)\colon f \in \Lambda(c)\}$ in probability as $n \to \infty$.

Theorem 1. Vapnik and Chervonenkis [1991] Suppose there exist constants A and B such that for all $f \in \mathcal{F}$, $A \le R(f) \le B$. Then ERM using $\mathcal{F}$ is strictly consistent if and only if $R_{\text{emp}}(f)$ converges uniformly to $R(f)$ in the sense that for all $\varepsilon > 0$

$$\lim_{n\to\infty} P\{\sup_{f\in\mathcal{F}}(R(f) - R_{\text{emp}}(f)) > \varepsilon\} = 0. \quad (2)$$

A necessary and sufficient condition for *uniform one-sided convergence* (2) is known, and will be described below. It is convenient however to first describe the necessary and sufficient conditions for *uniform two-sided convergence*

$$\lim_{n\to\infty} P\{\sup_{f\in\mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \varepsilon\} = 0 \quad (3)$$

for all $\varepsilon > 0$. In order to state the conditions, we need to introduce a swag of notation.

Let A be a set in a metric space (X, ρ). A set U is an ε*-covering* of A if for all $a \in A$, there is a $u \in U$ with $\rho(a,u) < \varepsilon$. The ε*-covering number* of A (with respect to ρ) $N_\rho^A(\varepsilon)$ is the number of elements in the smallest ε-cover of A. If f and g are functions defined on X and $x_1, \ldots, x_n \in X$, then the ℓ_p^n metric induced by the points is defined by $\rho_p(f,g) := \left(\frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)|^p\right)^{1/p}$ (for $p \in [1,\infty)$) and $\max\{|f(x_i) - g(x_i)|\colon i = 1,\ldots,n\}$ for $p = \infty$. If A is a class of functions on X we write $N_p^A(\varepsilon)$ for $N_{\rho_p}^A(\varepsilon)$. Let $\Phi := \{\phi = \phi(z,f) = L(y, f(x)) : z = (x,y), f \in \mathcal{F}\}$ be the *loss-function induced class*. Let $N_p^\Phi(\varepsilon; z_1, \ldots, z_n)$ denote the ε-covering number of Φ with respect to $z_1, \ldots, z_n$ in the ℓ_p metric. Let $H_p^\Phi(\varepsilon; z_1, \ldots, z_n) := \ln N_p^\Phi(\varepsilon; z_1, \ldots, z_n)$. The quantity $H_\infty^\Phi(\varepsilon, n) := \mathbf{E} H_\infty^\Phi(\varepsilon, z_1, \ldots, z_n)$ is known as the *VC-entropy* of Φ. The expectation is taken with $\underline{z} = (z_1, \ldots, z_n)$ drawn according to P^n. Vapnik and Chervonenkis [1981] have shown that if for all $\varepsilon > 0$, $\lim_{n\to\infty} \frac{1}{n} H_1^\Phi(\varepsilon, n) = 0$ then (3) holds. Furthermore if (3) holds, then for all $\varepsilon > 0$, $\lim_{n\to\infty} \frac{1}{n} H_\infty^\Phi(\varepsilon, n) = 0$. This is the "best" way of stating the result in the sense that from the definitions, it follows that $H_\infty^\Phi(\varepsilon, n) \ge H_1^\Phi(\varepsilon, n)$. (Use of ℓ_1 covering numbers can lead to better finite sample size bounds.) Traditionally the result has been stated in the more compact form:

Theorem 2. Equation 3 holds if and only if $\forall\varepsilon > 0$

$$\lim_{n\to\infty} \tfrac{1}{n} H_\infty^\Phi(\varepsilon, n) = 0. \quad (4)$$

It can equally well be stated in terms of $H_1^\Phi(\varepsilon, n)$; [see Vidyasagar, 1997, page 127].

It would seem however that we have made little progress since condition (4) still depends on the (unknown) probability distribution because of the expectation. Furthermore, even if we did know P, the actual calculation of $H_p^\Phi(\varepsilon, n)$ could be very difficult. Both of these difficulties can be overcome by asking a harder question. If instead of requiring uniform convergence just for a particular P, we can require it to hold uniformly *for all* P. We now set

$$G_\infty^\Phi(\varepsilon, n) := \sup_{z_1,\ldots,z_n} H_\infty^\Phi(\varepsilon; z_1, \ldots, z_n)$$

which is (in principle) calculable knowing only $\mathcal{F}$ and L. Clearly, for any distribution P, and any $\varepsilon > 0$, we have $H^\Phi(\varepsilon, n) < G^\Phi(\varepsilon, n)$. In fact ERM is consistent for $\mathcal{F}$ for any probability measure if and only if for all $\varepsilon > 0$, $\lim_{n\to\infty} \frac{1}{n} G_\infty^\Phi(\varepsilon, n) = 0$ where as above Φ is the loss-function class induced from $\mathcal{F}$. (The sufficiency in this result follows immediately; necessity follows from the construction of an appropriate P.)

