Stochastic complexity, channel capacity, and universal portfolio

Takeuchi, Jun'ichi
Faculty of Informatics, Kyushu University

出版情報：Journal of Math-for-Industry (JMI). 2 (B), pp.213-225, 2010-10-04. 九州大学大学院数理学研究院
バージョン：
権利関係：

Journal of Math-for-industry, Vol.2(2010B-10), pp.213–225

Math-for-industry
Education & Research Hub

Stochastic complexity, channel capacity, and universal portfolio

Jun'ichi Takeuchi

Received on September 19, 2010

Abstract. We pick up and discus three topics from information theory and learning theory: stochastic complexity, communication channel capacity, and portfolio theory in finance. At first glance, they seem very different ones, but they have common game theoretic profiles. The purpose of this article is to present brief introductions to each problem and describe the relation between them.

Keywords. stochastic complexity, minimax problem, channel capacity, universal portfolio

1. Introduction

In information theory and learning theory, many problems such as data compression, sequential prediction, determination of communication channel capacity, and portfolio theory, can be discussed in similar game theoretic manners. It is well known that the notion of the stochastic complexity (SC), which is a central notion of the minimum description length (MDL) principle, is closely related to the minimax code with coding regret and redundancy in universal data compression. Since code length is translated into cumulative logarithmic loss in sequential prediction, such minimax codes are equivalent to minimax prediction strategies. Moreover, such prediction strategies can be used in portfolio theory in finance. Since the minimax redundancy is equal to the maximin redundancy, which is translated as the channel capacity in communication, the minimax redundancy is equal to the channel capacity. In this paper, we review the formulation of SC and some recent results. Then we discuss determination of channel capacity and universal portfolio, with seeing their relations to the minimax regret and redundancy.

In the rest of this section, we briefly describe each topic and their relations. In Section 2, we give the formal definition of regret and redundancy, and discuss its basic properties. And then, we discuss MDL principle, channel capacity, and portfolio theory.

1.1. MDL principle

The MDL principle [21] primarily concerns information criteria for model selection. The information criterion is used as a measure to select the best parametric model given a data sequence $x^n = x_1 x_2 \dots x_n$. Usually, the model which induces the minimum value of the information criterion is selected as the best model. In the context of the MDL principle, stochastic complexity (SC) (or MDL criterion) plays a role of information criterion. Here, the SC for a given parametric model is defined as the shortest code length of the given data achievable with the help from the model [23]. The problem of finding such shortest code length is closely related to universal data compression, where it is assumed that the information source is unknown.

Before describing the universal data compression, we will mention about the Shannon code, which is a code for a known information source. Suppose that a data sequence $x^n \in \mathcal{X}^n$ $(n = 1, 2, \dots)$ is drawn according to a probability mass function $p(x^n)$, where $\mathcal{X}$ denotes a finite set called alphabet. Consider a problem to minimize the expected code length $E_p l(x)$, where E_p denote the expectation with respect to p and $l(x)$ denotes a code length (bit) assigned to x by a code φ. The code φ is a function which maps $x \in \mathcal{X}$ to a binary sequence $\varphi(x)$, where $\varphi(x)$ for each x is referred to as a code word. It is known that the solution is given by a code with $l(x) = -\log_2 p(x)$ (we neglect the value less than 1 bit), provided that $\prod_{t=1}^n \varphi(x_t)$, the code word for x^n, is uniquely decodable in an efficient manner. Note that this condition, referred to as 'prefix condition', requires that the Kraft's inequality $\sum_{x\in\mathcal{X}} 2^{-l(x)} \le 1$ is satisfied (see [10] [19] for example). The codes satisfying 'prefix condition' are referred to as *prefix code*, and the code with $l(x) = -\log_2 p(x)$ is referred to as Shannon code with respect to the information source p. Note that the Shannon code satisfies the Kraft's inequality. By this fact, we realize that there is essentially one to one correspondence between a code for data compression and a probability distribution of the data. Let us consider the universal coding. In particular in the context of the MDL, we assume that the information source is an element of a parametric model $S = \{p(\cdot|u) : u \in U \subset \mathbb{R}^d\}$. Imagine the situation in which we prepare a code q but the true information source $p(\cdot|u)$ is different from q. Then, the expected code length is larger than the one which is achieved by the code $p(\cdot|u)$. The difference is referred to as *redundancy*. Denote it by $R_n(q, u)$,

which is defined as

$$(1) \quad R_n(q,u) = E_u \log \frac{1}{q(x^n)} - E_u \log \frac{1}{p(x^n|u)},$$

where E_u denotes the expectation with respect to $p(\cdot|u)$. The redundancy is equal to the Kullback-Leibler divergence from $p(\cdot|u)$ to q with random variable x^n:

$$D^{(n)}(p(\cdot|u)|q) = E_u \log \frac{p(x^n|u)}{q(x^n)}.$$

Our main concern is the minimax redundancy

$$\bar{R}_n(S) = \min_q \max_{u \in U} R_n(q,u).$$

The code achieving the minimax redundancy is referred to as the minimax code with respect to redundancy. Concerning this notion, Clarke & Barron [7] proved that for fairly general models satisfying usual regularity conditions, Bayes mixture with Jeffreys prior [16] asymptotically achieves the minimax redundancy, and it equals

$$\frac{d}{2} \log \frac{n}{2\pi e} + \log \int |J(u)|^{1/2} du + o(1),$$

where d is the dimension of u, $J(u)$ is the Fisher information matrix of u, and $|J(u)|$ is its determinant, and $o(1)$ is a quantity converging to 0 as n goes to infinity. The Jeffreys prior is the prior density proportional to $|J(u)|^{1/2}$.

The code length of the minimax code for redundancy is a candidate of SC, since it can measure a performance of the model S, but it is not sufficient, since we need the shortest code length for the particular data sequence for the purpose of model selection, that is, we concerns not averaged code length but pointwise shortest code length. In [23], Rissanen gave the final definition of SC as the optimal code length for another minimax problem in which *regret* (or pointwise redundancy) is employed as a loss function. The regret of a code q with respect to a data sequence x^n and the model S, denoted as $r(q,x^n)$, is defined as

$$(2) \quad \begin{aligned} r(q,x^n) &= \log \frac{1}{q(x^n)} - \min_{u \in U} \log \frac{1}{p(x^n|u)} \\ &= \log \frac{1}{q(x^n)} - \log \frac{1}{p(x^n|\hat{u})}, \end{aligned}$$

where $\hat{u} = \hat{u}(x^n)$ is the maximum likelihood estimate given x^n. Note that $\log(1/p(x^n|\hat{u}))$ is the code length of the most suitable code to x^n in S. As for this problem, Shtar'kov showed that the *normalized maximum likelihood* (NML), which is defined as

$$(3) \quad m_n(x^n) = \frac{p(x^n|\hat{u}(x^n))}{\sum_{x^n \in \mathcal{X}^n} p(x^n|\hat{u}(x^n))},$$

is the exact minimax code [24]. Note that the regret of m_n does not depend on x^n. This kind of solution is referred to as the equalizer. It is interesting that the regret of the NML is shown to be

$$\frac{d}{2} \log \frac{n}{2\pi} + \log \int |J(u)|^{1/2} du + o(1),$$

under certain regularity conditions [23]. This almost equals the minimax redundancy. Consequently, SC of a sequence x^n typically equals

$$(4) \quad \log \frac{1}{p(x^n|\hat{u})} + \frac{d}{2} \log \frac{n}{2\pi} + \log \int |J(u)|^{1/2} du + o(1).$$

Further, if the model S is an exponential family [6], then the Bayes mixture with (slightly modified) Jeffreys prior achieves the minimax regret asymptotically.

Efficiency of SC as an information criterion has been shown in many literatures, and also its relation to the other topics such as data compression, sequential prediction, gambling, etc, has been discussed. See [21, 3, 4, 35, 13, 14] for example.

1.2. Channel Capacity

In the above discussions, the maximin risk other than the minimax risk is also important. The maximin redundancy is defined as

$$\begin{aligned} \underline{R}_n(S) &= \max_w \min_q \int R_n(q,u) w(du) \\ &= \max_w \int R_n(q_w,u) w(du), \end{aligned}$$

where the maximum is taken for all prior measures. By definition, $\underline{R}_n(S) \le \bar{R}_n(S)$ holds in general. In particular for the redundancy, $\underline{R}_n(S) = \bar{R}_n(S)$ holds [12, 15]. The quantity $\int R_n(q,u)w(du)$ is referred to as the Bayes risk with prior w. The prior which achieves the maximin risk is referred to as *the least favorable prior*. Here, an important point is that Bayes risk for redundancy equals *mutual information* between u and x^n. It implies that the maximin redundancy can be interpreted as channel capacity of the channel with input u and output x^n.

