# TRANSIENT DETECTION WITH UNKNOWN STATISTICS VIA SOURCE CODING

*Andrew Finelli, Peter Willett, Yaakov Bar-Shalom* — ECE Dept., Univ. of Connecticut

*Stefano Marano* — DIEM, Univ. of Salerno

## *Abstract*

**Quickest detection problems are fairly common in surveillance applications, as framing surveillance alerts as a change in an observation sequence's statistics is often apt. In this work, we consider the scenario where an appropriate statistical description of our observations is not available, neither before nor after the transient we are trying to detect. In this vein, we explore the use of the database Lempel-Ziv, or LZ77, procedure, to detect this transient in the observation data. This algorithm is known to have phrase lengths that are asymptotically distributed as Gaussian random variables, which allows us to form a quickest detection problem around statistics of the coded output. This work specifies procedures to perform source-agnostic transient detection using Locally Optimal (LO) statistic to augment a Page CUSUM test. The work also shows an application to acoustic data.**

***Index Terms***— Change Detection, Transient, Lempel-Ziv

## 1. INTRODUCTION

Change detection is a common goal in statistical analysis, one in which a test is employed to identify a change in the properties of a time series or stochastic process. It can model scenarios where there is either: (i) a change that must be detected as soon as possible, (ii) a change where the switch time must be estimated as accurately as possible, or (iii) a temporary change that needs to be caught quickly before it disappears. The latter issue can be done either as the data becomes available ("online"), or operating on a set of data after the whole set has become available ("offline"). In many cases, however, these problem classifications can be re-contextualized and solution algorithms can be modified to operate within multiple different regimes [19]. Most algorithms follow similar optimization and probabilistic derivations, such as the Generalized Likelihood Ratio Test [13, 11], that rely on knowing a great amount about the data statistics under the null and alternate hypotheses. Even tests designed as locally optimal – which mostly rely on the statistics in the ambient state but look for small changes – still require knowledge that may not be available [17].

One application of change detection can be found in surveillance applications. Here, a sensor will be collecting data while observing some surveillance area and the goal is to detect the signature of the threat of interest in the data. The change may be represented in the data as some change in the statistics of the time series, i.e. a change in distribution (e.g., Gaussian to some other distribution), or in some moment of the data (mean, variance, etc.). Online canonical solutions to the quickest detection problem are the Page test [14] and the Wald SPRT [17], the former of which can be updated recursively each time a new sample becomes available, for "on demand" results.

In some scenarios, such as in underwater theaters, there is a lack of information about the data statistics either before and/or after the change [1] – not just *univariate* (first-order) statistics, but the *full dependency structure*. Such source-agnostic applications present unique challenges, as many statistical tests rely on an accurate knowledge of the data statistics before *and* after a change. Some works employ a universal source coding algorithm to replace the need to know the statistics under the alternative hypothesis (anomalous state) [18, 20, 7]. The works approximate the statistics of a distribution after a change and offer bear-optimal performance; however, knowledge of the ambient source statistics is required for the methods derived in these works to be successful.

Universal source coding algorithms are most often found in communications and data storage as the goal is to compress (shorten) the number of symbols needed to express the information contained in the data stream [8]. Some of the more famous algorithms that achieve this for an ergodic source are the Lempel-Ziv algorithms. These algorithms come in two flavors: the tree-based LZ78 algorithm [24] and the database reliant LZ77 algorithm [23]. The algorithms form a basis for many modern data compression algorithms, i.e., those used to create .zip files, DEFLATE [9]. The latter of these algorithms, LZ77, also known as the Database Lempel-Ziv algorithm (DBLZ), has useful properties related to the number of data symbols that are coded together [21, 15, 22, 10].

The goal of this paper is to develop a source-agnostic change-detection algorithm that makes as few assumptions as possible. We wish to exploit changes in marginal statistics (such as a change in power), and we also intend to use structured changes (such as of bandwidth or the appearance of a tone). But we also want to be able to detect a change

THIS WORK WAS SUPPORTED AFOSR UNDER CONTRACT FA9500-18-1-0463.

in distribution – perhaps a temporary one, such as a transient signal that can by a "bell-ringer" for surveillance – even when the univariate statistics do not change, and when the change has no commonly-modeled structure. The algorithm will then be tested on quantized acoustic data to show its effectiveness. We will be describing the algorithm, justifying its operation mathematically, and finally presenting some results and conclusions about its effectiveness.

