On how complexity affects the stability of a predictor

(Extended Abstract)

Joel Ratsaby
Department of Electrical and Electronics Engineering
Ariel University
ISRAEL

Abstract

Given a finite random sample from a Markov chain environment, we select a predictor that minimizes a criterion function and refer to it as being calibrated to its environment. If its prediction error is not bounded by its criterion value, we say that the criterion fails. We define the predictor's complexity to be the amount of uncertainty in detecting that the criterion fails given that it fails. We define a predictor's stability to be the discrepancy between the average number of prediction errors that it makes on two random samples. We show that complexity is inversely proportional to the level of adaptivity of the calibrated predictor to its random environment. The calibrated predictor becomes less stable as its complexity increases or as its level of adaptivity decreases.

1 Introduction

Let $\{X_t : t \in \mathbb{Z}\}$ be a sequence of binary random variables possessing the following Markov property,

$$P\left(X_t = x_t \mid X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \ldots\right) = P\left(X_t = x \mid X_{t-1} = x_{t-1}, \ldots, X_{t-k^*} = x_{t-k^*}\right)$$

where $x_{t-k^*}, \ldots, x_{t-1}$, x_t take a binary value of -1 or 1. This sequence is known as a discrete-time Markov stochastic process, or Markov *chain*, of order k^*. Let the *environment* be a stationary homogeneous Markov chain of order k^*. We assume that k^* is unknown. Define a *state space* by a set $\mathbb{S}_{k^*}$ of states $s^{(i)}$, $i = 0, 1, \ldots, 2^{k^*} - 1$, where $s^{(0)} := [s^{(0)}_{k^*-1}, \ldots, s^{(0)}_0] = [-1, \ldots, -1, -1]$, $s^{(1)} := [s^{(1)}_{k^*-1}, \ldots, s^{(1)}_0] = [-1, \ldots, -1, 1]$, $\ldots$,$s^{(2^{k^*}-1)} := [1, \ldots, 1]$.

We consider systems that predict the environment. They are based on binary functions that are defined on a state space $\mathbb{S}_k$ where $k > 0$ is an integer which may be different from k^*. Given a sample of consecutive values of the environment that form a finite Markov chain

$$X^{(m)} := \{X_t\}_{t=-\max\{k,k^*\}+1}^{m}, \tag{1}$$

we define a criterion function based on this sample and choose a predictor which minimizes the criterion. The criterion function is an upper bound on the prediction error and is commonly referred to as a penalized empirical error estimate [4].

Statistical theory of empirical processes [7] guarantees (up to a certain level of confidence) that the error of a predictor which minimizes the criterion function is no larger than the minimum value of the criterion over all predictors. Roughly speaking, this means that the prediction error of the chosen predictor is as close as possible to the minimum error over all predictors. If this holds, we say that the criterion *succeeds* (otherwise it *fails*).

Using upper bounds as criteria for learning classification or prediction is a mainstay of statistical learning theory [11, 10, 5, 1]. In the current paper, we approach the problem of prediction in a context of systems analysis and study how a predictor's complexity affects its stability. (We use the term 'system' and 'predictor' interchangeably.)

We define stability as the difference in system-performance on two random samples, one on which the predictor is calibrated, and another which serves as a sample estimate of the future behavior of the environment. Since we are interested in understanding how complexity influences stability (which is based on

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

system performance), in the context of [9], as a functional requirement of a predictor, we require that if the above criterion fails on a random test sample, then this failure must be detected. We define system complexity to be the uncertainty in meeting this requirement, namely, the uncertainty in detecting a failure of the criterion given that such a failure occurs. As the main result of the paper, we derive an expression that shows how the complexity of a predictor influences its stability. One consequence of the result is that the predictor's stability depends on its level of adaptivity to the random environment.

Because of space limitation, the proofs of the statements in the paper are excluded and are available in [8].

2 Setup

Based on $\mathbb{S}_{k^*}$, the chain $X^{(m)}$ can be represented as a sequence

$$S^{*(m)} = \{S_t^*\}_{t=1}^m \tag{2}$$

of random states where

$$S_t^* \quad := \quad \left(X_{t-(k^*-1)}, X_{t-(k^*-2)}, \ldots, X_t\right) \in \mathbb{S}_{k^*} \tag{3}$$

defines the random state at time t. With respect to $\mathbb{S}_{k^*}$, a state transition occurs from S_t^* to S_{t+1}^* by shifting left the sequence of bits in (3), to obtain $S_{t+1}^* := (X_{t-(k^*-2)}, \ldots, X_t, X_{t+1})$. There are two possible transitions that can occur from S_t^* into S_{t+1}^*: a *negative* transition, where the lower bit X_{t+1} is -1 and *positive* transition where X_{t+1} is 1.

