Chapter 16

Nonparametric Prediction

László Györfi and Dominik Schäfer

Abstract. In this chapter we consider the prediction of stationary time series for various loss functions: squared loss (as it arises in the regression problem), $0-1$ loss (pattern recognition) and log utility (portfolio selection). The focus is on the construction of universal prediction rules, which are consistent for all possible stationary processes. Such rules can be obtained by combining elementary rules (experts) in a data dependent way.

16.1 Introduction

The problem of prediction of stationary time series arises in numerous fields. It is particularly desirable to construct **universal prediction rules** for the next output of a stationary time series given past observations. These are prediction rules which are asymptotically optimal, but do not require a priori knowledge about the underlying distribution of the time series (Haussler, Kivinen, Warmuth [18], Merhav, Feder [21]).

Depending upon the context of the prediction problem different loss functions are appropriate. The three most important loss functions are the squared loss (for real valued time series, i.e., the regression problem), the $0-1$ loss (for time series taking on values in a finite set, i.e., pattern recognition) and logarithmic utility (for time series of asset returns in portfolio selection).

Prediction rules that are asymptotically optimal can be constructed by combining elementary rules (experts) in a data dependent way. The key idea is simple: Roughly speaking, the worse an expert predicted in the past, the less credible he is, i.e., the less weight he is assigned in current decision taking (Cesa-Bianchi et al. [7], Littlestone, Warmuth [20], Vovk [26], [27], [28], Weinberger, Merhav and Feder [29]). The main purpose of this chapter is to present universal prediction rules with data dependent combination of experts in the three prototypical fields of regression, of pattern recognition and of portfolio selection.

16.2 Prediction for Squared Error

This section is devoted to the problem of sequential prediction of a real valued sequence. Let $y_1, y_2, \ldots$ be a sequence of real numbers, and let $x_1, x_2, \ldots$ be a sequence of d-dimensional vectors. At each time instant $i = 1, 2, \ldots$, the predictor is asked to guess the value of the next outcome y_i with knowledge of the past $(x_1, \ldots, x_i, y_1, \ldots y_{i-1}) = (x_1^i, y_1^{i-1})$. Thus, the predictor's estimate, at time i, is based on the value of (x_1^i, y_1^{i-1}). Formally, the strategy of the predictor is a sequence $g = \{g_i\}_{i=1}^{\infty}$ of decision functions, and the prediction formed at time i is $g_i(x_1^i, y_1^{i-1})$. After n time instants, the *normalized cumulative prediction error* on the string x_1^n, y_1^n is

$$L_n(g) = \frac{1}{n}\sum_{i=1}^{n}(g_i(x_1^i, y_1^{i-1}) - y_i)^2.$$

The main aim is to make $L_n(g)$ small (Haussler, Kivinen, Warmuth [18], Merhav, Feder [21]).

One possible means of prediction is to combine several predictors which will be called experts. Assume there are K experts: $h^{(1)}, \ldots, h^{(K)}$ and the prediction error $L_n(h^{(k)})$ of expert k is available from observation. At time instant $n+1$ we combine the experts according to their past performances. For this, a probability distribution on the set of experts is generated, where a good expert has relatively large weight, then the average of the experts' predictions is taken with respect to this distribution

(Cesa-Bianchi et al. [7], Littlestone, Warmuth [20], Vovk [26], Weinberger, Merhav and Feder [29]).

A "static" variant of this problem is regression estimation. Let Y be a real valued random variable and let X be a d dimensional random vector (observation). The aim of regression analysis is to approximate Y for given X, i.e., to find a function g such that $g(X)$ is "close" to Y. In particular, regression analysis aims to minimize the mean squared error

$$\min_g \mathbf{E}\{(g(X) - Y)^2\}.$$

It is well known that the solution of this minimization problem is given by the regression function

$$m(x) = \mathbf{E}\{Y|X = x\},$$

since for any function g

$$\mathbf{E}\{(g(X) - Y)^2\} = \mathbf{E}\{(m(X) - Y)^2\} + \mathbf{E}\{(m(X) - g(X))^2\}.$$

The second term on the right hand side is the L_2 error of g and will be denoted by $J(g)$:

$$J(g) = \mathbf{E}\{(m(X) - g(X))^2\}.$$

Obviously, the mean square error is close to its minimum if the L_2 error $J(g)$ is close to 0.

For the regression estimation problem we are given data

$$D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

which are i.i.d. copies of (X, Y). On the basis of this data, we want to construct estimates of $m(x)$ of the form

$$m_n(x) = m_n(x, D_n)$$

such that $J(m_n)$ is small, i.e., m_n tends to m for all distributions of (X, Y) with $\mathbf{E}\{Y^2\} < \infty$ (cf. Györfi, Kohler, Krzyżak, Walk [14]).

Stone [24] showed that there are universally consistent regression estimates. He considered local averaging estimates:

$$m_n(x) = \sum_{i=1}^{n} W_{ni}(x; X_1, \dots, X_n) Y_i,$$

where the weights W_{ni} are usually nonnegative and sum up to 1, moreover W_{ni} is "large", if x and X_i are "close" to each other, otherwise W_{ni} is "small". Common local averaging estimators comprise nearest neighbor, partitioning and kernel estimators.

For the k **nearest neighbor estimate**, $W_{ni}(x; X_1, \dots, X_n) = 1/k$, if X_i is one the k nearest neighbors of x from $X_1, \dots, X_n$, otherwise $W_{ni} = 0$. If

$$k_n \to \infty, \quad k_n/n \to 0,$$

then there are various consistency results.