## 2. PROBLEM FORMULATION

The data that we are considering is a time-series generated by an ergodic source whose statistics, i.e., time-dependence, distribution, moments, etc., are unknown a priori. The data are denoted as the set

$$\mathcal{Z} = \{z(n)\}_{n=1}^{N_s} \tag{1}$$

where $n$ is the sample index and $N_s$ is the total number of samples. The set is observed and processed sequentially. The data is either natively quantized with $b$ bit, or will be re-quantized (i.e., with a Llloyd-Max algorithm [12]) to the desired number of bits. We do have the restriction that the number of quantization levels can only ever be smaller than that of the data natively. The set of (re)quantized data is then denoted

$$\mathcal{Z}_b^* = \{z_b^*(n)\}_{n=1}^{N_s} \tag{2}$$

where, for example, $b = 2$ indicates binary quantized data.

The measurement (data) distribution, while not known a priori, will change at an unknown sample time $n_0$ to a different (similarly unknown) distribution. The distribution of the quantized measurements before the change occurs will be called the ambient, null state, or $\theta_{\mathcal{H}}$ distribution. After the change has occurred, the data will come from a source with the anomalous, alternate state, or $\theta_{\mathcal{K}}$ distribution.

$$z_b^*(n) \sim f\left(z_b^*(n) \middle| \theta_{\mathcal{H}}\right) \quad n = 1, \cdots, n_0 \tag{3}$$

$$z_b^*(n) \sim f\left(z_b^*(n) \middle| \theta_{\mathcal{K}}\right) \quad n = n_0 + 1, n_0 + 2, \cdots \tag{4}$$

Our goal is to determine when this change has occurred as quickly as possible under these source-agnostic conditions. Explicitly, we must define a stopping rule $\phi_n(z^*) \in \{0, 1\}$ such that

$$N_\phi \triangleq \left\{\phi_{N_\phi}(z_b^*) = 1 \cap \phi_n(z_b^*) = 0, \forall n \in \{1, N_\phi - 1\}\right\} \tag{5}$$

Thus, we say that the data has changed from its ambient state to its anomalous state at sample $N_\phi$. The goal is to select a stopping rule that minimizes the average time to detection. That is, we wish to minimize

$$\overline{T}_{\mathrm{D}} = \mathbf{E}\left[N_\phi - n_0 \middle| \mathcal{K}\right] \tag{6}$$

and maximize the average time between false alarms, defined as

$$\overline{T}_{\mathrm{FA}} = \lim_{n_0 \to \infty} \mathbf{E}\left[N_\phi \middle| \mathcal{H}\right] \tag{7}$$

This is a quickest detection problem.

## 3. SOLUTION FRAMEWORK

A common solution to quickest detection problems is to use Page's Cumulative Sum, or CUSUM, test [14]. This test recursively builds a test statistic as new measurements become available and makes a stopping rule decision after each update/measurement. The canonical form of the CUSUM test, operating on data $y(n)$, begins with $S_0 = 0$ and progresses for $n \geq 1$ as

$$S_n = \max\left\{S_{n-1} + g\left[y(n)\right], 0\right\} \tag{8}$$

$$\phi_n(z) = \begin{cases} 1 & S_n > \tau \\ 0 & \text{else} \end{cases} \tag{9}$$

Here, $g[\cdot]$ is generally taken to be the log-likelihood ratio (LLR), however any test statistic is valid as long as

$$\mathbf{E}\left\{g\left[y(n)\middle|\theta_{\mathcal{H}}\right]\right\} < 0 \tag{10}$$

$$\mathbf{E}\left\{g\left[y(n)\middle|\theta_{\mathcal{K}}\right]\right\} > 0 \tag{11}$$

In the source-agnostic case, however, we have no way of calculating the log-likelihood ratio as this requires knowledge of the data distribution in the anomalous and ambient states. In order to efficiently detect changes, we intend to process the data using a universal source coding algorithm in order to prescribe a more predictable structure to data regardless of originating distribution. The universal source coding algorithm used in this work is the Database Lempel-Ziv (DBLZ) or LZ77 algorithm [23]. This algorithm takes in the source symbols and produces a coded output with the goal of lossless compression.