4. GLIVENKO-CANTELLI CLASSES

Whilst in principle $G_\infty^\Phi(\varepsilon, n)$ is now calculable, without some further tricks, it is difficult to do so directly from the definition. In order to make the logical structure of the problem clearer and to link in with some of the literature on the topic, we introduce some further terminology. A key idea is the *dimension* of a class of hypotheses $\mathcal{F}$. That hypothesis classes with larger "dimensions" require more empirical data to work with is an old idea; for example it was quite explicitly discussed over 60 years ago in [Popper, 1980, Sections 38–39].

Suppose P is a probability distribution on the input space $\mathcal{X}$. Let $\underline{x}^n := (x_1, \ldots, x_n) \in \mathcal{X}^n$. Write $P(f) := \int f(x) \mathrm{d}P(x)$ and $P_n(f) := \frac{1}{n}\sum_{i=1}^n f(x_i)$. A class of real-valued functions $\mathcal{F}$ defined on $\mathcal{X}$ is called a *P-Glivenko-Cantelli* (P-GC) class [van der Vaart and Wellner, 1996] if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P^n\{\underline{x}^n \in \mathcal{X}^n \colon \sup_{f\in\mathcal{F}} |P(f) - P_n(f)| > \varepsilon\} = 0.$$

Thus the condition (3) is equivalent to Φ being P-GC. We say that $\mathcal{F}$ is an *ε-uniform GC class* (ε-UGC) if

$$\lim_{n\to\infty} \sup_P P^n\{\underline{x}^n \in \mathcal{X}^n \colon \sup_{f\in\mathcal{F}} |P(f) - P_n(f)| > \varepsilon\} = 0$$

where the outer supremum is taken over all probability measures P on $\mathcal{X}$. We say $\mathcal{F}$ is a *uniform GC class* if it is ε-UGC for all $\varepsilon > 0$. If the absolute value signs in these definitions are removed, we prepend "one-sided". Thus condition (2) requires that Φ be one-sided P-GC.

4.1 *Bracket Covers*

A relatively easily proved sufficient condition for $\mathcal{F}$ to be P-GC can be stated in terms of bracket covers. For this subsection, let $\|\cdot\|$ denote the $L_1(P)$ norm: $\|f\| := \int |f| \mathrm{d}P$. We say U is a *ε-bracket cover* of $\mathcal{F}$ if for all $f \in \mathcal{F}$, there exists a *bracket* $b := [l, u]$ such that 1) $\|l - u\| \leq \varepsilon$ and 2) $l \leq f \leq u$ (pointwise). Let $N_{[]}^{\mathcal{F}}(\varepsilon)$ denote the size of the smallest ε-bracket cover of $\mathcal{F}$ (with respect to $L_1(P)$). The following theorem is stated in van der Vaart and Wellner [1996] where its history is given.

Theorem 3. If $\mathcal{F}$ is such that for all $\varepsilon > 0$, $N_{[]}^{\mathcal{F}}(\varepsilon) < \infty$, then $\mathcal{F}$ is P-GC.

Whilst $N_{[]}^{\mathcal{F}}(\varepsilon)$ can be determined for a number of interesting classes $\mathcal{F}$ (see [van der Vaart and Wellner, 1996, section 2.7]), the condition in this theorem is "nowhere close to being necessary" Talagrand [1996].

A more complex result which *characterises* the *one-sided-P*-GC property has been presented in Vapnik and Chervonenkis [1991]. Call $B_l(\delta)$ a *lower δ-bracket* for $\mathcal{F}$ if for all $f \in \mathcal{F}$, there exists $l \in B_l(\delta)$ such that 1) $\|l - f\| \leq \delta$ and 2) $l \leq f$ (pointwise). Let $N_{[,\delta}^{\mathcal{F}}(\varepsilon, \underline{x}^n)$ denote the size of the smallest ε-cover of $B_l(\delta)$ in the ℓ_∞^n-metric (relative to some sequence of points $\underline{x}^n = (x_1, \ldots, x_n)$. Let $H_{[,\delta}^{\mathcal{F}}(\varepsilon, n) := \mathbf{E} \ln N_{[,\delta}^{\mathcal{F}}(\varepsilon, \underline{x}^n)$.

Theorem 4. Suppose $\mathcal{F}$ is totally bounded. Then $\mathcal{F}$ is one-sided P-GC if and only if for all $\delta, \varepsilon, \eta > 0$

$$\lim_{n\to\infty} \frac{1}{n} H_{[,\delta}^{\mathcal{F}}(\varepsilon, n) < \eta.$$

At present, it is unclear to what extent this mixture of bracket cover and VC entropy can give better practical results.