We denote by Q a $2^{k^*} \times 2^{k^*}$ transition probability matrix of the environment's Markov chain. Its $(ij)^{th}$ entry is denoted by

$$Q[i,j] := p\left(s^{(j)} \,\middle|\, s^{(i)}\right). \tag{4}$$

We denote by $p(1|i)$ and $p(-1|i)$, the probability of the two possible transitions from state $s^{(i)}$ and we assume that for all $0 \leq i \leq 2^{k^*} - 1$, $p(1|s^{(i)}) > 0$, thus the environment's Markov chain is irreducible. Let

$$\pi := \left[\pi_0, \ldots, \pi_{2^{k^*}-1}\right] \tag{5}$$

denote the stationary probability distribution where π_i is the probability that $S_t^* = s^{(i)}$. That a stationary probability distribution exists follows from the fact that the Markov chain is irreducible and the state space $\mathbb{S}_{k^*}$ is finite (Corollary 8.2, [6]).

Denote by $S^{(m)}$ the sequence of states of $\mathbb{S}_k$ that corresponds to $X^{(m)}$, that is,

$$S^{(m)} = \{S_t\}_{t=1}^m \tag{6}$$

and

$$S_t \quad := \quad \left(X_{t-(k-1)}, X_{t-(k-2)}, \ldots, X_t\right) \in \mathbb{S}_k. \tag{7}$$

We let the space $\mathbb{S}_k$ have a metric as follows: start with an undirected graph $G_k = (V_k, E_k)$ where V_k and E_k represent the vertex and edge sets. The vertices correspond to the states of $\mathbb{S}_k$. An edge exists between two distinct vertices if the transition probability (4) from at least one of the corresponding states to the other, is positive. This graph is known as an undirected De Bruijn graph of dimension k and is 2-connected (maximum degree 4).

Define the distance $\mathsf{d}(s, s')$ between states $s, s' \in \mathbb{S}_k$ to be the length of the shortest path between the corresponding vertices. (G_k is a connected graph so there is always a path between any two vertices.) Define the diameter of $\mathbb{S}_k$ as $\mathrm{diam}(\mathbb{S}_k) := \max_{s,s' \in \mathbb{S}_k} \mathsf{d}(s, s')$. The diameter equals k.

A γ-*cover* of $\mathbb{S}_k$ with respect to the metric d is a set $C \subseteq \mathbb{S}_k$ such that for every element $s \in \mathbb{S}_k$ there exists an $s' \in C$ such $\mathsf{d}(s, s') \leq \gamma$. The size of the smallest γ-cover of $\mathbb{S}_k$ is defined as the γ-*covering number* of $\mathbb{S}_k$ with respect to d, and is denoted by N_γ.

3 Prediction rules and margin

Denote by $\mathcal{H}$ the class of all binary functions $h : \mathbb{S}_k \to \{-1, 1\}$. We let $\mathcal{H}$ serve as the class of possible predictors of the environment. For a subset $R \subseteq \mathbb{S}_k$ let

$$\mathrm{dist}(s, R) := \min_{s' \in R} \mathsf{d}(s, s').$$

From [2], we use a notion of *width* of h at s which is defined by

$$w_h(s) := \mathrm{dist}\left(s, R_{\overline{h}(s)}\right) \tag{8}$$

where R_+, $R_- \subseteq \mathbb{S}_k$ are regions classified as 1 and -1, respectively, by h, and $\overline{h}(s)$ is the complement of $h(s)$. Because $s \notin R_{\overline{h}(s)}$ then $w_h(s) > 0$. Define $f_h : \mathbb{S}_k \to \mathbb{R}$ by

$$f_h(s) := h(s) w_h(s) \tag{9}$$

to be a *margin* function associated with h. We can evaluate the width and margin functions because k is known and thus the edges of the De Bruijn graph on $\mathbb{S}_k$ are known (the De Bruijn graph of the environment's space $\mathbb{S}_{k^*}$ and its corresponding transition matrix Q are not needed).

