Statistical mechanical aspects of joint source-channel coding

Ido Kanter and Haggai Kfir

Minerva Center and the Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel

An MN-Gallager Code over Galois fields, q, based on the Dynamical Block Posterior probabilities (DBP) for messages with a given set of autocorrelations is presented with the following main results: (a) for a binary symmetric channel the threshold, f_c, is extrapolated for infinite messages using the scaling relation for the median convergence time, $t_{med} \propto 1/(f_c - f)$; (b) a degradation in the threshold is observed as the correlations are enhanced; (c) for a given set of autocorrelations the performance is enhanced as q is increased; (d) the efficiency of the DBP joint source-channel coding is slightly better than the standard gzip compression method; (e) for a given entropy, the performance of the DBP algorithm is a function of the decay of the correlation function over large distances.

With the rapid growth of information content in today's wire and wireless communication, there is an increasing demand for efficient transmission systems. A significant gain in the transmission performance can be achieved by the application of the joint source-channel coding technique, which has attracted much attention during the recent past, see for instance [1–6]. Roughly speaking, source coding is mainly a data compression process that aims at removing as much redundancy as possible from the source signal, whereas channel coding is the process of intelligent redundancy insertion so as to be robust against channel noise. These two processes, source coding and channel coding, seem to act in opposition, where the first/second process shrinks/expands the transmitted data. For illustration, assume that our compression *shrinks* the size of the source signal by a factor 2 and in order to be robust against channel noise we have to *expend* our compressed file by a factor 4. Hence, the length of the transmitted sequence is only twice the length of the uncompressed source.

The source-channel coding theorem of Shannon [7] indicates that if the minimal achievable source coding rate of a given source is below the capacity of the channel, then the source can be reliably transmitted through the channel, assuming an infinite source sequence. This theorem implies that source coding and channel coding can be treated *separately* without any loss of overall performance, hence they are fundamentally separable. Practically, the source can be first efficiently compressed and then an efficient error correction method can be used.

The objective of joint source-channel coding is to combine both source (compression) and channel (error correction) coding into one mechanism in order to reduce the overall complexity of the communication while maintaining satisfactory performance. Another possible advantage of the joint source-channel coding is the reduction of the sensitivity to a bit error in a compressed message.

In a recent paper [8] a particular scheme based on a statistical mechanical approach for the implementation of the joint source-channel coding was presented and the main steps are briefly summarized below. The original boolean source is first mapped to a binary source [9,10] $\{x_i \pm 1\}$ $i = 1, ..., L$, and is characterized by a finite set, k_0, of autocorrelations

$$C_k = \frac{1}{L} \sum_{i=1}^{L} x_i x_{(i+k)\ \mathbf{mod\ L}} \tag{1}$$

where $k \leq k_0$ is the highest autocorrelation taken. The number of sequences oeying these k_0 constraints is given by

$$\Omega = Tr \prod_{k=1}^{k_0} \delta\left(\sum_j x_j x_{j+k} - C_k L\right) \tag{2}$$

Introducing the Fourier representation of the delta functions and then using the standard transfer matrix (of size $2^{k_0} \times 2^{k_0}$) method, [11] one finds $\Omega = \int dy_k \exp\{-L[\sum y_k C_k - \ln \lambda_{max}(\{y_k\})]\}$, where λ_{max} is the maximal eigenvalue of the corresponding transfer matrix. For large L, using the saddle point method the entropy, $H_2(\{C_k\})$, is given in the leading order by

$$H_2\left(\{C_k\}\right) = \frac{\frac{1}{k_0} \ln \lambda_{max}\left(\{y_k\}\right) - \sum_{i=1}^{k_0} y_k C_k}{\ln 2} \tag{3}$$

where $\{y_k\}$ are determined from the saddle point equations of Ω. [8] Assuming Binary Symmetric Channel (BSC) and using Shannon's lower bound, the channel capacity of sequences with a given set of k_0 autocorrelations is given by