For the **partitioning estimate** we are given a partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2} \dots\}$ of $\mathcal{R}^d$, and set

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K_n(x, X_i)}{\sum_{i=1}^n K_n(x, X_i)},$$

where $K_n(x,u) = \sum_{j=1}^\infty I_{[x \in A_{n,j}, u \in A_{n,j}]}$ (I_A is the indicator of the set A).

The **kernel estimate** is given by

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)},$$

where $h = h_n > 0$ is the bandwidth and K is an integrable function, called kernel, and $K_h(x) = K(x/h)$.

The other important concept for estimating regression functions is the least squares principle. It is based on the simple idea to estimate the L_2 risk of f

$$\mathbf{E}\left\{(f(X) - Y)^2\right\}$$

by the empirical L_2 risk

$$\frac{1}{n}\sum_{j=1}^n |f(X_j) - Y_j|^2, \tag{16.1}$$

and to choose as a regression function estimate a function which minimizes the empirical L_2 risk. More precisely, for **least squares estimates** one first chooses a "suitable" class of functions $\mathcal{F}_n$ (maybe depending on the data, but at least depending on the sample size n) and then selects a function from this class which minimizes the empirical L_2 risk, i.e. one defines the estimate m_n by

$$m_n \in \mathcal{F}_n \quad \text{with} \quad \frac{1}{n}\sum_{j=1}^n |m_n(X_j) - Y_j|^2 = \min_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{j=1}^n |f(X_j) - Y_j|^2. \tag{16.2}$$

The class of candidate functions grows as the sample-size n grows. Examples of possible choices of the set $\mathcal{F}_n$ are sets of piecewise polynomials with respect to a partition $\mathcal{P}_n$ of $\mathcal{R}^d$, or sets of smooth piecewise polynomials (splines).

The other framework in which the need for universal prediction arises is the case of time series where the data $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ are **dependent**. Here we assume long-range dependence, i.e., we assume that the data form a stationary and ergodic process with unknown autocovariance structure.

For given n, the problem is the following minimization:

$$\min_g \mathbf{E}\{(g(X_{n+1}, D_n) - Y_{n+1})^2\}.$$

¿From this one easily verifies that the best predictor is the conditional expectation

$$\mathbf{E}\{Y_{n+1} | X_{n+1}, D_n\}.$$

This, however, cannot be learned from data, i.e., there is no prediction sequence with

$$\lim_{n\to\infty}(g_n(X_{n+1},D_n)-\mathbf{E}\{Y_{n+1}|X_{n+1},D_n\})=0$$

a.s. for all stationary and ergodic sequence (cf., e.g., Györfi et al. [17]).

In general, our aim is to achieve the optimum

$$L^*=\lim_{n\to\infty}\min_g \mathbf{E}\{(g(X_{n+1},D_n)-Y_{n+1})^2\},$$

which again is impossible. However, there are universal Cesáro consistent prediction sequence g_n, i.e.,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(g_i(X_{i+1},D_i)-Y_{i+1})^2=L^*$$

a.s. for all stationary and ergodic sequence. Such prediction sequences are called **universally consistent**. We show how to construct universally consistent predictors by combination of predictor experts.

One of the main ingredients of the construction is the following lemma, whose proof is a straightforward extension of standard arguments in prediction theory of individual sequences, see, for example, Kivinen and Warmuth [19], Singer and Feder [23].

Lemma 1 *Let $\widetilde{h}_1,\widetilde{h}_2,\dots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $\widetilde{h}_i(y_1^{n-1})\in[-B,B]$ and $y_1^n\in[-B,B]^n$. Define*

$$w_{t,k}=q_k e^{-(t-1)L_{t-1}(\widetilde{h}_k)/c}$$

with $c\geq 8B^2$, and the experts' weights by

$$v_{t,k}=\frac{w_{t,k}}{\sum_{i=1}^{\infty}w_{t,i}}.$$

Then the prediction strategy $\widetilde{g}$ defined by

$$\widetilde{g}_t(y_1^{t-1})=\sum_{k=1}^{\infty}v_{t,k}\widetilde{h}_k(y_1^{t-1})\qquad(t=1,2,\dots)$$

has the property that for every $n\geq 1$,

$$L_n(\widetilde{g})\leq\inf_k\left(L_n(\widetilde{h}_k)-\frac{c\ln q_k}{n}\right).$$

Here $-\ln 0$ is treated as ∞.

We return to the problem of stationary and ergodic data $(X_1,Y_1),\dots,(X_n,Y_n)$. Assume that $|Y_0|\leq B$. The elementary predictors (experts) will be denoted by $h^{(k,\ell)}$, $k,\ell=1,2,\dots$. Each of the $h^{(k,\ell)}$ works as follows: Let G_ℓ be a quantizer of $\mathcal{R}^d$ and H_ℓ

be a quantizer of $\mathcal{R}$. For given k, ℓ, let I_n be the set of time instants $k < i < n$, for which a match of the k-length quantized sequences

$$G_\ell(x_{i-k}^i) = G_\ell(x_{n-k}^n)$$

and

$$H_\ell(y_{i-k}^{i-1}) = H_\ell(y_{n-k}^{n-1})$$

occurs. Then the prediction of expert $h^{(k,\ell)}$ is the average of the y_i's for which $i \in I_n$:

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) := \frac{\sum_{i \in I_n} y_i}{|I_n|}.$$

These elementary predictors are not universally consistent since for small k the bias tends to be large and for large k the variance grows considerably because of the few matchings. The same is true for the quantizers. The problem is how to choose k, ℓ in a data dependent way such as to obtain a universally consistent predictor. The solution is the combination of experts.