We can express the decision of h at s as $h(s) = \mathrm{sgn}(f_h(s))$ thus the function f_h not only contains the binary decision information of h but, more importantly, the absolute value of $f_h(s)$ is a form of confidence in the decision $h(s)$. We use this fact to consider errors made at confident predictions.

Given any binary function $h \in \mathcal{H}$, the predictor based on h decides at time t according to the following rule: if $h(S_{t-1}) = 1$ it predicts for X_t the value 1, otherwise it predicts -1.

The probability that h makes a prediction error is constant with respect to time t because the environment is stationary. We denote it by

$$L(h) := \mathbb{P}\left(X_t f_h\left(S_{t-1}\right) < 0\right). \tag{10}$$

Denote the l_∞ -norm of f_h by $\|f_h\| := \max_{s\in\mathbb{S}_k} |f_h(s)|$. Denote the class of margin functions by $\mathcal{F} := \{f_h : h \in \mathcal{H}\}$. An α-cover of $\mathcal{F}$ with respect to the l_∞ norm on $\mathbb{S}_k$ is a set $\hat{F}_\alpha := \left\{f_j^{(\alpha)}\right\}_{j=1}^{r}$ such that for every element $f \in \mathcal{F}$ there exists an $f_j^{(\alpha)} \in \hat{F}_\alpha$ such $\left\|f - f_j^{(\alpha)}\right\|_{l_\infty} \leq \alpha$. We denote by $h_j := \operatorname{sgn}\left(f_j^{(\alpha)}\right)$ the binary function that corresponds to $f_j^{(\alpha)}$ (note that $j := j(\alpha)$ and we omit the dependence on α for brevity). The size r of the smallest α-cover of $\mathcal{F}$ is defined as the α-covering number of $\mathcal{F}$ with respect to l_∞ norm on $\mathbb{S}_k$ and is denoted by $\mathcal{N}_\alpha$. From [2] (proof of Theorem 4.1), it follows that

$$\mathcal{N}_\alpha \leq \left(2\left\lceil\frac{3\mathrm{diam}\left(\mathbb{S}_k\right)}{\alpha}\right\rceil + 1\right)^{N_{\alpha/3}} \tag{11}$$

where N_α is the α-covering number of $\mathbb{S}_k$ with respect to the metric d. It follows that

$$\log_2 \mathcal{N}_{\alpha/2} \leq N_{\alpha/6} \log_2\left(\frac{15k}{\alpha}\right). \tag{12}$$

The notion of margin error has been useful in statistical learning (see [3, 1] and references within). The *margin error* of h at time t is defined as

$$L^{(\gamma)}(h) := \mathbb{P}\left(X_t f_h(S_{t-1}) < \gamma\right). \tag{13}$$

The *empirical margin error* based on $X^{(m)}$ is defined as follows,

$$L_m^{(\gamma)}(h) := \frac{1}{m}\sum_{t=1}^{m} \mathbb{I}\left\{X_t f_h(S_{t-1}) < \gamma\right\}. \tag{14}$$

4 Calibrated predictor

In [8], section A.1, it is shown that there exists a finite integer $\mathfrak{l}_0$, such that for $\mathfrak{l} \geq \mathfrak{l}_0$, the transition matrix Q in (4) satisfies $Q^{\mathfrak{l}} > 0$, that is, every entry of $Q^{\mathfrak{l}}$, denoted by $p^{(\mathfrak{l})}(s^{(j)}|s^{(i)})$, is positive. We choose $\mathfrak{l}_0 := \min\{l : Q^l > 0\}$ and in theory, if Q was known then $\mathfrak{l}_0$ can be evaluated by computing Q^l for a sequence $l \geq 1$ until the first l is found such that $Q^l > 0$. Denote by μ_0 the minimum entry of $Q^{\mathfrak{l}_0}$.

We henceforth make the following assumption:

Assumption 1. The environment's transition matrix Q satisfies one of the following conditions: (i) the minimum entry of $Q^{\mathfrak{l}_0}$ is $\mu_0 \neq 2^{-k^*}$ or (ii) $\mu_0 = 2^{-k^*}$ and for all $0 \leq i \leq 2^{k^*} - 1$, the transition probabilities $p(1|i) = p(-1|i) = \frac{1}{2}$.