The LZ77 coded output is determined online and the $m$-th codeword denoted as

$$\{c(m)\}_{m=1}^{N_c(n)} \tag{12}$$

where $N_c(n)$ is the number of new codewords that exist up to sample $n$. It is important to note that this a new codeword is not produced every time a new datum becomes available. Using the coded output, that implies a decision can only occur on a sample where a new codeword can be declared. We define a function $t(m)$ that produces the sample number $(n)$ that corresponds to the final data sample represented by codeword $c(m)$. This allows us to define the phrase length of codeword $m$ as

$$L(m) = \text{Length}\left[c(m)\right] = t(m) - t(m-1) \tag{13}$$

and we define $t(0) = 0$. As it will be useful later, we define an indicator function for whether or not a new codeword has been declared after processing sample $n$ as

$$\mathcal{I}(n) = \begin{cases} 1 & \text{if } \exists m \leq N_c(n) \text{ s.t. } t(m) = n \\ 0 & \text{else} \end{cases} \tag{14}$$

The DBLZ algorithm searches a database every time a new codeword is declared. Naïve implementations of this algorithm can thus have a runtime of $\mathcal{O}\left(n^2\right)$. Our implementation of the DBLZ algorithm draws inspiration from the tree-based Lempel-Ziv algorithm [24] to create a tree of previously discovered codewords which will speed up the search process. The implementation shortens the runtime of the algorithm (on average) to $\mathcal{O}\left(n\log(n)\right)$. The worst-case runtime is still $\mathcal{O}\left(n^2\right)$, but the worst case is of little practical interest to this work. As having a larger database strengthens the convergence of the following trend, we highly suggest universal source coding algorithms with at most an $\mathcal{O}\left(n\log(n)\right)$ run time, which most modern implementations can achieve.

The length of source phrases coded by the DBLZ algorithm are random for stochastic sources. The distribution is asymptotically Gaussian, with mean and variance that depend upon the statistics of the sources for the data stream and database [21]. Specifically, it has been shown that if an ergodic source has memory $M$, alphabet $\mathcal{A}$, and entropy $H < \log|\mathcal{A}|$, then

$$\frac{L(m) - \frac{\log(N_{\mathrm{DB}})}{H}}{\sigma_M\sqrt{\log(N_{\mathrm{DB}})/H^3}} \sim \mathcal{N}\left(0,1\right) \tag{15}$$

where

$$\sigma_M^2 = \mathrm{Var}\left\{\log \mathrm{P}\left[z_b^*(n)|z_b^*(n-1),\cdots,z_b^*(n-M)\right]\right\} \tag{16}$$

and $N_{\mathrm{DB}}$ is the size of the database. When the database is built on data whose distribution does not match that of the observations, the mean and variance of this Gaussian change depending on the Kullback-Leibler (KL) divergence between the two distributions.

In a practical application of the LZ77 algorithm, these phrase lengths are integers and thus their distribution is discrete. In order to avoid issues that may arise when dealing with integer numbers and working with continuous distributions, we will introduce a small amount of noise to the phrase length sequence output by the LZ77 coder. This is akin to dithering in other digital audio applications [16]. We use a Gaussian noise (to not change the distribution) with a standard deviation of 0.05 in our dithering procedure.

Now that we know the phrase lengths output from the DBLZ algorithm regularizes to have a distribution whose form we know, albeit with two unknown parameters, we can attempt to detect a change in both or either parameter of a Gaussian distribution (mean and/or variance) to detect a change in the coded outputs. We will therefore use the CUSUM test (8)-(11) with $y(n)$ replaced by $L(m)$ and $g\left[y(n)\right]$ replaced by a statistic based on detecting a change in the parameters of the Gaussian phrase length distribution.

## 4. MODIFIED PAGE TEST STATISTICS

We have developed two different statistics for detecting a change in the phrase length distribution parameters. The test are classified by the detection paradigm they fall into, Locally Optimal (LO), as well as how they consider the data, either using the phrase lengths or via the difference of successive phrase lengths. The use of successive phrase length differences changes the distribution to have twice the variance, but an approximate zero mean both before and after the change.