The combination of predictors can be derived according to the previous lemma. Let $\{q_{k,\ell}\}$ be a probability distribution on the set of all pairs (k, ℓ) of positive integers, and for $c = 8B^2$ put

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/c}$$

and

$$v_{t,k,\ell} = \frac{w_{t,k,\ell}}{\sum_{i,j=1}^\infty w_{t,i,j}} .$$

Then for the combined prediction rule

$$g_t(x_1^t, y_1^{t-1}) = \sum_{k,\ell=1}^\infty v_{t,k,\ell} h^{(k,\ell)}(x_1^t, y_1^{t-1})$$

the following universal consistency result holds:

Theorem 1 *(Györfi, Lugosi [15]). If the quantizers G_ℓ and H_ℓ "are asymptotically fine", and $\mathbf{P}\{Y_i \in [-B, B]\} = 1$, then the combined predictor g is universally consistent.*

16.3 Prediction for $0-1$ Loss: Pattern Recognition

In pattern recognition y_i takes on values in the finite set $\{1, 2, \ldots M\}$. At time instant i the classifier (predictor) decides on y_i based on the past observation (x_1^i, y_1^{i-1}).

After n rounds the empirical error for x_1^n, y_1^n is

$$L_n(g) = \frac{1}{n} \sum_{i=1}^n I_{\{g(x_1^i, y_1^{i-1}) \neq y_i\}}.$$

The natural loss is given by the $0-1$ loss, and $L_n(g)$ is the relative frequency of errors.

In the "static" version of pattern recognition the random variable Y takes on values in $\{1, 2, \dots M\}$, and based on the random observation vector X one has to decide on Y. The decision rule (classifier) is defined by a decision function

$$g : \mathcal{R}^d \to \{1, 2, \dots M\}.$$

The classifier has an error probability

$$L(g) = \mathbf{P}\{g(X) \neq Y\}.$$

As is well known, the error probability is minimized by the Bayes decision,

$$g^*(x) = i, \quad \text{if } \mathbf{P}\{Y = i|X = x\} = \max_j \mathbf{P}\{Y = j|X = x\}.$$

In pattern recognition we want to approach the Bayes decision if data $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ are given, which are i.i.d. copies of (X, Y). It is of considerable interest to find a pattern recognition rule

$$g_n(x) = g_n(x, D_n)$$

such that

$$L(g_n) = \mathbf{P}\{g_n(X) \neq Y|D_n\}$$

is close to $L(g^*)$ for all possible distributions of (X, Y). Similarly to the regression estimation problem, this may be achieved (cf. Devroye, Györfi and Lugosi [12]).

Clearly, this should be generalized to the case of dependent data $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, where the data form a stationary and ergodic process. For given n, the problem is the following minimization:

$$\min_g \mathbf{P}\{g(X_{n+1}, D_n) \neq Y_{n+1}\},$$

which –as in the general regression estimation case– cannot be learned from data. Nor can there be a strategy achieving the optimum

$$R^* = \lim_{n\to\infty} \min_g \mathbf{P}\{g(X_{n+1}, D_n) \neq Y_{n+1}\}.$$

However, there are universal Cesáro consistent classifier sequences $g = \{g_n\}$, i.e., for the notation

$$L_n(g) = \frac{1}{n}\sum_{i=1}^{n} I_{\{g_i(X_{i+1}, D_i) \neq Y_{i+1}\}}$$

there exists g such that

$$\lim_{n\to\infty} L_n(g) = R^*$$

a.s. for all stationary and ergodic sequence. Such classifier sequences are called **universally consistent**. Györfi, Lugosi and Morvai [16] have constructed universally consistent classifiers by randomized combination of classifiers (experts).

The main ingredient of the proof is a beautiful result of Cesa-Bianchi et al. [7]. It states that, given a set of N experts, and a sequence of fixed length n, there exists a randomized predictor whose number of mistakes is not greater than that of the best classifier plus $\sqrt{(n/2)\ln N}$ for *all possible* sequences y_1^n. The simpler algorithm and statement cited below is due to Cesa-Bianchi [6]:

Lemma 2 *Let $\widetilde{h}^{(1)}, \dots, \widetilde{h}^{(N)}$ be a finite collection of classifier strategies (experts). The classifier strategy $\widetilde{g}$ is defined by*

$$\widetilde{g}_t(y_1^{t-1}, x_1^t, u) = \begin{cases} 0 & \text{if } u > \dfrac{\sum_{k=1}^N I_{\{\widetilde{h}^{(k)}(y_1^{t-1}, x_1^t)=1\}} \widetilde{w}_t(k)}{\sum_{k=1}^N \widetilde{w}_t(k)} \\ 1 & \text{otherwise,} \end{cases}$$

$(t = 1, 2, \dots, n)$, where for all $k = 1, \dots, N$ and $t > 1$

$$\widetilde{w}_1(k) = 1 \quad \text{and} \quad \widetilde{w}_t(k) = e^{-\sqrt{8 \ln N / n} L_{t-1}(\widetilde{h}^{(k)})}.$$