In both parts (i) and (ii) of the above assumption, Q may have a uniform stationary distribution $\pi^T = \left[2^{-k^*}, \ldots, 2^{-k^*}\right]$, which means Q is doubly stochastic and $\lim_{l\to\infty} Q^l$ is a matrix U, of the same size as Q, with all its entries identical to 2^{-k^*}. Part (ii) treats the special case where this limit U is reached exactly at time $\mathfrak{l}_0$, that is, $Q^{\mathfrak{l}_0} = U$. According to the cases of Assumption 1, define

$$\rho(k^*, \mathfrak{l}_0) := \begin{cases} \frac{1-2^{k^*}\mu_0}{2\mu_0} & \text{if case (i)} \\ & \text{and } \mathfrak{l}_0 = 1, \\ \frac{2^{k^*-1}}{\left(1-2^{k^*}\mu_0\right)^{(\mathfrak{l}_0-1)/\mathfrak{l}_0}\left(1-\left(1-2^{k^*}\mu_0\right)^{1/\mathfrak{l}_0}\right)} & \text{if case (i)} \\ & \text{and } \mathfrak{l}_0 \geq 2, \\ 2^{k^*-1} & \text{if case (ii).} \end{cases} \tag{15}$$

Let

$$\xi(m,\gamma,\delta) := r(k,k^*)\rho(k^*,\mathfrak{l}_0) \sqrt{\frac{2}{m}\left(\left(1 + \left(N_{\gamma/12}+1\right)\log_2\left(\frac{30k}{\gamma}\right)\right)\ln 2 + \ln\left(\frac{1}{\delta}\right)\right)} \tag{16}$$

where

$$r(k,k^*) := \begin{cases} 1 & \text{if } k^* \geq k+1 \\ k - k^* + 2 & \text{if } k^* \leq k. \end{cases} \tag{17}$$

We define the *penalized margin error* of h as

$$\hat{L}_m^{(\gamma)}(h) := L_m^{(\gamma)}(h) + \xi(m,\gamma,\delta) \tag{18}$$

which is a random variable since it depends on $X^{(m)}$ through $L_m^{(\gamma)}(h)$. The following is a concentration bound for a Markov chain which holds uniformly over the class $\mathcal{H}$ and over the range of values for γ.

Lemma 1. *For $\gamma > 0$ let N_γ be the γ-covering number of $\mathbb{S}_k$ with respect to the metric d. Let $X^{(m)}$ be a Markov chain sampled from the environment. For any $0 < \delta \leq 1$, with probability at least $1-\delta$, for all $h \in \mathcal{H}$ and for every $0 < \gamma \leq diam(\mathbb{S}_k)$, the following holds*

$$L(h) \leq \hat{L}_m^{(\gamma)}(h). \tag{19}$$

The proof is provided in [8], section A.2. Next, we use $\hat{L}_m^{(\gamma)}(h)$ as a *criterion function* for selecting a good

predictor. Given a random sequence $X^{(m)}$ let (h', γ') be any pair that satisfies the following:

$$\hat{L}_m^{(\gamma')}(h') = \min_{h\in\mathcal{H},\gamma\in(0,\mathrm{diam}(\mathbb{S}_k)]} \hat{L}_m^{(\gamma)}(h). \tag{20}$$

Let

$$\gamma_m := \max\{\gamma' : (h', \gamma') \text{ satisfies (20)}\} \tag{21}$$

and denote by h_m its corresponding function. Define (h_m, γ_m) as a *calibrated predictor*, that is, a predictor which is calibrated to its random environment based on a sample $X^{(m)}$. It is shown in [8], section A.3, that the calibrated predictor (h_m, γ_m) always exists.

Remark 2. The calibrated predictor (h_m, γ_m) minimizes the penalized margin error over $h \in \mathcal{H}$ and over the range of values of γ. This together with Lemma 1, means that with probability at least $1-\delta$, the upper bound on the error of h_m, $L(h_m) \leq \hat{L}_m^{(\gamma_m)}(h_m)$, is minimum over all $h \in \mathcal{H}$. If this occurs, we say that the criterion *succeeds*.

Note that while γ_m is not used in the predictor's decision, its choice influences which $h \in \mathcal{H}$ is selected to be h_m.

Since (16) decreases with γ then the higher the value of γ_m, the lower the upper bound on the error of h_m and the better that h_m fits the general (or typical) behavior of the environment rather than fit a particular realization of the random sample $X^{(m)}$. This motivates the following definition of adaptivity to the environment.