The two tests use a LO philosophy. In the first LO test, the phrase lengths are modeled as Gaussian with zero mean (i.e., we operate on the difference of two successive phrase lengths, considered to be independent).

$$\mathcal{H}: \quad L(m) \sim \mathcal{N}\left(0, \sigma_0^2\right) \tag{17}$$

$$\mathcal{K}: \quad L(m) \sim \mathcal{N}\left(0, (\sigma_0-\theta)^2\right), \quad \theta > 0 \tag{18}$$

LO tests linearize the measurement distribution around some parameter $\theta \in \theta_{\mathcal{H}}$, the parameter space for the ambient distribution. The resulting test statistic for the CUSUM test based on the application of this procedure to (17) is

$$S_m = \max_{1\leq j\leq n}\sum_{k=j}^{m}\frac{\frac{d}{d\theta}f(w_m|\theta)}{f(w_m|\theta)} = \max_{1\leq j\leq m}\sum_{k=j}^{m}\frac{\sigma_0^2 - w_m^2}{\sigma_0^3} \tag{19}$$

This test (and the following) are combined with slight variations that look for changes in other directions (i.e., both parameters increase, one increases and the other decreases, etc.), as well as an added bias, $\mathcal{B}$, to satisfy (10) and (11).

The second LO test looks for a simultaneous decrease or increase in mean and variance.

$$\mathcal{H}: \quad L(m) \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right) \tag{20}$$

$$\mathcal{K}: \quad L(m) \sim \mathcal{N}\left((1-\theta)\mu_0, [(1-\theta)\sigma_0]^2\right), \quad \theta > 0 \tag{21}$$

This formulation results in the following Double Locally Optimal sequential test

$$S_m = \max_{1\leq j\leq m}\sum_{k=j}^{m}\frac{\sigma_0^2 + \mu_0 x_m - x_m^2}{\sigma_0^2} \tag{22}$$

These tests also require a bias and a second test for an increase in variance and/or mean. These alternate tests are the same with the summand changed to the negative of itself.

Each of these sequential tests operates on the phrase lengths, but we must translate this to operation on the raw data ($\mathcal{Z}_b^*$) as it becomes available. Thus, we create a stopping rule for the CUSUM test based on the coded phrase lengths ($\phi_m(L)$) and translate that to the raw data inputs ($\phi_n(L)$) as

$$\phi_m(L) = \begin{cases} 1 & S_m > \tau \\ 0 & \text{else} \end{cases} \tag{23}$$

$$\phi_n(L) = \phi_m(L)\mathcal{I}(n) \tag{24}$$

We note that this means we can only declare a change has been made *after* a new codeword has been created.

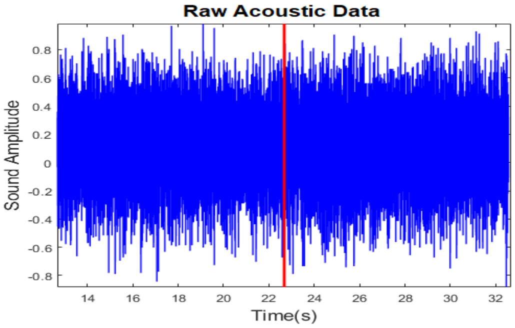

**Fig. 1**. Raw (unprocessed) acoustic file in which we wish to detect the addition of AWGN. The red line at $t \simeq 26.675$ seconds denotes the (undiscernable?) time a change occurs.

## 5. RESULTS AND CONCLUSIONS

To display the effectiveness of our approach, we present a scenario with real data from an underwater hydrophone, sampled at 8kHz and using an 8-bit quantization. The first $10^6$ samples are used for the DBLZ and are not used for detection (to preserve independence of the database). The change occurs at $t \simeq 26.675$ seconds into processing. The change we wish to detect is the addition of white Gaussian noise (AWGN) to the signal with a standard deviation of 0.005.