$$C = \frac{1 - H_2(f)}{H_2(\{C_k\}) - H_2(P_b)} \tag{4}$$

where f is the channel bit error rate and p_b is a bit error rate. The saddle point solutions derived from eq. 3 indicate that the equilibrium properties of the one dimensional Ising spin system

$$H = -\sum_i \sum_{k=1}^{k_0} \frac{y_k}{\beta} x_i x_{i+k} \tag{5}$$

obey in the leading order the autocorrelation constraints of eq. 2. Note that in the typical scenario of statistical mechanics, one of the main goals is to calculate the partition function and the equilibrium properties of a *given Hamiltonian.* Here we present a prescription of how to solve the reverse question. Given the desired macroscopic properties, the set of the autocorrelations, the goal is to find the *appropriate Hamiltonian* obeying these macroscopic constraints. This property of the effective Hamiltonian, eq. 5, is used in simulations below to generate an ensemble of signals (source messages) with the desired set of autocorrelations.

The decoding of symbols of k_0 successive bits is based on the standard message passing introduced for the MN decoder over Galois fields with $q = 2^{k_0}$ [12] and with the following modification. The horizontal pass is left unchanged, *but a dynamical set of probabilities assigned for each block are used in the vertical pass.* The Dynamical Block Probabilities (DBP), $\{P_n^c\}$, are determined following the current belief regarding the neighboring blocks and are given by

$$\begin{aligned} \gamma_n^c &= S_I(c) \left(\sum_{l=1}^{q} q_L^l S_L(l,c) \right) \left(\sum_{r=1}^{q} q_R^r S_R(c,r) \right) \\ P_n^c &= \frac{\gamma_n^c}{\sum_{j=1}^{q} \gamma_n^j} \end{aligned} \tag{6}$$

where $l/r/c$ denotes the state of the left/right/center $(n{-}1\,/\,n{+}1\,/\,n)$ block respectively and q_L^l/q_R^r are their posterior probabilities. $S_I(c) = e^{-\beta H_I}$, where H_I is the inner energy of a block of k_0 spins at a state c, see eq. 5. Similarly $S_L(l,c)$ ($S_R(c,r)$) stands for the Gibbs factor of consecutive Left/Center (Center/Right) blocks at a state l,c (c,r). [8]

Note that the complexity of the calculation of the block prior probabilities is $O(Lq^2/\log q)$ where $L/\log q$ is the number of blocks. The decoder complexity per iteration of the MN codes over a finite field q can be reduced to order $O(Lqu)$ [13–15], where u stands for the average number of checks per block. Hence the total complexity of the DBP decoder is of the order of $O(Lqu + Lq^2/\log q)$.

In this Letter we examine the efficiency of the DBP-MN decoder as a function of the maximal correlation length taken, k_0, the strength of the correlations, the size of the finite fields q, and we compare this efficiency with the standard gzip compression procedure. A direct answer to the above questions is to implement exhaustive simulations on increasing message length, various finite fields q, and sets of autocorrelations, which result in the bit error probability versus the flip rate f. Besides the enormous computational time required , the conclusions would be controversial since it is unclear how to compare, for instance, the performance as a function of q; with the same number of transmitted blocks or with the same number of transmitted bits.

In order to overcome these difficulties, for a given MN-DBP code over GF(q) and a set of autocorrelations, the threshold f_c is estimated from the scaling argument of the convergence time, which was previously observed for $q = 2$ [16,17]. The median number of message passing steps, t_{med}, necessary for the convergence of the MN-DBP algorithm is assumed to diverge as the level of noise approaches f_c from below. More precisely, we found that the scaling for the divergence of t_{med} is independent of q and is consistent with

$$t_{med} = \frac{A}{f_c - f} \tag{7}$$

where for a given set of autocorrelations and q, A is a constant. Moreover, for a given set of autocorrelations and a finite field q, the extrapolated threshold f_c is independent of L, as demonstrated in Fig. 1. This observation is essential to determine the threshold of a code based on the above scaling behavior. Note that the estimation of t_{med} is a simple computational task in comparison with the estimation of low bit error probabilities for large L, especially close to the threshold. We also note that the analysis is based on t_{med} instead of the average convergence time, t_{av}, [16] since we wish to prevent the dramatic effect of a small fraction of finite samples with slow convergence or no convergence. [19,18]

