Entropy **2009**, *11*, 634-642; doi:10.3390/e11040634

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

A Lower-Bound for the Maximin Redundancy in Pattern Coding

Aurélien Garivier

CNRS, Telecom ParisTech, Laboratoire Laboratoire Traitement et Communication de l'Information, 75013 Paris, France; E-Mail: aurelien.garivier@telecom-paristech.fr

Received: 1 September 2009 / Accepted: 20 October 2009 / Published: 22 October 2009

Abstract: We show that the maximin average redundancy in pattern coding is eventually larger than $1.84 \left(\frac{n}{\log n}\right)^{1/3}$ for messages of length n. This improves recent results on pattern redundancy, although it does not fill the gap between known lower- and upper-bounds. The pattern of a string is obtained by replacing each symbol by the index of its first occurrence. The problem of pattern coding is of interest because strongly universal codes have been proved to exist for patterns while universal message coding is impossible for memoryless sources on an infinite alphabet. The proof uses fine combinatorial results on partitions with small summands.

Keywords: universal coding; pattern; minimax

1. Introduction

1.1. Universal Coding

Let $\mathbb{P}$ be a stationary *source* on an alphabet A, both known by the coder and the decoder. Let $X = (X_n)_{n\in\mathbb{N}}$ be a random process with distribution $\mathbb{P}$. For a positive integer n, we denote by X_1^n the vector of the n first components of X and by $\mathbb{P}^n$ the distribution of X_1^n on A^n. We denote the logarithm with base 2 by $\log$ and the natural logarithm by $\ln$. Shannon's classical bound [1] states the average bit length of codewords for any coding function is lower-bounded by the *n-th order entropy* $H(X_1^n) = \mathbb{E}[-\log \mathbb{P}^n(X_1^n)]$; moreover, this codelength can be nearly approached, see [2]. One important idea in the proof of this result is the following: every code on the strings of length n is associated with a *coding distribution* q_n on A^n in such a way that the code length for x is $-\log q_n(x)$, *and reciprocally* any distribution q_n on A^n can be associated with a coding function whose code length is approximately $-\log q_n(x)$. When $\mathbb{P}$ is ergodic, its *entropy rate* $H(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1^n)$ exists. It is a tight lower bound on the number of bits required per character.

If $\mathbb{P}$ is only known to be an element $\mathbb{P}_\theta$ of some class $\mathcal{C} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, *universal coding* consists in finding a single code, or equivalently a single sequence of coding distributions $(q_n)_n$, approaching the entropy rate for all sources $\mathbb{P}_\theta \in \mathcal{C}$ at the same time. Such versatility has a price: for any given source $\mathbb{P}_\theta$, there is an additional cost called the *(expected) redundancy* $R(q_n, \theta)$ of the coding distribution q_n that is defined as the difference between the expected code length $\mathbb{E}_\theta\left[-\log q_n(X_1^n)\right]$ and the n-th order entropy $H(X_1^n)$. Two criteria measure the *universality* of q_n:

- First, a deterministic approach judges the performance of q_n in the worst case by the *maximal redundancy* $R^+(q_n, \Theta) = \sup_{\theta\in\Theta} R(q_n, \theta)$ The lowest achievable maximal redundancy is called *minimax redundancy*:
$$R^+(n, \Theta) = \min_{q_n} \max_{\theta} R(q_n, \theta)$$
- Second, a Bayesian approach consists in providing Θ with a prior distribution π, and then considering the expected redundancy $\mathbb{E}_\pi[R(q_n, \theta)]$ (the expectation is here taken over θ). Let q_n^π be the coding distribution minimizing $\mathbb{E}_\pi[R(q_n, \theta)]$ The *maximin redundancy* $R^-(n, \Theta)$ of class $\mathcal{C}$ is the supremum of all $\mathbb{E}_\pi[R(q_n^\pi, \theta)]$ over all possible prior distributions π:
$$R^-(n, \Theta) = \max_{\pi} \min_{q_n} \mathbb{E}_\pi[R(q_n, \theta)]$$