4.2 *Fat-Shattering Dimensions*

We now introduce two so-called scale-sensitive dimensions which we will use below. Let $\mathcal{F}$ be a set of real valued functions. We say that a set of points X is *γ-shattered by $\mathcal{F}$ relative to* $r = (r_x)_{x\in X}$ if there are real numbers r_x indexed by $x \in X$ such that for all binary vectors b indexed by X, there is a function $f_b \in \mathcal{F}$ satisfying

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

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

If X can be γ-shattered choosing $r_x = r$ the same for all $x \in \underline{x}$, we say the points X are *level γ-shattered* (at level r) and denote the largest number of such points by the *level fat shattering function* $\mathrm{LFat}_{\mathcal{F}}(\gamma)$. Fat and LFat can not be too different as shown in Alon et al. [1997]: If $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$, then for all $\gamma > 0$, $\mathrm{LFat}_{\mathcal{F}}(\gamma) \leq \mathrm{Fat}_{\mathcal{F}}(\gamma) \leq (2\lceil 1/2\gamma\rceil - 1)\, \mathrm{Fat}_{\mathcal{F}}(\gamma/2)$.

The fat-shattering dimension for a number of function classes has been computed. An easy example is the set $BV \subseteq \mathbf{R}^{[0,1]}$ of functions of bounded variation for which $\mathrm{Fat}_{BV}(\gamma) = O(1/\gamma)$. It has been determined for neural networks in Gurvits and Koiran [1995] and a better bound based on an approximation result will be stated in Section 6.

The idea of fat-shattering was introduced into the statistical learning theory community in Kearns and Schapire [1994], although it has been used in approximation theory since the late 1950's where it was apparently first proposed by Kolmogorov [see Tikhomirov, 1960, page 103]. The intuitively appealing thing about $\mathrm{Fat}_{\mathcal{F}}(\gamma)$ is that it measures the complexity of $\mathcal{F}$ *at the accuracy scale one is interested in working at.* This is in contrast to previous generalisations of the "Vapnik-Chervonenkis dimension" (VC-dim) to treat the problem of learning real-valued functions. For example, the quantity known as the Pollard *pseudo-dimension* (see [Haussler, 1992]) $\mathrm{Pdim}(\mathcal{F})$ can be expressed as $\mathrm{Pdim}(\mathcal{F}) = \lim_{\gamma\to 0} \mathrm{Fat}_{\mathcal{F}}(\gamma)$. There are many function classes for which is $\mathrm{Pdim}(\mathcal{F})$ is infinite or very large whilst $\mathrm{Fat}_{\mathcal{F}}(\gamma)$ is small for all γ of the appropriate scale (which depends on the accuracy to which one wishes to learn; see the sample complexity results in the next Section).

We say a triple (A, γ, r), $A \subseteq \mathcal{X}$, $P(A) > 0$, $\gamma > 0$, $r \in \mathbf{R}$, is a *witness of irregularity* if 1) $P|_A$ has no atoms (i.e. $\forall x \in A, P(x) = 0$) and 2) for all $n \geq 1$,

$$P^n\{\underline{x}^n \in A^n \colon \mathcal{F}\ \gamma\text{-shatters } \underline{x}^n \text{ at level } r\} = 1.$$

The second condition can be stated in words as for all $n \geq 1$, almost all subsets of size n are γ-shattered by $\mathcal{F}$ at level r.

4.3 *Characterisations via Fat-Shattering Dimensions*

Theorem 5. Talagrand [1987, 1996] $\mathcal{F}$ is P-*GC* if and only if there is no witness of irregularity.

The idea of a witness of regularity has apparently not been studied in the statistical learning theory literature. However note that both Sontag Sontag [192] and Kowalczyk Kowalczyk [1997] have considered a very similar question for the related (and simpler) concept of VC dimension. (Specifically they consider when the set of points that can be (VC)-shattered is dense.)

In hindsight, consideration of Talagrand's result along with Vapnik and Chervonenkis' characterisation of uniform GC classes in terms of G^{Φ} might lead one to conjecture that $\mathcal{F}$ is *uniform*-GC iff $\forall\gamma > 0, \forall r \in \mathbf{R}, \exists n', \forall n > n'$, there is no sequence of n points $x_1, \ldots, x_n$ that is γ-shattered at level r. That is indeed the case and although not formally derived from Talagrand's result, was proven in Alon et al. [1997] where it is stated in the following more elegant form:

Theorem 6. The following are equivalent:
1) $\mathcal{F}$ is uniform-GC.
2) $\mathrm{Fat}_{\mathcal{F}}(\gamma) < \infty$ for all $\gamma > 0$.
3) $\mathrm{LFat}_{\mathcal{F}}(\gamma) < \infty$ for all $\gamma > 0$.

One of the attractive things about this result is that it solely depends on "combinatorial" properties of $\mathcal{F}$. Furthermore, as we shall see in Section 5, these dimensions can be used to make detailed statements about the performance of learning algorithms on a finite number of sample points.

A key technical tool in proving these characterisations is the upper bounding of a ℓ_{∞}^{n} covering number in terms of $\mathrm{Fat}_{\mathcal{F}}(\gamma)$. We state the result slightly differently to Alon et al. [1997]; the proof that it is equivalent is given in Lee [1996].