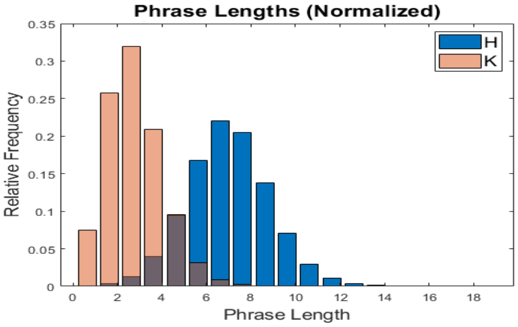

**Fig. 2**. Distribution of phrase lengths output from the LZ77 algorithm before and after the change ($\mathcal{H}$ and $\mathcal{K}$, respectively).

Figure 1 shows the quantized sound file before any processing with a red line denoting the point where a change we wish to detect occurs. We see no discernible difference in the data at first glance. Figure 2 shows histograms of the phrase lengths before and after the change. There is a noticeable decrease in mean and standard deviation. Figure 3 shows the modified CUSUM tests discussed above applied to this data. We see that each test is capable of detecting the change with little to no false alarms (depending on threshold selected).

Finally, Figure 4 shows Monte Carlo statistics of $\overline{T}_{\mathrm{FA}}$ vs. $\overline{T}_{\mathrm{D}}$. These statistics were taken over 250 Monte Carlo runs where the ambient and anomalous states are binary Markov chains, with similar stationary distributions, and each with memory $M = 1$. We use a database of size $10^6$ bits for testing and find that for shorter length disturbances, the single LO test will perform more reliably, however if the disturbance is longer (i.e., the test statistic has more codewords to evaluate than in the anomalous state) then using a larger threshold and leveraging information from both parameters with the Doubly LO test provides better detection performance (for these Markov and short memory sources).

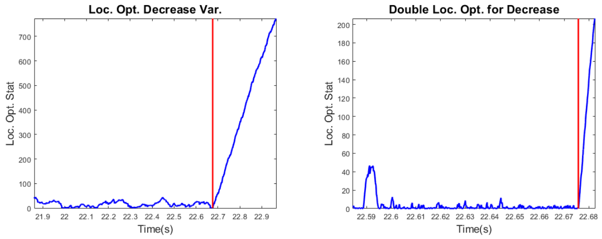

**Fig. 3**. Output of the modified CUSUM test applied to underwater hydrophone data to detect the addition of a small amount of AWGN.The red line indicates the time when the change (addition of the AWGN) occurs.

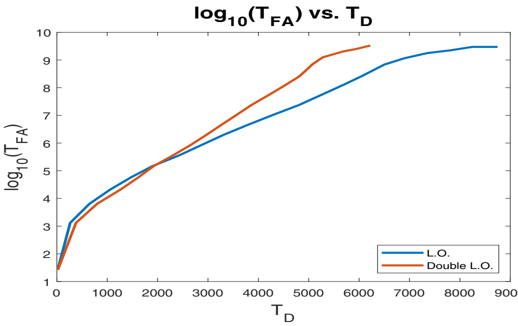

**Fig. 4**. A plot of the logarithm of average time between false alarms ($\log_{10}\left(\overline{T}_{\mathrm{FA}}\right)$) and the average time to detection ($\overline{T}_{\mathrm{D}}$) over 250 Monte Carlo runs for a change between two memory length $M = 1$ Markov Chains with similar stationary distributions. We see that the performance is better initially for the single LO test, but for higher thresholds the doubly LO test quickly performs better.

## 6. CONCLUSION

This work motivated, developed, and justified several related statistical test for preforming source-agnostic change detection based on universal source coding. We then developed several tests on the standardized, coded phrase lengths and showed their effectiveness on real data gathered from underwater hydrophones. The results show promise for a non-bespoke change detection algorithm. Future work will explore quantization as well as different CUSUM versions, especially those based on the MAST formalism [5, 6].

## 7. REFERENCES

[1] D. A. Abraham, *Underwater Acoustic Signal Processing: Modeling, Detection, and Estimation (Modern Acoustics and Signal Processing)*, Springer International Publishing, Basel, Switzerland, 2019.

[2] Y. Bar-Shalom, X. Li and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*, John Wiley and Sons, 2001.