In information theory, a communication channel with input x and output y is modeled by a conditional probability distribution of y given x. Let $p(y|x)$ denote its density function. We can think of $p(y|x)$ as a "probabilistic function", where y is not determined by x, but y contains certain information about x. Hence the conditional distribution $p(y|x)$ can model a communication channel which has a certain noisy behaviour. If the input x is drawn from a probability distribution, then we can measure the average amount of the information about x possessed by y using *mutual information* between x and y. Let $w(x)$ be its density function, then we have a joint probability density function $p(x,y) = w(x)p(y|x)$ and the mutual information between x and y is defined as

$$I_{x;y} = I_{x;y}(w) = \int p(x,y) \log \frac{p(x,y)}{w(x)p_w(y)} dxdy,$$

where $p_w(y)$ is the marginal density of y determined by w, and given as $p_w(y) = \int p(y|x)w(x)dx$. The capacity C of the channel modeled by $p(y|x)$ is defined as the maximum of $I_{x;y}(w)$ with respect to w:

$$C = \max_w I_{x;y}(w).$$

The operational meaning of channel capacity is given by Shannon's channel coding theorem as follows: assume that a channel code for which the block size n is sufficiently large, then nC bits is the supremum of the average information which can be sent by one code word. In other words, C is the supremum transfer rate which can be achieved for the communication channel. The problem of determining the capacity of a given communication channel is of special importance for communication engineering.

The mutual information $I_{x;y}$ can be rewritten as

$$\begin{aligned} I_{x;y} &= \int w(x)p(y|x)\log\frac{p(y|x)}{p_w(y)}dydx \\ &= \int w(x)R_1(p_w,x)dx \\ &= \min_q \int w(x)R_1(q,x)dx, \end{aligned}$$

which turns out to be the minimum Bayes risk of redundancy with respect to the prior w. This confirms that finding the channel capacity is equivalent to finding the maximin redundancy, which is equal to the minimax redundancy.

1.3. Portfolio Theory

The last topic concerns portfolio theory. It has been studied in several fields, such as finance, Bayes theory, game theory, information theory, etc. Related to information theory, Cover [8], Cover & Ordentlich [9], Cross & Barron [11], etc. have studied this topic. In [10] (Cover & Thomas's text book on information theory), one chapter is devoted to portfolio theory. In particular, Cover [8] introduced a model of portfolio referred to as *universal portfolio*, which has the similar formulation as universal data compression with coding regret.

Let us consider a stock market which consists of m stocks. Describe the market by a series of m-dimensional vectors $\{\boldsymbol{x}_i\}$, where jth element x_{ij} of $\boldsymbol{x}_i$ denotes the price relative of jth stock. The price relative x_{ij} denotes a ratio of jth stock's price at the end of the ith day to that at the beginning of the day. The investor's strategy (portfolio) is specified by a portfolio vector, which is a non-negative valued vector $\boldsymbol{b}$ normalized as $\sum_{j=1}^m b_j = 1$. Its jth element denotes a fraction of the capital invested in the jth stock to the whole capital invested in all m stocks. Then, $\boldsymbol{b}\cdot\boldsymbol{x}_i$ denotes the growth rate of the whole capital (wealth relative) at the ith day.

Define $S(\boldsymbol{x}^n|\boldsymbol{b}) = \prod_{i=1}^n \boldsymbol{b}\cdot\boldsymbol{x}_i$, referred to as wealth growth rate, then it denotes the wealth the investor has at the nth day, provided he had the wealth of the amount 1 at the beginning of ith day and he dealt everyday so as to keep his portfolio as $\boldsymbol{b}$. This strategy is referred to as *constantly rebalanced portfolio* (CRP). If he knew the behaviour of stock market $\boldsymbol{x}^n$ in advance, he could optimize $\boldsymbol{b}$ so that $S(\boldsymbol{x}^n|\boldsymbol{b})$ was maximized. Let $\boldsymbol{b}^*$ denote the optimal value of $\boldsymbol{b}$ given $\boldsymbol{x}^n$. The portfolio $\boldsymbol{b}^*$ is referred to as *optimal constantly rebalanced portfolio*. Of course, to find $\boldsymbol{b}^*$ in advance is impossible. Figure 1 shows the performance of the $\boldsymbol{b}^*$. In the figure, daily prices of 2 stocks Mobil and Kinar over a period of 4500 days from 1962.07.03 until 1980.09.12 are indicated by blue and red lines respectively, where the prices are normalized as they were 1 at the beginning of the first day. Also, the black line shows the wealth growth rate achieved by the optimal constantly rebalanced portfolio applied to these two stocks. By this, we can see, if the

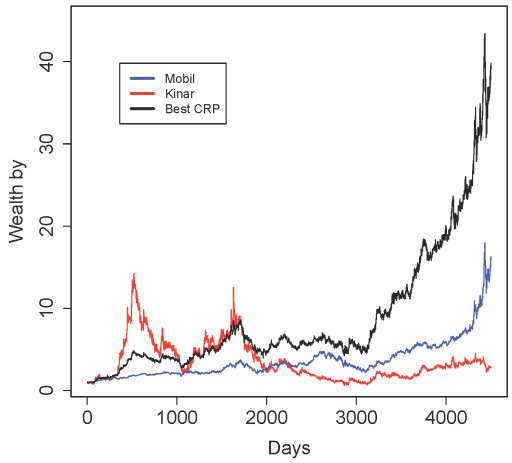

Figure 1: Performance of optimal CRP

investor knew these stocks' behaviour in advance, it was possible to increase the capital 5 times greater than the case in which he keeps his capital in either one stock over the period.

The goal of universal portfolio is to approach the optimal constantly rebalanced portfolio. The real portfolio should be determined by the past observations. Such portfolio strategies are referred to as *causal portfolio*, and strictly described as follows: it is a series of portfolio vectors $\{\tilde{\boldsymbol{b}}(\boldsymbol{x}^i)\}$, where $\tilde{\boldsymbol{b}}(\boldsymbol{x}^i)$ is determined based on the past observations $\boldsymbol{x}^i = \boldsymbol{x}_1 \ldots \boldsymbol{x}_i$. This situation is similar to universal coding game.

In fact, Cover considered the minimax problem for the following risk (wealth ratio):

$$\frac{S(\boldsymbol{x}^n|\boldsymbol{b}^*)}{\tilde{S}(\boldsymbol{x}^n|\tilde{\boldsymbol{b}})} = \frac{\prod_{i=1}^n \boldsymbol{b}^*\cdot\boldsymbol{x}_i}{\prod_{i=1}^n \tilde{\boldsymbol{b}}(\boldsymbol{x}^{i-1})\cdot\boldsymbol{x}_i}, \tag{5}$$

where we let $\tilde{S}(\boldsymbol{x}^n|\tilde{\boldsymbol{b}}) = \prod_{i=1}^n \tilde{\boldsymbol{b}}(\boldsymbol{x}^{i-1})\cdot\boldsymbol{x}_i$, which is the wealth growth rate by $\{\tilde{\boldsymbol{b}}(\boldsymbol{x}^i)\}$.

Here the optimal constantly rebalanced portfolio $\boldsymbol{b}^*$ is similar to the maximum likelihood estimate (MLE) in statistical inference and plays a similar role of the MLE in universal coding with coding regret. It is known that the logarithm of minimax wealth ratio in m stocks case is equal to the minimax regret of universal coding of the multinomial Bernoulli model with alphabet size m [9, 10]. Further, the minimax portfolio is similar to the one for the universal coding.

2. Preliminaries

We introduce the detailed definition of redundancy and regret, and discuss the relation between them.

Let $p(\cdot|u)$ $(u \in U \subset \mathbb{R}^k)$ denotes a probability mass function of a stochastic process with finite alphabet $\mathcal{X}$, and we regard it as an information source. We denote a model of information sources as $S = \{p(\cdot|u) : u \in U\}$. Let $\hat{u} = \hat{u}(x^n)$ denote the maximum likelihood estimate of u given x^n. It is possible to discuss the case in which $\mathcal{X}$ is continuous, for which we only have to introduce a density function with respect to appropriate reference measure, instead of a mass function.

For a space of parameters U, we assume $\bar{U^\circ} = \bar{U}$. (For $A \subset \mathbb{R}^k$, we let A° denote the interior of A and $\bar{A}$ the closure of A).

Let K denote a subset of U such that $\bar{K^\circ} = \bar{K}$. We let $S(K) = \{p(\cdot|u) : u \in K\}$ and $\mathcal{X}^n(K) = \{x^n : \hat{u}(x^n) \in K\}$.

We can construct a code (source code) based on a stochastic process q, whose code length for x^n is $-\log q(x^n)$. Hence we regard q itself as a code. Redundancy of q with respect to $p(\cdot|u)$ is defined by (1). We can write

$$R_n(q, u) = E_u \log \frac{p(x^n|u)}{q(x^n)} = D^{(n)}(p(\cdot|u)|q).$$

Here $D^{(n)}(p(\cdot|u)|q)$ denotes the Kullback Leibler divergence (KL divergence for short) of q from p regarding x^n as a random variable.