Let $U_1, U_2, \dots$ be i.i.d. uniformly distributed random variables on $[0, 1]$. Then at time moment t the randomized classification is

$$\widetilde{g}_t(y_1^{t-1}, x_1^t, U_t)$$

and for any $y_1^n \in \{0, 1\}^n$ and $x_1^n \in \mathcal{R}^{nd}$

$$\mathbf{E} L_n(\widetilde{g}) \le \min_{k=1,\dots,N} L_n(\widetilde{h}^{(k)}) + \sqrt{\frac{\ln N}{2n}}.$$

16.4 Prediction for Log Utility: Portfolio Selection

Consider investment in the stock market. We follow Breiman [5], Algoet and Cover [3], Cover [9] and Cover and Thomas [11]. The market consists of d stocks, and during one investment period (e.g., one day), it will be described by a return vector $x = (x^{(1)}, \dots x^{(d)})$, where the j-th component $x^{(j)}$ is the factor by which capital invested in stock j grows during the market period. The investor is allowed to diversify his capital at the beginning of each day of trading according to a portfolio vector $b = (b^{(1)}, \dots b^{(d)})$, the j-th component $b^{(j)}$ of which gives the proportion of the investor's capital invested in stock j. Assume that the portfolio vector $b = (b^{(1)}, \dots b^{(d)})$ is a probability distribution, i.e. consumption of capital and short selling of stocks are excluded. If S_0 denotes the initial capital, then at the end of the day the investor will be left with a wealth of

$$S_1 = S_0 \sum_{j=1}^d b^{(j)} x^{(j)} = S_0 (b, x),$$

where $(\cdot, \cdot)$ stands for the inner product.

For long term investment, assume the investor starts with an initial capital S_0 and let x_i be the return vector on day i. If $b = b_1$ is the portfolio vector the investor chooses for the first day of trading, he will accumulate a wealth of

$$S_1 = S_0 \cdot (b_1, x_1)$$

by the end of this day. For the second day, S_1 becomes his new initial capital and the portfolio vector for day two, b_2, may depend on x_1: $b_2 = b(x_1)$. Then

$$S_2 = S_0 \cdot (b_1, x_1) \cdot (b_2, x_2) = S_0 \cdot (b, x_1) \cdot (b(x_1), x_2).$$

In general, after the nth day of trading using a nonanticipating portfolio strategy $b_i = b(x_1^{i-1})$ $(i = 1, ..., n)$ the investor achieves

$$S_n = S_0 \prod_{i=1}^{n} (b(x_1^{i-1}), x_i) = S_0 e^{\sum_{i=1}^{n} \log(b(x_1^{i-1}), x_i)} = S_0 e^{nW_n(B)}.$$

The portfolio strategy $B = \{b(x_1^{i-1})\}$ is a sequence of functions, the quality of which is characterized by the average growth rate

$$W_n(B) = \frac{1}{n} \sum_{i=1}^{n} \log(b(x_1^{i-1}), x_i).$$

Obviously, the maximization of $S_n = S_n(B)$ and the maximization of $W_n(B)$ are equivalent.

Throughout, we assume that $x_1, x_2, \ldots$ are realizations of the random vectors X_1, X_2, $\ldots$ drawn from the vector valued stationary and ergodic process $\{X_n\}_{-\infty}^{\infty}$ (note that by Kolmogorov's Theorem any stationary and ergodic process $\{X_n\}_1^{\infty}$ can be extended to a bi-infinite stationary process on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$, such that ergodicity holds for both, $n \to \infty$ and $n \to -\infty$).

The fundamental limits for investment are delineated by results of Algoet and Cover [3], Algoet [1, 2], who showed that the so called log-optimum portfolio $B^* = \{b^*(\cdot)\}$ is the best possible choice. More precisely, on day n let $b^*(\cdot)$ be such that

$$\mathbf{E}\{\log(b^*(X_1^{n-1}), X_n) | X_1^{n-1}\} = \mathbf{E}\{\max_{b(\cdot)} \log(b(X_1^{n-1}), X_n) | X_1^{n-1}\}.$$

If $S_n^* = S_n(B^*)$ denotes the capital after day n achieved by a log-optimum portfolio strategy B^*, then for any portfolio strategy B with capital $S_n = S_n(B)$ and for any stationary ergodic process $\{X_n\}_{-\infty}^{\infty}$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{S_n}{S_n^*} \le 0 \quad \text{almost surely}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \log S_n^* = W^* \quad \text{almost surely,}$$

where

$$W^* = \mathbf{E}\left\{ \max_{b(\cdot)} \mathbf{E}\{\log(b(X_{-\infty}^{-1}), X_0) | X_{-\infty}^{-1}\} \right\}$$

is the maximal growth rate of any portfolio.

These limit relations give rise to the following definition:

Definition 1 *A portfolio strategy B is called* **universal with respect to a class $\mathcal{C}$ of stationary and ergodic processes** $\{X_n\}_{-\infty}^{\infty}$, *if for each process in the class,*

$$\lim_{n\to\infty} \frac{1}{n} \log S_n(B) = W^* \quad \textit{almost surely.}$$

Universal strategies asymptotically achieve the best possible growth rate for all ergodic processes in the class. Algoet [1] introduced two portfolio strategies, and proved that the more complicated one is universal. The purpose of this section is to prove the universality of a strategy B similar to his second portfolio.

