Identification in Prediction Theory

Lars Bäumer

Bielefeld 2000

Acknowledgment

I wish to thank Professor R. Ahlswede for his support. I feel grateful that he shared the idea that led to this research with me.
I also thank Dr. U. Tamm and Dr. C. Kleinewächter for the discussion and their constructive remarks.

Contents

Chapter 1

Introduction

In this work the concept of identification is applied in the theory of prediction. This approach was suggested to us by our advisor Professor R. Ahlswede. This and other directions of research can be found also in [2]. Well known is Shannon's theory of transmission of messages over a noisy channel ([15]). Using the framework of Shannon's channel model a new concept of information transfer - called identification - was introduced by Ahlswede and Dueck in [1].

In the classical transmission model a sender wants to inform a receiver about a message by sending codewords over a channel. The channel may induce some errors and the goal is to have a large number of possible messages such that with sufficiently high probability the receiver should be able to decide which message had been sent. In identification via channels the receiver is no longer interested in what the actual message is, rather he is concerned about one particular message and only wants to know whether this message has occurred or not. However the sender does not know which message the receiver is interested in. Alternatively one can also think of several receivers, one for each message. Each receiver is interested whether his message has occurred or not. This modification of the problem actually leads to a general solution concept in mathematics. Whenever there is a problem in which the question has to be answered “What is the solution?” one can also formulate the corresponding identification problem by asking the questions “Is the solution equal to ...? ”. We are going to apply this solution concept of identification to prediction problems.

In a typical prediction problem a person who has made observations $x_1, \dots, x_t$ at time t has to answer the question “What is x_{t+1} ?”. The starting point of the analysis here is to modify this problem by considering for every possible x a person that asks “Is $x_{t+1} = x$?”.

In the formulation of the prediction problem it has to be specified how the data $x_1, x_2, \dots$ is generated. Basically there are two different cases. In the probabilistic setting the sequence is generated by a random process. We will be mainly concerned with the deterministic setting where the sequence is thought to be arbitrary. This is the framework of the award winning paper by Feder, Merhav and Gutman ([8]). In this setting one wishes to deal with all sequences simultaneously. At first glance it may be surprising that if the sequence is arbitrary that the past can be helpful in predicting the future as they are not necessarily related and some care in defining the desired goals is necessary. The prediction scheme one is looking for shall use the past whenever it is helpful.

Information theorists have been concerned about prediction from the very beginning. Two ideas of Shannon shall be noted. In [16] he estimated the entropy of a language by giving persons who speak this language some text with gaps and asking them to make predictions about how to fill the gaps. In this way the persons use their enormous (unconscious) knowledge of the language and it is possible to get good estimates. In [17], inspired by Hagelbarger, he designed a *mind reading machine*. This machine is developed to play the game of matching pennies against human opponents. So it tries to predict human decisions between two alternatives at every time instant. The success of this machine is explained by the fact that "untrained" human opponents are not able to draw completely random bits. In our terminology the mind reading machine is a finite-state machine with eight states. The predictor presented in Chapter 2.1 is in this way a better mind reading machine as it outperforms for any sequence the best finite-state predictor, for that particular sequence. The price for this, apart from the complexity of the scheme, is the amount of information memorized from the past. In fact this predictor has infinite memory.

The thesis is organized as follows. In Chapter 2 we introduce the finite-state predictability of an individual sequence. This is the minimal asymptotic relative frequency of prediction errors made by the best finite-state predictor for that sequence. A predictor that achieves this performance simultaneously for all sequences in the long run (this will be called a universal predictor) is developed in Section 2.1. Section 2.2 deals with the generalization of the problem to general loss functions. In Chapter 3 we begin to work out the new approach of identification in prediction problems. We define the finite-state identifiability of a sequence. Actually we distinguish here two quantities the strong identifiability and the identifiability which differ in the way how restrictive the definitions are done. Then we show that the universal predictor that attains the finite-state predictability can also be used to derive a universal identification scheme (Section 3.1). Furthermore we compare the new notion of identifiability of a sequence with the predictability and derive relations between these quantities (Section 3.2). The analysis of a special class of finite-state machines, the Markov machines, enables us to show that asymptotically strong identifiability and identifiability coincide (Section 3.3).

Motivated by the identification theory for channels where the consideration of randomized codes brought a big advantage we analyze the effects of randomized finite-state machines for identification. In Section 3.4 we show that asymptotically randomization does not increase the performance here.

Chapter 2

Finite-State Predictability

We assume that there is a finite number of possibilities for the observations made at each time instant. Therefore we work throughout the thesis with a finite alphabet

$$\mathcal{X} = \{0, \ldots, M-1\}$$

of size $M \geq 2$. The set of all words of length n is denoted by $\mathcal{X}^n$. Words of length n are denoted as

$$x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n.$$

The set of all infinite sequences of letters from $\mathcal{X}$ is denoted by $\mathcal{X}^\infty$ and a typical element of $\mathcal{X}^\infty$ will be denoted by $x^\infty \in \mathcal{X}^\infty$.
A deterministic predictor with infinite memory is a family $(b_t)_{t\geq 1}$ of functions $b_t : \mathcal{X}^{t-1} \to \mathcal{X}$. If x^{t-1} has been observed at time t so far then $b_t(x^{t-1})$ is the predicted letter. The performance criterion for a predictor of this form is the asymptotic relative frequency of prediction errors:

$$\frac{1}{n}\sum_{t=1}^{n} d(x_t, b_t(x^{t-1})),$$

where $d(x,y) = 0$ if $x = y$ and $d(x,y) = 1$ if $x \neq y$ (d is the Hamming distance).

If the sequence is thought to be an arbitrary individual sequence some care in defining the universal prediction problem has to be employed. Let $\mathcal{B} \triangleq \{(b_t)_{t\geq 1} : b_t : \mathcal{X}^{t-1} \to \mathcal{X}\}$ be the class of all deterministic predictors. Observe the following two facts.

1. For every individual sequence $x_1, x_2, \ldots$ there is one predictor $(b_t)_{t\geq 1} \in \mathcal{B}$ which makes no errors at all for that sequence, $b_t(x^{t-1}) = x_t$ for all $t \in \mathbb{N}$.

2. For every predictor $(b_t)_{t\geq 1} \in \mathcal{B}$ there is a sequence $\bar{x}_1, \bar{x}_2, \ldots$ for which this predictor makes errors at all time instants. Such a sequence is defined inductively by $\bar{x}_t \triangleq \bar{x}$ with $\bar{x} \neq b_t(\bar{x}^{t-1})$ for all $t \in \mathbb{N}$.

Therefore the search for a universal predictor that for all sequences is nearly as good as the best predictor from $\mathcal{B}$ for that particular sequence cannot be successful. To avoid these trivialities we will restrict the class $\mathcal{B}$ to some class $\mathcal{B}' \subset \mathcal{B}$ in a reasonable way and then try to achieve the performance of the best predictor from $\mathcal{B}'$. This class $\mathcal{B}'$ will be denoted as comparison class. But notice that, because of 2., every predictor from $\mathcal{B}$ is very bad for some sequences. Therefore we cannot hope to find a universal predictor in $\mathcal{B}$. This difficulty is avoided by allowing the predictors to be randomized.
Let us now describe how we restrict the class $\mathcal{B}$. The comparison class $\mathcal{B}'$ that we use will be the class of all *finite-state predictors.*

Definition 1. A finite-state predictor is a triple $(\mathcal{S}, g, f)$ consisting of

$$\mathcal{S} = \{1, \ldots, S\} \text{ a finite set of states,}$$

$$g : \mathcal{S} \times \mathcal{X} \to \mathcal{S} \text{ a next-state function,}$$

$$f : \mathcal{S} \to \mathcal{X} \text{ a prediction rule.}$$

An finite-state predictor works as follows. At time t it predicts the value of x_{t+1} depending on its current state s_t by

$$\hat{x}_{t+1} = f(s_t).$$

Then x_{t+1} is revealed and the machine changes its state to

$$s_{t+1} = g(s_t, x_{t+1})$$

according to the next-state function.

The specific labels of the states do not matter, therefore we assume without loss of generality that at the beginning the machine is always in state 1, i.e., $s_0 = 1$.

In this way, if g and x^n are given, a sequence $s_0, s_1, \ldots, s_{n-1}$ of states is generated. For this we use the following abbreviations.