We introduce the minimax redundancy. We define q's maximum redundancy for $S(K)$ as

$$\bar{R}_n(q, S(K)) = \sup_{u \in K} R_n(q, u),$$

and minimax redundancy with respect to $S(K)$ as

$$\bar{R}_n(S(K)) = \inf_q \sup_{u \in K} R_n(q, u),$$

where inf is taken for all probability mass functions on $\mathcal{X}^n$. We refer to the q which achieves the minimax redundancy as the minimax code for redundancy.

We also introduce the maximin redundancy. We define minimum redundancy for a prior measure w and $S(K)$ as

$$\underline{R}_n(w, S(K)) = \inf_q \int_K R_n(q, u) w(du). \tag{6}$$

Then we have

$$\begin{aligned} \underline{R}_n(w, S(K)) &= \inf_q \int D^{(n)}(p(\cdot|u)|q) w(du) \\ &= \int D^{(n)}(p(\cdot|u)|p_w) w(du), \end{aligned}$$

where we let $p_w(\cdot) = \int p(\cdot|u) w(du)$. The maximin redundancy is defined as

$$\begin{aligned} \underline{R}_n(S(K)) &= \sup_w \inf_q \int_K R_n(q, u) w(du) \\ &= \sup_w \int_K R_n(p_w, u) w(du), \end{aligned}$$

where sup is taken for all prior measures on K. We refer to the w achieving maximin redundancy as the least favorable prior. Note that

$$\underline{R}_n(w, S(K)) = \inf_q \int_K R_n(q, u) w(du) \le \inf_q \sup_{u \in K} R_n(q, u)$$

holds for any w. Hence we have

$$\underline{R}_n(S(K)) \le \bar{R}_n(S(K)).$$

In fact the equality holds for redundancy [12, 15].

Here we introduce the minimax and maximin regret. Regret of code q for x^n and S is defined by (2). The minimax regret for $W_n \subset \mathcal{X}^n$and S is defined by

$$\bar{r}(W_n) = \inf_q \sup_{x^n \in W_n} r(q, x^n) = \inf_q \sup_{x^n \in W_n} \log \frac{p(x^n|\hat{u})}{q(x^n)}.$$

The maximin regret is defined as

$$\underline{r}(W_n) = \sup_v \inf_q \sum_{x^n \in W_n} r(q, x^n) v(x^n),$$

where sup is taken for all probability mass functions over W_n. Note that we have

$$\begin{aligned} \underline{r}(W_n) &= \sup_v \inf_q \sum_{x^n \in W_n} r(q, x^n) v(x^n) \\ &\le \sup_v \inf_q \sup_{x^n \in W_n} r(q, x^n) \\ &= \inf_q \sup_{x^n \in W_n} r(q, x^n) = \bar{r}(W_n). \end{aligned} \tag{7}$$

In fact, the equality holds for regret [24, 34].

We define empirical Fisher information $\hat{J}(u, x^n)$, and Fisher information $J(u)$ by

$$\begin{aligned} \hat{J}(u, x^n) &= \hat{J}(u) \\ &= \Big(\frac{-1}{n} \frac{\partial^2 \log p(x^n|u)}{\partial u_i \partial u_j}\Big), \\ J(u) &= \lim_{n \to \infty} E_u[\hat{J}(u, x^n)]. \end{aligned}$$

Then the Jeffreys prior w_K on K is defined by

$$w_K(u) = \frac{|J(u)|^{1/2}}{C_J(K)}, \tag{8}$$

$$C_J(K) = \int_K |J(u)|^{1/2} du. \tag{9}$$

3. MDL principle

3.1. Foundation of MDL principle

The MDL criterion is given as follows [21, 3, 32, 26]. For a model $S = \{p(\cdot|u) : u \in U \subset \mathbb{R}^d\}$, the total code length for a data sequence $x^n = x_1 x_2 ... x_n$ with respect to S is given as

$$-\log p(x^n|\hat{u}(x^n)) + \frac{d}{2} \log n + O(1), \tag{10}$$

where $\hat{u}(x^n)$ is the maximum likelihood estimate of u given x^n. The first term is the data description length and the second term is the parameter description length. Note that the data description length is given by the Shannon code with respect to $p(\cdot|\hat{u})$ and is the minimum of $\{-\log p(x^n|u) : u \in U\}$. Recall that the Shannon code with respect to p is a prefix code which minimizes the expected code length with respect to p. Hence, $-\log p(x^n|\hat{u})$ is the optimal code length, but we cannot prepare such code prior to seeing the data. In other words, the code whose code length equals $-\log p(x^n|\hat{u}(x^n))$ for all $x^n \in \mathcal{X}^n$, does not satisfy Kraft's inequality:

$$\sum_{x^n} 2^{-(-\log p(x^n|\hat{u}(x^n)))} = \sum_{x^n} p(x^n|\hat{u}(x^n)) > 1.$$

The parameter description length is added in order to make the code to satisfy Kraft's inequality. This encoding scheme can be seen as follows:

1. Encode $\hat{u}(x^n)$ (the model $p(\cdot|\hat{u}(x^n))$) with appropriate precision.

2. Encode the data x^n with the Shannon code with respect to $p(\cdot|\hat{u}(x^n))$.

This type of source coding is referred to as two stage coding. Here, 'appropriate precision is of order $1/\sqrt{n}$, which is determined by minimizing the total code length and of the same order as standard deviation of MLE, i.e. the parameter description length itself is determined by MDL principle.

Performance of MDL criterion is theoretically shown, but the code length (10) can be improved, if we employ Bayes code rather than two stage coding, as pointed out by Rissanen [22]. Bayes code with a model S is the Shannon code with respect to the Bayes mixture

$$q_w(x^n) = \int p(x^n|u)w(du),$$

where $w(du)$ is a prior measure on U. The code length of Bayes code depends on the prior in general, hence choice of prior is important.

For this problem, Clarke & Barron [7] gave an answer in terms of expected code length. Assuming that the data x^n is drawn from an i.i.d. source $p(\cdot|u)$ ($u \in U \subset \mathbb{R}^d$), they proved that the Bayes mixture m_U with Jeffreys prior $w_U(u)$ (see (8)) satisfies

$$(11)\ R_n(m_U, u) = \frac{d}{2}\log\frac{n}{2\pi e} + \log C_J(U) + o(1),$$

where $o(1)$ converges to 0 as $n \to \infty$. Letting K denote a compact set interior to U, (11) holds uniformly for all $u \in K$. Using the equation (11), we can show that redundancy $R_n(m_U, u)$ does not depend on u asymptotically, if $u \in K$. Actually, the following holds for a sequence of Jeffreys mixtures $\{m_{K_n}\}$, for a sequence of parameter set $\{K_n\}$, where the interior of K_n includes K and K_n slowly converges to K [7].

$$(12)\max_{u\in K} R_n(m_{K_n}, u) = \frac{d}{2}\log\frac{n}{2\pi e} + \log C_J(K) + o(1).$$

The codes with this property is referred to as asymptotic minimax codes (with respect to redundancy). When K is U and U essentially corresponds to the whole space, e.g. the Bernoulli model $p(x|u) = u^x(1-u^x)$ with $U = [0,1]$, then the above treatment is impossible. Some of such cases can be handled with a sequence of prior densities which have higher density at near the boundary than the Jeffreys prior and weakly converge to the Jeffreys prior over U. In fact, Xie & Barron [33] treats the multinomial Bernoulli model and [29] treats the one-dimensional exponential family case. Note that, when $\mathcal{X}$ is not finite, $\int_U |J(u)|^{1/2}du$ with U being the essential whole space is usually infinite.

After that, Rissanen founded the final definition of stochastic complexity in [23]. It is known that the code based on the normalized maximum likelihood (NML) [24] achieves the minimax regret. The NML, denoted by $m_n(x^n)$, is given by normalizing the maximum likelihood $p(x^n|\hat{u}(x^n))$ (see (3)). Let C_n denote the normalization constant:

$$C_n = \sum_{x^n\in\mathcal{X}^n} p(x^n|\hat{u}(x^n)).$$

Then we have

$$r(m_n, x^n) = \log C_n,$$

i.e. the regret of m_n does not depend on x^n. Note that we can write

$$\begin{aligned} r(q, x^n) &= \log\frac{1}{q(x^n)} - \log\frac{1}{C_n m_n(x^n|\hat{u}(x^n))} \\ &= \log C_n + \log\frac{m_n(x^n)}{q(x^n)}. \end{aligned}$$

Hence we have

$$\max_{x^n} r(q, x^n) = \log C_n + \max_{x^n}\log\frac{m_n(x^n)}{q(x^n)} \geq \log C_n,$$

where equality holds, iff $q = m_n$. This implies m_n is the minimax solution with respect to regret.

Rissanen [23] defined stochastic complexity by code length of x^n based on m_n, and shown that its value equals

$$(13)\ \log\frac{1}{p(x^n|\hat{u}(x^n))} + \frac{d}{2}\log\frac{n}{2\pi} + \log C_J(U) + o(1).$$

This uniformly holds for x^n satisfying $\hat{u}(x^n) \in K$, where K is a compact set interior to U.