B is constructed as follows. We first define an infinite array of elementary portfolios $H^{(k,\ell)} = \{h^{(k,\ell)}(\cdot)\}$, $k,\ell = 1,2,\ldots$ To this end, let $\mathcal{P}_\ell = \{A_{\ell,j}, j = 1,2,\ldots,m_\ell\}$ be a sequence of finite partitions of the feature space $\mathcal{R}^d$, and let G_ℓ be the corresponding quantizer:

$$G_\ell(x) = j, \text{ if } x \in A_{\ell,j}.$$

With some abuse of notation, for any n and $x_1^n \in \mathcal{R}^{dn}$, we write $G_\ell(x_1^n)$ for the sequence $G_\ell(x_1),\ldots,G_\ell(x_n)$. Now, fix positive integers k,ℓ, and for each k-long string s of positive integers, define the partitioning portfolio

$$b^{(k,\ell)}(x_1^{n-1},s) = \arg\max_b \prod_{\{k<i<n: G_\ell(x_{i-k}^{i-1})=s\}} (b,x_i), \qquad n > k+1,$$

if the product is nonvoid, and uniform b otherwise. If the product is nonvoid then

$$b^{(k,\ell)}(x_1^{n-1},s) = \arg\max_b \frac{\sum_{\{k<i<n: G_\ell(x_{i-k}^{i-1})=s\}} \log(b,x_i)}{\left|\{k<i<n: G_\ell(x_{i-k}^{i-1})=s\}\right|}, \qquad n > k+1.$$

¿From this we define the elementary portfolio $h^{(k,\ell)}$ by

$$h^{(k,\ell)}(x_1^{n-1}) = b^{(k,\ell)}(x_1^{n-1}, G_\ell(x_{n-k}^{n-1})), \qquad n = 1,2,\ldots$$

That is, $h_n^{(k,\ell)}$ quantizes the sequence x_1^{n-1} according to the partition $\mathcal{P}_\ell$, and browses through all past appearances of the last seen quantized string $G_\ell(x_{n-k}^{n-1})$ of length k. Then it designs a fixed portfolio vector according to the returns on the days following the occurence of the string.

Finally, let $\{q_{k,\ell}\}$ be a probability distribution on the set of all pairs (k,ℓ) of positive integers such that for all k,ℓ, $q_{k,\ell} > 0$. The strategy B then arises from weighing the elementary portfolio strategies $H^{(k,\ell)}$ according to their past performances and $\{q_{k,\ell}\}$:

$$b(x_1^{n-1}) := \frac{\sum_{k,\ell} q_{kl} S_{n-1}(H^{(k,\ell)}) h^{(k,\ell)}(x_1^{n-1})}{\sum_{k,\ell} q_{kl} S_{n-1}(H^{(k,\ell)})},$$

where $S_n(H^{(k,\ell)})$ is the capital accumulated after n days when using the portfolio strategy $H^{(k,\ell)}$ with initial capital S_0. Thus, after day n, the investor's capital becomes

$$S_n(B) = \sum_{k,\ell} q_{k,\ell} S_n(H^{(k,\ell)}).$$

The strategy B asymptotically achieves the best possible growth rate of wealth:

Theorem 2 *Assume that*

(a) the sequence of partitions is nested, that is, any cell of $\mathcal{P}_{\ell+1}$ is a subset of a cell of $\mathcal{P}_\ell$, $\ell = 1,2,\ldots$;

(b) if $diam(A) = \sup_{x,y \in A} \|x - y\|$ denotes the diameter of a set, then for any sphere S centered at the origin

$$\lim_{\ell \to \infty} \max_{j: A_{\ell,j} \cap S \neq \emptyset} diam(A_{\ell,j}) = 0 \ .$$

Then the portfolio scheme B defined above is universal with respect to the class of all ergodic processes such that $\mathbf{E}\{|\log X^{(j)}|\} < \infty$, for $j = 1, 2, \ldots d$.

The first tool in the proof of Theorem 2 is known as Breiman's generalized ergodic theorem [4, 5], see also Algoet [2].

Lemma 3 (Breiman, [4]). *Let $Z = \{Z_i\}_{-\infty}^{\infty}$ be a stationary and ergodic process. Let T denote the left shift operator, shifting any sequence $\{..., z_{-1}, z_0, z_1, ...\}$ one digit to the left. Let f_i be a sequence of real-valued functions such that for some function f, $f_i(Z) \to f(Z)$ almost surely. Assume that $\mathbf{E} \sup_i |f_i(Z)| < \infty$. Then*

$$\lim_{t \to \infty} \frac{1}{n} \sum_{i=1}^{n} f_i(T^i Z) = \mathbf{E} f(Z) \qquad a.s.$$

The second tool is a theorem due to Algoet and Cover ([3], Theorems 3 and 4).

Theorem 3 *(*Algoet and Cover, [3]).
(a) Let $\mathbf{Q}_{n \in \mathcal{N} \cup \{\infty\}}$ be a family of regular probability distributions on $(0, \infty)^d$ such that $\mathbf{E}\{|\log U_n^{(j)}|\} < \infty$ for any coordinate of a return vector $U_n = (U_n^{(1)}, ..., U_n^{(d)})$ distributed according to $\mathbf{Q}_n$. In addition, let $B^*(\mathbf{Q}_n)$ be the set of all log-optimal portfolios w.r.t. $\mathbf{Q}_n$, i.e. of all portfolios b that attain $\max_b \mathbf{E}\{\log(b, U_n)\}$. Consider an arbitrary sequence $b_n \in B^*(\mathbf{Q}_n)$. If

$$\mathbf{Q}_n \to \mathbf{Q}_\infty \quad \text{weakly as } n \to \infty$$

then, for $\mathbf{Q}_\infty$-almost all u,

$$(b_n, u) \to (b^*, u) \quad (n \to \infty)$$

where the right hand side is constant as b^* ranges over $B^*(\mathbf{Q}_\infty)$.

