Sampling with Walsh Transforms

Yi LU
Institute of Software,
Chinese Academy of Sciences,
Beijing, P.R. China, 100190
luyi666@gmail.com

Abstract

With the advent of massive data outputs at a regular rate, admittedly, signal processing technology plays an increasingly important role. Nowadays, signals are not merely restricted to the physical world, they have been extended to cover a much wider range of statistical sources (including financial signals, graph signals).

Under the general assumption of discrete statistical signal sources, in this paper, we propose a practical problem of sampling incomplete signals for which we do not know a priori, with bounded sample size. We approach this sampling problem by Shannon’s communication theory. We use an extremal binary channel with high probability of transmission error, which is rare in communication theory. Nonetheless, the channel communication theory translates it into a very useful statistical result. Our main result demonstrates that it is the large Walsh coefficient(s) that characterize(s) discrete statistical signals, regardless of the signal sources. In particular, when sampling incomplete signals of the same source multiple times, one can expect to see repeatedly those large Walsh coefficient(s) of same magnitude(s) at the fixed frequency position(s). By the connection of channel communication theory, we establish the necessary and sufficient condition for our bounded sampling problem.

1 Introduction

With the advent of massive data outputs at a regular rate, we are confronted by the challenge of big data processing and analysis. Admittedly, signal processing \[\text{[2, 6]}\] has become an increasingly important multi-disciplinary research area \[\text{[12]}\]. One open question is the sampling problem with the signals, for which we assume that we do not know \textit{a priori}. Further, due
to reasons of practical consideration, the signals are *incomplete* in the sense that either they are disturbed by possibly strong noises and/or the precision of the sampling measurements is limited. Assuming that the signal source is not restricted to a particular application domain (e.g., images, voices), we are concerned with a practical problem of sampling these incomplete signals.

In this paper, we show that we can sample incomplete signals without knowing *a priori*. We approach this sampling problem by Shannon’s communication theory [10]. We use the basic theorem of information theory (cf. [11]), i.e., Shannon’s channel coding theorem (a.k.a. Shannon’s second theorem), which establishes the achievability of channel capacity. In the channel communication problem, one wants to send information through the channel with probability of error arbitrarily small. Shannon’s channel coding theorem states that we can achieve that as long as the transmission rate does not exceed the channel capacity; otherwise, the error is bounded away from zero. Clearly, the concept of channel capacity uniquely characterizes a channel.

Our work uses the binary channel, which is common in digital communications. Nonetheless, our channel is assumed to have extremely high probability of transmission error (we call it the extremal binary channel), which is of rare usage in communication theory [11]. In particular, for the Binary Symmetric Channel (BSC) with crossover probability \((1 - d)/2\) and \(d\) is small (i.e., \(|d| \ll 1\)), the channel capacity is approximately \(d^2/(2 \log 2)\), i.e., on the order of \(d^2\). For this extremal BSC, we translate Shannon’s channel coding theorem to a very useful statistical result. Further, we also consider a non-symmetric binary channel with crossover probability \((1 - d)/2\) and \(1/2\) respectively (\(d\) is small). And the channel capacity is approximately one fourth of the former BSC case, i.e., \(d^2/(8 \log 2)\).

Our main result is that, for this extremal non-symmetric binary channel, Shannon’s channel coding theorem can be translated to an even more useful statistical result. This is applicable to our sampling problem with incomplete signals under the general assumption of statistical signal sources (i.e., no further assumption is made about the signal source). With the connection of Shannon’s channel coding theorem, we demonstrate that we can sample the discrete statistical signals. We give necessary and sufficient condition to sample these incomplete signals with bounded sample size. It is interesting to observe that the classical signal processing tool of Walsh transform [27] is essential, *because*, regardless of the real signal source, the large Walsh coefficient(s) characterize(s) discrete statistical signals. That is, when sampling incomplete signals of the same source multiple times, one can expect to see *repeatedly* those large Walsh coefficient(s) of same magni-
tude(s) at the fixed frequency position(s).