Definition 2. If x^n and a next-state function g are given and $s_0, s_1, \ldots, s_{n-1}$ is the generated state sequence then let

$\langle x^n|s, x\rangle \triangleq |\{t : s_t = s, x_{t+1} = x\}|$,

$\langle x^n|x\rangle \triangleq |\{t : x_t = x\}|$,

$\langle x^n|s\rangle \triangleq |\{t : s_t = s\}|$.

The symbols for these counts do not indicate the dependence on the specific next-state function g but it should always be clear from the context which g is meant.

We can also allow probabilistic prediction rules f, i.e., we select $\hat{x}_{t+1}$ randomly with respect to a conditional probability distribution, given s_t. There are always optimal deterministic prediction rules meaning that if the next-state function g and the initial state s_0 are fixed then for given x^n a prediction rule that minimizes the relative frequency of prediction errors of the finite-state predictor is deterministic and given by

$$f(s) = \hat{x}, \text{ where } \hat{x} \text{ maximizes } \langle x^n|s, x\rangle \text{ over all } x \in \mathcal{X}. \tag{2.1}$$

This optimal rule for fixed g depends on the whole sequence x^n and in general cannot be determined while the data are observed or, as we shall call it, in a sequential way. The best prediction rule may depend on the whole sequence but anyway for each sequence there is a best finite-state predictor and although it cannot be determined sequentially it will serve us as a comparison for our sequential predictors.

Applying the optimal rule, as described in (2.1), to the sequence x^n yields a fraction of prediction errors equal to

$$\pi_S(x^n, g) \triangleq \frac{1}{n} \sum_{s=1}^{S} \left[\langle x^n|s\rangle - \max_{x\in\mathcal{X}} \{\langle x^n|s, x\rangle\} \right].$$

Definition 3. The *S-state-predictability of x^n* is given by

$$\pi_S(x^n) \triangleq \min_{g\in\mathcal{G}_S} \pi_S(x^n, g),$$

where $\mathcal{G}_S$ is the set of all $S^{|\mathcal{X}|\cdot S}$ next-state functions.

Definition 4. The *asymptotic S-state predictability of x^∞* is given by

$$\pi_S(x^\infty) \triangleq \limsup_{n\to\infty} \pi_S(x^n).$$

Example 1. Consider the sequence $x^\infty = 01010101\ldots$.
Then clearly $\pi_1(x^\infty) = \frac{1}{2}$ and $\pi_2(x^\infty) = 0$.

Definition 5. The *finite-state predictability of x^∞* is given by

$$\pi(x^\infty) \triangleq \lim_{S\to\infty} \pi_S(x^\infty).$$

The limit in Definition 5 always exists because $\pi_S(x^\infty)$ is monotonically non-increasing in S.

2.1 A Universal Predictor

In this section, based on the results of Feder, Merhav and Gutman ([8]), we present a slightly generalized predictor that attains the finite-state predictability for all binary sequences. The first main step is to develop a predictor that attains the 1-state predictability universally, i.e., the predictor has to compete for each sequence with the best constant predictor. Our predictor works as follows: At time t it predicts

$$\hat{x}_{t+1} \triangleq \begin{cases} 0, & \text{with probability}\, \phi_t(\frac{\langle x^t|0\rangle+\gamma}{t+2\gamma}) \\ 1, & \text{with probability}\, \phi_t(\frac{\langle x^t|1\rangle+\gamma}{t+2\gamma}) \end{cases}$$

where $\gamma > 0$ is a constant and

$$\phi_t(\alpha) \triangleq \begin{cases} 0, & 0 \le \alpha < \frac{1}{2} - \epsilon_t \\ \frac{1}{2\epsilon_t}(\alpha - \frac{1}{2}) + \frac{1}{2}, & \frac{1}{2} - \epsilon_t \le \alpha \le \frac{1}{2} + \epsilon_t \\ 1, & \frac{1}{2} + \epsilon_t < \alpha \le 1 \end{cases}$$

and $(\epsilon_t)_{t \ge 0}$ is a sequence of parameters with $\epsilon_t > 0$ that will be specified later.

Let $\hat{\pi}(x^n)$ be the expected fraction of errors made by this predictor on the sequence x^n.

The following theorem shows that $\hat{\pi}(x^n)$ approaches $\pi_1(x^n)$ universally for all sequences.

Theorem 1. *Let $\gamma > 0$. For any sequence $x^n \in \{0,1\}^n$ and for $\epsilon_t = \frac{1}{2\sqrt{t+2\gamma}}$ it holds*

$$\begin{aligned} \hat{\pi}(x^n) &\le \pi_1(x^n) + \delta_1(n,\gamma), \\ \text{where } \delta_1(n,\gamma) &= O(\tfrac{1}{\sqrt{n}}). \end{aligned} \tag{2.2}$$

Furthermore, for any sequence $x^n \in \{0,1\}^n$ and for constant $\epsilon_t = \epsilon$, $0 < \epsilon < \frac{1}{2}$, it holds

$$\begin{aligned} \hat{\pi}(x^n) &\le \pi_1(x^n) + \tfrac{\epsilon}{1-2\epsilon} + \nu(n,\epsilon), \\ \text{where } \nu(n,\epsilon) &= O(\tfrac{\log n}{n}). \end{aligned} \tag{2.3}$$

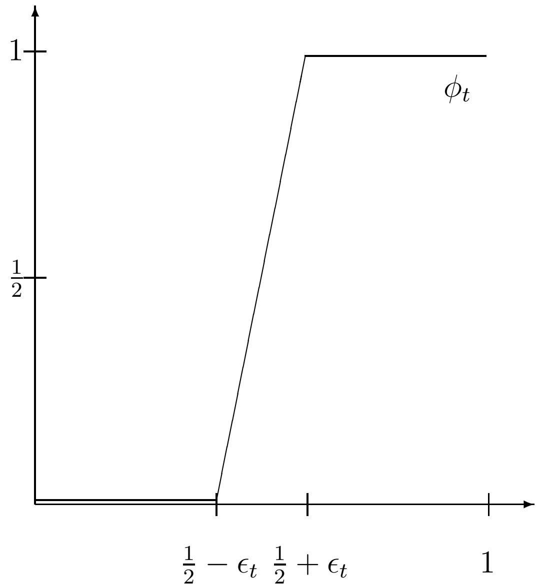

Figure 2.1: The function ϕ_t

Remark 1.

1. A natural choice of $\phi = \phi_t$ could have been

$$\phi(\alpha) = \begin{cases} 0, & \alpha < \frac{1}{2} \\ \frac{1}{2}, & \alpha = \frac{1}{2} \\ 1, & \alpha > \frac{1}{2}. \end{cases}$$

   This means we do majority voting and only if the number of ones and zeros is equal we flip a fair coin. But this is problematic for some sequences, e.g., $x^n = 0101\ldots0101$. $\pi_1(x^n) = \frac{1}{2}$ but the predictor would make 75% errors. The reason for this gap lies in the fact that $\frac{\langle x^t|0\rangle+\gamma}{t+2\gamma}$ converges from above to $\frac{1}{2}$ which is a discontinuity point of ϕ. Thus it is crucial to make ϕ continuous.

2. It was shown in [7] that the convergence rate of $O(\frac{1}{\sqrt{n}})$ is best possible.

3. As mentioned before it is essential that the universal predictor is randomized. There is no deterministic universal predictor.

Proof of Theorem 1: Observe that $\pi_1(x^n) = \frac{1}{n}\min\{\langle x^n|0\rangle, \langle x^n|1\rangle\}$ depends only on the type of the sequence x^n, that is on the total number of 0's and 1's in the sequence. Let us show first that among all sequences of the same type the one for which our predictor performs worst is

$$\tilde{x}^n \triangleq \overbrace{0101\ldots01}^{2\langle x^n|1\rangle}\ \overbrace{00\ldots00}^{\langle x^n|0\rangle-\langle x^n|1\rangle} \tag{2.4}$$

where we assume without loss of generality that $\langle x^n|0\rangle \geq \langle x^n|1\rangle$.

For a sequence of some given type consider the sequence of absolute differences $C_t \triangleq |\langle x^t|0\rangle - \langle x^t|1\rangle|$. Then $C_0 = 0$ and $C_n = \langle x^n|0\rangle - \langle x^n|1\rangle$. We can think of these C_t as states in a state diagram. Let us call a pattern $(C_t = k, C_{t+1} = k+1, C_{t+2} = k)$ (for some integer k) an *upward loop* and similarly a *downward loop* as $(C_t = k, C_{t+1} = k-1, C_{t+2} = k)$. If we change an upward loop into a downward loop this corresponds to changing at some point of the sequence a 01 into a 10 or vice versa. So this operation does not change the type of the sequence but as we shall show next the expected number of errors made by our predictor is increased.