(b) Let X be a return vector on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ satisfying $\mathbf{E}\{|\log X^{(j)}|\} < \infty$. If $\mathcal{F}_k$ is an increasing sequence of sub-σ-fields of $\mathcal{F}$,

$$\mathcal{F}_k \nearrow \mathcal{F}_\infty \subseteq \mathcal{F},$$

then

$$\mathbf{E}\left\{ \max_{b \ \mathcal{F}_k-measurable} \mathbf{E}[\log(b, X) | \mathcal{F}_k] \right\} \nearrow \mathbf{E}\left\{ \max_{b \ \mathcal{F}_\infty-measurable} \mathbf{E}[\log(b, X) | \mathcal{F}_\infty] \right\}$$

as $k \to \infty$.

Proof of Theorem 2. We have to prove that

$$\liminf_{n\to\infty} W_n(B) = \liminf_{n\to\infty} \frac{1}{n}\log S_n(B) \geq W^* \quad \text{a.s.}$$

W.l.o.g. we may assume $S_0 = 1$, so that

$$\begin{aligned} W_n(B) &= \frac{1}{n}\log S_n(B) \\ &= \frac{1}{n}\log\left(\sum_{k,\ell} q_{k,\ell} S_n(H^{(k,\ell)})\right) \\ &\geq \frac{1}{n}\log\left(\sup_{k,\ell} q_{k,\ell} S_n(H^{(k,\ell)})\right) \\ &= \frac{1}{n}\sup_{k,\ell}\left(\log q_{k,\ell} + \log S_n(H^{(k,\ell)})\right) \\ &= \sup_{k,\ell}\left(W_n(H^{(k,\ell)}) + \frac{\log q_{k,\ell}}{n}\right). \end{aligned}$$

Thus

$$\begin{aligned} \liminf_{n\to\infty} W_n(H) &\geq \liminf_{n\to\infty}\sup_{k,\ell}\left(W_n(H^{(k,\ell)}) + \frac{\log q_{k,\ell}}{n}\right) \\ &\geq \sup_{k,\ell}\liminf_{n\to\infty}\left(W_n(H^{(k,\ell)}) + \frac{\log q_{k,\ell}}{n}\right) \\ &\geq \sup_{k,\ell}\liminf_{n\to\infty} W_n(H^{(k,\ell)}). \end{aligned} \tag{16.3}$$

In order to evaluate the lim inf on the right hand side we investigate the performance of the $b^{(k,\ell)}(\cdot,\cdot)$ on the stationary and ergodic sequence $X_0, X_{-1}, X_{-2}, \ldots$ First let k, ℓ and s be fixed. $\mathbf{P}^{(k,\ell)}_{j,s}$ denotes the (random) measure concentrated on $\{X_i : 1-j+k \leq i \leq 0, G_\ell(X_{i-k}^{i-1}) = s\}$ with

$$\mathbf{P}^{(k,\ell)}_{j,s}(A) := \frac{\sum_{i:1-j+k\leq i\leq 0, G_\ell(X_{i-k}^{i-1})=s} I_A(X_i)}{|\{i : 1-j+k \leq i \leq 0, G_\ell(X_{i-k}^{i-1}) = s\}|}.$$

If the above set of X_i's is void, then let $\mathbf{P}^{(k,\ell)}_{j,s} := \delta_{(1,\ldots,1)}$ be the probability measure concentrated on $(1, \ldots, 1)$.

Observe that for all s with probability one

$$\mathbf{P}^{(k,\ell)}_{j,s} \to \begin{cases} \mathbf{P}_{X_0|G_\ell(X_{-k}^{-1})=s} & \text{if } \mathbf{P}(G_\ell(X_{-k}^{-1}) = s) > 0, \\ \delta_{(1,\ldots,1)} & \text{if } \mathbf{P}(G_\ell(X_{-k}^{-1}) = s) = 0 \end{cases} \tag{16.4}$$

weakly as $j \to \infty$. Indeed, let f be a bounded continuous function. By the ergodic theorem:

$$\begin{aligned} \int f(x)\mathbf{P}^{(k,\ell)}_{j,s}(dx) &= \frac{\frac{1}{|1-j+k|}\sum_{i:1-j+k\leq i\leq 0, G_\ell(X_{i-k}^{i-1})=s} f(X_i)}{\frac{1}{|1-j+k|}|\{i : 1-j+k \leq i \leq 0, G_\ell(X_{i-k}^{i-1}) = s\}|} \\ &\to \frac{\mathbf{E}\{f(X_0) I_{\{G_\ell(X_{-k}^{-1})=s\}}\}}{\mathbf{P}\{G_\ell(X_{-k}^{-1}) = s\}} \\ &= \mathbf{E}\{f(X_0)|G_\ell(X_{-k}^{-1}) = s\} \\ &= \int f(x)\mathbf{P}_{X_0|G_\ell(X_{-k}^{-1})=s}(dx) \quad \text{a.s.,} \end{aligned}$$

if $\mathbf{P}(G_\ell(X_{-k}^{-1}) = s) > 0$. If $\mathbf{P}(G_\ell(X_{-k}^{-1}) = s) = 0$, then with probability one $\mathbf{P}_{j,s}^{(k,\ell)}$ is concentrated on $(1, ..., 1)$ for all j, and