Theorem 7. Suppose $\mathcal{F}$ is a class of $[0,1]$-valued functions on $\mathcal{X}$ and $0 < \varepsilon \leq 1$. Then for all $\underline{x}^n \in \mathcal{X}^n$,

$$G_{\infty}^{\mathcal{F}}(\varepsilon, n) \leq 3\,\mathrm{Fat}_{\mathcal{F}}(\varepsilon/4)\ln^2 \frac{16n}{\varepsilon^2}$$

(This result plays a role directly analogous to the Sauer-Shelah lemma for the theory of learning $\{0,1\}$-valued functions based on VCdim.) There are various refinements possible for this; see Alon et al. [1997] for details and pointers to the literature.

5. SAMPLE COMPLEXITY

The characterisations of various notions of GCness are but a starting point. They actually lead to insightful bounds on the sample complexity $m(\varepsilon, \delta)$ of learning problems, which is defined as the number of samples m needed to learn to accuracy ε with probability $1 - \delta$. The precise definitions vary according to the exact setting of the learning problem, and we will only consider the two settings known as regression and agnostic learning. The significance of these results is that they give a *finite* bound on the number of samples need to learn to some pre-specified accuracy. Their practical significance is not so much the specific numbers they indicate, but in the scaling laws (in terms of ε) *which are valid for "small" sample sizes.*

5.1 *Regression*

The setting here is that we are given a random sample $\{(x_i, y_i) : i = 1, \ldots, n\}$ where the random variables X, Y are such that $\mathbf{E}[Y|X = x] = f^*(x)$ for some $f^* \in \mathcal{F}$. This would be the case for example if $y_i = f^*(x_i) + e_i$ for some zero-mean noise e_i. (If there is no noise, this setting is called "function learning" in the Statistical Learning Theory literature.) The results stated below rely on bounding the quantity $G^{\Phi}(\varepsilon, n)$ in terms of various dimensions, such as $\mathrm{Fat}_{\mathcal{F}}(\gamma)$. In order to view the forest rather than the trees we will state the results using $O(\cdot)$ notation, but note that all of the constants can be (and have been) explicitly determined. We will only state the results for the squared loss function too; many others can be handled as well. All that is required is that for whichever $p \in [1, \infty]$ one wishes to work with, $N_p^{\Phi}(\varepsilon, n)$ can be expressed in terms of $N_p^{\mathcal{F}}(c\varepsilon, n)$ for some constant c. This is done quite explicitly for example in Alon et al. [1997], Bartlett et al. [1996]. Sometimes one can in fact express the "dimension" of Φ in terms of that of $\mathcal{F}$ (such an approach is adopted in Haussler [1992]).

Theorem 8. Suppose $\mathcal{F}$ is uniformly bounded. Let $d = \mathrm{Pdim}(\mathcal{F})$, the pseudo-dimension of $\mathcal{F}$. Suppose that for some $r > 0$, $\mathrm{Fat}_{\mathcal{F}}(\gamma) = O(\gamma^{-r})$ (this assumption is true of any function class one would use in practice). The following expressions for the sample complexity of regression with squared loss hold.

$$m(\varepsilon, \delta) = O\left(\frac{1}{\varepsilon}\left(d \ln \frac{1}{\varepsilon} + \ln \frac{1}{\delta}\right)\right)$$
$$m(\varepsilon, \delta) = O\left(\frac{1}{\varepsilon}\left(\mathrm{Fat}_{\mathcal{F}}(\varepsilon)\ln^2 \frac{\mathrm{Fat}_{\mathcal{F}}(\varepsilon)}{\varepsilon} + \ln \frac{1}{\delta}\right)\right).$$

The result in terms of pseudo-dimension is given in Haussler [1992]. The second result follows from Theorem 7; see Lee [1996] for a proof. Although the second result "looks" worse (in terms of how it scales with the "dimension" of $\mathcal{F}$), it can be much better since

$\mathrm{Fat}_{\mathcal{F}}(\gamma)$ can be finite for all $\gamma > 0$ even when d is infinite. A detailed general discussion of this setting (with some generalisations) can be found in Haussler [1992]. We will not dwell on it here though since, as we argue in the next subsection, it is not the most suitable model for system identification problems. The role of fat-shattering in regression problems was the focus of Bartlett et al. [1996].

5.2 *Agnostic Learning*

Whilst the traditional regression framework certainly provides some insight, it is unsatisfying in one important respect: it requires one to assume that the "true" underlying function $f^*(x) = \mathbf{E}[Y|X = x]$ is in the class $\mathcal{F}$. This is obviously something that can not be verified. However if we relax the assumption we are faced with another problem: there will now be *two* sources of error. As well as the "statistical" error incurred as before due to only having a finite sample of data, there will also be an "approximation" error between the "truth" and the best model in the class. This decomposition is additive in the case of a square loss function, and we will restrict our attention mainly to that case. It is often explained in terms of bias plus variance. Whilst one can not theoretically quantify the approximation error without (unverifiable) assumptions on the truth, it is still possible to make theoretical claims about the sampling error.