Assume first that $\langle x^t|0\rangle > \langle x^t|1\rangle$. Denote the expected number of errors incurred along an upward loop by

$$\alpha \triangleq 1 - \phi_t\left(\frac{\langle x^t|0\rangle + \gamma}{t+2\gamma}\right) + \phi_{t+1}\left(\frac{\langle x^t|0\rangle + \gamma + 1}{t+2\gamma+1}\right)$$

and the expected number of errors incurred along a downward loop by

$$\beta \triangleq \phi_t\left(\frac{\langle x^t|0\rangle + \gamma}{t+2\gamma}\right) + 1 - \phi_{t+1}\left(\frac{\langle x^t|0\rangle + \gamma}{t+2\gamma+1}\right).$$

Now we consider the difference

$$\alpha - \beta = \phi_{t+1}\left(\frac{\langle x^t|0\rangle + \gamma + 1}{t+2\gamma+1}\right) + \phi_{t+1}\left(\frac{\langle x^t|0\rangle + \gamma}{t+2\gamma+1}\right) - 2\phi_t\left(\frac{\langle x^t|0\rangle + \gamma}{t+2\gamma}\right).$$

For the arguments in the equation above the following relations hold

$$\frac{\langle x^t|0\rangle + \gamma + 1}{t+2\gamma+1} > \frac{\langle x^t|0\rangle + \gamma}{t+2\gamma} > \frac{\langle x^t|0\rangle + \gamma}{t+2\gamma+1} \geq \frac{1}{2}.$$

Now we distinguish two cases.
Case 1: $\frac{\langle x^t|0\rangle+\gamma}{t+2\gamma} \geq \frac{1}{2}+\epsilon_t$
Then $\phi_t\left(\frac{\langle x^t|0\rangle+\gamma}{t+2\gamma}\right) = 1$ and therefore $\alpha-\beta \leq 0$.
Case 2: $\frac{1}{2} \leq \frac{\langle x^t|0\rangle+\gamma}{t+2\gamma} < \frac{1}{2}+\epsilon_t$
Then using for the first two terms of the difference $\alpha-\beta$ a continuation of the sloping part of ϕ_t as an upper bound we get

$$\alpha-\beta \leq \frac{1}{2\epsilon_{t+1}}\left(\frac{\langle x^t|0\rangle+\gamma+1}{t+2\gamma+1}-\frac{1}{2}+\frac{\langle x^t|0\rangle+\gamma}{t+2\gamma+1}-\frac{1}{2}\right)-\frac{2}{2\epsilon_t}\left(\frac{\langle x^t|0\rangle+\gamma}{t+2\gamma}-\frac{1}{2}\right)$$

$$= \frac{1}{2\epsilon_{t+1}}\left(\frac{2\langle x^t|0\rangle-t}{t+2\gamma+1}\right)-\frac{1}{2\epsilon_t}\left(\frac{2\langle x^t|0\rangle-t}{t+2\gamma}\right).$$

Therefore $\alpha-\beta \leq 0$ if

$$\epsilon_t(t+2\gamma) \leq \epsilon_{t+1}(t+2\gamma+1).$$

So the function w given by $w(t)=\epsilon_t(t+2\gamma)$ should be monotonically non-decreasing in t. This means that ϵ_t chosen to be constant or $\epsilon_t = \frac{1}{2\sqrt{t+2\gamma}}$ as in the theorem is possible. The case when $\langle x^t|0\rangle < \langle x^t|1\rangle$ is completely analogous and if $\langle x^t|0\rangle = \langle x^t|1\rangle$, then $\alpha-\beta=0$.
So we have shown that if we are given a sequence of some type and we replace an upward loop by a downward loop we get a sequence of the same type for which the predictor makes a bigger expected number of errors. If we now iterate this process we will finally end up with the sequence of (2.4).
The expected number of errors the predictor makes on the sequence $\tilde{x}^n$ of (2.4) is therefore a uniform upper bound on $\hat{\pi}_1(x^n)$. Let $l_t \triangleq 1-\phi_t$ then

$$n\hat{\pi}(\tilde{x}^n) = \sum_{k=1}^{\langle x^n|1\rangle} l_{2k-2}\left(\frac{k}{2k}\right) + \underbrace{\sum_{k=1}^{\langle x^n|1\rangle} l_{2k-1}\left(\frac{k-1+\gamma}{2k-1+2\gamma}\right)}_{\triangleq A}$$

$$+\underbrace{\sum_{k=1}^{\langle x^n|0\rangle-\langle x^n|1\rangle} l_{k+2\langle x^n|1\rangle-1}\left(\frac{\langle x^n|1\rangle+k-1+\gamma}{2\langle x^n|1\rangle-1+k+2\gamma}\right)}_{\triangleq B}=\frac{\langle x^n|1\rangle}{2}+A+B.$$

Let us consider first the case when ϵ is fixed ($l_t = l$ for all t). In order to upperbound A observe that from the definition of l follows that

$$l\left(\frac{k-1+\gamma}{2k-1+2\gamma}\right)\le\frac{1}{2}+\frac{1}{2\epsilon}\left(\frac{1}{2}-\frac{k-1+\gamma}{2k-1+2\gamma}\right)$$
$$=\frac{1}{2}+\frac{1}{4\epsilon}\cdot\frac{1}{2k-1+2\gamma}.$$

Therefore

$$A\le\frac{\langle x^n|1\rangle}{2}+\frac{1}{4\epsilon}\sum_{k=1}^{\langle x^n|1\rangle}\frac{1}{2k-1+2\gamma}$$
$$\le\frac{\langle x^n|1\rangle}{2}+\frac{1}{4\epsilon}\int_1^{\langle x^n|1\rangle}\frac{du}{2u-1+2\gamma}+\frac{1}{4\epsilon}\cdot\frac{1}{2\gamma+1}$$
$$=\frac{\langle x^n|1\rangle}{2}+\frac{1}{8\epsilon}\ln\left(2\langle x^n|1\rangle-1+2\gamma\right)-\frac{1}{8\epsilon}\ln\left(2\gamma+1\right)+\frac{1}{4\epsilon}\frac{1}{2\gamma+1}$$
$$\le\frac{\langle x^n|1\rangle}{2}+\frac{1}{8\epsilon}\ln\left(2n-1+2\gamma\right)+\frac{1}{4\epsilon}\frac{1}{2\gamma+1},$$

where we used the fact that $2\langle x^n|1\rangle\le n$ in the last inequality.

Now we consider the sum B. For the argument of l it is true that it is always larger than $\frac{1}{2}$ and that $\frac{\langle x^n|1\rangle+k-1+\gamma}{2\langle x^n|1\rangle-1+k+2\gamma}\ge\frac{1}{2}+\epsilon$ if

$$k\ge\frac{1+4\epsilon\langle x^n|1\rangle-2\epsilon+4\epsilon\gamma}{1-2\epsilon}\triangleq K.$$

For these k's l is zero and otherwise we can upperbound it by $\frac{1}{2}$. Therefore

$$B\le\sum_{k=1}^{\lfloor K\rfloor}\frac{1}{2}\le\frac{2\epsilon\langle x^n|1\rangle}{1-2\epsilon}+\frac{1}{2}\cdot\frac{1-2\epsilon+4\epsilon\gamma}{1-2\epsilon}\le\frac{n\epsilon}{1-2\epsilon}+\frac{1}{2}\cdot\frac{1-2\epsilon+4\epsilon\gamma}{1-2\epsilon}.$$

If we combine the estimates for A and B we get

$$\hat{\pi}(x^n) \leq \underbrace{\frac{\langle x^n|1\rangle}{n}}_{\pi_1(x^n)} + \frac{\epsilon}{1-2\epsilon} + \frac{\ln(2n-1+2\gamma)}{8\epsilon n} + \frac{1}{n}\left(\frac{1}{4\epsilon(2\gamma+1)} + \frac{1-2\epsilon+4\epsilon\gamma}{2-2\epsilon}\right),$$

which is the result claimed in (2.3). Now let us consider the case when ϵ is variable. We start by estimating the sum A. Since

$$l_{2k-1}\left(\frac{k-1+\gamma}{2k-1+2\gamma}\right) \leq \frac{1}{2} + \frac{1}{2\epsilon_{2k-1}}\left(\frac{1}{2} - \frac{k-1+\gamma}{2k-1+2\gamma}\right) = \frac{1}{2} + \frac{1}{2}\frac{1}{\sqrt{2k-1+2\gamma}},$$

we get

$$\begin{aligned} A &\leq \frac{\langle x^n|1\rangle}{2} + \frac{1}{2}\sum_{k=1}^{\langle x^n|1\rangle} \frac{1}{\sqrt{2k-1+2\gamma}} \\ &\leq \frac{\langle x^n|1\rangle}{2} + \frac{1}{2}\int_1^{\langle x^n|1\rangle} \frac{du}{\sqrt{2u-1+2\gamma}} + \frac{1}{2\sqrt{2\gamma+1}} \\ &= \frac{\langle x^n|1\rangle}{2} + \frac{1}{2}\sqrt{2\langle x^n|1\rangle - 1 + 2\gamma} + \frac{1}{2}\left(\frac{1}{\sqrt{2\gamma+1}} - \sqrt{2\gamma+1}\right) \\ &\leq \frac{\langle x^n|1\rangle}{2} + \frac{1}{2}\sqrt{n-1+2\gamma} + \frac{1}{2}\left(\frac{1}{\sqrt{2\gamma+1}} - \sqrt{2\gamma+1}\right). \end{aligned}$$