The rest of the paper is organized as follows. In Section 2, we give preliminaries on Walsh transforms. In Section 3, we briefly review Shannon’s channel coding theorem. In Section 4, we translate Shannon’s theorem into statistical results for an extremal binary channel. We present main sampling results in Section 5, which shows strong connection between Shannon’s theorem and Walsh transforms. We give concluding remarks in Section 6.

2 Walsh Transforms in Statistics

Given a real-valued function \( f : \text{GF}(2)^n \rightarrow \mathbb{R} \), which is defined on an \( n \)-tuple binary vector of input, the Walsh transform of \( f \), denoted by \( \hat{f} \), is another real-valued function defined as

\[
\hat{f}(x) = \sum_{y \in \text{GF}(2)^n} (-1)^{<x,y>} f(y),
\]

for all \( x \in \text{GF}(2)^n \), where \( <x,y> \) denotes the inner product between two \( n \)-bit vectors \( x, y \). For basic properties and references on Walsh transforms, we refer the reader to [7, 9].

As earlier as 1930’s, in [13], a fast implementation algorithm of Walsh transforms was given, and Walsh transforms were used in statistics to find dependencies within a multi-variable data set, in the context of factorial experiments (see [5] for a general discussion).

Let \( f \) be a probability distribution of an \( n \)-bit random variable \( \mathcal{X} = (X_n, X_{n-1}, \ldots, X_1) \), where each \( X_i \in \{0, 1\} \). Then, \( \hat{f}(m) \) is the bias of the Boolean variable \( <m, \mathcal{X}> \) for any fixed \( n \)-bit vector \( m \), which is often called the output pattern or mask. Recall that a Boolean random variable \( \mathcal{A} \) has bias \( \epsilon \), which is defined by \( \epsilon = \Pr(\mathcal{A} = 0) - \Pr(\mathcal{A} = 1) \). Note that if \( \mathcal{A} \) is uniformly distributed, \( \mathcal{A} \) has bias 0. Obviously, the pattern \( m \) should be nonzero.

In the multi-variable tests, each \( X_i \) indicates the presence or absence (represented by ‘1’ or ‘0’) of a particular factor in an agricultural experiment, or it indicates the presence or absence of a particular feature in a pattern recognition experiment. Fast Walsh Transform (FWT) is used to obtain all coefficients \( \hat{f}(m) \) in one shot. By checking the Walsh coefficients one by one and identifying the large ones, given that the linear relation \( <m, \mathcal{X}> = 0 \) holds with bias \( \hat{f}(m) \), which stands out in magnitude from the rest, we are able to tell the dependencies among \( X_i \)’s.
In a more typical and general setting of modern applied statistics, Walsh transform is commonly used to test whether or not the $n$-bit random variable $X$ is uniformly distributed. Usually, $X$ is the output of a function $F$. One needs to collect the probability distribution $f$ of $X$ by exhaustively going through all the possible inputs of $F$. After that, exactly the same procedure follows as aforementioned multi-variable tests in order to see how uniformly distributed $X$ is by spotting out large coefficients $\hat{f}(m)$. Obviously, the larger $\hat{f}(m)$ is, the less uniformly distributed $X$ is.

Traditionally, $F$ is a deterministic function with small or medium input size. Consequently, it is computationally easy to collect the complete and precise probability distribution $f$. And classical statistical techniques are established to analyze $X$, given its probability distribution a priori. Nevertheless, it becomes increasingly prevailing that $F$ in reality have large input size, which is beyond limit of exhaustive enumeration. More often than not, the old assumption that $F$ is a deterministic function with small or medium input size simply does no longer hold. As a matter of fact, $F$ might be such a function that we do not have complete description, or $F$ might be a non-deterministic function, or it might just have large input size. Thus, it is infeasible to collect the complete and precise probability distribution $f$ in order to perform statistical analysis. Before we present statistical sampling technique with incomplete signals as reliable as possible, we introduce the necessary tools from Shannon’s communication theory [10] in next section.