Definition 3. *(Level of adaptivity)* Let (h_m, γ_m) be a predictor system calibrated to the environment based on a sample $X^{(m)}$. Its level of adaptivity to the environment is defined to be γ_m.

5 Complexity of calibrated predictor

For a fixed $m \geq 1$ and $0 < \delta \leq 1$, and for any $h \in \mathcal{H}$, $0 < \gamma \leq \mathrm{diam}(\mathbb{S}_k)$ let us define

$$E_h^{(\gamma)} := \left\{x^{(m)} : L(h) > L_m^{(\gamma)}(h) + \xi(m, \gamma, \delta)\right\}$$

as the set of bad samples on which the upper bound (19) fails to hold for predictor system (h, γ). Let the class of such sets be defined as

$$\mathcal{E}_{\mathcal{H}} := \left\{E_h^{(\gamma)} : h \in \mathcal{H},\, 0 < \gamma \leq \mathrm{diam}\,(\mathbb{S}_k)\right\}.$$

Next we approximate $\mathcal{E}_{\mathcal{H}}$ by a finite class of sets that are defined in a similar way. Let l be a non-negative integer. Consider a minimal $(1/2)^{l+2}k$-cover $\hat{F}_{k(1/2)^{l+2}}$ of $\mathcal{F}$ (the factor k is the diameter of $\mathbb{S}_k$). For $f_j^{(k(1/2)^{l+2})} \in \hat{F}_{k(1/2)^{l+2}}$ denote by $h_j = \mathrm{sgn}(f_j^{(k(1/2)^{l+2})})$. Define a set of bad samples associated with h_j as

$$B_j^{(\gamma)} := \left\{x^{(m)} : L^{(\gamma)}(h_j) > L_m^{(\gamma)}(h_j) + \xi(m, 4\gamma, \delta)\right\}.$$

For $0 \leq l < \infty$ denote by $B_{j,l} := B_j^{((1/2)^{l+2})}$ and define the class

$$C^{(l)} := \{B_{j,l}\}_{j=1}^{\mathcal{N}_{k(1/2)^{l+2}}}$$

where $\mathcal{N}_\gamma$ is the γ-covering number of $\mathcal{F}$ with respect to the l_∞-norm on $\mathbb{S}_k$. The next lemma states that given (h_m, γ_m) we can approximate the set $E_{h_m}^{(\gamma_m)}$ by an element of the class $C^{(l_m)}$ where l_m is directly obtained from γ_m by checking in which interval γ_m is contained.

Lemma 4. *For $m \geq 1$ let (h_m, γ_m) be a predictor calibrated based on $X^{(m)}$. Define l_m as a non-negative integer that satisfies $(1/2)^{l_m+1}k \leq \gamma_m \leq (1/2)^{l_m}k$. Then there exists a $1 \leq j \leq \mathcal{N}_{k(1/2)^{l_m+2}}$, which is denoted j_m, such that $E_{h_m}^{(\gamma_m)} \subseteq B_{j_m,l_m}$ where $B_{j_m,l_m} \in C^{(l_m)}$.*

The proof is in [8], section A.4.

Next we define a notion of complexity of the calibrated predictor. In the context of [9], we set the functional requirement of the calibrated predictor (selected by the criterion) to be as follows: if the bound (19) fails to hold for (h_m, γ_m) (the criterion fails), then this must be detected. We define the complexity of the system (h_m, γ_m) to be the level of uncertainty in detecting that the criterion fails given that it fails.

The event that represents failure of the criterion is $X^{(m)} \in E_{h_m}^{(\gamma_m)}$. From Lemma 4, if $X^{(m)} \in E_{h_m}^{(\gamma_m)}$ then $X^{(m)} \in B_{j_m,l_m}$ therefore it is possible to detect failure of the criterion by detecting that $X^{(m)}$ falls in at least one element B_{j,l_m} of $C^{(l_m)}$.