In order to estimate B observe that the nonzero components must satisfy

$$\frac{\langle x^n|1\rangle + k - 1 + \gamma}{2\langle x^n|1\rangle - 1 + k + 2\gamma} \leq \frac{1}{2} + \frac{1}{2\sqrt{2\langle x^n|1\rangle - 1 + k + 2\gamma}}.$$

The largest k satisfying this condition denoted as K can be upperbounded by

$$K \leq \frac{3}{2} + \sqrt{\frac{1}{4} + 2\langle x^n|1\rangle + 2\gamma}.$$

Since all non-zero terms of B are less than $\frac{1}{2}$ we get

$$B \leq \frac{K}{2} \leq \frac{3}{4} + \frac{1}{2}\sqrt{\frac{1}{4} + 2\langle x^n|1\rangle + 2\gamma} \leq \frac{3}{4} + \frac{1}{2}\sqrt{\frac{1}{4} + n + 2\gamma}.$$

Combining the estimates for A and B we derive that

$$\hat{\pi}(x^n) \leq \underbrace{\frac{\langle x^n|1\rangle}{n}}_{\pi_1(x^n)} + \frac{1}{2n}\left(\sqrt{n-1+2\gamma} + \sqrt{\frac{1}{4}+n+2\gamma}\right) + \frac{C_\gamma}{n},$$

where $C_\gamma \triangleq \frac{1}{2}\left(\frac{1}{\sqrt{2\gamma+1}} - \sqrt{2\gamma+1}\right) + \frac{3}{4}$.
This is the desired result of (2.2) and thus the proof of the theorem is complete.

□

Next we deal with the problem how to achieve universally the performance $\pi_S(x^n, g)$ for a given next-state function g with a sequential predictor.
For each state $s \in \mathcal{S}$ the optimal prediction rule $\hat{x}_{t+1} = f(s)$ is fixed and thus we can extend Theorem 1 by considering S sequential predictors of the previously described form. For simplicity we choose $\gamma = 1$. Specifically let

$$\hat{p}_t(x|s) \triangleq \frac{\langle x^t|s,x\rangle + 1}{\langle x^t|s\rangle + 2} \qquad x \in \{0,1\}, s \in \mathcal{S}$$

and consider the predictor

$$\hat{x}_{t+1} = \begin{cases} 0, & \text{with probability } \phi_t(\hat{p}_t(0|s_t)) \\ 1, & \text{with probability } \phi_t(\hat{p}_t(1|s_t)), \end{cases}$$

where ϕ is as before with $\epsilon = \epsilon_{\langle x^t|s_t\rangle}$.
Now we can apply Theorem 1 to each subsequence of x^n which corresponds to a state $s \in \mathcal{S}$ and get

$$\hat{\pi}(x^n, g) \leq \frac{1}{n}\sum_{s=1}^{S} \min\{\langle x^n|s,0\rangle, \langle x^n|s,1\rangle\} + \langle x^n|s\rangle\; \delta_1(\langle x^n|s\rangle)$$

$$= \pi_S(x^n, g) + \sum_{s=1}^{S} \frac{\langle x^n|s\rangle}{n}\; \delta_1(\langle x^n|s\rangle)$$

$$\leq \pi_S(x^n, g) + \underbrace{\frac{S}{n}\sqrt{\frac{n}{S}+1} + \frac{S}{2n}}_{\delta_S(n)}. \tag{2.5}$$

Observe that, as there are less samples in each state, the convergence rate slows down (from $O(\frac{1}{\sqrt{n}})$ to $O(\sqrt{\frac{S}{n}})$).

The next problem we deal with is how to achieve sequentially the S-state predictability for fixed S.

Definition 6. A refinement of a finite-state machine with next-state function g and S states is a finite-state machine with $\tilde{S} \geq S$ states and next-state function $\tilde{g}$ such that there exists a function $h : \tilde{\mathcal{S}} \to \mathcal{S}$ with the property that at each time instant $s_t = h(\tilde{s}_t)$ where s_t and $\tilde{s}_t$ are the states at time t generated by g, $\tilde{g}$ and any $x^n \in \mathcal{X}^n$.

The next lemma shows that a refinement of a finite-state machine can only increase the performance of the finite-state predictor.

Lemma 1. *If the finite-state machine corresponding to $\tilde{g}$ is a refinement of the finite-state machine corresponding to g then for all $x^n \in \{0,1\}^n$ it holds*

$$\pi_S(x^n, g) \geq \pi_{\tilde{S}}(x^n, \tilde{g}).$$

Proof:

$$\pi_S(x^n, g) = \frac{1}{n}\sum_{s=1}^{S} \min\{\langle x^n|s,0\rangle, \langle x^n|s,1\rangle\}$$

$$= \frac{1}{n}\sum_{s=1}^{S} \min\{\sum_{\tilde{s}:h(\tilde{s})=s} \langle x^n|\tilde{s},0\rangle, \sum_{\tilde{s}:h(\tilde{s})=s} \langle x^n|\tilde{s},1\rangle\}$$

$$\geq \frac{1}{n}\sum_{s=1}^{S} \sum_{\tilde{s}:h(\tilde{s})=s} \min\{\langle x^n|\tilde{s},0\rangle, \langle x^n|\tilde{s},1\rangle\}$$

$$= \frac{1}{n}\sum_{\tilde{s}=1}^{\tilde{S}} \min\{\langle x^n|\tilde{s},0\rangle, \langle x^n|\tilde{s},1\rangle\} = \pi_{\tilde{S}}(x^n, \tilde{g}).$$

□

Consider now a refinement $\tilde{g}$ of all S^{2S} possible S-state machines. The state $\tilde{s}_t$ of $\tilde{g}$ is the vector $(s_t^1, \ldots, s_t^M)$, where $s_t^i, i = 1, \ldots, S^{2S}$, is the state at time t of the i-th S-state machine g_i. From Lemma 1 it follows that for all g

$$\pi_{\tilde{S}}(x^n, \tilde{g}) \leq \pi_S(x^n, g)$$

and therefore also

$$\pi_{\tilde{S}}(x^n, \tilde{g}) \leq \pi_S(x^n).$$

Thus the sequential scheme based on $\tilde{g}$ asymptotically universally attains $\pi_S(x^n)$.

The disadvantages of this scheme are obviously that it is very complex, furthermore it attains the predictability only for a fixed given value of S. The rate of convergence also is not best possible.

In order to develop a predictor that universally attains the finite-state predictability and overcomes the disadvantages mentioned above we introduce Markov predictors and the Markov predictability of a sequence and show that it is equal to the finite-state predictability of the sequence. This enables us to design the desired prediction scheme.

Definition 7. A Markov-Predictor of order $k \geq 1$ is a finite-state predictor with 2^k possible states where

$$s_t = (x_{t-k+1}, \ldots, x_t).$$

The initial state $(x_{-k+1}, \ldots, x_0)$ does not affect the asymptotic performance of the Markov predictor. Therefore the choice of s_0 is irrelevant for our purposes. For instance it can be chosen to give the smallest possible value in (2.6) below (in [8] for technical reasons the cyclic convention $x_{-i} = x_{n-i}$ for $i \in \{0, \ldots, k-1\}$ was used).