A classical minimax theorem (see [3]) states that mild hypotheses are sufficient to ensure that $R^-(n, \Theta) = R^+(n, \Theta)$. Class $\mathcal{C}$ is said to be *strongly universal* if $R^+(n, \Theta) = o(n)$: then universal coding is possible uniformly on $\mathcal{C}$. An important result by Rissanen [4] asserts that if the parameter set Θ is k-dimensional, and if there exists a $\sqrt{n}-$consistent estimator for θ, then

$$R^-(n, \Theta) = R^+(n, \Theta) = \frac{k}{2} \log n + O(1) \tag{1}$$

This well-known bound has many applications in information theory, often related to the Minimum Description Length Principle. It is remembered as a "rule of thumb" that redundancy is $1/2 \log n$ for each parameter of the model. This result actually covers a large variety of cases, among others: memoryless processes, Markov chains, Context tree sources, hidden Markov chains. However, further generalization have been investigated. Shields (see [5]) proved that no coder can achieve a non-trivial redundancy rate on all stationary ergodic processes. Csiszár and Shields [6] gave an example of a non-parametric, intermediate complexity class, known as *renewal processes*, for which $R^-(n, \Theta)$ and $R^+(n, \Theta)$ are both of order $O(\sqrt{n})$. If alphabet A is not known, or if its size is not insignificant compared to n, Rissanen's bound (1) is uninformative. If the alphabet A is infinite, Kieffer [7] showed that no universal coding is possible even for the class of memoryless processes.

1.2. Dictionary and Pattern

Those negative results prompted the idea of coding separately the *structure* of string x and the symbols present in x. It was first introduced by Åberg in [8] as a solution to the *multi-alphabet* coding problem, where the message x contains only a small subset of the known alphabet A. It was further studied and motivated in a series of articles by Shamir [9–12] and by Jevtić, Orlitsky, Santhanam and Zhang [13–16] for practical applications: the alphabet is unknown and has to be transmitted separately anyway (for instance, transmission of a text in an unknown language), or the alphabet is very large in comparison to the message (consider the case of images with $k = 2^{24}$ colors, or texts when taking words as the alphabet units).

To explain the notion of pattern, let us take the example of [9]: string $x =$ "abracadabra" is made of $n = 11$ characters. The information it conveys can be separated in two blocks:

- a *dictionary* $\Delta = \Delta(x)$ defined as the sequence of different characters present in x in order of appearance; in the example $\Delta = (a, b, r, c, d)$.
- a *pattern* $\psi = \psi(x)$ defined as the sequence of positive integers pointing to the *indices* of each letter in Δ; here, $\psi = 12314151231$.

Let $\mathcal{P}^n$ be the set of all possible patterns of n-strings. For instance, $\mathcal{P}^1 = \{1\}$, $\mathcal{P}^2 = \{11, 12\}$, $\mathcal{P}^3 = \{111, 112, 121, 122, 123\}$. Using the same notations as in [15], we call multiplicity $\mu_j(\psi)$ of symbol j in pattern $\psi \in \mathcal{P}^n$ the number of occurrences of j in ψ; the *multiplicity of pattern* ψ is the vector made of all symbol's multiplicities: $\mu(\psi) = (\mu_j(\psi))\, 1 \leqslant j \leqslant n$—in the former example, $\mu = (5, 2, 2, 1, 1, 0, \ldots)$. Note that $\sum_{j=1}^{n} \mu_j = n$. Moreover, the *profile* $\phi = (\phi_\mu)_{\mu \geqslant 1}$ of pattern ψ provides, for every multiplicity μ, its frequency in $\mu(\psi)$. It can be formally defined as the multiplicity of ψ's multiplicity: $\mu\left(\mu(\psi)\right)$. The profile of string "abracadabra" is $(2, 2, 0, 0, 1, 0, \ldots)$ as two symbols (c and d) appear once, two symbols (b and r) appear twice and one symbol (a) appears five times. We denote by Φ^n the set of possible profiles for patterns of length n, so that $\Phi^1 = \{(1)\}$, $\Phi^2 = \{(2, 0), (0, 1)\}$, $\Phi^3 = \{(3, 0, 0), (1, 1, 0), (0, 0, 1)\}$. Note that $\sum_{\mu=1}^{n} \mu\phi_\mu = n$. As explained in [15], there is one-to-one mapping between Φ^n and the set of *unordered partitions* of integer n. In Section 3., this point will be used and specified.