Example 1 (Bernoulli model)**.** Assume the binary alphabet $\mathcal{X} = \{0, 1\}$. Let

$$\mathrm{Ber}(x|u) = u^x(1-u)^{1-x}$$

denote a probability mass function on $\mathcal{X}$. In this paper, we refer to

$$S = \{\mathrm{Ber}(\cdot|u) : u \in [0,1]\}.$$

as the Bernoulli model.

For evaluation of the code length of the NML, we need to compute $\sum_{x^n} \mathrm{Ber}(x^n|\hat{u}(x^n))$, where $\hat{u} = \hat{u}(x^n)$ is the MLE of u. Define

$$E(n,k) = \Big\{x^n : \sum_{i=1}^{n} x_i = k\Big\},$$

which we call a type class of type (n,k). We have

$$|E(n,k)| = \frac{n!}{k!(n-k)!}.$$

Since $\mathrm{Ber}(x^n|u) = u^k(1-u)^{n-k}$, we have

$$(14)\quad \sum_{x^n \in \mathcal{X}^n} \mathrm{Ber}(x^n|\hat{u}) = \sum_{k=0}^{n} \hat{u}^k(1-\hat{u})^{n-k}\frac{n!}{k!(n-k)!}.$$

By the Stirling's formula

$$n! = \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n e^{\lambda_n} \quad (\frac{1}{12n+1} \le \lambda_n \le \frac{1}{12n}),$$

we have

$$\begin{aligned}\sum_{x^n} \mathrm{Ber}(x^n|\hat{u}) &\approx \sum_{k=0}^{n} \sqrt{\frac{1}{2\pi n\hat{u}(1-\hat{u})}} \\ &\approx \sqrt{\frac{n}{2\pi}}\int_0^1 \frac{1}{\sqrt{\hat{u}(1-\hat{u})}}d\hat{u} \\ &= \sqrt{\frac{n}{2\pi}}\int_0^1 |J(u)|^{1/2}du,\end{aligned}$$

where $J(u)$ is the Fisher information of u. Hence the code length of m_n equals (4).

The evaluation (4) is an analytic form, but it holds only asymptotically. For the Bernoulli case, the strict value of SC can be efficiently computed, since we can compute (14) in linear time with respect to n. Recently, an efficient algorithm was found for the multinomial Bernoulli model and some extension [17, 18]. It provides an efficient method to compute the SC and the joint probability $m_n(x^n)$ for those models.

Here we review the outline of the proof of (13) by Rissanen. The proof is given under the following assumptions, where K is assumed to be a bounded open set and $\mathcal{X}^n(K)$ denotes the set $\{x^n : \hat{u}(x^n) \in K\}$.

i) $\forall u \in K$, $0 < c_1 \le |J(u)| \le c_2 < \infty$.

ii) The central limit theorem for MLE $\hat{u}(x^n)$ holds, uniformly for $u \in K$: Let $\xi = \sqrt{n}(\hat{u}-u)$ and let R_r denote a cube with side length $2r$ centered at the origin. The following uniformly holds for all $u \in U$.

$$\mathrm{P}_u(\xi \in R_r) = \frac{|J(u)|^{1/2}}{(2\pi)^{d/2}}\int_{R_r} e^{-\xi^T J(u)\xi/2}d\xi + o(1).$$

iii) For all $n \ge 1$, for all $x^n \in \mathcal{X}^n(K)$, $\hat{J}(\hat{u}, x^n) < M < \infty$ holds, where M is a positive definite matrix. Also, for all $n \ge 1$, for all $x^n \in \mathcal{X}^n(K)$, for all i, j, a family of functions $J_{ij}(u(\xi), x^n)$ is equicontinuous at $\xi = 0$, where $u(\xi) = \hat{u} + \xi/\sqrt{n}$.

Let $C_n(K)$ denote the normalization constant of the NML for $\mathcal{X}^n(K)$:

$$C_n(K) = \sum_{x^n \in \mathcal{X}^n(K)} p(x^n|\hat{u}(x^n)).$$

Note that it suffices to show

$$C_n(K) = \frac{C_J(K)n^{d/2}(1+o(1))}{(2\pi)^{d/2}}.$$

First, we quantize the parameter space K into cubes with side length $2r/\sqrt{n}$. Let b_i $(i = 0, \pm1, \pm2, ...)$ denote such cubes. Then we have

$$C_n(K) = \sum_i \sum_{x^n : \hat{u} \in b_i} p(x^n|\hat{u}).$$

For each i, we let $\bar{u}$ denote b_i's center. Then, by the conditions iii) etc, for $\hat{u} \in b_i$ we have $p(x^n|\hat{u}) \sim p(x^n|\bar{u})$. Hence we have

$$\begin{aligned}\sum_{x^n : \hat{u} \in b_i} p(x^n|\hat{u}) &\sim \sum_{x^n : \hat{u} \in b_i} p(x^n|\bar{u}) \\ &= \mathrm{P}_{\bar{u}}(\hat{u} \in b_i) \\ &= \mathrm{P}_{\bar{u}}(\xi \in R_r).\end{aligned}$$

This together with Condition ii) yields the following.

$$C_n(K) \sim n^{d/2}\int_K \frac{|J(u)|^{1/2}}{(2\pi)^{d/2}}du = \frac{C_J(K)n^{d/2}}{(2\pi)^{d/2}}.$$

Note that (13) is relevant to (11), since the following holds

$$(15)\qquad E_u[\log\frac{p(x^n|\hat{u}(x^n))}{p(x^n|u)}] = \frac{d}{2} + o(1)$$

uniformly for $u \in K$ [7], where K is a compact set interior to U. This equation and (13) yield (11). Hence, the normalized maximum likelihood is asymptotically minimax for redundancy, too. However, (11) does not always imply (13). Then, Bayes code with slightly modified Jeffreys prior (modified Jeffreys code for short) is not minimax with respect to regret in general. Actually, it is known that the modified Jeffreys code achieves minimax regret, if and only if the target class is an exponential family.

3.2. Bayes procedure achieving Stochastic Complexity

In this section, we see that the slightly modified Jeffreys mixture asymptotically achieves SC for exponential families including classes of Markov sources. Moreover, we will see its modification to more general cases.

Let m_w denote the Bayes mixture with a prior $w(u)$ for a target class $S = \{p(\cdot|u) : u \in U \subset \mathbb{R}^d\}$:

$$m_w(x^n) = \int_U p(x^n|u)w(u)du.$$

First note that Taylor expansion of $\log p(x^n|u)$ around $\hat{u}$:

$$\log p(x^n|u) = \log p(x^n|\hat{u}) - \frac{1}{2}(u-\hat{u})^T n\hat{J}(u')(u-\hat{u}).$$

where u' is a certain point between u and $\hat{u}$, and we used $\partial \log p(x^n|u)/\partial u^i = 0$ at $u = \hat{u}$. (Recall the definition of $\hat{J}$, which is the Hessian of $(-1/n)\log p(x^n|u)$). Hence we have

$$\frac{p(x^n|u)}{p(x^n|\hat{u})} = \exp\big(-\frac{1}{2}(u-\hat{u})^T n\hat{J}(u')(u-\hat{u})\big).$$

Let B_n denote a sphere with radius $\log n/\sqrt{n}$ centered at $\hat{u}$, then we have

$$\begin{aligned}\frac{m_w(x^n)}{p(x^n|\hat{u})} &\sim \int_{B_n} \exp(\frac{-nu^T\hat{J}(\hat{u})u}{2})w(u)du \\ &\sim \frac{(2\pi)^{d/2}w(\hat{u})}{n^{d/2}|\hat{J}(\hat{u})|^{1/2}}.\end{aligned}$$

Hence, we have the following asymptotic evaluation for the regret of m_w:

$$\frac{d}{2}\log\frac{n}{2\pi} + \log\frac{|\hat{J}(\hat{u})|^{1/2}}{w(\hat{u})} + o(1). \tag{16}$$

In particular, for the Jeffreys prior $w = w_U$ (recall (8)), its regret $r(m_U, x^n)$ is evaluated to be

$$\frac{d}{2}\log\frac{n}{2\pi} + \log\int_U |J(u)|^{1/2}du + \frac{1}{2}\log\frac{|\hat{J}(\hat{u})|}{|J(\hat{u})|} + o(1).$$

This asymptotic holds uniformly for all sequences such that the MLE $\hat{u}$ is contained in a compact subset K interior to U.

Here, note that $J(\hat{u}) = \hat{J}(\hat{u})$ holds always, when the target class S is an exponential family defined as [6, 2]

$$\begin{aligned}S &= \{p(x|\theta) : \theta \in \Theta \subset \mathbb{R}^d\}, \\ p(x|\theta) &= \exp\big(\theta\cdot x - \psi(\theta)\big).\end{aligned}$$

Hence for that case, the above asymptotic implies that the Jeffreys mixture is asymptotically an equalizer for sequences such that $\hat{\theta} \in K$. The situation is similar to the case of redundancy treated by Clarke & Barron [7]. Actually, the same sequence of Jeffreys mixtures for the sequence of the parameter sets $\{K_n\}$, which is used to obtain (12), provides the following bound for the case in which S is an exponential family.

$$\max_{x^n:\hat{\theta}\in K} r(m_{K_n}, x^n) = \frac{d}{2}\log\frac{n}{2\pi} + \log C_J(K) + o(1). \tag{17}$$

Since the Markov model, which is not an exponential family in strict sense, can be treated as an exponential family when the sample size is large [20, 30], the bound holds for the Jeffreys mixture for the Markov model [30].