Then the k-th order Markov predictability of the finite sequence x^n is given by

$$\mu_k(x^n) \triangleq \frac{1}{n} \sum_{x^k \in \{0,1\}^k} \min\{\langle x^n | x^k, 0\rangle, \langle x^n | x^k, 1\rangle\}. \tag{2.6}$$

The asymptotic k-th order Markov predictability of the infinite sequence x^∞ is given by

$$\mu_k(x^\infty) \triangleq \limsup_{n\to\infty} \mu_k(x^n).$$

Finally the Markov predictability of x^∞ is given by

$$\mu(x^\infty) \triangleq \lim_{k\to\infty} \mu_k(x^\infty).$$

As the class of finite-state machines contains as a subclass the class of Markov machines it follows

$$\mu(x^\infty) \geq \pi(x^\infty).$$

The following theorem from [8, Theorem 2] establishes a converse inequality from which follows that Markov predictability and finite-state predictability are equivalent.

Theorem 2. *For all integers $k, S \geq 1$ and any finite sequence $x^n \in \{0,1\}^n$ it holds*

$$\mu_k(x^n) \leq \pi_S(x^n) + \sqrt{\frac{\ln S}{2(k+1)}}. \tag{2.7}$$

Remark 2. The inequality of the theorem is meaningful only if the second term on the right hand side is small, i.e., if k is big compared to $\ln S$. Thus the theorem shows that no matter how clever a finite-state machine is chosen for a given sequence, if k is big enough the Markov predictor of the corresponding order will be almost as good.

Now if in (2.7) we take the limit supremum as $n \to \infty$, then the limit $k \to \infty$ and finally the limit $S \to \infty$ we end up with $\mu(x^\infty) \leq \pi(x^\infty)$ which implies

$$\mu(x^\infty) = \pi(x^\infty).$$

Now it is clear how we can derive a sequential universal prediction scheme that attains $\mu(x^\infty)$ and thus $\pi(x^\infty)$.

We know that for fixed k we can achieve the k-th order Markov predictability by the predictor

$$\hat{x}_{t+1} = \begin{cases} 0, & \text{with probability } \phi_t(\hat{p}_t(0|x_{t-k+1}, \ldots, x_t)) \\ 1 & \text{with probability } \phi_t(\hat{p}_t(1|x_{t-k+1}, \ldots, x_t)), \end{cases} \tag{2.8}$$

where for $x \in \{0, 1\}$

$$\hat{p}_t(x|x_{t-k+1}, \ldots, x_t) = \frac{\langle x^n|(x_{t-k+1}, \ldots, x_t), x\rangle + 1}{\langle x^n|(x_{t-k+1}, \ldots, x_t)\rangle + 2}.$$

To attain $\mu(x^\infty)$ the order k must grow the more data are available. There are two conflicting goals.

- Increasing the order fast in order to attain the Markov predictability as soon as possible.
- Increasing the order slowly in order to ensure that there are enough counts for each state.

It turns out that the order k is not allowed to grow faster than $O(\log t)$ in order to satisfy both requirements.

Let us denote by $\hat{\mu}_k(x^n)$ the expected fraction of errors made by the predictor (2.8) on the sequence x^n.

Then we know that

$$\hat{\mu}_k(x^n) \leq \mu_k(x^n) + \delta_{2^k}(n),$$

with δ_{2^k} as defined in (2.5) and $\delta_{2^k}(n) = O(\sqrt{\frac{2^k}{n}})$.

Divide the observed data into non-overlapping segments

$$x^\infty = x^{(1)}, x^{(2)}, \ldots$$

and apply the k-th order sequential predictor (2.8) to the k-th segment $x^{(k)}$. Let the length n_k of the k-th segment be at least $\alpha_k\, 2^k$, where $(\alpha_k)_k$ is a monotonically increasing sequence such that $\lim_{k\to\infty} \alpha_k = \infty$. Then

$$\begin{aligned}
\hat{\mu}_k(x^{(k)}) &\le \mu_k(x^{(k)}) + \delta_{2^k}(n_k) \\
&\le \mu_k(x^{(k)}) + \frac{\sqrt{\alpha_k + 1}}{\alpha_k} + \frac{1}{2\alpha_k} \\
&= \mu_k(x^{(k)}) + \xi(k),
\end{aligned}$$

where $\xi(k) = O(\frac{1}{\sqrt{\alpha_k}})$.

On an arbitrary long finite sequence x^n, where $n = \sum_{k=1}^{k_n} n_k$ and k_n denotes the number of segments in x^n, the above predictor achieves an average fraction of errors denoted by $\hat{\mu}(x^n)$ which satisfies

$$\hat{\mu}(x^n) = \sum_{k=1}^{k_n} \frac{n_k}{n} \hat{\mu}_k(x^{(k)}) \le \sum_{k=1}^{k_n} \frac{n_k}{n} \mu_k(x^{(k)}) + \sum_{k=1}^{k_n} \frac{n_k}{n} \xi(k).$$

Now for any fixed $k' < k_n$ we obtain

$$\hat{\mu}(x^n) \le \sum_{k=1}^{k'-1} \frac{n_k}{n} \mu_k(x^{(k)}) + \sum_{k=k'}^{k_n} \frac{n_k}{n} \mu_k(x^{(k)}) + \sum_{k=1}^{k_n} \frac{n_k}{n} \xi(k)$$

$$\le \frac{1}{2} \sum_{k=1}^{k'-1} \frac{n_k}{n} + \sum_{k=1}^{k_n} \frac{n_k}{n} \mu_{k'}(x^{(k)}) + \sum_{k=1}^{k_n} \frac{n_k}{n} \xi(k).$$

From Lemma 1 it follows that

$$\sum_{k=1}^{k_n} \frac{n_k}{n} \mu_{k'}(x^{(k)}) \le \mu_{k'}(x^n).$$

Since $\xi(k)$ is monotonically decreasing and the lengths of the segments are monotonically increasing it follows that

$$\sum_{k=1}^{k_n} \frac{n_k}{n} \xi(k) \leq \frac{1}{k_n} \sum_{k=1}^{k_n} \xi(k) \triangleq \bar{\xi}(k_n),$$

where by the Cesaro theorem $\lim_{n\to\infty} \bar{\xi}(k_n) = 0$.

Theorem 3. *For all sequences $x^\infty \in \mathcal{X}^\infty$*

$$\hat{\mu}(x^\infty) = \limsup_{n\to\infty} \hat{\mu}(x^n) = \mu(x^\infty) = \pi(x^\infty).$$

In summary we have shown that a sequential Markov predictor whose order is increased from k to $k+1$ after observing at least $n_k = \alpha_k 2^k$ data samples asymptotically achieves the performance of any finite-state predictor.

2.2 General Loss Functions

In this section we present a more general formulation of the prediction problem treated so far and give some references to related work.

It is possible to generalize our problem in the following way. Given is a finite set $\mathcal{B}$ of so called *strategies* and a *loss function* $l : \mathcal{B} \times \mathcal{X} \to \mathbb{R}$. At time t after having observed $x_1, \ldots, x_t$ one has to decide for a strategy, that is, select an element $b_{t+1} \in \mathcal{B}$. Then x_{t+1} is revealed and a loss of $l(b_{t+1}, x_{t+1})$ is incurred. Again the time average $\frac{1}{n}\sum_{t=1}^{n} l(b_t, x_t)$ is tried to be kept small and again it can be defined how good this can be done for a sequence by a finite-state machine.

Examples

1. If we set $\mathcal{B} = \mathcal{X}$ and l to be the Hamming distance then we are back to our original prediction problem.

2. If $\mathcal{B} = (0, 1]$, $\mathcal{X} = \{0, 1\}$ and $l(b, 0) = -\log b$ and $l(b, 1) = -\log(1 - b)$ then we have the lossless coding problem. Here b_{t+1} has the interpretation of the estimated probability of the next letter to be a zero. The time average $\frac{1}{n}\sum_{t=1}^{n} l(b_t, x_t)$ then is the normalized length of a codeword of the sequential Shannon encoder based on the current letter probabilities from the data observed so far. This length can be attained using arithmetic coding techniques.

3. $\mathcal{B} = (0, 1]$, $\mathcal{X} = \{0, 1\}$. A sequential gambling problem can be formulated in this framework in the following way. At round t the player has to divide his capital. The share wagered on the next outcome is then doubled, i.e., if S_t is the player's capital after round t then $S_{t+1} = 2b_{t+1}S_t$ if $x_{t+1} = 0$ and $S_{t+1} = 2(1 - b_{t+1})S_t$ if $x_{t+1} = 1$. If l is as in 2., then the exponential growth rate of the player's capital $\frac{\log S_n}{n}$ is the time average of $1 - l(b_t, x_t)$.

4. There are also continuous alphabet applications. For instance prediction under the mean squared error criterion, i.e., $l(b, x) = (x - b)^2$.