1.3. Pattern Coding

Any process X from a source $\mathbb{P}_\theta$ induces a pattern process $\Psi = (\Psi_n)_{n\in\mathbb{N}}$ with marginal distributions on $\mathcal{P}^n$ defined by $\mathbb{P}_\theta\left(\Psi_1^n = \psi\right) = \sum_{\psi(x)=\psi} \mathbb{P}_\theta\left(X_1^n = x\right)$. Thus, we can define a *n-th block pattern entropy* $H(\Psi_1^n) = \mathbb{E}_\theta\left[-\log \mathbb{P}_\theta(\Psi_1^n)\right]$. For stationary ergodic $\mathbb{P}_\theta$, Orlitsky & al. [16] prove that the *pattern entropy rate* $H(\Psi) = \lim_{n\to\infty} \frac{1}{n} H(\Psi_1^n)$ exists and is equal to $H(X)$ (whether this quantity is finite or not). This result was independently discovered by Gemelos and Weissman [17].

In the sequel, we shall consider only the case of *memoryless* sources $\mathbb{P}_\theta$, with marginal distributions p_θ on a (possibly infinite) alphabet $\mathcal{A}$. Hence, Θ will be the set parameterizing all probability distributions on $\mathcal{A}$.

Obviously, the process they induce on $(\mathcal{P}^n)_{n\in\mathbb{N}}$ is not memoryless. But as patterns convey less information than the initial strings, coding them seems to be an easier task. The *expected pattern redundancy* of a coding distribution q_n on $\mathcal{P}^n$ can be defined by analogy as the difference between the expected code length under distribution $\mathbb{P}_\theta$ and the n-th block pattern entropy:

$$\begin{aligned} R_\Psi(q_n,\theta) &= \mathbb{E}_\theta\left[-\log q_n(\Psi_1^n)\right] - H(\Psi_1^n) \\ &= \sum_{\psi\in\Psi^n} \mathbb{P}_\theta(\psi) \log \frac{\mathbb{P}_\theta(\psi)}{q_n(\psi)} \end{aligned}$$

As the alphabet is unknown, the *maximal pattern redundancy* $R_\Psi^+(q_n,\Theta)$ must be defined as the maximum of $R_\Psi^+(q_n,\theta)$ over *all alphabets* A *and all memoryless distributions on* A. Of course, the *minimax pattern redundancy* $R_\Psi^+(n,\Theta)$ is defined as the lower-bound of $R_\Psi^+(q_n,\Theta)$ in q_n. Similarly, the *maximin pattern redundancy* $R_\Psi^-(n,\Theta)$ is defined as the supremum with respect to all possible alphabets A and all prior distributions π of the lowest achievable average redundancy, that is:

$$R_\Psi^-(n,\Theta) = \sup_{A,\pi} \inf_{q_n} \mathbb{E}_\pi[R_\Psi(q_n,\theta)]$$

2. Theorem

There is still uncertainty on the true order of magnitude of $R_\Psi^-(n,\Theta)$ and $R_\Psi^+(n,\Theta)$. However, Orlistky & *et al.* in [15] and Shamir in [11] proved that for some constants c_1 and c_2 it holds that $c_1 n^{1/3-\epsilon} \leqslant R_\Psi^-(n,\Theta) \leqslant R_\Psi^+(n,\Theta) \leqslant c_2\sqrt{n}$. There is hence a gap between upper- and lower-bounds. This gap has been reduced in an article by Shamir [10] where the upper-bound is improved to $O\left(n^{2/5}\right)$. The following theorem contributes to the evaluation of $R_\Psi^-(n,\Theta)$, by providing a slightly better and more explicit lower-bound, the proof of which is particularly elegant.