$$\int f(x)\mathbf{P}_{j,s}^{(k,\ell)}(dx) \quad = \quad f(1, ..., 1).$$

By definition, $b^{(k,\ell)}(X_{1-j}^{-1}, s)$ is a log-optimal portfolio w.r.t. $\mathbf{P}_{j,s}^{(k,\ell)}$. Let $b_{k,\ell}^*(s)$ be a log-optimal portfolio w.r.t. the limit distribution of $\mathbf{P}_{j,s}^{(k,\ell)}$. Then, using Theorem 3(a), we infer from (16.4) that as j tends to infinity the almost surely the following convergence holds:

$$(b^{(k,\ell)}(X_{1-j}^{-1}, s), x_0) \to (b_{k,\ell}^*(s), x_0))$$

for $\mathbf{P}_{X_0|G_\ell(X_{-k}^{-1})=s}$- and hence $\mathbf{P}_{X_0}$-almost all values of x_0. Here and in the following we exploit the fact that there are only finitely many values of s to be considered. In particular, we obtain

$$(b^{(k,\ell)}(X_{1-j}^{-1}, G_\ell(X_{-k}^{-1})), X_0) \to (b_{k,\ell}^*(G_\ell(X_{-k}^{-1})), X_0) \quad \text{a.s.} \tag{16.5}$$

as $j \to \infty$.

We are now in a position to apply Lemma 3. For $x = (..., x_{-1}, x_0, x_1, ...)$, set

$$f_i(x) := \log(h^{(k,\ell)}(x_{1-i}^{-1}), X_0) = \log(b^{(k,\ell)}(x_{1-i}^{-1}, G_\ell(X_{-k}^{-1})), X_0).$$

Note that

$$f_i(X) = |\log(h^{(k,\ell)}(X_{1-i}^{-1}), X_0)| \leq \sum_{j=1}^{d} |\log X_0^{(j)}|,$$

the right hand side of which has finite expectation, and

$$f_i(X) \to (b_{k,\ell}^*(G_\ell(X_{-k}^{-1})), X_0) \quad \text{a.s. as } i \to \infty$$

from (16.5). As $n \to \infty$, Lemma 3 yields

$$\begin{aligned} W_n(H^{(k,\ell)}) \quad &= \quad \frac{1}{n}\sum_{i=1}^{n} \log(h^{(k,\ell)}(X_1^{i-1}), X_i) \\ &\to \quad \mathbf{E}\{\log(b_{k,\ell}^*(G_\ell(X_{-k}^{-1})), X_0)\} \\ &= \quad \mathbf{E}\{\max_{b(\cdot)} \mathbf{E}\{\log(b(G_\ell(X_{-k}^{-1})), X_0)|G_\ell(X_{-k}^{-1})\}\} \\ &= \quad \epsilon_{k,\ell} \quad \text{a.s.} \end{aligned}$$

Therefore, by virtue of (16.3)

$$\liminf_{n\to\infty} W_n(B) \geq \sup_{k,\ell} \epsilon_{k,\ell} \qquad \text{a.s.}$$

Since the partitions $\mathcal{P}_\ell$ are nested, we have $\sigma(G_\ell(X_{-k}^{-1})) \subseteq \sigma(G_{\ell'}(X_{-k'}^{-1}))$ for all $\ell' \geq \ell, k' \geq k$, and the sequence

$$\begin{aligned} &\max_{b(\cdot)} \mathbf{E}\{\log(b(G_\ell(X_{-k}^{-1})), X_0)|G_\ell(X_{-k}^{-1})\} \\ &\quad = \max_{b \text{ is } \sigma(G_\ell(X_{-k}^{-1}))-\text{measurable}} \mathbf{E}\{\log(b, X_0)|G_\ell(X_{-k}^{-1})\} \end{aligned}$$

becomes a sub-martingale indexed by the pair (k,ℓ). This sequence is bounded by

$$\max_{b(\cdot)} \mathbf{E}\{\log(b(X_{-\infty}^{-1}), X_0))|X_{-\infty}^{-1}\},$$

which has finite expectation. The sub-martingale convergence theorem (see, e.g., Stout (1974) implies that this sub-martingale is convergent a.s., and $\sup_{k,\ell} \epsilon_{k,\ell}$ is finite. In particular, by the submartingale property, $\epsilon_{k,\ell}$ is a bounded double sequence increasing in k and ℓ, so that

$$\sup_{k,\ell} \epsilon_{k,\ell} = \lim_{k\to\infty} \lim_{\ell\to\infty} \epsilon_{k,\ell}.$$

Assumption (b) for the sequence of partitions implies that for fixed k

$$\sigma(G_\ell(X_{-k}^{-1})) \nearrow \sigma(X_{-k}^{-1})$$

as $\ell \to \infty$. Hence, by Theorem 3(b)

$$\begin{aligned}
\lim_{l\to\infty} \epsilon_{k,\ell} &= \lim_{l\to\infty} \mathbf{E}\left\{ \max_{b \text{ is } \sigma(G_\ell(X_{-k}^{-1}))-\text{measurable}} \mathbf{E}\{\log(b, X_0)|G_\ell(X_{-k}^{-1})\}\right\} \\
&= \mathbf{E}\left\{ \max_{b \text{ is } \sigma(X_{-k}^{-1})-\text{measurable}} \mathbf{E}\{\log(b, X_0)|X_{-k}^{-1}\}\right\}.
\end{aligned}$$