The agnostic learning model used here is based on that in Kearns et al. [1994]. We assume as before that we are presented with a set of examples $\{(x_i, y_i) : i = 1, \ldots, n\}$ with $y_i \in \mathcal{Y}$ which is a bounded subset of $\mathbf{R}$. The examples are drawn independently from some arbitrary distribution P on $\mathcal{X} \times \mathcal{Y}$. A class $\mathcal{F}$ of real-valued functions defined on $\mathcal{X}$ is said to be *agnostically learnable* with *sample complexity* $m(\varepsilon, \delta)$ if for any probability distribution P on $\mathcal{X} \times \mathcal{Y}$, given $0 < \delta \leq 1$ and $\varepsilon > 0$, using $m(\varepsilon, \delta)$ examples one can find a hypothesis $h \in \mathcal{F}$ such that with probability at least $1 - \delta$,

$$R(h) \leq \inf_{f \in \mathcal{F}} R(f) + \varepsilon.$$

Theorem 9. Suppose $\mathcal{F} \subseteq [0, 1]^{\mathcal{X}}$, $\mathcal{Y} \subseteq [0, 1]$ and that for some $r > 0$, $\mathrm{Fat}_{\mathcal{F}}(\gamma) = O(\gamma^{-r})$. Then the sample complexity of agnostically learning with $\mathcal{F}$ satisfies

$$m(\varepsilon, \delta) = O\left(\frac{1}{\varepsilon^2}\left(\mathrm{Fat}_{\mathcal{F}}(\varepsilon) \ln^2 \frac{\mathrm{Fat}_{\mathcal{F}}(\varepsilon)}{\varepsilon} + \ln \frac{1}{\delta}\right)\right).$$

See [Lee, 1996] for a proof. This result is proved in the above form in Lee [1996]. Most of the technical tools needed for the proof are given in Lee et al. [1996] where further references to the literature are given. Those results combined with the bound on covering numbers in Theorem 7 give the result.

The key thing to note is that the sample complexity for agnostic learning scales roughly as $1/\varepsilon^2$ compared with $1/\varepsilon$ for regression. We shall see below that this gap is essential in certain cases. The sample complexity of agnostic learning with the absolute loss function is treated in Bartlett and Long [1995], Bartlett et al. [1996].

5.3 *Application to System Identification*

Results along the lines of the ones presented can be applied to the overall system identification task. Rather than catalog the existing results, let us be satisfied with a guide to the literature, and some remarks on future directions.

To the best of my knowledge the first application of results of the flavour presented above to problems of System Identification were in Weyer [1992], Weyer et al. [1992, 1993, 1996] where results of Vapnik were used in a setting where the *models* were linear systems, although the true plant was not necessarily. Since then these sorts of results have been applied by various authors Meir [1997], Fiechter [1997], Campi and Kumar [1997].One difficulty in applying the above (or related) results directly is the independence assumption required for the above results to hold. This was overcome in Weyer et al. [1996] by making assumptions about the length of the tail of the impulse response of the unknown system (which does not have to be linear). There is a now a growing body of work extending the characterisation results (and hence opening the way for sample complexity results) to dependent processes. A sample of literature includes Nobel and Dembo [1993], Nobel [1995], Peškir and Yukich [1994]. The usual methods of theoretically dealing with dependent processes (by reducing the problem to an approximately independent process, with fewer observations) have been used.

The *point* of applying the general techniques to System Identification problems is severalfold. The key ones are 1) *Finite* sample complexity bounds can be obtained in a probabilistic setting where all of the traditional results are asymptotic; 2) The results tend to show more clearly what it is about a problem that makes it hard; 3) The further development of the general framework, and in particular the possibility of new inductive principles, could provide a range of new algorithms for System Identification — See the penultimate Section of the this paper.

6. APPROXIMATION THEORETIC ISSUES

In the agnostic learning model the issue of approximation error can not be ignored. Whilst there are many classical results in the theory of approximation that are applicable to the problems under consideration in this paper, we will restrict ourselves to one particular type of result that has attracted considerable interest in the statistical learning theory community.

6.1 *Approximation by Convex Combinations*

Let $P_{\mathcal{X}}$ be a probability distribution on $\mathcal{X}$ and H be the Hilbert space with inner product $\langle f, g\rangle = \int fg\mathrm{d}P_{\mathcal{X}}$. Let $\|\cdot\|$ denote the induced norm, $\|g\| := \langle g, g\rangle^{1/2}$. Recall a convex combination of $\Psi = \{\psi_i\}_i$ is an element of $\mathrm{co}_n(\Psi) := \{\sum_{i=1}^n \alpha_i\psi_i \colon \psi_i \in \Psi, \alpha_i \geq 0, i = 1, \ldots, n, \sum_{i=1}^n |\alpha_i| \leq 1\}$. If $G \subset H$, let $\mathrm{co}(G) := \bigcup_{n\in\mathbf{N}} \mathrm{co}_n(\Psi)$ denote the *convex hull* of G and $\overline{\mathrm{co}}(G)$ the closure of $\mathrm{co}(G)$.