3 Review on Shannon’s Channel Coding Theorem

We will briefly recall Shannon’s famous channel coding theorem (cf. [4]). First, we review some basic definitions of Shannon entropy. The entropy $H(X)$ of a discrete random variable $X$ with alphabet $\mathcal{X}$ and probability mass function $p(x)$ is defined by

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x).$$

The joint entropy $H(X_1, \ldots, X_n)$ of a collection of discrete random variables $(X_1, \ldots, X_n)$ with a joint distribution $p(x_1, x_2, \ldots, x_n)$ is defined by

$$H(X_1, \ldots, X_n) = - \sum_{x_1, x_2, \ldots, x_n} p(x_1, x_2, \ldots, x_n) \log_2 p(x_1, x_2, \ldots, x_n).$$
Define the conditional entropy $H(Y|X)$ of a random variable $Y$ given another $X$ as

$$H(Y|X) = \sum_x p(x)H(Y|X=x).$$

The mutual information $I(X;Y)$ between two random variables $X, Y$ is equal to $H(Y) - H(Y|X)$, which always equals $H(X) - H(X|Y)$. A communication channel is a system in which the output $Y$ depends probabilistically on its input $X$. It is characterized by a probability transition matrix that determines the conditional distribution of the output given the input.

**Theorem 1** (Shannon’s Channel Coding Theorem). Given a channel, denote the input, output by $X, Y$ respectively. We can send information at the maximum rate $C$ bits per transmission with an arbitrarily low probability of error, where $C$ is the channel capacity defined by

$$C = \max_{p(x)} I(X;Y),$$

and the maximum is taken over all possible input distributions $p(x)$.

For the binary symmetric channel (BSC) with crossover probability $p$ (i.e., the input symbols are complemented with probability $p$), $C$ can be simplified to the following result (cf. [4]):

$$C = 1 - H(p) \text{ bits/transmission.} \quad (3)$$

Let $p = (1 + d)/2$ and so $|d| \leq 1$. We have a useful numerical approximation result of $H((1 + d)/2)$ for $|d| < 1$ (see Appendix for proof):

$$H\left(\frac{1+d}{2}\right) = 1 - \left(\frac{d^2}{2} + \frac{d^4}{12} + \frac{d^6}{30} + \frac{d^8}{56} + \cdots\right) \times \frac{1}{\log 2}. \quad (4)$$

Note that if $|d| \ll 1$, (4) reduces to the following simplified form of approximation

$$H\left(\frac{1+d}{2}\right) = 1 - d^2/(2 \log 2) + O(d^4). \quad (5)$$

For BSC with crossover probability $p = (1 + d)/2$ and $d$ is small (i.e., $|d| \ll 1$), which we call an extremal BSC, we can calculate $C$ in (3) by

$$C = 1 - H\left(\frac{1+d}{2}\right) = \left(d^2 + O(d^4)\right)/(2 \log 2) \approx \frac{d^2}{2 \log 2}.$$ 

Thus, we have shown the result of the channel capacity of an extremal BSC:

$$5$$
Corollary 1 (extremal BSC). Given a BSC channel with crossover probability \( p = (1 + d)/2 \). If \( d \) is small (i.e., \( |d| \ll 1 \)), then, \( C \approx c_0 \cdot d^2 \), where the constant \( c_0 = 1/(2 \log 2) \).

Therefore, for an extremal BSC, we can send one bit with an arbitrarily low probability of error with the minimum number of transmissions \( 1/C = (2 \log 2)/d^2 \), i.e., \( O(1/d^2) \). In next section, we will translate Corollary 1 to useful statistical results, which deals with the case \( |d| \ll 1 \). Interestingly, in communication theory, this extremal BSC is rare because of the low efficiency \[1\] and we typically have \( |d| \gg 0 \).