In a manner parallel to the case of redundancy, the similar modification for the boundary problem is also possible for the regret case. Xie & Barron [34] treats the multinomial Bernoulli case, [29] treats the one-dimensional exponential families, and [30] treats the Markov model case.

When the target model is not an exponential family, the Jeffreys prior is no longer asymptotically minimax for regret, since for such cases, there exist sequences x^n such that $\log(|\hat{J}(\hat{u})|/|J(\hat{u})|)$ is significantly larger than 0. This is an implication from the fact that $\hat{J}(u) - J(u)$ is essentially embedding exponential curvature of S in the space of all probability densities on $\mathcal{X}$ [1, 2]. Here, note that, the embedding exponential curvature is equal to 0 everywhere, if and only if the model is an exponential family. It seems to imply that the Bayes procedure cannot achieve the SC, even with the restriction to data x^n so that the MLE being away from the boundary, because the regret bound of NML implies that $\log(|\hat{J}(\hat{u})|/|J(\hat{u})|)$ must vanish for the minimax procedure.

However, it is still possible to overcome this difficulty with the Bayes procedure [29, 31]. Consider the case in which the target model is a curved exponential family embedded in an exponential family $M = \{\bar{p}(\cdot|\theta) : \theta \in \Theta \subset \mathbb{R}^{\bar{d}}\}$.

$$\bar{p}(x^n|\theta) = \exp(\theta\cdot x - \psi(\theta)).$$

Then, define a curved exponential family as

$$S = \{p(x|u) = \bar{p}(x|\phi(u)) : u \in U \subset \mathbb{R}^d\},$$

where d is less than $\bar{d}$ and ϕ is a function from U to Θ. The key is the usage of priors not on the target model S, but on the full family M. Let K denote a compact subset of U interior to U. We will consider the minimax problem for $\{x^n : \hat{u} \in K\}$ with the target class M. Define a sequence of Bayes mixtures $\{\bar{m}_n\}$ as

$$\begin{aligned}\bar{m}_n(x^n) = (1-\epsilon_n)\int_{K_n} p(x^n|u)w_{K_n}(u)du \\ +\epsilon_n\int_{\bar{K}} \bar{p}(x^n|\theta)\bar{w}(\theta)d\theta,\end{aligned}$$

where K_n is a compact subset of U such that $K \subset K_n^\circ$, $\bar{K}$ is a compact set interior to Θ such that $\bar{K} \supset \{\phi(u) : u \in K\}$, and $\bar{w}$ is the uniform prior over $\bar{K}$. Further assume that $\epsilon_n = o(1)$, and that K_n converges to K at appropriately slow rate. Then, it is possible to show $\bar{m}_n$ is asymptotically minimax. The point is that, there is a point in M whose likelihood is higher than the MLE $\hat{u}$ in S, whenever $\hat{J}(\hat{u}) - J(\hat{u})$ is not equal to 0. This follows from the identity

$$\hat{J}_{ij}(\hat{u}) - J_{ij}(\hat{u}) = -\frac{\partial^2\phi}{\partial u_i\partial u_j}\cdot(\eta(\phi(\hat{u})) - \eta(\hat{\theta})),$$

which is easily derived by the definition of the curved family $p(x^n|u)$. By this, if $\hat{J}_{ij}(\hat{u}) - J_{ij}(\hat{u})$ is not equal to 0, then $\phi(\hat{u})$ is not equal to $\hat{\theta}$. Hence there exists a point $\tilde{\theta} \in \bar{K}$ whose likelihood $p(x^n|\tilde{\theta})$ is higher than $p(x^n|\hat{u})$, so the contribution from around $\tilde{\theta}$ increases likelihood of $\bar{m}_n$ to the value corresponding to the minimax regret.

3.3. Sequential Prediction

Let us consider the sequential prediction problem, in which based on the past observations x^t we want to predict the next symbol x_{t+1} at each time t. Here, we assume probabilistic prediction, i.e. at time t, we output the probability distribution of x_{t+1}, which we denote by $q(x_{t+1}|x^t)$. Note that $q(x_{t+1}|x^t)$ is thought of as a conditional probability of x_{t+1} given x^t. Hence, we can reconstruct the probability mass function of x^n as

$$q(x^n) = \prod_{t=1}^{n} q(x_t|x^{t-1}).$$

Here we let $q(x_1|x^0) = q(x_1)$ denote the initial prediction. We regard $q(x^n)$ as a prediction strategy. Then, we can define the regret of prediction strategy using regret of the code based on the probability mass function $q(x^n)$. We have

$$r(q, x^n) = \sum_{t=1}^{n-1} \log \frac{1}{q(x_{t+1}|x^t)} - \sum_{t=1}^{n-1} \log \frac{1}{p(x_{t+1}|x^t, \hat{u}(x^n))},$$

where $\log(1/q(x_{t+1}|x^t))$ is a logarithmic loss of prediction $q(\cdot|x^t)$ when the actual observation is x_{t+1}. The summation $\sum_{t=1}^{n-1} \log(1/q(x_{t+1}|x^t))$ is regarded as cumulative logarithmic loss. Then, we see that $p(\cdot|\hat{u}(x^n))$ is the prediction strategy which minimizes the cumulative logarithmic regret for x^n among $q \in S$. Since we do not know $\hat{u}(x^n)$ prior to seeing x^n, we cannot use $p(\cdot|\hat{u}(x^n))$ for prediction. In other words, $p(\cdot|\hat{u}(x^n))$ is the ideal prediction strategy which we cannot actually use. Then, the regret $r(q, x^n)$ measures how worse q compared to the ideal prediction strategy for x^n.

With employing this measure, if we know the length n of the sequence x^n in advance, the normalized maximum likelihood m_n performs as minimax prediction strategy. As mentioned before, if the target class is an exponential family, the modified Jeffreys code m_{K_n} achieves the minimax regret asymptotically, hence it performs as the minimax prediction strategy, asymptotically.

Since the normalized maximum likelihood is exactly minimax, it may seem that the modified Jeffreys mixture is not needed. However, the NML code cannot be used when we do not know n in advance. Further, the cost to compute the conditionals $m_n(x_{t+1}|x^t) = m_n(x^{t+1})/m_n(x^t)$ is expensive. In fact, we have to marginalize the joint distribution $m_n(x^n)$ as

$$m_n(x^t) = \sum_{x_{t+1}^n \in \mathcal{X}^{n-t}} m_n(x^t x_{t+1}^n).$$

It means that we cannot use it for prediction. Further it is fatal when we are to use for data compression. In order to use a probability mass function $q(x^n)$ for data compression, we usually utilize arithmetic coding algorithm (see [10] for example), which is computed based on conditionals $q(x_{t+1}|x^t)$ for each $t = 0, 1, \cdots$. In contrast for the Bayes procedure including the modified Jeffreys mixture, computation of conditionals is rather easier in many cases, since we have

$$m_w(x_{t+1}|x^t) = \frac{m_w(x^{t+1})}{m_w(x^t)} = \int p(x_{t+1}|x^t, u) w(du|x^t),$$

where $w(du|x^t)$ is the posterior measure given x^t.

4. Channel Capacity

In this section, we review the derivations of communication channel capacities of basic channels, following standard text books of information theory such as [10], [19]. We also argue about the minimax redundancy for those channels, which is not usually treated in the standard text books.

4.1. Discrete Channel

For a general discrete channel, it is impossible to find the channel capacity in closed form. However, as well known, it is easy to find the channel capacity when the channel has certain symmetries. First we describe the simplest case of the binary symmetric channel.

Let x and y be binary random variables. Let $p(y|x)$ be a conditional probability distribution modeling a communication channel with input x and output y. We assume $\epsilon = p(0|1) = p(1|0)$, i.e. ϵ represents the error probability of this channel. This channel is referred to as *binary symmetric channel*. Let w denote the probability mass function of x. The mutual information can be written as

$$\begin{aligned} I_{x;y}(w) &= \sum_{x,y} p(y|x) w(x) \log \frac{p(y|x)}{p_w(y)} \\ &= -\sum_{y} p_w(y) \log p_w(y) \\ &+ \sum_{x} w(x) \sum_{y} p(y|x) \log p(y|x). \end{aligned}$$

Define the binary entropy function as

$$h(u) = -u \log u - (1-u) \log(1-u).$$

Then, we have for $x = 0, 1$,

$$\sum_{y} p(y|x) \log p(y|x) = -h(\epsilon).$$

Hence we have

$$I_{x;y}(w) = -\sum_{y} p_w(y) \log p_w(y) - h(\epsilon).$$

The second term does not depend on w. The first term is the entropy of y, which is maximized when $p_w(y)$ is uniform. Since the symmetry of $p(y|x)$, it can be achieved by letting $w(x)$ be the uniform distribution $u(x) = 1/2$. Finally, we have found

$$C = 1 - h(\epsilon).$$

Since p_u is uniform, we have for $x = 0, 1$

$$\max_x R_1(p_u, x) = R_1(p_u, x) = C = 1 - h(\epsilon),$$

that is, p_u is the equalizer and the minimax solution.