Theorem 1 *For all integers n large enough, the maximin pattern redundancy is lower-bounded as:*

$$R_\Psi^-(n,\Theta) \geqslant 1.84 \left(\frac{n}{\log n}\right)^{1/3}$$

Gil Shamir [18] suggests that a bound of similar order can be obtained by properly updating (B12) in [11]. The proof provided in this paper was elaborated independently; both of them use the channel capacity inequality described in Section 3.. However, it is interesting to note that they rely on different ideas (unordered partitions of integers and Bernstein's inequality here, sphere packing arguments or inhomogeneous grids there). An important difference appears in the treatment of the quantization, see Equation 2. [11] provides fine relations between the minimax average redundancy and the alphabet size. The approach presented here does not discriminate between alphabet sizes; in a short and elegant proof, it leads to a slightly better bound for infinite alphabets.

3. Proof

We use here standard technique for lower-bounds (see [19]): the n-th order maximin redundancy is bounded from below by (and asymptotically equivalent to) the capacity of the channel joining an input variable W with distribution π on Θ to the output variable Ψ_1^n with conditional probabilities $\mathbb{P}_\theta(\Psi_1^n)$. Let $H(\Psi_1^n|W)$ be the conditional entropy of Ψ_1^n given W, and let $I\left(\Psi_1^n;W\right) = H(\Psi_1^n) - H(\Psi_1^n|W)$ denote the mutual information of these two random variables, see [2]. Then from [19] and [4] we know that inequality

$$R_\Psi^-(n,\Theta) \geqslant I\left(\Psi_1^n;W\right)$$

holds for all alphabets $\mathcal{A}$ and all prior distributions π on the set of memoryless distributions on $\mathcal{A}$: it is sufficient to give a lower-bound for the mutual information $I\left(\Psi_1^n; W\right)$ between parameter W and observation Ψ. In words, $R_{\Psi}^{-}(n, \Theta)$ is larger than the logarithm of the number of memoryless sources that can be *distinguished* from one observation of Ψ_1^n.

Given the positive integer n, let $c = c_n$ be an integer growing with n to infinity in a way defined later, let λ be a positive constant to be specified later, let $d = \lambda\sqrt{c}$ and let $\mathcal{A} = \{1, \ldots, c\}$ We denote by $\Theta^{c,d}$ the set of all unordered partitions of c made of summands at most equal to d:

$$\Theta^{c,d} = \left\{\theta = (\theta_j)_{j\in\mathbb{N}^+} : d \geqslant \theta_1 \geqslant \theta_2 \geqslant \ldots \text{ and } \sum_{j=1}^{\infty} \theta_j = c\right\}$$

Then $\Theta^c \triangleq \Theta^{c,c}$ is the set of all unordered partitions of c. Let also $\Phi^{c,d}$ be the subset of Φ^c containing the profiles of all patterns $\psi \in \mathcal{P}^c$ whose symbols appear at most d times:

$$\Phi^{c,d} = \left\{\phi = (\phi_1, \ldots, \phi_d) \in \mathbb{N}^d : \sum_{\mu=1}^{d} \mu\phi_\mu = c\right\}$$

There is a one-to-one mapping χ_c between Θ^c and Φ^c defined by

$$\left\{\begin{array}{lll} \chi_c(\theta)_\mu & = & \left|\{i : \theta_i = \mu\}\right|; \\ \chi_c^{-1}(\phi)_j & = & \left\{\begin{array}{ll} 0 & \text{if } \sum_{i=1}^{d} \phi_i < j, \\ \max\{\mu : \sum_{i=\mu}^{d} \phi_i \geqslant j\} & \text{else.} \end{array}\right. \end{array}\right.$$