Theorem 10. **(Barron Barron [1993])** Suppose for all $g \in G$, $\|g\| \leq 1$. Let $f^* \in \overline{\mathrm{co}}(G)$. Then for every $n > 1$, there exists $f_n \in \mathrm{co}_n(G)$ such that

$$\|f^* - f_n\| \leq \frac{1}{\sqrt{n}}.$$

The interesting thing about the result is that it does not depend on the dimension of G or H. As well as giving insight into the approximation error that may be incurred for certain situations (by explictly working out $\overline{\mathrm{co}}(G)$ for certain cases), it can also be used to bound $\mathrm{Fat}_{\mathcal{F}}(\gamma)$ for interesting $\mathcal{F}$. That is how the following result applicable to neural networks was proved in Bartlett [1996].

Theorem 11. Let F be a class of functions mapping $\mathcal{X}$ to $[-1, 1]$. Suppose $0 \in F$, $f \in F \Rightarrow -f \in F$ and that $\mathrm{Fat}_F(\gamma)$ grows at most polynomially in $1/\gamma$. Let $C > 0$ and let

$$\mathcal{F} := \left\{\sum_{i=1}^n w_i f_i \colon i \in \mathbf{N}, f_i \in F, \sum_{i=1} |w_i| \leq C\right\}.$$

Then

$$\mathrm{Fat}_{\mathcal{F}}(\gamma) = O\left(\frac{\mathrm{Fat}_F(\gamma)\ln^2(1/\gamma)}{\gamma^2}\right).$$

Theorem 10 is originally due to Maurey. As well as the probabilistic non-constructive proof given in Barron [1993], can be proved in a constructive fashion by exhibiting an explicit iterative algorithm that achieves the same rate Jones [1992].

There are two obvious objections to Barron's result (raised for example in Juditsky et al. [1995]): it does not take account of smoothness of the elements of G, and is not applicable directly to a learning problem. The first objection has been answered by Makovoz **?**] who proved the following result. Let $N^G(\varepsilon)$ be the covering number of G (in the Hilbert space norm $\|\cdot\|$). Let $\varepsilon_n(G) := \inf\{\varepsilon > 0 \colon N^G(\varepsilon) \leq n\}$.

Theorem 12. Under the hypotheses of Theorem 10,

$$\|f^* - f_n\| \leq \frac{2\varepsilon_n(G)}{\sqrt{n}}.$$

For sets G of smooth functions, this gives a better rate than $1/\sqrt{n}$ (and in fact it gives the optimal rate).

The second objection was effectively answered in Lee et al. [1996] where an "agnostic" version of Barron's result (with an explicit iterative approximation scheme) was presented. We state the result solely in terms of the approximation achieved:

Theorem 13. Let $G \subseteq H$, $\|h\| \leq 1$ for all $h \in H$, and suppose $f^* \in H$. Let $d_f := \min_{g\in\overline{\mathrm{co}}(G)} \|f^* - g\|$. Then for all $n > 1$,

$$\|f^* - f_n\|^2 - d_f^2 \leq \frac{4}{n}.$$

Note we have stated this in terms of *squared*-norm; but the rate is the same as in Theorem 10. The advantage here is that f^* need not be in G. Instead one measures the rate of decrease of the approximation error relative to the error of the best approximant from $\overline{\mathrm{co}}(G)$. This result was used in Lee et al. [1996] in order to construct a polynomial time algorithm for agnostically learning single hidden layer neural networks with a bound on the the sum of the absolute value of their output weights. (Whilst polynomial time, the algorithm would not be considered a practical one.)

A number of variations of the Barron result have been presented. One can use L_p norms instead of L_2 to measure the error. The first result along these lines was in Carl [1985] although that was not widely noticed. Some further results are given in Donahue et al. [1997] and **?**]. There are still some open problems. The most obvious two are to find 1) An iterative algorithm that achieves the Makovoz rate; 2) An agnostic version of the result that achieves that Makovoz rate.

6.2 *The Importance of Convexity*

In general the sample complexity of agnostic learning with squared loss is $O(1/\varepsilon^2)$ as discussed previously. The approximation results in the previous subsection indicate there is something special about convex classes. In the present subsection we show more precisely the extent to which this is true.

We will say that $\mathcal{F}$ is *closure-convex* if for all $P_{\mathcal{X}}$ on $\mathcal{X}$, the closure of $\mathcal{F}$ in the corresponding Hilbert space H is convex. Note that the closure of a convex function class is convex, hence convex function classes are closure-convex. For this subsection, we will assume $\mathcal{F}$ has finite pseudo-dimension; this restriction can be lifted, but the results don't look quite as clean then. Furthermore assume $\mathcal{F}$ is uniformly bounded ($\|f\|_\infty < c < \infty$ for all $f \in \mathcal{F}$). The following results are proved in Lee et al. [1998] (see W.S. Lee and Williamson [1996] for an earlier version already in print).