General loss functions in the probabilistic setting were studied in [3]. There it was shown that if the data $x_1, x_2, \ldots$ are generated by a stationary ergodic source which is known and $\mathcal{B}$ consists of any measurable functions of the past $(x_1, x_2, \ldots, x_t)$ then the best strategy in order to minimize the *expected* time average loss is the one that attains the minimal conditional expectation of $l(b_{t+1}, x_{t+1})$ given the past. Furthermore, it was shown that this minimal loss is achievable almost surely under certain regularity conditions on the loss function even if the source is unknown a priori.
In the deterministic setting general loss functions were studied in [10]. Older work was devoted to another, in a way slightly more general problem, the so

called *sequential compound decision problem* which was initiated by Robbins ([14]) and this was further studied by various authors ([4],[5],[13],[12]). In our language the problem is restricted to the case $S = 1$, i.e., the comparison class is only that of all constant predictors or strategies. It is more general because the observer has access only to *noisy* versions of the data $x_1, x_2, \ldots, x_t$.

Chapter 3

Finite-State Identifiability

Now consider for every $x \in \mathcal{X}$ a person x who at time t has to answer the question "Is $x_{t+1} = x$?" .
We start by defining how good a sequence can be identified using a finite-state machine.

Definition 8. A finite-state identification scheme is a triple $(\mathcal{S}, g, f)$ consisting of

$$\mathcal{S} \text{ a set of } S \text{ states,}$$

$$g : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S} \text{ a next-state function,}$$

$$f = (f_0, \ldots, f_{|\mathcal{X}|-1}) : \mathcal{S} \rightarrow \{0,1\}^{|\mathcal{X}|} \text{ a decision rule.}$$

As before we can assume without loss of generality that the initial state is always 1, i.e., $s_0 = 1$.
The interpretation is that $f_x(s_t) = 1$ means that person x predicts that $x_{t+1} = x$ and $f_x(s_t) = 0$ means that person x predicts that $x_{t+1} \neq x$. Applied to some sequence x^n the fraction of errors person x makes is then given by

$$\eta_S(f, g, x^n, x) \triangleq \frac{1}{n} \sum_{t=1}^{n} (1 - f_x(s_{t-1}))\delta_{x_t,x} + f_x(s_{t-1})(1 - \delta_{x_t,x}),$$

where δ is the Kronecker symbol.

For a fixed next-state function g an optimal decision rule f is given by

$$f_x(s) = \begin{cases} 1, & \langle x^n|s,x\rangle > \langle x^n|s\rangle - \langle x^n|s,x\rangle \\ 0, & \langle x^n|s,x\rangle \leq \langle x^n|s\rangle - \langle x^n|s,x\rangle \end{cases}$$

for all $x \in \mathcal{X}$ and $s \in \mathcal{S}$.

If we apply this optimal f to the sequence x^n the fraction of errors person x makes is given by

$$\eta_S(g, x^n, x) \triangleq \frac{1}{n} \sum_{s=1}^{S} \min\{\langle x^n|s,x\rangle, \langle x^n|s\rangle - \langle x^n|s,x\rangle\}.$$

We can now define an average error criterion and a maximal error criterion. Furthermore we can distinguish the case where each person can use its own finite-state machine on the sequence (1. and 2. of Definition 9) and the more restrictive case where the persons have to use one finite-state machine (3. and 4. of Definition 9).

Definition 9. 1. The maximal S-state identifiability of the sequence x^n is given by

$$\eta_S(x^n) \triangleq \max_{x\in\mathcal{X}} \min_{g} \eta_S(g, x^n, x).$$

2. The average S-state identifiability of the sequence x^n is given by

$$\bar{\eta}_S(x^n) \triangleq \frac{1}{|\mathcal{X}|} \sum_{x\in\mathcal{X}} \min_{g} \eta_S(g, x^n, x).$$

3. The strong maximal S-state identifiability of the sequence x^n is given by

$$\eta'_S(x^n) \triangleq \min_{g} \max_{x\in\mathcal{X}} \eta_S(g, x^n, x).$$

4. The strong average S-state identifiability of the sequence x^n is given by

$$\bar{\eta}'_S(x^n) \triangleq \min_{g} \frac{1}{|\mathcal{X}|} \sum_{x\in\mathcal{X}} \eta_S(g, x^n, x).$$

Definition 10. The asymptotic maximal S-state identifiability of the sequence x^∞ is given by

$$\eta_S(x^\infty) \triangleq \limsup_{n\to\infty} \eta_S(x^n).$$

The corresponding values of the asymptotic S-state identifiability are defined analogously in the other cases of Definition 9.

Definition 11. The maximal finite-state identifiability of the sequence x^∞ is given by

$$\eta(x^\infty) \triangleq \lim_{S\to\infty} \eta_S(x^\infty).$$

The corresponding values of the finite-state identifiability are defined analogously in the other cases of Definition 9.

The following relations follow easily from the definitions.

Lemma 2. *For all sequences $x^n \in \mathcal{X}^n$*

$$\eta'_S(x^n) \geq \eta_S(x^n) \geq \bar{\eta}_S(x^n), \tag{3.1}$$

$$\eta'_S(x^n) \geq \bar{\eta}'_S(x^n) \geq \bar{\eta}_S(x^n). \tag{3.2}$$

Remark 3. In the binary case, $\mathcal{X} = \{0,1\}$, we have

$$\eta_S(g, x^n, x) = \frac{1}{n}\sum_{s=1}^{S} \min\{\langle x^n|s, x\rangle, \langle x^n|s, 1-x\rangle\} \tag{3.3}$$

$$= \eta_S(g, x^n, 1-x) = \pi_S(g, x^n) \tag{3.4}$$

and this implies

$$\eta_S(x^n) = \bar{\eta}_S(x^n) = \eta'_S(x^n) = \bar{\eta}'_S(x^n) = \pi_S(x^n). \tag{3.5}$$

Thus, in the binary case identification of sequences gives no advantage over prediction.

3.1 A Universal Identification Scheme

Definition 12. For a sequence $x^n \in \mathcal{X}^n$ and a letter $x \in \mathcal{X}$ let $\mathbf{1}_x x^n \in \{0,1\}^n$ be the sequence with

$$(\mathbf{1}_x x^n)_t = \begin{cases} 1, & \text{if } x_t = x \\ 0, & \text{if } x_t \neq x. \end{cases}$$

Then it holds

$$\eta_1(g, x^n, x) = \frac{1}{n} \min\{\langle x^n|x\rangle, n - \langle x^n|x\rangle\} = \pi_1(\mathbf{1}_x x^n), \tag{3.6}$$

where the argument g of η_1 is the only possible constant next-state function as $S = 1$.

This suggests the following sequential identification scheme. At time t person x applies the predictor analyzed in Theorem 1 to the sequence $\mathbf{1}_x x^t$. If we denote by $\hat{\eta}_1(x^n, x)$ the expected fraction of identification errors person x makes using this scheme for the sequence x^n then Theorem 1 implies that

$$|\eta_1(g, x^n, x) - \hat{\eta}_1(x^n, x)| = O(\frac{1}{\sqrt{n}}) \text{ for all } x \in \mathcal{X}. \tag{3.7}$$

This means we know how to achieve sequentially the 1-state identifiability universally for all sequences.

In the case when $S = 1$ we can actually derive a formula for η_1 in terms of π_1.

Theorem 4. *For all sequences $x^n \in \mathcal{X}^n$ it holds*

$$\eta_1(x^n) = \min\left\{\pi_1(x^n), 1 - \pi_1(x^n)\right\}.$$

Proof: Note that $\eta_1(x^n) = \frac{1}{n} \max_x \min\{\langle x^n|x\rangle, n - \langle x^n|x\rangle\}$ and $\pi_1(x^n) = 1 - \max_x \frac{\langle x^n|x\rangle}{n}$.

Case 1: There exists $\bar{x} \in \mathcal{X}$ with $\langle x^n|\bar{x}\rangle \geq \frac{n}{2}$.

Then $\pi_1(x^n) = 1 - \frac{\langle x^n|\bar{x}\rangle}{n} \leq \frac{1}{2}$ and

$$\eta_1(x^n) = \frac{1}{n}\max\{n - \langle x^n|\bar{x}\rangle, \max_{x\neq\bar{x}}\langle x^n|x\rangle\}$$

$$= \frac{1}{n}(n - \langle x^n|\bar{x}\rangle) = \pi_1(x^n) \leq 1 - \pi_1(x^n),$$

where we used that $n - \langle x^n|\bar{x}\rangle = \sum_{x\neq\bar{x}}\langle x^n|x\rangle$.