It is immediately verified that $\chi\left(\Theta^{c,d}\right) = \Phi^{c,d}$. In [20], citing [21], Dixmier and Nicolas show the existence of an increasing function $f : \mathbb{R}^+ \to \left[0, \pi\sqrt{\frac{2}{3}}\right[$ such that $\ln\left|\Theta^{c,d}\right| = f(\lambda)\sqrt{c}\,(1 + o(1))$ as $c \to \infty$, where $\lambda = d/\sqrt{c}$. Numerous properties of function f, and numerical values, are given in [20]; notably, f is an infinitely derivable and concave function which satisfies $f(\lambda) = -2\lambda\log\lambda + 2\lambda + O(\lambda^3)$ when $\lambda \to 0$ and $f(\lambda) = \pi\sqrt{2/3} - \sqrt{6}/\pi\exp(-\pi\lambda/\sqrt{6})$ when $\lambda \to \infty$.

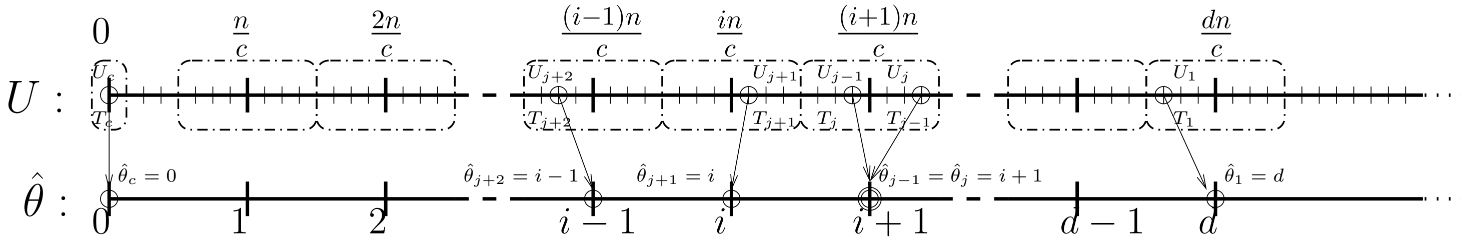
For $\theta \in \Theta^{c,d}$, let p_θ be the distribution on $\mathcal{A}$ defined by $p_\theta(i) = \frac{\theta_i}{c}$, and let $\mathbb{P}_\theta$ be the memoryless process with marginal distribution p_θ. Let W be a random variable with uniform distribution on the set $\Theta^{c,d}$. Let $X = (X_n)_{n\in\mathbb{N}^+}$ be a random process such that conditionally on the event $\{W = \theta\}$, then the distribution of X is $\mathbb{P}_\theta$, and let $\Psi = (\Psi_n)_{n\in\mathbb{N}^+}$ be the induced pattern process.

We want to bound $I(\Psi_1^n; W) = H(W) - H(W|\Psi_1^n)$ from below. As

$$H(W) = \log\left|\Theta^{c,d}\right| = f(\lambda)\log e\sqrt{c}\,(1 + o(1))$$

we need to find an upper-bound for $H(W|\Psi_1^n)$. The idea of the proof is the following. ¿From Fano's inequality, upper-bounding $H(W|\Psi_1^n)$ reduces to finding a good estimator $\hat{\theta}$ for W: conditionally on $W = \theta$, string X_1^n is a memoryless process with distribution $\mathbb{P}_\theta$ and we aim at recovering parameter θ from its pattern Ψ_1^n. Each parameter $\theta = (\theta_j)_{j\geqslant 1}$ is here an unordered partition with small summands of integer c. Let T_j be the number of occurrences of j-th most frequent symbol in ψ. Then $T = (T_j)_{j\geqslant 1}$ constitutes a random unordered partition of n. We show that by "shrinking" T by a factor c/n we build a unordered partition $\hat{\theta}$ of c that is equal to parameter θ with high probability, see Figure 1. Note that only partitions with small summands are considered: this allows to have a better uniform control on the probabilities of deviation of each symbol's frequency, while the cardinality of $\Theta^{c,d}$ remains of same (logarithmic) order as that of Θ^c. Parameters c and d are chosen in order to optimize the rate in Theorem 1, while the value of $\lambda = d/\sqrt{c}$ is chosen at the end to maximize the constant.