Applying Theorem 3(b) again with

$$\sigma(X_{-k}^{-1}) \nearrow \sigma(X_{-\infty}^{-1}) \quad \text{as } k \to \infty$$

finally yields

$$\begin{aligned}
\sup_{k,\ell} \epsilon_{k,\ell} &= \lim_{k\to\infty} \mathbf{E}\left\{ \max_{b \text{ is } \sigma(X_{-k}^{-1})-\text{measurable}} \mathbf{E}\{\log(b, X_0)|X_{-k}^{-1}\}\right\} \\
&= \mathbf{E}\left\{ \max_{b \text{ is } \sigma(X_{-\infty}^{-1})-\text{measurable}} \mathbf{E}\{\log(b, X_0)|X_{-\infty}^{-1}\}\right\} \\
&= \mathbf{E}\left\{ \max_{b(\cdot)} \mathbf{E}\{\log(b(X_{-\infty}^{-1}), X_0)|X_{-\infty}^{-1}\}\right\} \\
&= W^*
\end{aligned}$$

and the proof of the theorem is finished. □

Bibliography

[1] P. Algoet, Universal schemes for prediction, gambling, and portfolio selection, *Annals of Probability* **20** (1992) 901–941.

[2] P. Algoet, The strong law of large numbers for sequential decisions under uncertainty, *IEEE Transactions on Information Theory* **40** (1994) 609–634.

[3] P. Algoet, T.M. Cover, Asymptotic optimality asymptotic equipartition properties of log-optimum investments, *Annals of Probability* **16** (1998) 876–898.

[4] L. Breiman, The individual ergodic theorem of information theory, *Annals of Mathematical Statistics* **31** (1957) 809–811. Correction. *Annals of Mathematical Statistics* **31** (1960) 809–810.

[5] L. Breiman, Optimal gambling systems for favorable games, *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, Univ. California Press, Berkeley 1 (1961) 65–78.

[6] N. Cesa-Bianchi, Analysis of two gradient-based algorithms for on-line regression, in *Proc. 10th Ann. Conf. Computational Learning Theory*, New York ACM Press (1997) 163–170.

[7] N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, D. Haussler, R. Schapire, M.K. Warmuth, How to use expert advice, *Journal of the ACM* **44**(3) (1997) 427–485.

[8] Y.S. Chow, Local convergence of martingales and the law of large numbers, *Annals of Mathematical Statistics* **36** (1965) 552–558.

[9] T. Cover, Universal Portfolios, *Mathematical Finance* **1** (1991) 1–29.

[10] T. Cover, E. Ordentlich, Universal Portfolios with Side Information, *IEEE Transactions on Information Theory* **42** (1996) 348–363.

[11] T. Cover, J. Thomas, *Elements of Information Theory,* John Wiley and Sons (1991).

[12] L. Devroye, L. Györfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer Verlag (1996).

[13] M. Feder, N. Merhav, M. Gutman, Universal prediction of individual sequences, *IEEE Transactions on Information Theory* **38** (1992) 1258–1270.

[14] L. Györfi, M. Kohler, A. Krzyżak, H. Walk, *A Distribution-Free Theory of Nonparametric Regression,* Springer Verlag (2002).

[15] L. Györfi, G. Lugosi, Strategies for sequential prediction of stationary time series, in *Modeling Uncertainty: An Examination of its Theory, Methods and Applications*, M. Dror, P. L'Ecuyer, F. Szidarovszky (Eds.), Kluwer Academic Publisher (2001).

[16] L. Györfi, G. Lugosi. G. Morvai, A simple randomized algorithm for consistent sequential prediction of ergodic time series, *IEEE Transactions on Information Theory* **45** (1999) 2642–2650.

[17] L. Györfi, G. Morvai, S. Yakowitz, Limits to consistent on-line forecasting for ergodic time series, *IEEE Transactions on Information Theory* **44** (1998) 886–892.

[18] D. Haussler, J. Kivinen, M. Warmuth, Sequential Prediction of Individual Sequences Under General Loss Functions, *IEEE Transactions on Information Theory* **44** (1998) 1906–1925.

[19] J. Kivinen, M.K. Warmuth, Averaging expert predictions, In H. U. Simon P. Fischer, editor, *Computational Learning Theory: Proceedings of the Fourth European Conference, EuroCOLT'99*, Springer, Berlin. Lecture Notes in Artificial Intelligence 1572 (1999) 153–167.

[20] N. Littlestone, M.K. Warmuth, The weighted majority algorithm, *Information and Computation* **108** (1994) 212–261.

[21] N. Merhav, M. Feder, Universal prediction, *IEEE Transactions on Information Theory* **44** (1998) 2124–2147.

[22] M. Opper, D. Haussler, Worst Case Prediction over Sequences under Log Loss, In: *The Mathematics of Information Coding, Extraction, and Distribution*, Springer Verlag (1997).

[23] A. Singer, M. Feder, Universal linear prediction by model order weighting, *IEEE Transactions on Signal Processing* **47** (1999) 2685–2699.

[24] C.J. Stone, Consistent nonparametric regression, *Annals of Statistics* **5** (1977) 595–645.

[25] W.F. Stout, *Almost sure convergence*, Academic Press, New York (1974).

[26] V.G. Vovk, Aggregating strategies, In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, Association of Computing Machinery, New York (1990) 372–383.

[27] V.G. Vovk, A Game of Prediction with Expert Advice, *Journal of Computer and System Sciences* **56**(2) (1998) 153–173.

[28] V.G. Vovk, Competitive on-line statistics, *International Statistical Review* **69**(2) (2001) 213–248.

[29] M. Weinberger, N. Merhav, M. Feder, Optimal Sequential Probability Assignment for Individual Sequences, *IEEE Transactions on Information Theory* **40** (1994) 384–396.