Theorem 14. Under the above conditions
1) If $\mathcal{F}$ is closure-convex, the sample complexity of agnostically learning $\mathcal{F}$ is $O(\frac{1}{\varepsilon}(\ln\frac{1}{\varepsilon} + \ln\frac{1}{\delta}))$.

2) The sample complexity of learning $\mathrm{co}(\mathcal{F})$ is $O(\frac{1}{\varepsilon}(\frac{1}{\varepsilon}\ln\frac{1}{\varepsilon}+\ln\frac{1}{\delta}))$.
3) If $\mathcal{F}$ is not closure-convex, then the sample complexity for agnostically learning $\mathcal{F}$ with squared loss is $\Omega\left(\frac{\ln(1/\delta)}{\varepsilon^2}\right)$.

(The $\Omega(\cdot)$ notation is analogous to $O(\cdot)$ except it is a *lower* bound.)

6.3 *Nonuniform Agnostic Learnability*

If one relaxes the requirement in the definition of sample complexity that m samples must suffice for *any* unknown "target" $f^* := \inf_{f\in\mathcal{F}} R(f)$ (that is for any distribution P on $\mathcal{X}\times\mathcal{Y}$), the sharp distinction drawn in the previous subsection disappears. Without stating the result formally, suffice it to say that it can be shown Bartlett et al. [1997] that for agnostic learning with squared loss:

(1) For any $\mathcal{F}$, and a generic $f^* \notin \mathcal{F}$, a *non-uniform* rate of $O(1/\varepsilon)$ (ignoring log factors) is possible for the nonuniform sample complexity (there is an implicit constant that depends on f^*).
(2) Even if $\mathcal{F}$ is not convex, as long as the boundary of $\mathcal{F}$ is sufficiently smooth in a certain sense, and the distance from $\mathcal{F}$ to f^* is not too large, then a *uniform* agnostic sample complexity of $O(1/\varepsilon)$ can be obtained (again ignoring log factors).

These results show the extent to which the precise manner in which the question is posed counts.

7. NEW INDUCTIVE PRINCIPLES

We have hopefully shown that to a large extent the results now in place for the performance of ERM in the setting adopted here leave little room for improvement on a large scale, although there is plenty of room for improving constants in many results. It is crucial to note though that ERM is just one choice of inductive principle. One can very legitimately ask whether there may be "better" ones. We will now outline, by way of some recent results for a classification problem, that there is in fact *very* significant potential here. The results are stated for the classification case for the simple reason that none have been worked out for regression yet; we definitely believe it is possible to do so, and the power of the classification results indicates why it is worthwhile attempting to do so.

[This section has not been completed yet; a skeleton outline is provided of what would be included.]

7.1 *Classification*

Set up classification problem.

Normal solution in terms of VCdim

New solution in terms of fat-shattering from Shawe-Taylor et al. [1998, 1996].

In order to understand, first need ordinary Structural Risk Minimzation.

But the hierarchy needs to be chosen before one sees the data. Interpretation of maximum margin hyperplane in terms of a *data-dependent SRM*.

Generalization: The luckiness framework (just describe in words). Interpretation: a way of precisifying the idea of "effective number of parameters" in the general setting adopted here.

Application to Bayesian evidence framework

Possibly mention Support Vector machines (a way of implementing nonlinear classifiers with a large number of implicit parameters which nevertheless can be implmented efficiently *and* have good statistical generalization performance.)

7.2 *Application to Regression?*

SV machines have been generalized to regression [refs]

There are some ideas relating to luckiness fo regression(Buescker/Kumar, Lugosi/Pinter, Cesa-Bianchi etal

Overall aim would be to develop new inducitve principles that give good generalization on a small number of samples (analogous to maximum margin classifier)

8. CONCLUSIONS

A number of results from statistical learning theory have been presented. They are of direct relevance to Nonlinear System Identification in so far as the general problem formulation adopted from Sjöberg et al. [1995], Juditsky et al. [1995] is valid, for once that setup is adopted, one is faced with a general regression problem. We have chosen to present the logical structure of the state of knowledge of the field, especially concentrating on the ERM inductive principle. An important conclusion is that the number of parameters in a model class is *not* a characterization of problem difficulty. This follows from the fact that as we have seen, $\mathrm{Fat}_{\mathcal{F}}(\gamma)$ *is* such a characterization, and one can readily construct a variety of models with the same numbers of parameters but vastly differing $\mathrm{Fat}_{\mathcal{F}}(\gamma)$. We have seen that geometric properties of $\mathcal{F}$ (such as its convexity) are very important, but that the exact way in which the questions are asked matters too.

Finally we have seen that whilst the situation for ERM is to a large extent understood and characterised, there is a lot of open ground in the search for new inductive principles, and furthermore that such inductive principles may be just the tool needed to capture the intuitive ideas people would like to apply to the task of Nonlinear System Identification.

References

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale sensitive dimensions, uniform convergence and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

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

P.L. Bartlett. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. Submitted to IEEE Transactions on Information Theory, 1996.

P.L. Bartlett, J. Baxter, P.M. Long, and R.C. Williamson. Reach, metric curvature, folding and grout: Agnostic learning nonconvex classes of functions. In preparation, 1997.