Figure 1. The profile of pattern ψ forms a partition of n that can be "shrunk" to θ, the parameter partition of c, with high probability.



Let us now give the details of the proof. If $W = \theta$ and if we observe string $X_1^n = x$ having pattern $\Psi_1^n = \psi \in \mathcal{P}^n$, we construct an estimator $\hat{\theta} = \left(\hat{\theta}_j\right)_{1\leqslant j\leqslant c}$ of θ in the following way: let $\phi(\psi)$ be the profile of ψ, and $T = (T_j)_{j\geqslant 1} = \chi_n^{-1}(\phi(\psi))$ be the corresponding partition of n. For $j \geqslant c$, let $\hat{\theta}_j = \left[\frac{T_j c}{n}\right]$, where $[x]$ denotes the nearest integer of x. Observe that as alphabet $\mathcal{A}$ contains only c different symbols, for all $j > c$ we have $T_j = \hat{\theta}_j = \theta_j = 0$.

The distribution of T is difficult to study, but is very related to much simpler random variables. For $1 \leqslant i \leqslant n$ and $j \geqslant 1$, let $U_j^i = \mathbb{1}_{X_i = j}$; as U_j^i has a Bernoulli distribution with parameter $\frac{\theta_j}{c}$, and as process X is memoryless, we observe that $U_j \triangleq \sum_{i=1}^n U_j^i$, the number of occurrences of symbol j in x, has a binomial distribution $\mathcal{B}\left(n, \frac{\theta_j}{c}\right)$. Let $\tilde{\theta}_j = \left[\frac{U_j c}{n}\right]$, and $\tilde{\theta} = \left(\tilde{\theta}_j\right)_{j\geqslant 1}$; $\tilde{\theta}$ would be an estimator of θ if we had access to x, but here estimators may only be constructed from ψ. However, there is a strong connection between $\hat{\theta}$ and $\tilde{\theta}$: the symbols in x are in one-to-one correspondence with the symbols in ψ. Hence, T is just the order statistics of U: $T_j = U_{(j)}$ and thus $\hat{\theta}_j = \tilde{\theta}_{(j)}$.

Now, if $\left|\frac{U_j c}{n} - \theta_j\right| < \frac{1}{2}$ then $\tilde{\theta}_j = \theta_j$. Thus, if for all j in the set $\{1, \dots, c\}$ it holds that $\left|\frac{U_j c}{n} - \theta_j\right| < \frac{1}{2}$, then $\tilde{\theta} = \theta$ and $\tilde{\theta}$, as an increasing sequence, is equal to its order statistics $\hat{\theta}$. It follows that

$$\bigcap_{j=1}^{c} \left\{ \left| \frac{U_j c}{n} - \theta_j \right| < \frac{1}{2} \right\} \subset \left\{ \hat{\theta} = \theta \right\} \tag{2}$$

and hence, using the union bound:

$$\mathbb{P}_\theta(\hat{\theta} \neq \theta) \leqslant \mathbb{P}_\theta\left(\bigcup_{j=1}^{c} \left\{ \left| \frac{U_j c}{n} - \theta_j \right| \geqslant \frac{1}{2} \right\} \right) \leqslant \sum_{j=1}^{c} \mathbb{P}_\theta\left(\left| \frac{U_j}{n} - \frac{\theta_j}{c} \right| \geqslant \frac{1}{2c} \right) \tag{3}$$