Given $X^{(m)} \in E_{h_m}^{(\gamma_m)}$, the index j_m of the set B_{j_m,l_m} that contains $X^{(m)}$ is random because the set $E_{h_m}^{(\gamma_m)}$ is random due to (h_m, γ_m). This index j_m takes values in the set $\left\{1, \ldots, \left|C^{(l_m)}\right|\right\}$ and its conditional entropy is bounded by the entropy of the uniform probability distribution on this set,

$$H\left(j_m \,\middle|\, X^{(m)} \in E_{h_m}^{(\gamma_m)}\right) \leq \log_2\left|C^{(l_m)}\right| \text{ bits.}$$

Therefore, the uncertainty in detecting that the criterion fails, given that it fails, is what we define as the complexity of the system. It is no more than $\log_2\left|C^{(l_m)}\right|$ bits and, from (11), is bounded from above as

$$\log_2\left|C^{(l_m)}\right| \leq N_{k(1/2)^{l_m+1}/6}\left(l_m + \log_2(30)\right). \tag{22}$$

By definition of l_m we have $l_m \leq \log_2\left(\frac{k}{\gamma_m}\right)$ and $\frac{\gamma_m}{2} \leq \left(\frac{1}{2}\right)^{l_m+1} k$ therefore (22) is no larger than $N_{\gamma_m/12}\log_2\left(\frac{30k}{\gamma_m}\right)$.

Definition 5. *(Complexity of calibrated predictor)* Let (h_m,γ_m) be the calibrated predictor for a Markov chain $X^{(m)}$. Define the complexity of (h_m,γ_m) as

$$\mathcal{C}(h_m,\gamma_m) := N_{\gamma_m/12}\log_2\left(\frac{30k}{\gamma_m}\right) \text{ bits.} \tag{23}$$

Note that the *larger* the adaptivity level γ_m of the calibrated predictor, the *lower* its complexity $\mathcal{C}(h_m,\gamma_m)$. Thus, a calibrated predictor which is better adapted to its random environment has a lower complexity.

6 Stability of calibrated predictor

As mentioned in section 1, the definition of system stability involves two samples of the environment. The first sample is defined in (1). We now sample $n+\max\{k,k^*\}$ consecutive bits from the environment to obtain a second sample

$$X^{(n)} \quad := \quad \{X_t\}_{t=-\max\{k,k^*\}+1}^{n}. \tag{24}$$

We say that a system is stable if its performance on the two sequences $X^{(m)}$, $X^{(n)}$ is not too different. That is, the first sample $X^{(m)}$ is taken to be the past behavior of the environment and the second sample $X^{(n)}$ is viewed as a sample estimate of the future behavior, thus stability of the predictor implies that its past and future response to the random environment is not too different.

We now describe this in details. Denote by $0<\sigma\leq 1$ a sensitivity parameter and let

$$\phi_\sigma(\gamma) := (\gamma/\sigma^2)(1+4\sigma).$$

Define the *performance discrepancy* of a system (h,γ) by

$$\Psi^{(\sigma)}(h,\gamma) := \Psi^{(\sigma)}_{m,n}(h,\gamma) = L_n^{(\gamma)}(h) - L_m^{(\phi_\sigma(\gamma))}(h), \tag{25}$$

where the subscript n in $L_n^{(\gamma)}$, and m in $L_m^{(\gamma)}$ shows the dependence on the sequences $X^{(n)}$ and $X^{(m)}$, respectively. (The need for ϕ_σ arises from technical reasons, see (88), (89), (98), (99), in the proof of Theorem 11, [8].) Define the following null hypothesis:

NULL HYPOTHESIS: For all h and γ, $\mathbb{E}\left[\Psi^{(\sigma)}(h,\gamma)\right] \leq 0$.

The null hypothesis is true because

$$\begin{aligned}\mathbb{E}\left[L_n^{(\gamma)}(h) - L_m^{(\phi_\sigma(\gamma))}(h)\right] &= L^{(\gamma)}(h) - L^{(\phi_\sigma(\gamma))}(h)\\ &\leq 0 \end{aligned} \tag{26}$$

where (26) follows from the fact that

$$\begin{aligned} L^{(\gamma)}(h) &:= \mathbb{P}\left(X_t f_h\left(S_{t-1}\right)<\gamma\right)\\ &\leq \mathbb{P}\left(X_t f_h\left(S_{t-1}\right)<\phi_\sigma(\gamma)\right) \\ &= L^{(\phi_\sigma(\gamma))}(h)\end{aligned} \tag{27}$$

and (27) holds by the fact that $\phi_\sigma(\gamma)>\gamma$ for every γ and $\sigma\in(0,1]$.