4 Statistical Translations of Shannon’s Theorem

Let \( X_0, X_1 \) denote the Boolean random variable with bias \(+d, -d\) respectively (and we restrict ourselves to \( |d| \ll 1 \) in this section). Denote the probability distribution of \( X_0, X_1 \) by \( D_0, D_1 \) respectively. Let \( D \in \{D_0, D_1\} \).

We are given a binary sequence of random bits with length \( N \), and each bit is independent and identically distributed (i.i.d.) following the distribution \( D \). As a consequence of Shannon’s channel coding theorem, we can answer the question of the minimum \( N \) required to decide whether \( D = D_0 \) or \( D = D_1 \) with an arbitrarily low probability of error.

We translate this problem into a BSC channel coding problem as follows. The inputs are transmitted through a BSC with error probability \( p = (1 - d)/2 \). By Shannon’s channel coding theorem, with a minimum number of \( N = 1/C \) transmissions, we can reliably (i.e., with an arbitrarily low probability of error) determine whether the input is ‘0’ or ‘1’. The former case implies that the received sequence corresponds to the distribution \( D_0 \) (i.e., a bit ‘1’ occurs in the output sequence with probability \( p \)), while the latter case implies that the received sequence corresponds to the distribution \( D_1 \) (i.e., a bit ‘0’ occurs in the output sequence with probability \( p \)). This solves the problem stated in last paragraph. Using Corollary 1 with \( p = (1 - d)/2 \) (for \( |d| \ll 1 \)), we have \( N = (2 \log 2)/d^2 \), i.e., \( O(1/d^2) \).

Thus, we have just shown that Shannon’s channel coding theorem can be translated to the following result in statistics:

Theorem 2. Assume that the boolean random variable \( A, B \) has bias \(+d, -d\) respectively and \( d \) is small. We are given a sequence of random samples, which are i.i.d. following the distribution of either \( A \) or \( B \). We can identify the sample source with an arbitrarily low probability of error, using the minimum number \( N \) of samples \( (2 \log 2)/d^2 \), i.e., \( O(1/d^2) \).
Further, the following variant is more frequently encountered in statistics, in which we have to deal with a biased distribution and a uniform distribution:

**Theorem 3.** Assume that the boolean random variable $A$ has bias $d$ and $d$ is small. We are given a sequence of random samples, which are i.i.d. following the distribution of either $A$ or a uniform distribution. We can identify the sample source with an arbitrarily low probability of error, using the minimum number $N$ of samples $(8\log 2)/d^2$, i.e., $O(1/d^2)$.

**Proof.** It is clear that previous construction of using a BSC does not work here, as the biases (i.e., $d, 0$) of the two sources are non-symmetric. Thus, we propose to use Shannon’s channel coding theorem with a non-symmetric binary channel rather than a BSC.

Assume the channel with the following transition matrix

$$p(y|x) = \begin{pmatrix} 1 - p_e & p_e \\ 1/2 & 1/2 \end{pmatrix},$$

where $p_e = (1 - d)/2$ and $d$ is small. The matrix entry in the $x$th row and the $y$th column denotes the conditional probability that $y$ is received when $x$ is sent. So, the input bit 0 is transmitted by this channel with error probability $p_e$ (i.e., the received sequence has bias $d$ if input symbols are 0) and the input bit 1 is transmitted with error probability $1/2$ (i.e., the received sequence has bias 0 if input symbols are 1).