[3] Y. Bar-Shalom, P. Willett and X. Tian, *Tracking and Data Fusion: A Handbook of Algorithms*, YBS Publishing, 2011.

[4] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*, Englewood Cliffs, NJ, USA:Prentice-Hall, 1993.

[5] P. Braca, D. Gaglione, S. Marano, L. M. Millefiori, P. Willett and K. R. Pattipati, "Quickest detection of COVID-19 pandemic onset", *IEEE Signal Processing Letters*, vol. 28, pp. 683-687, 2021, doi: 10.1109/LSP.2021.3068072.

[6] P. Braca, D. Gaglione, S. Marano, L. M. Millefiori, P. Willett and K. Pattipati, "Decision support for the quickest detection of critical COVID-19 phases", *Sci. Rep*, [online] Available: https://arxiv.org/abs/2011.11540.

[7] D. K. Chittam, R. Bansal and R. Srivastava, "Universal Compression of a Piecewise Stationary Source Through Sequential Change Detection," *2018 Twenty Fourth National Conference on Communications (NCC)*, 2018, pp. 1-6, doi: 10.1109/NCC.2018.8600011.

[8] T. Cover, J. Thomas, *Elements of Information Theory, 2nd Edition*, John Wiley and Sons, Inc., New York, 2006.

[9] P. Deutsch, "DEFLATE Compressed Data Format Specification version 1.3", IETF. p. 1, doi:10.17487/RFC1951. RFC 1951.

[10] P. Grassberger, "Estimating the information content of symbol sequences and efficient codes," in *IEEE Transactions on Information Theory*, vol. 35, no. 3, pp. 669-675, May 1989, doi: 10.1109/18.30993.

[11] S. M. Kay, *Fundamentals of Statistical Signal Processing, Vol. II: Detection Theory*. Upper Saddle River, NJ, USA: Prentice Hall, 1998.

[12] S. Lloyd, "Least squares quantization in PCM", *IEEE Transactions on Information Theory*, 28 (2): 129–137, doi:10.1109/TIT.1982.1056489.

[13] J. Neyman and E. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 231*, pp. 289–337, 1933.

[14] E. Page, "Continuous inspection schemes", Biometrika, vol. 41, no. 1/2, pp. 100-115, Jun. 1954.

[15] N.T. Plotkin, A.J. Wyner, "An Entropy Estimator Algorithm and Telecommunications Applications". In: *Heidbreder G.R. (eds) Maximum Entropy and Bayesian Methods. Fundamental Theories of Physics*, vol 62. Springer, Dordrecht. doi: 10.1007/978-94-015-8729-7 29

[16] K. C. Pohlmann, *Principles of Digital Audio*, McGraw-Hill Professional, ISBN 978-0-07-144156-8.

[17] H. V. Poor and O. Hadjiliadis, *Quickest Detection*, Cambridge, U.K.:Cambridge Univ. Press, 2009.

[18] R. Srivastava and R. K. Bansal, "Sequential Change Detection Through Universal Compression - An Asymptotic Study," *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 2396-2400, doi: 10.1109/ISIT.2018.8437708.

[19] H. L. Van Trees, *Detection, Estimation, and Modulation Theory: Part 1. Detection, Estimation, and Filtering Theory*, John Wiley and Sons, Inc., New York, 1971.

[20] A. Verma and R. K. Bansal, "Sequential Change Detection Based on Universal Compression for Markov Sources," *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 2189-2193, doi: 10.1109/ISIT.2019.8849595.

[21] A. D. Wyner, J. Ziv and A. J. Wyner, "On the role of pattern matching in information theory," in *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2045-2056, Oct. 1998, doi: 10.1109/18.720530.

[22] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," in *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1270-1279, July 1993, doi: 10.1109/18.243444.

[23] J. Ziv and A. Lempel, "A Universal Algorithm for Sequential Data Compression", *IEEE Transactions on Information Theory*, 23 (3): 337–343. doi:10.1109/TIT.1977.1055714.

[24] J. Ziv and A. Lempel, "Compression of Individual Sequences via Variable-Rate Coding", *IEEE Transactions on Information Theory*, 24 (5): 530–536, doi:10.1109/TIT.1978.1055934.