We chose parameter set $\Theta^{c,d}$ so that all summands in partition θ are small with respect to c. Consequently, the variance of the $\left(U_j^i\right)_{i,j}$ is uniformly bounded: $\mathrm{Var}[U_j^i] = \frac{\theta_j}{c}\left(1 - \frac{\theta_j}{c}\right) \leqslant \frac{d}{c}$. Recall the following Bernstein inequality [22]: if $Y_1, \dots, Y_n$ are independent random variables such that Y_i takes its values in $[-b, b]$ and such that $\mathrm{Var}[Y_i] \leqslant v$, and if $S = Y_1 + \cdots + Y_n$, then for any positive x it holds that:

$$\mathbb{P}\left(S - \mathbb{E}[S] \geqslant x\right) \leqslant \exp\left(-\frac{x^2/2}{n(v + x/3)} \right)$$

Using this inequality for the $\left(U_j^i\right)_{1\leqslant i\leqslant n}$, we obtain:

$$\mathbb{P}_\theta\left(\left|\frac{U_j}{n}-\frac{\theta_j}{c}\right|\geqslant\frac{1}{2c}\right)\leqslant 2e^{-\frac{n/4c^2}{2(d/c+1/6c)}}=2e^{-\frac{n}{8c(d+1/6)}}$$

Thus, we obtain from (3):

$$\mathbb{P}(\hat{\theta}\neq\theta)=\frac{1}{|\Theta^{c,d}|}\sum_{\theta\in\Theta^{c,d}}\mathbb{P}_\theta(\hat{\theta}\neq\theta)\leqslant 2ce^{-\frac{n}{8c(d+1/6)}}$$

Now, using Fano's inequality [2]:

$$\begin{aligned}H(W|\Psi_1^n) &\leqslant H(W|\hat{\theta})\\ &\leqslant \mathbb{P}(W\neq\hat{\theta})\log\left|\Theta^{c,d}\right|+\log 2\\ &\leqslant 2ce^{-\frac{n}{8\lambda c^{3/2}}}f(\lambda)\sqrt{c}\,(1+o(1))\end{aligned}$$

Hence,

$$\begin{aligned}R_\Psi^-(n,\Theta) &\geqslant I(\Psi_1^n;W)=H(W)-H(W|\Psi_1^n)\\ &\geqslant f(\lambda)\log e\sqrt{c}\,(1+o(1))-2ce^{-\frac{n}{8\lambda c^{3/2}}}f(\lambda)\log e\sqrt{c}\,(1-o(1))\\ &= f(\lambda)\log e\sqrt{c}\left(1-2ce^{-\frac{n}{8\lambda c^{3/2}}}-o(1)\right)\end{aligned}$$

By choosing $c=\left(\frac{n}{\frac{16}{3}\lambda\log n}\right)^{2/3}$ we get:

$$\begin{aligned}R_\Psi^-(n,\Theta) &\geqslant f(\lambda)\log e\left(\frac{n}{\frac{16}{3}\lambda\log n}\right)^{1/3}\left(1-2\left(\frac{n}{\frac{16}{3}\lambda\log n}\right)^{2/3}e^{-\frac{2}{3}\log n}-o(1)\right)\\ &= \frac{f(\lambda)}{\lambda^{1/3}}\log e\left(\frac{3n}{16\log n}\right)^{1/3}(1-o(1))\end{aligned}$$

By looking at the table of f given at page 151 of [20], we see that function $\lambda\to f(\lambda)/\lambda^{1/3}$ reaches its maximum around $\lambda=0.8$; for that choice, $f(\lambda)\approx 2.07236$ and we obtain:

$$R_\Psi^-(n,\Theta)\geqslant 1.843\left(\frac{n}{\log n}\right)^{1/3}(1-o(1))$$

Acknowledgment

I would particularly like to thank professor Jean-Louis Nicolas from Institut Girard Desargues, who very kindly helped for the combinatorial arguments.