The main result, Theorem 7 in section 7, shows that with a probability less than δ the discrepancy is large, that is, $\Psi^{(\sigma)}_{m,n}(h_m,\gamma_m)>\epsilon$.

Next, we define the following significance test of level δ and critical value ϵ (the value of ϵ is stated in the theorem):

SIGNIFICANCE TEST: if $\Psi^{(\sigma)}(h_m,\gamma_m)>\epsilon$, then reject the null hypothesis.

We use this test to decide if the calibrated predictor is stable as follows: draw $X^{(m)}$ and obtain the calibrated predictor (h_m,γ_m). Then draw $X^{(n)}$ and evaluate $L_n^{(\gamma_m)}(h_m)$. Calculate the critical value ϵ (which depends on γ_m) using Theorem 7 and apply the above significance test to (h_m,γ_m). If the difference between the corresponding empirical errors $L_n^{(\gamma_m)}(h_m)$ and $L_m^{(\phi_\sigma(\gamma_m))}(h_m)$ is larger than ϵ then we reject the null hypothesis. This means that the performance discrepancy value $\Psi^{(\sigma)}_{m,n}(h_m,\gamma_m)$ deviates by a significant amount from its expected value and thus $L_n^{(\gamma_m)}(h_m)$ is atypically random for what is expected of the calibrated predictor (h_m,γ_m). We take this to mean that the predictor is *unstable*. This is formalized in the next definition.

Definition 6. *(Stability of calibrated system)* Let (h_m,γ_m) be a calibrated predictor for a randomly drawn sequence $X^{(m)}$ from a Markov environment. Evaluate its empirical margin error $L_m^{(\phi_\sigma(\gamma_m))}(h_m)$ at margin-level $\phi_\sigma(\gamma_m)$. Let $X^{(n)}$ be a random sequence generated by the same probability distribution and evaluate the empirical margin error $L_n^{(\gamma_m)}(h_m)$ of h_m on $X^{(n)}$. For any $0<\delta\leq 1$, $0<\sigma<1$, we say that the calibrated predictor (h_m,γ_m) is $\epsilon(m,n,\gamma_m,\delta,\sigma)$-*stable with sensitivity value* σ if

$$\mathbb{P}\left(\Psi^{(\sigma)}_{m,n}(h_m,\gamma_m)>\epsilon(m,n,\gamma_m,\delta,\sigma)\right)\leq\delta \tag{28}$$

where $\Psi^{(\sigma)}_{m,n}(h_m,\gamma_m)$ is the performance discrepancy of the predictor which is defined in (25).

The above definition means that with a confidence of at least $1-\delta$, the calibrated predictor (h_m,γ_m) is stable if the random deviation between its empirical margin error on the two sequences is no larger than a value

ϵ that depends on δ, m, n, γ_m and on a sensitivity parameter value σ. In the next section, after stating the main result, we explain how the complexity of the predictor influences ϵ and hence affects the stability of the predictor.

For a fixed γ, $\phi_\sigma(\gamma)$ is decreasing with respect to increasing σ thus for any $\sigma_1 \leq \sigma_2$ we have,

$$L_m^{(\phi_{\sigma_2}(\gamma))}(h) \quad \leq \quad L_m^{(\phi_{\sigma_1}(\gamma))}(h).$$

If h is ϵ-stable with sensitivity value σ_2 then it is also ϵ-stable with sensitivity value σ_1. In the main result, Theorem 7, we show that σ has a multiplicative effect on the level of adaptivity γ_m. This means that a larger value of σ allows for γ_m to be smaller while still keeping ϵ fixed, that is, $\Psi^{(\sigma)}$ is restricted to the same range ϵ. So a larger σ means that ϵ is less sensitive to the value of γ_m in the sense that even if γ_m is low (which means that the calibrated predictor is less adapted to its environment), still, the discrepancy level is kept within the same range of ϵ.

Hence setting the parameter σ to a higher value, means that we want the stability value of the calibrated predictor to be less sensitive to the adaptivity level. In this sense, we can say that the higher σ the stronger the ϵ-stability.

7 Main result

As explained in the end of section 4, a significance test with a critical value ϵ determines when to reject the null hypothesis and therefore when to decide that the calibrated predictor is unstable. The next result gives the critical value ϵ.