Case 2: For all $x \in \mathcal{X}$ $\langle x^n|x\rangle < \frac{n}{2}$.

In this case $\pi_1(x^n) > \frac{1}{2}$ and

$$\eta_1(x^n) = \frac{1}{n}\max_x\langle x^n|x\rangle = 1 - \pi_1(x^n) < \pi_1(x^n).$$

□

If $S > 1$ then η_S is not a function of π_S any longer. Nevertheless it is possible to determine some relations between these quantities and this will be done in the next section.

3.2 Relations between Predictability and Identifiability

Theorem 5. *For all $S \geq 1$, for all sequences $x^n \in \mathcal{X}^n$ and all next-state functions g*

$$\pi_S(x^n, g) \geq \max_{x\in\mathcal{X}} \eta_S(g, x^n, x)$$

and

$$\pi_S(x^n) \geq \eta'_S(x^n).$$

Proof: Let f be the optimal prediction rule for g and x^n. Consider the following decision rule $\tilde{f} : \mathcal{S} \to \{0,1\}^{|\mathcal{X}|}$ with

$$\tilde{f}_x(s) = \begin{cases} 1, & \text{if } f(s) = x \\ 0, & \text{if } f(s) \neq x \end{cases}$$

for all $x \in \mathcal{X}$ and $s \in \mathcal{S}$.

Now observe that if there is no prediction error at some time instant then also no identification error occurs for all persons. As $\tilde{f}$ is not necessarily optimal the first inequality is proved. Let $\tilde{g}$ be a next-state function such that $\pi_S(x^n, \tilde{g}) = \min_g \pi_S(g, x^n)$. Then it holds

$$\pi_S(x^n) = \pi_S(\tilde{g}, x^n) \geq \max_{x \in \mathcal{X}} \eta_S(\tilde{g}, x^n, x) \geq \min_g \max_{x \in \mathcal{X}} \eta_S(g, x^n, x) = \eta'_S(x^n),$$

which is the second inequality.

□

Note that η' is the biggest of all η-quantities.

A converse inequality is obtained by the following theorem.

Theorem 6. *For all $S \geq 1$ and for all sequences $x^n \in \mathcal{X}^n$*

$$\frac{1}{|\mathcal{X}|} \pi_{S^{|\mathcal{X}|}}(x^n) \leq \bar{\eta}_S(x^n).$$

Proof: Let $g_0, \ldots, g_{|\mathcal{X}|-1}$ be the optimal next state functions for person $0, \ldots, |\mathcal{X}| - 1$, respectively. Let $f_0, \ldots, f_{|\mathcal{X}|-1}$ be the corresponding optimal decision rules.

Let $\tilde{\mathcal{S}} = \mathcal{S}^{|\mathcal{X}|}$ and choose

$$\tilde{g} : \tilde{\mathcal{S}} \times \mathcal{X} \to \tilde{\mathcal{S}}$$

such that

$$\tilde{g}(s_0, \ldots, s_{|\mathcal{X}|-1}, x) = (g_0(s_0, x), \ldots, g_{|\mathcal{X}|-1}(s_{|\mathcal{X}|-1}, x))$$

and consider the following prediction rule $\tilde{f} : \tilde{\mathcal{S}} \to \mathcal{X}$

$$\tilde{f}(s_0, \ldots, s_{|\mathcal{X}|-1}) = \begin{cases} x, & \text{if } f_x(s_x) = 1, \text{ if } x \text{ is not unique,} \\ & \text{choose arbitrarily any of these,} \\ & \\ \text{arbitrary}, & \text{if } f_x(s_x) = 0 \text{ for all } x \in \mathcal{X}. \end{cases}$$

Then

$$\pi_{S|\mathcal{X}|}(x^n) \le \pi_{S|\mathcal{X}|}(\tilde{g}, \tilde{f}, x^n) \le \sum_{x\in\mathcal{X}} \min_g \eta_S(g, x^n, x) = |\mathcal{X}|\, \bar{\eta}_S(x^n).$$

□

Note that $\bar{\eta}$ is the smallest of all η-quantities.

Theorem 7. *For all $S \ge 1$ and for all sequences $x^n \in \mathcal{X}^n$*

$$\bar{\eta}'_S(x^n) \le \frac{2}{|\mathcal{X}|}\pi_S(x^n).$$

Proof: For given $S \ge 1$ and $x^n \in \mathcal{X}^n$ let g and f be the optimal next-state function and prediction rule, respectively. Then we can define the following identification rule $\tilde{f} : \mathcal{S} \to \{0, 1\}^{|\mathcal{X}|}$ with

$$\tilde{f}_x(s) = \begin{cases} 1, & \text{if } f(s) = x \\ 0, & \text{if } f(s) \ne x \end{cases}$$

for all $x \in \mathcal{X}$ and $s \in \mathcal{S}$.

Now observe that if at some time instant there is no prediction error induced by the finite-state predictor given by g and f then there will be also no identification error induced by g and $\tilde{f}$. But if g and f produce a prediction error then there will be exactly two persons making an identification error if we use g and $\tilde{f}$. Therefore

$$\begin{aligned} 2\pi_S(x^n) = 2\pi_S(g, f, x^n) &= \sum_{x\in\mathcal{X}} \eta_S(g, \tilde{f}, x^n, x) \\ &\ge \min_g \sum_{x\in\mathcal{X}} \eta_S(g, x^n, x) = |\mathcal{X}|\, \bar{\eta}'_S(x^n). \end{aligned}$$

□

Corollary 1. *For all sequences $x^\infty \in \mathcal{X}^\infty$*

$$\frac{1}{|\mathcal{X}|}\pi(x^\infty) \le \bar{\eta}(x^\infty) \le \frac{2}{|\mathcal{X}|}\pi(x^\infty).$$

Proof: Combining Theorem 6 and 7 and taking the $\limsup_{n\to\infty}$ and the $\lim_{S\to\infty}$ gives the desired result.

$\square$

Corollary 1 characterizes the average identifiability of any sequence in terms of the predictability of that sequence where upper and lower bound differ by a factor of 2.

3.3 Markov Machines for Identification

Similar to Definition 7 in Section 2.1 we now examine a special class of finite-state machines the class of Markov machines.

Definition 13. For any $k \geq 1, x^n \in \mathcal{X}^n, x \in \mathcal{X}$ denote by

$$\mu_k^I(x^n, x) \triangleq \frac{1}{n} \sum_{x^k \in \mathcal{X}^k} \min\{\langle x^n|x^k, x\rangle, \langle x^n|x^k\rangle - \langle x^n|x^k, x\rangle\}$$

the Markov identifiability of order k of the sequence x^n with respect to x. Furthermore let

$$\mu_k^I(x^\infty, x) \triangleq \limsup_{n\to\infty} \mu_k^I(x^n, x),$$

$$\mu^I(x^\infty, x) \triangleq \lim_{k\to\infty} \mu_k^I(x^\infty, x),$$

$$\mu^I(x^\infty) \triangleq \max_{x\in\mathcal{X}} \mu^I(x^\infty, x),$$

$$\bar{\mu}^I(x^\infty) \triangleq \frac{1}{|\mathcal{X}|} \sum_{x\in\mathcal{X}} \mu^I(x^\infty, x).$$

The result of [10, Theorem 2] which was derived for general loss functions and which is similar to Theorem 2 leads in our case to the following proposition.

Proposition 8. *For all $k \geq 1, S \geq 1$ and all sequences $x^n \in \mathcal{X}^n$*

$$\mu_k^I(x^n, x) \leq \min_g \eta_S(g, x^n, x) + \sqrt{\frac{2\ln S}{k+1}}.$$

Theorem 9. *For all sequences $x^\infty \in \mathcal{X}^\infty$ it holds that*

$$\eta(x^\infty) = \eta'(x^\infty),$$

$$\bar{\eta}(x^\infty) = \bar{\eta}'(x^\infty).$$

Proof: Taking in Proposition 8 the limit supremum $n \to \infty$, the limit $k \to \infty$ and the limit $S \to \infty$ it follows that $\mu^I(x^\infty) \leq \eta(x^\infty)$. Therefore

$$\eta'(x^\infty) \geq \eta(x^\infty) \geq \mu^I(x^\infty) = \lim_{k\to\infty} \limsup_{n\to\infty} \underbrace{\max_{x\in\mathcal{X}} \mu_k^I(x^n, x)}_{\geq \eta'_{|\mathcal{X}|^k}(x^n)}$$

$$\geq \lim_{k\to\infty} \limsup_{n\to\infty} \eta'_{|\mathcal{X}|^k}(x^n) = \eta'(x^\infty).$$

□

Remark 4. Only the asymptotic values for $S \to \infty$ of η, η' and $\bar{\eta}, \bar{\eta}'$ coincide. The values of η_S and η'_S do differ in general.
If we compare the definitions of η and η' we see that the difference is the order of min and max. Therefore Theorem 9 can be interpreted that asymptotically we have here a Minimax-Theorem.