References

1. Shannon, C.E. A mathematical theory of communication. *Bell System Tech. J.* **1948**, *27*, 379–423, 623–656.
2. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons Inc.: New York, NY, USA, 1991.

3. Haussler, D. A general minimax result for relative entropy. *IEEE Trans. Inform. Theory* **1997**, *43*, 1276–1280.
4. Rissanen, J. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory* **1984**, *30*, 629–636.
5. Shields, P.C. Universal redundancy rates do not exist. *IEEE Trans. Inform. Theory* **1993**, *39*, 520–524.
6. Csiszár, I.; Shields, P.C. Redundancy rates for renewal and other processes. *IEEE Trans. Inform. Theory* **1996**, *42*, 2065–2072.
7. Kieffer, J.C. A unified approach to weak universal source coding. *IEEE Trans. Inform. Theory* **1978**, *24*, 674–682.
8. Åberg, J.; Shtarkov, Y.M.; Smeets, B.J. Multialphabet Coding with Separate Alphabet Description. In *Proceedings of Compression and complexity of sequences*; Press, I.C.S., Ed., IEEE: Palermo, Italy, 1997; pp. 56–65.
9. Shamir, G.I.; Song, L. On the entropy of patterns of i.i.d. sequences. In *Proceedings of 41st Annual Allerton Conference on Communication, Control and Computing*; Curran Associates, Inc.: Monticello, IL, USA, 2003; pp. 160–169.
10. Shamir, G.I. A new redundancy bound for universal lossless compression of unknown alphabets. In *Proceedings of the 38th Annual Conference on Information Sciences and Systems - CISS*; IEEE: Princeton, NJ, USA, 2004; pp. 1175–1179.
11. Shamir, G.I. Universal lossless compression with unknown alphabets-the average case. *IEEE Trans. Inform. Theory* **2006**, *52*, 4915–4944.
12. Shamir, G.I. On the MDL principle for i.i.d. sources with large alphabets. *IEEE Trans. Inform. Theory* **2006**, *52*, 1939–1955.
13. Orlitsky, A.; Santhanam, N.P. Speaking of infinity. *IEEE Trans. Inform. Theory* **2004**, *50*, 2215–2230.
14. Jevtić, N.; Orlitsky, A.; Santhanam, N.P. A lower bound on compression of unknown alphabets. *Theoret. Comput. Sci.* **2005**, *332*, 293–311.
15. Orlitsky, A.; Santhanam, N.P.; Zhang, J. Universal compression of memoryless sources over unknown alphabets. *IEEE Trans. Inform. Theory* **2004**, *50*, 1469–1481.
16. Orlitsky, A.; Santhanam, N.P.; Viswanathan, K.; Zhang, J. Limit Results on Pattern Entropy of Stationary Processes. In *Proceedings of the 2004 IEEE Information Theory workshop*; IEEE: San Antonio, TX, USA, 2004; pp. 2954–2964.
17. Gemelos, G.; Weissman, T. On the entropy rate of pattern processes. Technical report hpl-2004-159; HP Laboratories Palo Alto: San Antonio, TX, USA, 2004.
18. Shamir, G.I. From University of Utah, Electrical and Computer Ingeneering. Private communication, 2006.
19. Davisson, L.D. Universal noiseless coding. *IEEE Trans. Inform. Theory* **1973**, *IT-19*, 783–795.
20. Dixmier, J.; Nicolas, J.L., Partitions sans petits sommants. In *A Tribute to Paul Erdös*. Cambridge University Press: New York, NY, USA, 1990; Chapter 8, pp. 121–152.
21. Szekeres, G. An asymptotic formula in the theory of partitions. *Quart. J. Math. Oxford* **1951**, *2*, 85–108.

22. Massart, P. *Ecole d'Eté de Probabilité de Saint-Flour XXXIII*. LNM. Springer-Verlag: London, UK, 2003; Chapter 2.