To compute the channel capacity $C$ (i.e., to find the maximum) defined in (2), no closed form solution exist in general. Nonlinear optimization algorithms [1, 3] are known to find a numerical solution. Below, we propose a simple method to give a closed form estimate $C$ for our extremal binary channel. As $I(X;Y) = H(Y) - H(Y|X)$, we first compute $H(Y)$ by

$$H(Y) = \frac{1}{2} \log \left( p_0 (1 - p_e) + (1 - p_0) \times \frac{1}{2} \right),$$

where $p_0$ denote $p(x = 0)$ for short. Next, we compute $H(Y|X)$ as follows,

$$H(Y|X) = \sum_x p(x) H(Y|X = x)$$

$$= p_0 \left( H(p_e) - 1 \right) + 1. \quad (7)$$

Combining (6) and (7), we have

$$I(X;Y) = H \left( p_0 \times \frac{1}{2} - p_0 p_e + \frac{1}{2} \right) - p_0 H(p_e) + p_0 - 1.$$
As $p_e = (1 - d)/2$, we have

$$I(X;Y) = H\left(\frac{1 + p_0d}{2}\right) - p_0\left(H\left(\frac{1 - d}{2}\right) - 1\right) - 1.$$  

We apply (5)

$$I(X;Y) = -\frac{p_0^2d^2}{2\log 2} - p_0\left(H\left(\frac{1 - d}{2}\right) - 1\right) + O(p_0^4d^4),$$  

(8)

for small $d$. Note that the last term $O(p_0^4d^4)$ on the right side of (8) is ignorable. Thus, $I(X;Y)$ approaches the maximum when

$$p_0 = -\frac{H(1 - \frac{d}{2}) - 1}{d^2/(\log 2)} \approx \frac{d^2/(2\log 2)}{d^2/(\log 2)} = \frac{1}{2}. $$

Consequently, we estimate the channel capacity from (8) by

$$C \approx -\frac{1}{4}d^2/(2\log 2) + \frac{1}{2}\left(1 - H\left(\frac{1 - d}{2}\right)\right) \approx -d^2/(8\log 2) + d^2/(4\log 2),$$

which is $d^2/(8\log 2)$.

5 Sampling with Incomplete Signals

Assume a general setting of discrete statistical signals. Our statistical signals $X$ can be modeled as follows. Let $X$ be the $n$-bit output of an arbitrary and fixed function $F$, assuming that the input is random and uniformly distributed. We are given a sequence of i.i.d. random samples $X$. The sample source comes from either the output of $F$ or a uniform distribution. Our first sampling problem is to try to identify the sample source with an arbitrarily low probability of error, using minimum samples. Then, we propose a more practical sampling problem. Assuming that it is infeasible to know $F$ a priori, we want to identify the sample source with an arbitrarily low probability of error and with bounded sample size.

We first consider the case of a Boolean function $F$ (i.e., $n = 1$).

5.1 Special Case: $F$ with 1-bit output

**Theorem 4.** Assume that the Boolean function $F$ has bias $d$. We can identify $F$ with an arbitrarily low probability of error, using minimum number $N$ of samples $N = (8\log 2)/d^2$, i.e., $O(1/d^2)$. 

8
Proof. Assuming that the input is random and uniformly distributed, we consider the output bit $X$ of $F$ as a Boolean random variable. Let $X$ take the distribution $f$. We apply Theorem 3 to conclude our proof.

Remarks. We point out that, Theorem 4 and the generalized version Theorem 5 (in next subsection) serve as a very useful tool in the area of cryptography [8]. Theorem 4 assumes that $F$ (and the bias $d$) is known a priori. In Corollary 2 below, we will answer a more practical sampling problem of identifying a Boolean function $F$ without knowing a priori by limited sampling measurements.

Corollary 2. Assume that the sample size is upper-bounded by $N$. Regardless of the input size of $F$, we can identify $F$ with an arbitrarily low probability of error, whose bias $d$ satisfies $|d| \geq c/\sqrt{N}$, where the constant $c = \sqrt{8 \log 2}$.