For the cases of general finite sets $\mathcal{X}$ and $\mathcal{Y}$, the similar argument holds. Assume that $\sum_y p(y|x) \log p(y|x)$ is a constant G independent of x, and $(1/k) \sum_x p(y|x)$ is the uniform distribution, where we let k denote the number of elements of $\mathcal{X}$. Then, for each x, the entropy of the conditional distribution $p(y|x)$ defined as $g(x) = -\sum_y p(y|x) \log p(y|x)$, which equals the constant G. Hence for the conditional entropy $H_{y|x}$ of y given x, $H_{y|x} = \sum_x w(x) g(x) = G$ does not depend on w. Hence, to maximize $I_{x;y}(w) = H_y - H_{y|x}$, it suffices to maximize H_y. The maximum is achieved, when w equals the uniform distribution $u(x) = 1/k$. Then the channel capacity is $\log k - H_{y|x}$. In this case, we have for all $x \in \mathcal{X}$,

$$R_1(p_u, x) = \sum_y p(y|x) \log \frac{p(y|x)}{p_u(y)} = -H_{y|x} + \log k = C.$$

Hence, $p_u(x)$ is the equalizer and the minimax solution.

Further for the general case, it is not difficult to show that the maximin solution equals the minimax solution. Let $w^*(x)$ denote the maximin input distribution. Define a subset $\mathcal{X}' \subset \mathcal{X}$ as

$$\mathcal{X}' = \{x \in \mathcal{X} : w^*(x) > 0\}.$$

By adopting Lagrange multiplier method to optimization of $I_{x;y}(w)$, w^* turns out to be a solution of the following equalities for $x' \in \mathcal{X}'$:

$$\frac{\partial}{\partial w(x')} \sum_{x,y} p(y|x) w(x) \log \frac{p(y|x)}{p_w(y)} = \lambda,$$

where λ is a Lagrange multiplier. This yields

$$\sum_y p(y|x') \log \frac{p(y|x')}{p_w(y)} - \sum_y p(y|x') = \lambda.$$

Hence we have

$$\sum_y p(y|x') \log \frac{p(y|x')}{p_w(y)} = \lambda + 1.$$

The left hand side equals $R_1(p_w, x')$. Hence, for all $x \in \mathcal{X}'$, $R_1(x, p_{w^*})$ is independent of x. Let $C = \lambda + 1$, then we have

$$\sum_{x \in \mathcal{X}} w^*(x) R_1(p_{w^*}, x) = \sum_{x \in \mathcal{X}'} w^*(x) R_1(p_{w^*}, x) = C.$$

Hence C equals the channel capacity.

As for $\bar{x} \in \mathcal{X} \setminus \mathcal{X}'$, the assumption $w(\bar{x}) = 0$ implies

$$\frac{\partial}{\partial w(\bar{x})} \sum_{x,y} p(y|x) w(x) \log \frac{p(y|x)}{p_w(y)} \le \lambda,$$

Hence we have

$$R_1(p_{w^*}, \bar{x}) \le \lambda + 1 = C.$$

Finally we have for all $x \in \mathcal{X}$,

$$R_1(p_{w^*}, x) \le C,$$

which implies that p_{w^*} is the minimax solution. Note that p_{w^*} is not the equalizer in general.

4.2. Gaussian Channel

The Gaussian channel is a model with continuous input and output. Here we consider a basic case of scalar input and output. It is represented as

$$y = x + z,$$

where x is an input signal, y an output signal, and z an additive Gaussian noise with mean 0. The noise z is assumed to be independent of the signal x. Let N be averaged power of the noise, then we can write the conditional probability density function given x as

$$p(y|x) = \frac{1}{\sqrt{2\pi N}} \exp\Big(-\frac{(y-x)^2}{2N}\Big).$$

The channel capacity of the Gaussian channel is usually given as

$$C = \frac{1}{2} \log\Big(1 + \frac{P}{N}\Big), \tag{18}$$

where P denotes the limitation on the average signal power. Note that the average power equals the variance of the random variable.

We review its derivation. Let $w(x)$ denote the probability density function of x. Then the mutual information between x and y is given by

$$I_{x;y}(w) = \int p(y|x) w(x) \log \frac{p(y|x)}{p_w(y)} dx dy,$$

where $p_w(y)$ denotes the marginal density of y determined by w;

$$p_w(y) = \int p(y|x) w(x) dx.$$

We have

$$I_{x;y} = \int p(y|x) w(x) \log p(y|x) dx dy - \int p_w(y) \log p_w(y) dy.$$

Since $p(y|x)$ is a Gaussian density with mean x and variance N, we have

$$\int p(y|x) \log p(y|x) dy = -\frac{1}{2} \log(2\pi e N),$$

which equals minus differential entropy of z. It is defined as

$$H_z = -\int p(z) \log p(z) dz.$$

where $p(z)$ is a probability density function of z. Then, letting H_y denote the differential entropy of y, we have

$$I_{x;y} = -H_z + H_y.$$

Since H_z is fixed, our task is to maximize H_y under the condition that x's variance is fixed to be P. Since x and z is independent, y's variance is $P+N$. It is known that the Gaussian density maximizes the differential entropy, when the variance is given. If $w(x)$ is Gaussian, then $p_w(y)$ is also Gaussian. Hence, the maximum of $I_{x;y}$ is achieved when $w(x)$ is Gaussian. Then, we have

$$\max_{w\in S_P} H_y = \frac{1}{2}\log(2\pi e(N+P)), \tag{19}$$

where S_P is the class of probability densities whose variance are P. This implies (18). We can write

$$C = \max_{w\in S_P} \int w(x) R_1(p_w, x).$$

This type of restriction on w is special one for analogue channel case. If we do not put any restriction, the channel capacity is infinite. In fact, it is trivial that we can send infinite information with infinite power.

However, by this speciality, we cannot argue the minimax problem in usual manner. Let w^* be the Gaussian density with mean 0 and variance P. Then the redundancy $R_1(p_{w^*}, x)$ is not bounded, hence the minimax redundancy is infinite. This is consistent with

$$\max_w \int w(x) R_1(p_w, x) = \infty.$$

However, as discussed in [25], it is possible to define the minimax problem by changing the risk function.

Here we assume that the mean of w equals 0 without loss of generality. Define a Lagrangian L as

$$\begin{aligned} L(w) &= \min_q \Big(\int w(x)(R_1(q,x) \\ &\quad -\lambda_1(x^2-P))dx - \lambda_2 \int w(x)dx \Big) \\ &= \min_q \int w(x)(R_1(q,x) - \lambda_1(x^2-P) - \lambda_2)dx \end{aligned}$$

where λ_1 and λ_2 are the Lagrange multipliers. In a similar manner to the discussion in the former section, w^* is thought of as the maximizer of $L(w)$ with a certain λ_1 and λ_2. With those λ_1 and λ_2, define a new risk $\tilde{R}(x,q)$ as

$$\tilde{R}(q,x) = R_1(q,x) - \lambda_1(x^2-P) - \lambda_2.$$

Then we have

$$w^* = \arg\max_w L(w) = \arg\max_w \min_q \int w(x)\tilde{R}(q,x)dx.$$

Hence, w^* is the maximin for $\tilde{R}$. When w equals w^* (Gaussian with variance P), $p_{w^*}(y)$ is Gaussian with variance $P+N$, we have

$$L(w^*) = \int w^*(x)(R_1(p_{w^*}, x) - \lambda_2)dx = C - \lambda_2.$$

As for the minimax risk, we have

$$\tilde{R}(x, p_{w^*}) = R_1(p_{w^*}, x) - \lambda_1(x^2-P) - \lambda_2.$$

Note that w^* satisfies $L(w^*+\delta w) = 0$ for any variation δw. Hence we have for all $x \in \mathbb{R}$,

$$R_1(p_{w^*}, x) - \lambda_1(x^2-P) - \lambda_2 - 1 = 0.$$

This implies

$$\tilde{R}(p_{w^*}, x) = 1$$

for all $x \in \mathbb{R}$, that is, p_{w^*} is the equalizer and the minimax solution. By this, we can determine λ_2 by $C - \lambda_2 = 1$. Hence, we have $\lambda_2 = C - 1$.

In the literature [25], a different setting to determine the Gaussian channel capacity is employed, where x is restricted as $|x| \le A$. The setting is analogous to the situation of source coding considered by Clarke & Barron [7], that is, they considered the case in which the parameter space is compact, while their problem concerns the asymptotic behaviour of redundancy. For the Gaussian channel case, it is interesting that the input distribution achieving the channel capacity is discrete.