Theorem 7. *Denote by N_γ the γ-covering number of $\mathbb{S}_k$ with respect to the metric d. Let $\rho(k^*, \mathfrak{l}_0)$ and $r(k, k^*)$ be defined in (15) and (17). Let $X^{(m)}$, $X^{(n)}$ be two random sequences sampled from the environment's Markov probability distribution of order k^*. Let (h_m, γ_m) be a calibrated predictor based on $X^{(m)}$ and defined according to (21). For any $0 < \sigma, \delta \leq 1$, (h_m, γ_m) is $\epsilon(m, n, \gamma_m, \delta, \sigma)$-*stable *with sensitivity σ where*

$$\epsilon(m,n,\gamma_m,\delta,\sigma) = r(k,k^*)\rho(k^*,\mathfrak{l}_0)$$
$$\left(\sqrt{\frac{2}{n}\left(N_{\sigma\gamma_m/3}\ln\left(2\left\lceil\frac{3k}{\sigma\gamma_m}\right\rceil+1\right)+\ln\left(\frac{2k}{\gamma_m\delta(1-\sigma)}\right)\right)}\right.$$
$$\left.+\sqrt{\frac{2}{m}\left(N_{\gamma_m/3}\ln\left(2\left\lceil\frac{3k}{\gamma_m}\right\rceil+1\right)+\ln\left(\frac{2k}{\gamma_m\delta(1-\sigma)}\right)\right)}\right) \tag{29}$$

The proof is in [8], section A.5.

An important feature of (29) is its dependence on the adaptivity level γ_m. As shown in the proof, we use sample-dependent concentration bounds to achieve this. From Definition 5 the complexity of a predictor (h_m, γ_m) is $O(N_{\gamma_m}\log(\frac{k}{\gamma_m}))$. The expression (29) involves two such factors. Hence Theorem 7 implies that the value of ϵ increases as the complexity of the calibrated predictor increases.

The larger the system's level of adaptivity γ_m to its random environment, the lower its complexity $\mathcal{C}(h_m, \gamma_m)$ and, from (29), the lower the value of ϵ. This means that if a less complex calibrated predictor (h_m, γ_m) (one which has a high value of γ_m) is stable, then its performance discrepancy value is restricted to a small range ϵ. In contrast, if a more complex predictor (h_m, γ_m) (which has a lower γ_m value) is stable, then its performance discrepancy may still be high because it is restricted to a larger range ϵ.

Hence, with the relationship between adaptivity level and complexity (see paragraph under Definition 5), it follows that a *less* complex calibrated-predictor is *better* adapted to its random environment, has a smaller possible range of discrepancy values and is *more* stable.

From (29), the factor $r(k, k^*)$ makes the value of ϵ increase (the performance discrepancy value can therefore be larger) as the mismatch between k and k^* increases. Thus the more that k and k^* are different, the more unstable that the calibrated-predictor can be. The factor $\rho(k^*, \mathfrak{l}_0)$ shows how ϵ depends on the properties of the environment's Markov chain.

8 Conclusions

We consider the question of how the complexity of a prediction system affects its stability. We define a penalized empirical margin error to be the criterion and a predictor which minimizes this criterion is said to be calibrated to its random environment. Its complexity is defined to be the uncertainty in detecting that the criterion fails to select a predictor with a minimum criterion value.

We then introduce a notion of stability which measures the difference in performance of the calibrated system on two samples of the random environment. We show that the possible range of this difference grows as the complexity of the system increases and conclude that the larger the system's level of adaptivity to its random environment, the lower its complexity and the higher its stability.

References

[1] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[2] M. Anthony and J. Ratsaby. Learning bounds via sample width for classifiers on finite metric spaces. *Theoretical Computer Science*, 529:2–10, 2014.

[3] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based learning methods*. Cambridge University Press, 2000.

[4] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer Verlag, 1996.

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

[6] P. Nicolas. *Understanding Markov Chains, Examples and Applications*. Springer, 2013.

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

[8] J. Ratsaby. Full version of the paper. http://www.ariel.ac.il/sites/ratsaby/Publications/PDF/ADV.pdf

[9] N. P. Suh. Complexity in engineering. *CIRP Annals - Manufacturing Technology*, 54(2):46 – 63, 2005.

[10] L. G. Valiant. A theory of the learnable. *Comm. ACM*, 27(11):1134–1142, 1984.

[11] V. N. Vapnik. *Estimations of dependences based on statistical data*. Springer, 1982.