Corollary 2 follows naturally from Theorem 4. It states the necessary condition on the bias $d$ in order to identify the unknown $F$, given the upper-bound on the sample size. In fact, this is the necessary and sufficient condition under the general assumption of statistical signal sources. The reason for being the sufficient condition follows from Shannon’s channel coding theorem, which states that when we send information at the rate above the channel capacity, the error is bounded away from zero. Therefore, if $|d| < c/\sqrt{N}$, following our proof to Theorem 3 our problem in this section translates back to sending information at the rate above the channel capacity and the error is bounded away from zero. Consequently, we have shown that the condition on the bias $d$ is sufficient.

It seems that Theorem 4 and Corollary 2 have nothing to do with Walsh transforms so far. We will see that this is not true next.

5.2 General Case: $F$ with $n$-bit output

Theorem 5. Assume that the largest Walsh coefficient of the output distribution $f$ of $F$ is $d = \hat{f}(m_0)$ for a nonzero $n$-bit vector $m_0$. We can identify $F$ with an arbitrarily low probability of error, using minimum number $N$ of samples $N = (8 \log 2)/d^2$, i.e., $O(1/d^2)$.

Proof. Given a fixed $n$-bit vector $m \neq 0$, we consider the Boolean function $F_m(\cdot) = \langle m, F(\cdot) \rangle$. Following Section 2 the bias of $F_m$ is equal to $\hat{f}(m)$. As we have a total of $(2^n - 1)$ nonzero vectors $m$, we apply Theorem 3 to
conclude that

\[ N = \min_{m \neq 0} \frac{8 \log 2}{( \hat{f}(m))^2} \]

As \( d \) is the largest Walsh coefficient \( \hat{f}(m) \) over all nonzero \( n \)-bit \( m \), we finish our proof.

Now, it is easy to see that for \( n = 1 \), the only nontrivial Walsh coefficient is \( \hat{f}(1) \), which is nothing but the bias of \( F \). Thus, Theorem 4 is a special case of Theorem 5.

In analogy to Corollary 2, Corollary 3 below follows naturally from Theorem 5. Under the general assumption of statistical signal sources, it states the necessary and sufficient condition on the Walsh coefficient(s) \( d \) of the output distribution of an unknown \( F \) in order to identify it with bounded sample size.

**Corollary 3.** Assume that the sample size is upper-bounded by \( N \). Regardless of the input size of \( F \), we can identify \( F \) with an arbitrarily low probability of error, whose output distribution has some nontrivial Walsh coefficient \( d \) with \(|d| \geq c/\sqrt{N}\), where the constant \( c = \sqrt{8 \log 2} \).

## 6 Concluding Remarks

We model a general discrete statistical signals as the output samples of an unknown arbitrary but fixed function (which is the signal source). We translate the channel communication theory in the extremal case of a binary channel into a very useful statistical result. Due to high probability of transmission error, this extremal binary channel is of rare usage in communication theory. Nonetheless, the translated statistical result allows to solve a certain class of sampling problems, where we know nothing about the signal source \textit{a priori} and we can only afford bounded sampling measurements. Our main result demonstrates that the classical signal processing tool of Walsh transform is essential. In particular, it is the large Walsh coefficient(s) that characterize(s) discrete statistical signals, regardless of the signal source. More specifically, when sampling \textit{incomplete} signals of the same source multiple times, one can expect to see \textit{repeatedly} those large Walsh coefficient(s) of same magnitude(s) at the fixed frequency position(s). By the connection of channel communication theory, we establish the necessary and sufficient condition for our bounded sampling problem under the general assumption of statistical signal sources.
References

[1] S. Arimoto, An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Trans. Inform. Theory, IT-18 14-20, 1972.

[2] J. M. Blackledge. Digital Signal Processing – Mathematical and Computational Methods, Software Development and Applications. Horwood Publishing, England, Second edition, 2006.

[3] R. Blahut, Computation of Channel capacity and rate distortion functions. IEEE Trans. Inform. Theory, IT-18: 460-473, 1972.

[4] T. M. Cover, J. A. Thomas. Elements of Information Theory. John Wiley & Sons, New York, 1991.