5. Universal Portfolio

In this section, we briefly review universal portfolio, following [10]. As mentioned in Subsection 1.3, the goal of universal portfolio is to find a causal portfolio which minimizes the worst case wealth ratio:

$$\max_{\boldsymbol{x}^n \in \mathcal{X}^n} \frac{\max_{\boldsymbol{b}} S(\boldsymbol{x}^n|\boldsymbol{b})}{\tilde{S}(\boldsymbol{x}^n|\tilde{\boldsymbol{b}})} = \max_{\boldsymbol{x}^n \in \mathcal{X}^n} \frac{\prod_{i=1}^n \boldsymbol{b}^* \cdot \boldsymbol{x}_i}{\prod_{i=1}^n \tilde{\boldsymbol{b}}(\boldsymbol{x}^{i-1}) \cdot \boldsymbol{x}_i}, \tag{20}$$

where $\boldsymbol{b}$ and $\tilde{\boldsymbol{b}}(\boldsymbol{x}^{i-1})$ are portfolio vectors, and $\boldsymbol{x}_i$ is a price relative vector. Recall that the range of $\boldsymbol{b}$ is the probability simplex and the range of each element of $\boldsymbol{x}_i$ is $[0,\infty)$, that is, $\mathcal{X} = [0,\infty)^m$.

In [10], it is shown that a strategy like the NML achieves the minimax risk for the $m = 2$ case. (As noted in [10], extension to general m is straightforward.) It is remarkable that the minimax risk is given as

$$\min_{\tilde{\boldsymbol{b}}} \max_{\boldsymbol{x}^n} \frac{S(\boldsymbol{x}^n|\hat{\boldsymbol{b}})}{S(\boldsymbol{x}^n|\tilde{\boldsymbol{b}})} = \sum_{y^n \in \{0,1\}^n} \mathrm{Ber}(y^n|\hat{u}(y^n)),$$

which equals and the normalization constant of NML for the Bernoulli model (recall (14)). This implies that log of the minimax risk of universal portfolio is equal to the minimax regret of universal coding for the Bernoulli model.

In the sequel, we assume $m = 2$ and let $\boldsymbol{b} = (b_0, b_1) = (1-b, b)$.

The key idea of the proof is the following modification of $S(\boldsymbol{x}^n|\boldsymbol{b})$.

$$S(\boldsymbol{x}^n|\boldsymbol{b}) = \prod_{i=1}^n \boldsymbol{b} \cdot \boldsymbol{x}_i = \sum_{y^n \in \{0,1\}^n} \prod_{i=1}^n b_{y_i} x_{iy_i}.$$

Since $\boldsymbol{b} = (b_0, b_1) = (1-b, b)$, we have

$$\mathrm{Ber}(y^n|b) = \prod_i b^{y_i}(1-b)^{1-y_i} = \prod_{i=1}^n b_{y_i}.$$

Further define the function mapping $y^n \in \{0,1\}^n$ to a real number as

$$x(y^n) = \prod_{i=1}^n x_{iy_i}. \tag{21}$$

Then we have

$$S(\boldsymbol{x}^n|\boldsymbol{b}) = \prod_{i=1}^n \boldsymbol{b}\cdot\boldsymbol{x}_i = \sum_{y^n\in\{0,1\}^n} \mathrm{Ber}(y^n|b)x(y^n). \tag{22}$$

As for $\tilde{S}(\boldsymbol{x}^n|\tilde{\boldsymbol{b}})$ we can do a similar treatment. Let $q(y^n)$ be an arbitrary probability mass function on $\{0,1\}^n$. We will show below, that there exists a series $\tilde{\boldsymbol{b}}(x^{i-1})$ such that the following holds.

$$\prod_{i=1}^n \tilde{\boldsymbol{b}}(\boldsymbol{x}^{i-1})\cdot\boldsymbol{x}_i = \sum_{y^n\in\{0,1\}^n} q(y^n)x(y^n). \tag{23}$$

Define

$$\tilde{b}_j(\boldsymbol{x}^{i-1}) = \frac{\sum_{y^{i-1}\in\{0,1\}^{i-1}} q(y^{i-1}j)x(y^{i-1})}{\sum_{y^{i-1}\in\{0,1\}^{i-1}} q(y^{i-1})x(y^{i-1})}.$$

Then, $\tilde{b}_0(\boldsymbol{x}^{i-1}) + \tilde{b}_1(\boldsymbol{x}^{i-1}) = 1$ and

$$\begin{aligned}\tilde{b}(\boldsymbol{x}^{i-1})\cdot\boldsymbol{x}_i &= \tilde{b}_0(\boldsymbol{x}^{i-1})x_{i0} + \tilde{b}_1(\boldsymbol{x}^{i-1})x_{i1}\\ &= \frac{\sum_{y^i\in\{0,1\}^i} q(y^i)x(y^i)}{\sum_{y^{i-1}\in\{0,1\}^{i-1}} q(y^{i-1})x(y^{i-1})}\end{aligned}$$

hold. This yields (23). Note that (22) is a special case of (23).

Since (22) and (23), for any q, there exists a $\tilde{\boldsymbol{b}}$ such that

$$\frac{S(\boldsymbol{x}^n|\boldsymbol{b}^*)}{S(\boldsymbol{x}^n|\tilde{\boldsymbol{b}})} = \frac{\sum_{y^n\in\{0,1\}^n}\mathrm{Ber}(y^n|b^*)x(y^n)}{\sum_{y^n\in\{0,1\}^n} q(y^n)x(y^n)}.$$

Hence for the upper bound, we have

$$\begin{aligned}\frac{S(\boldsymbol{x}^n|\boldsymbol{b}^*)}{S(\boldsymbol{x}^n|\tilde{\boldsymbol{b}})} &\le \max_{y^n\in\{0,1\}^n}\frac{\mathrm{Ber}(y^n|b^*)x(y^n)}{q(y^n)x(y^n)}\\ &= \frac{\mathrm{Ber}(\hat{y}^n|b^*)}{q(\hat{y}^n)},\end{aligned} \tag{24}$$

where $\hat{y}^n$ is defined as

$$\hat{y}^n = \arg\max_{y^n\in\{0,1\}^n}\frac{\mathrm{Ber}(y^n|b^*)x(y^n)}{q(y^n)x(y^n)}.$$

Then we have

$$\frac{S(\boldsymbol{x}^n|\boldsymbol{b}^*)}{S(\boldsymbol{x}^n|\tilde{\boldsymbol{b}})} \le \frac{\mathrm{Ber}(\hat{y}^n|b^*)}{q(\hat{y}^n)} \le \frac{\mathrm{Ber}(\hat{y}^n|\hat{b}(\hat{y}^n))}{q(y^n)}, \tag{25}$$

where $\hat{b}(\hat{y}^n)$ denotes the MLE of b given $\hat{y}^n$ for the Bernoulli model $\mathrm{Ber}(y^n|b)$. Hence, letting q be the normalized maximum likelihood (3) for the Bernoulli model:

$$q(y^n) = \frac{\mathrm{Ber}(y^n|\hat{b}(y^n))}{\sum_{y^n\in\{0,1\}^n}\mathrm{Ber}(y^n|\hat{b}(y^n))},$$

an upper bound on the worst case wealth ratio as

$$\max_{\boldsymbol{x}^n\in\mathcal{X}^n}\frac{S(\boldsymbol{x}^n|\boldsymbol{b}^*)}{S(\boldsymbol{x}^n|\tilde{\boldsymbol{b}})} \le \sum_{y^n\in\{0,1\}^n}\mathrm{Ber}(y^n|\hat{b}(y^n)). \tag{26}$$

To obtain the matching lower bound, we utilize the maximin log wealth ratio, whose lower bound provides a lower bound on the log minimax wealth ratio (recall (7)). Actually, as we will see soon, the maximin log wealth ratio with respect to the set $K_n = \{(0,1),(1,0)\}^n \subset \mathcal{X}^n$ is sufficient. (Recall that $\mathcal{X} = [0,\infty)^2$.) The set K_n consists of sequences such as $(0,1)(1,0)(0,1)\ldots(1,0)$. In [10], such sequences are referred to as *extremal stock market vector*. It corresponds to a series of events that one stock is lost (the company fails) every day. For every $\boldsymbol{x}^n \in K_n$, assign $y^n \in \{0,1\}^n$ by $y_i = x_{i1}$. ($\boldsymbol{x}_i = (1-y_i, y_i)$). Then define for $y^i \in \{0,1\}^i$,

$$q(y_i|y^{i-1}) = \tilde{b}_{y_i}\Big(\prod_{t=1}^{i-1}(y_t, 1-y_t)\Big) = \tilde{b}_{y_i}(\boldsymbol{x}^{i-1}).$$