3.4 Effects of Randomization

In the theory of identification via channels one discovery was that randomized codes are tremendously superior compared with non-randomized codes whereas in the classical transmission model it doesn't affect the capacity (of the discrete memoryless channel).
In this section we consider randomized finite-state machines, i.e., we replace the next-state function $g : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ by a family

$$\mathcal{G} = \{G(\cdot|s, x) : s \in \mathcal{S}, x \in \mathcal{X}\} \cup G_0$$

of conditional probability distributions $G(\cdot|s,x)$ on $\mathcal{S}$ and an initial probability distribution G_0 on $\mathcal{S}$. The interpretation is that the initial state is chosen according to G_0 and then at each following time instant, if the machine is in state s and letter x occurs, the machine changes its state to s' with probability $G(s'|s,x)$. We consider randomized decision rules f where $f = (f_0, \ldots, f_{|\mathcal{X}|-1}) : \mathcal{S} \to [0,1]^{|\mathcal{X}|}$ with the interpretation that $f_x(s)$ is the probability that person x decides that the next symbol will be equal to x if the machine is in state s. Without loss of generality we can again restrict ourselves to deterministic decision rules, i.e., $f_x(s) = 0$ or 1 for all x and s. In order to see this, suppose we are given $\mathcal{G}$ and x^n. Then let for $t = 0, \ldots, n-1$ S_t be the random variable for the state at time t. The joint distribution of $S_0, \ldots, S_{n-1}$ is uniquely determined by $\mathcal{G}$ and x^n. Then the expected fraction of errors person x will make is given by

$$\eta_S^R(\mathcal{G}, f, x^n, x) \triangleq \frac{1}{n}\sum_{t=1}^{n}\sum_{s\in\mathcal{S}} P_{S_{t-1}}(s)(f_x(s)(1-\delta_{x,x_t}) + (1-f_x(s))\delta_{x,x_t})$$

$$= \frac{1}{n}\sum_{s\in\mathcal{S}}(f_x(s)\sum_{t=1}^{n} P_{S_{t-1}}(s)(1-\delta_{x,x_t}) + (1-f_x(s))\sum_{t=1}^{n} P_{S_{t-1}}(s)\delta_{x,x_t})$$

from which we see that $f_x(s) = 0$ or 1 is always an optimal choice resulting in an expected fraction of errors equal to

$$\eta_S^R(\mathcal{G}, x^n, x) \triangleq \frac{1}{n}\sum_{s\in\mathcal{S}}\min\{\sum_{t=1}^{n} P_{S_{t-1}}(s)(1-\delta_{x,x_t}), \sum_{t=1}^{n} P_{S_{t-1}}(s)\delta_{x,x_t}\}.$$

Definition 14. 1. $\eta_S^R(x^n, x) \triangleq \inf_{\mathcal{G}} \eta_S^R(\mathcal{G}, x^n, x)$,

2. $\eta_S^R(x^n) \triangleq \max_{x\in\mathcal{X}} \eta_S^R(x^n, x)$,

3. $\eta_S'^R(x^n) \triangleq \inf_{\mathcal{G}} \max_{x\in\mathcal{X}} \eta_S^R(\mathcal{G}, x^n, x)$,

4. $\bar{\eta}_S^R(x^n) \triangleq \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \eta_S^R(x^n, x)$,

5. $\bar{\eta}'^{R}_{S}(x^n) \triangleq \inf_{\mathcal{G}} \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \eta^R_S(\mathcal{G}, x^n, x)$.

The asymptotic quantities, $\eta^R_S(x^\infty)$, $\eta^R(x^\infty)$ etc., are defined analogously to Definitions 10 and 11.

Theorem 10. *For all sequences $x^\infty \in \mathcal{X}^\infty$*

$$\eta(x^\infty) = \eta^R(x^\infty) = \eta'^R(x^\infty).$$

Proof: From [10, Theorem 4] we can derive that

$$\mu^I_k(x^n, x) \leq \eta^R_S(x^n, x) + \sqrt{\frac{2\ln S}{k+1}}.$$

Taking the limit supremum as $n \to \infty$ and the limit as $k \to \infty$ and finally the limit $S \to \infty$ we obtain that $\mu^I(x^\infty, x) \leq \eta^R(x^\infty, x)$ and therefore

$$\mu^I(x^\infty) \leq \eta^R(x^\infty).$$

Together with Theorem 9 it follows

$$\eta'(x^\infty) = \eta(x^\infty) = \mu^I(x^\infty) \leq \eta^R(x^\infty) \leq \eta'^R(x^\infty) \leq \eta'(x^\infty).$$

□

Theorem 10 shows that asymptotically randomization does not help here. The reason for this observation lies in the fact that deterministic Markov machines outperform asymptotically, as the number of states increases, any randomized finite-state machine.

Bibliography

[1] Ahlswede R. and Dueck G., *"Identification via channels"*, IEEE Trans. Inform. Theory, vol. 35, no. 1, pp. 15-29, 1989.

[2] Ahlswede R., *"General theory of information transfer"*, Preprintreihe SFB 343 "Diskrete Strukturen in der Mathematik", Nr. 97-118, 1997.

[3] Algoet P.H., *"The strong law of large numbers for sequential decisions under uncertainty"*, IEEE Trans. Inform. Theory, vol. 40, no. 3, pp. 609-633, 1994.

[4] Blackwell D., *"An analog to the minimax theorem for vector payoffs"*, Pac. J. Math., vol. 6, pp. 1-8, 1956.

[5] Blackwell D., *"Controlled random walks"*, Proc. Int. Congr. Mathematicians, 1954, Vol. III, Amsterdam, North Holland, pp.336-338, 1956.

[6] Cover T.M. and Shenhar A., *"Compound Bayes predictors for sequences with apparent Markov structure"*, IEEE Trans. Syst. Man Cybern., vol. SMC-7, pp. 421-424, 1977.

[7] Cover T.M., *"Behavior of sequential predictors of binary sequences"*, Proc. 4th Prague Conf. Inform. Theory, Statistical Decision Functions, Random Processes, 1965, Prague: Publishing House of the Czechoslovak Academy of Sciences, Prague, pp. 263-272, 1967.

[8] Feder M., Merhav N. and Gutman M., *"Universal prediction of individual sequences"*, IEEE Trans. Inform. Theory, vol. 38, no. 4, pp. 1258-1270, 1992.

[9] Feder M., Merhav N. and Gutman M., *"Some properties of sequential predictors for binary Markov sources"*, IEEE Trans. Inform. Theory, vol. 39, no. 3, pp. 887-892, 1993.

[10] Feder M. and Merhav N., *"Universal schemes for sequential decision from individual data sequences"*, IEEE Trans. Inform. Theory, vol. 39, no. 4, pp. 1280-1292, 1993.

[11] Feder M. and Merhav N., *"Relations between entropy and error probability"*, IEEE Trans. Inform. Theory, vol. 40, no. 1, pp. 259-266, 1994.

[12] Hannan J.F. and Robbins H., *"Asymptotic solutions of the compound decision problem for two completely specified distributions"*, Ann. Math. Statist., vol. 26, pp. 37-51, 1957.

[13] Hannan J.F., *"Approximation to Bayes risk in repeated plays"*, Contributions to the Theory of Games, Vol. III, Annals of Mathematics Studies, Princeton, NJ, no. 39, pp. 97-139, 1957.

[14] Robbins H., *"Asymptotically subminimax solutions of compound statistical decision problems"*, in Proc. 2nd Berkeley Symp. Math. Stat. Probab., pp. 131-148, 1951.

[15] C.E. Shannon, *"A mathematical theory of communication"*, Bell System Tech. J., vol. 27, pp. 379-423, pp. 623-656, 1948.

[16] C.E. Shannon, *"Prediction and entropy of printed English"*, Bell Sys. Tech. J., vol. 30, pp. 5-64, 1951.

[17] C.E. Shannon, *"The mind reading machine"*, Bell Laboratories Memorandum, 1953, in *Shannon's Collected Papers*, A.D. Wyner and N.J.A. Sloane Eds., IEEE Press, pp. 688-689, 1993.