[5] M. F. Driscoll, On Pairwise and Mutual Independence: Characterization of Rectangular Distributions, Journal of the American Statistical Association, Vol. 73, pp. 432-433, 1978.

[6] R. M. Gray, L. D. Davisson. An Introduction to Statistical Signal Processing. Cambridge University Press, 2004. http://www-ee.stanford.edu/~gray/sp.pdf

[7] K. J. Horadam. Hadamard Matrices and Their Applications. Princeton University Press, 2007.

[8] A. Joux. Algorithmic Cryptanalysis. Chapman & Hall/CRC, Cryptography and Network Security, 2009.

[9] Y. Lu, Y. Desmedt, Walsh transforms and cryptographic applications in bias computing, submitted. http://sites.google.com/site/yilusite/Lu-Desmedt_preprint2014.pdf

[10] C. E. Shannon. A mathematical theory of communication. Bell Sys. Tech. Journal, 27: 379-423, 623-656, 1948.

[11] M. A. Shokrollahi, personal communication.

[12] Signal Processing for Big Data (Special Issue). IEEE Signal Processing Magazine, vol. 31, No. 5, September, 2014.

[13] F. Yates. The Design and Analysis of Factorial Experiments, Imperial Bureau of Soil Science, Harpenden, England, 1937.
Appendix: Proof of (4)

\[ H \left( \frac{1 + d}{2} \right) = 1 - \left( \frac{d^2}{2} + \frac{d^4}{12} + \frac{d^6}{30} + \frac{d^8}{56} + \cdots \right) \times \frac{1}{\log 2}. \]

Proof. We have

\[ -H \left( \frac{1 + d}{2} \right) = \frac{1 + d}{2} \log_2 \frac{1 + d}{2} + \frac{1 - d}{2} \log_2 \frac{1 - d}{2} \]

(9)

\[ = \frac{1}{\log 2} \left( \frac{1 + d}{2} \log \frac{1 + d}{2} + \frac{1 - d}{2} \log \frac{1 - d}{2} \right) \]

(10)

\[ = \frac{1}{\log 2} \left( \frac{1 + d}{2} \log(1 + d) + \frac{1 - d}{2} \log(1 - d) \right) \]

(11)

\[ = \frac{1}{\log 2} \left( \frac{1}{2} \log(1 - d^2) + \frac{d}{2} \log \frac{1 + d}{1 - d} - \log 2 \right) \]

(12)

by definition of entropy. Using Taylor expansion series for \(0 \leq d < 1\), we have

\[ \log(1 - d^2) = - \left( d^2 + \frac{d^4}{2} + \frac{d^6}{3} + \frac{d^8}{4} + \cdots \right) \]

(13)

\[ \log \frac{1 + d}{1 - d} = 2 \left( d + \frac{d^3}{3} + \frac{d^5}{5} + \frac{d^7}{7} + \cdots \right) \]

(14)

Putting (13) and (14) into (12), we have

\[ -H \left( \frac{1 + d}{2} \right) = \frac{1}{\log 2} \left( -\frac{1}{2} \left( d^2 + \frac{d^4}{2} + \frac{d^6}{3} + \frac{d^8}{4} + \cdots \right) + \left( d^2 + \frac{d^4}{3} + \frac{d^6}{5} + \frac{d^8}{7} + \cdots \right) - \log 2 \right) \]

\[ = \frac{1}{\log 2} \left( \frac{d^2}{2} + \frac{d^4}{12} + \frac{d^6}{30} + \frac{d^8}{56} + \cdots \right) - 1, \]

which leads to (4) for \(0 \leq d < 1\). For \(-1 < d \leq 0\), we use symmetry of entropy \(H \left( \frac{1 - d}{2} \right) = H \left( \frac{1 + d}{2} \right) \) and apply above result to complete the proof for (4) for \(|d| < 1\). \(\square\)