Note that $q(y_i|y^{i-1})$ is a conditional probability mass function of y_i given y^{i-1}. Then we have for any $\boldsymbol{x}^n \in K_n$,

$$\tilde{S}(\boldsymbol{x}^n|\tilde{\boldsymbol{b}}) = \prod_{i=1}^n \tilde{\boldsymbol{b}}(\boldsymbol{x}^{i-1})\cdot\boldsymbol{x}_i = \prod_{i=1}^n q(y_i|y^{i-1}) = q(y^n),$$

which is a probability mass function on $\{0,1\}^n$. Note that q can be an arbitrary probability mass function by appropriately defining $\tilde{\boldsymbol{b}}$. In particular for $S(\boldsymbol{x}^n|\boldsymbol{b})$, we have for $\boldsymbol{x}^n \in K_n$,

$$S(\boldsymbol{x}^n|\boldsymbol{b}) = \mathrm{Ber}(y^n|b).$$

Hence we have for any $x^n \in K_n$,

$$\frac{\max_{\boldsymbol{b}} S(\boldsymbol{x}^n|\boldsymbol{b})}{\tilde{S}(\boldsymbol{x}^n|\tilde{\boldsymbol{b}})} = \frac{\max_b \mathrm{Ber}(y^n|b)}{q(y^n)} = \frac{\mathrm{Ber}(y^n|\hat{b}(y^n))}{q(y^n)}$$

holds. Note that $b^* = \hat{b}(y^n)$ in this case. Therefore, we have

$$\begin{aligned}&\max_{v\in\mathcal{P}(\mathcal{X}^n)}\min_{\tilde{b}}\int\log\frac{S(\boldsymbol{x}^n|\boldsymbol{b}^*)}{\tilde{S}(\boldsymbol{x}^n|\tilde{\boldsymbol{b}})}v(d\boldsymbol{x}^n)\\ &\quad\ge \max_{v\in\mathcal{P}(K_n)}\min_{\tilde{b}}\int\log\frac{S(\boldsymbol{x}^n|\boldsymbol{b}^*)}{\tilde{S}(\boldsymbol{x}^n|\tilde{\boldsymbol{b}})}v(d\boldsymbol{x}^n)\\ &\quad= \max_{v\in\mathcal{P}(\{0,1\}^n)}\min_{q}\int\log\frac{\mathrm{Ber}(y^n|\hat{b}(y^n))}{q(y^n)}v(dy^n),\end{aligned}$$

where $\mathcal{P}(A)$ denotes the set of all probability measures on A. Since the last expression equals the maximin regret of

the Bernoulli model, we have

$$\max_{v\in\mathcal{P}(\mathcal{X}^n)} \min_{\tilde{b}} \int \log \frac{S(\boldsymbol{x}^n|\boldsymbol{b}^*)}{\tilde{S}(\boldsymbol{x}^n|\tilde{\boldsymbol{b}})} v(d\boldsymbol{x}^n) \geq \log \sum_{y^n\in\{0,1\}^n} \mathrm{Ber}(y^n|\hat{b}(y^n)).$$

Since this provides a lower bound on the minimax log wealth ratio, together with (26), we have

$$\min_{\tilde{b}} \max_{x^n} \frac{S(\boldsymbol{x}^n|\boldsymbol{b}^*)}{\tilde{S}(\boldsymbol{x}^n|\tilde{b})} = \sum_{y^n\in\{0,1\}^n} \mathrm{Ber}(y^n|\hat{b}(y^n)).$$

By this, we see that log of the minimax wealth ratio equals the minimax regret of universal coding of the Bernoulli model.

Acknowledgments

The author expresses his sincere gratitude to Andrew R. Barron, Hiroshi Nagaoka, and Shiro Ikeda for fruitful discussions with them. This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 19300051, 2007, 2008, 2009, 2010, and Global COE Program, Education-and-Research Hub for Mathematics-for-Industry at Kyushu University.

References

[1] S. Amari, *Differential-geometrical methods in statistics (2nd pr.)*, Springer-Verlag, 1990.

[2] S. Amari & H. Nagaoka, *Methods of Information Geometry*, AMS & Oxford University Press, 2000.

[3] A. R. Barron, *Logically smooth density estimation*, Ph.D. thesis, Stanford Univ., 1985.

[4] A. R. Barron & T. M. Cover, "Minimum complexity density estimation," *IEEE trans. IT,* vol. 37, no. 4, pp. 1034-1054, 1991.

[5] A. R. Barron, J. Rissanen & B. Yu, "The minimum description length principle in coding and modeling," *IEEE trans. Inform. Theory,* Vol. 44 No. 6, pp. 2743 - 2760, 1998.

[6] L. Brown, *Fundamentals of statistical exponential families*, Institute of Mathematical Statistics, 1986.

[7] B. Clarke & A. R. Barron, "Jeffreys prior is asymptotically least favorable under entropy risk," *JSPI,* 41:37-60, 1994.

[8] T. M. Cover, "Universal Portfolios," *Mathematical Finance*, 1(1): 1-29, January 1991.

[9] T. M. Cover & E. Ordentlich, "Universal Portfolios with Side Information." *IEEE Transactions on Information Theory,* 42(2):348-363, March 1996.

[10] T. M. Cover & J. A. Thomas, *Elements of Information Theory, second edition (Wiley Series in Telecommunications and Signal Processing) Wiley-Interscience*, 2006.

[11] J. E. Cross & A. R. Barron, "Efficient universal portfolios for past dependent target classes," *Mathematical Finance,* Vol.13, Issue 2, pp.245-276 2003.

[12] L. Davisson & A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory,* vol. 26, pp. 166-174, March 1980.

[13] P. Grünwald, I. J. Myung, & M. Pitt, *Advances in Minimum Description Length Theory and Applications*, MIT Press, 2005.

[14] P. Grünwald, *The Minimum Description Length Principle,* MIT Press, 2007.

[15] D. Haussler, "A general minimax result for relative entropy," *IEEE trans. Inform. Theory,* vol. 43, no. 4, pp. 1276-1280, 1997.

[16] H. Jeffreys, *Theory of probability, 3rd ed.*, Univ. of California Press, Berkeley, Cal, 1961.

[17] P. Kontkanen & P. Myllymaki, "A linear-time algorithm for computing the multinomial stochastic complexity," *Information Processing Letters,* 103 pp. 227-233, 2007.

[18] P. Kontkanen and P. Myllymaki, "MDL histogram density estimation," *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics,* March 2007.

[19] R. J. McEliece, *Theory of Information and Coding,* Cambridge Univ. Pr. (Sd) Student Edition, 2004.

[20] H. Nagaoka, "The exponential family of Markov chains and its information geometry," *Proc. of the 28th Symposium on Information Theory and its Applications* (SITA2005), 2005.

[21] J. Rissanen, "Modeling by shortest data description," *Automatica.* 14, pp. 465-471, 1978.

[22] J. Rissanen, "Stochastic complexity," *J. R. Statist. Soc. B,* vol. 49, No. 3. pp. 223-239 and 252-265, 1987.

[23] J. Rissanen, "Fisher information and stochastic complexity," *IEEE trans. IT,* vol. 40, no. 1, pp. 40-47, 1996.

[24] Y. M. Shtar'kov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3-17, July 1988.

[25] J. G. Smith, "The information capacity of amplitude and variance-constrained scalar gaussian channels," *Information and Control* 18, pp. 203-219, 1971.

[26] J. Takeuchi, "Characterization of the Bayes estimator and the MDL estimator for exponential families," *IEEE trans. IT,* vol. 43, no. 5, pp. 1165-1174, 1997.

[27] J. Takeuchi, "Fisher information determinant and stochastic complexity for Markov models," *Proc. of 2009 IEEE Int. Symposium on Inform. Theory*, 2009.

[28] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret for exponential families," *Proc. of SITA'97,* pp. 665-668, 1997.

[29] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," *Proc. of 1998 IEEE ISIT,* 1998.

[30] J. Takeuchi, T. Kawabata, & A. R. Barron, "Properties of Jeffreys mixture for Markov sources," *Proc. of Workshop on Information Based Induction Sciences* (*IBIS2001*), pp. 327-332, 2001. Full paper was accepted for publication, *IEEE trans. Inform. Theory*, 2007.

[31] J. Takeuchi, A. R. Barron, & T. Kawabata, "Statistical curvature and stochastic complexity," *Proc. of the 2nd Symposium on Information Geometry and Its Applications*, pp. 29–36, 2006.

[32] C. Wallace & P. Freeman, "Estimating and inference by compact coding," *J. Roy. Statist. Soc. B,* vol. 49, no. 3, pp. 240-265, 1987.

[33] Q. Xie & A. R. Barron, "Minimax redundancy for the class of memoryless sources", *IEEE trans. IT,* vol. 43, no. 2, pp. 646-657, 1997.

[34] Q. Xie & A. R. Barron, "Asymptotic minimax regret for data compression, gambling and prediction," *IEEE trans. Inform. Theory,* vol. 46, no. 2, pp. 431-445, 2000.

[35] K. Yamanishi, "A learning criterion for stochastic rules," *Machine Learning,* a special issue for COLT'90, 9(2/3), 1992.

Jun'ichi Takeuchi
Faculty of Informatics, Kyushu University, Fukuoka, Japan
E-mail: tak(at)inf.kyushu-u.ac.jp