P.L. Bartlett and P.M. Long. More theorems about scale-sensitive dimensions and learning. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 392–401, New York, 1995. ACM Press.

P.L. Bartlett, P.M. Long, and R.C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.

M.C. Campi and P.R. Kumar. Learning dynamical systems in a stationary environment. Preprint, University of Illinois, 1997.

Bernd Carl. Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces. *Annales de l'Institut Fourier*, 35(3): 79–118, 1985.

M.J. Donahue, L. Gurvits, C. Darken, and E. Sontag. Rates of convex approximation in non-hilbert spaces. *Constructive Approximation*, 13(2):187–220, 1997.

C.-N. Fiechter. Pac adaptive control of linear systems. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 72–80, New York, 1997. ACM Press.

L. Gurvits and P. Koiran. Approximation and learning of convex superpositions. In Paul Vitanyi, editor, *Proceedings of EUROCOLT95*, volume 904 of *Lecture Notes in Artifical Intelligence*, pages 222–236. Springer Verlag, Berlin, 1995.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. Comput.*, 100(1):78–150, September 1992.

L. K. Jones. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, 20:608–613, 1992.

A. Juditsky, H. Hjalmarasson, A. Benveniste, B. Delyon, L. Ljung, J. Sjöberg, and Q. Zhang. Nonlinear black-box models in system identification: Mathematical foundations. *Automatica*, 31(12):1725–1750, 1995.

M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17 (2):115–141, 1994.

M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

A. Kowalczyk. Dense shattering and teaching dimensions for differentiable. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 143–151. ACM Press, 1997.

W. S. Lee, P. L. Bartlett, and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

W.S. Lee. *Agnostic Learning and Single Hidden Layer Neural Networks*. PhD thesis, Australian National University, 1996. http://routh.ee.adfa.oz.au/˜weesun/thesis.ps.Z.

W.S. Lee, P.L. Bartlett, and R.C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 1998. to appear.

R. Meir. Performance bounds for nonlinear time series prediction. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 122–129, 1997.

A. Nobel. A counterexample concerning uniform ergodic theorems for a class of functions. *Statistics and Probability Letters*, pages 165–168, 1995.

A. Nobel and A. Dembo. A note on uniform laws of averages for dependent processes. *Statistics and Probability Letters*, 17:169–172, 1993.

G. Peškir and J.E. Yukich. Uniform ergodic theorems for dynamical systems under vc entropy conditions. In *Probability in Banach Spaces 9*, pages 105–128. Birkäuser, Boston, 1994.

K.R. Popper. *The Logic of Scientific Discovery*. Hutchinson, London, 1980. Originally published in 1934.

K.R. Popper. *Realism and the Aim of Science*. Rowman and Littlefield, Totowa, New Jersey, 1981.

J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. A framework for structural risk minimisation. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 68–76. ACM Press, 1996.

J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 1998. to appear.

J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.-Y.Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box models in system identification: Mathematical foundations. *Automatica*, 31 (12):1691–1724, 1995.

E.D. Sontag. Feedforward nets for interpolation and classification. *Journal of Computer and System Sciences*, 45(1):20–48, 192.

M. Talagrand. The Glivenko-Cantelli problem. *Annals of Probability*, 6:837–870, 1987.

Michel Talagrand. The Glivenko-Cantelli problem, ten years later. *Journal of Theoretical Probability*, 9(2):371–384, 1996.

V.M. Tikhomirov. Diameters of sets in function spaces and the theory of best approximations. *Russian Mathematical Surveys*, 15(3):75–111, 1960.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

V.N. Vapnik. *Estimation of Dependences from Empirical Data*. Springer, New York, 1982.

V.N. Vapnik. Three fundamental concepts of the capacity of learning machines. *Physica A*, 200:838–844, 1993.

V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

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

V.N. Vapnik and A.Ya. Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk method. *Pattern Recognition and Image Analysis*, 1(3):283–305, 1991.

M. Vidyasagar. *A Theory of Learning and Generalization*. Springer, London, 1997.

E. Weyer. *System Identification in the Behavioural Framework*. PhD thesis, The Norwegian Institute of Technology, 1992.

E. Weyer, I.M.Y. Mareels, and R.C. Williamson. Sample complexity of least squares identification of FIR and ARX models. In *Proceedings of the 13th IFAC World Congress*, volume J, pages 239–244, 1996.

E. Weyer, R.C. Williamson, and I.M.Y. Mareels. An approach to system identification based on risk minimization and behaviours. In *Proceedings of the 31st Conference on Decision and Control*, pages 927–932. IEEE Press, 1992.

E. Weyer, R.C. Williamson, and I.M.Y. Mareels. A principle for system identification in the behavioural framework. In *Proceedings of the 12th World Congress of the International Federation of Automatic Control*, pages 387–390, 1993.

P.L. Bartlett W.S. Lee and R.C. Williamson. The importance of convexity in learning with squared loss. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 1400–146. ACM Press, 1996.