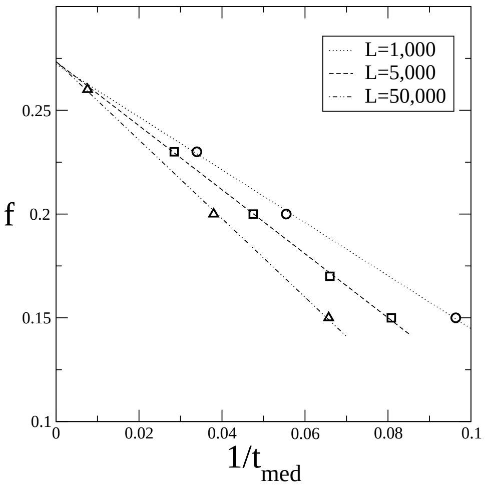

FIG. 1. The flip rate f as a function of $1/t_{med}$ for GF(4) with $C_1 = C_2 = 0.8$ and $L = 1,000,\ 5,000\ ,50,000$. The lines are a result of a linear regression fit. The threshold, $f_c \sim 0.272$, extrapolated from the scaling behavior eq. 7, is independent of N.

All simulation results presented below are derived for rate 1/3 and the construction of the matrices A and B of the MN code are taken from [16]. In all examined sets of autocorrelations, $10^3 \leq L \leq 5\times 10^4$ and $4 \leq q \leq 64$, the scaling for the median convergence time was indeed recovered. At this stage an efficient tool to estimate the threshold of an MN-DBP decoder exists and we are ready to examine the efficiency of the DBP decoder as a function of $\{C_k\}$ and q.

Results of simulations for $q = 4,\ 8,\ 16$ and 32 and selected sets of autocorrelations are summarized in Table I, and the definition of the symbols is: $\{C_k\}$ stand for the imposed values of autocorrelations as defined in eqs. 1-2; $\{y_k\}$ are the interaction strengths, eqs. 3 and 5; H represents the entropy of sequences with the given set of autocorrelations, eq. 2; f_c is the estimated threshold of the MN-DBP decoder derived from the scaling behavior of t_{med}; f_{Sh} is the Shannon's lower bound, eq. 4; Ratio is the efficiency of our code f_c/f_{Sh}; Z_R indicates the gzip compression rate averaged over files of the sizes 10^6 bits with the desired set of autocorrelations. We assume that the compression rate with $L = 10^6$ achieves its asymptotic ratio, as was indeed confirmed in the compression of files with different L; $1/R^\star$ indicates the ideal (minimal) ratio between the transmitted message and the source signal after implementing the following two steps: compression of the file using gzip and then using an ideal optimal encoder/decoder, for a given BSC with f_c. A number greater than (less than) 3 in this column indicates that the MN-DBP joint source-channel coding algorithm is more efficient (less efficient) in comparison to the channel separation method using the standard gzip compression. The last four columns of Table I are devoted for the comparison of our DBP algorithm to advanced compression methods. PPM_R and AC_R stand for the compression rate of files of the size 10^6 bits with the desired autocorrelations using the Prediction by Partial Match [20] and for the Arithmetic Coder [21], respectively. Similarly to the gzip case, $1/R_{PPM}$ and $1/R_{AC}$ stand for the optimal (minimal) rate required for the separation process (first a compression and then an ideal optimal encoder/decoder) assuming a BSC with f_c.

q	c_1	c_2	c_3	c_4	c_5	y_1	y_2	y_3	y_4	H	f_c	f_{Sh}	Ratio	Z_R	$\frac{1}{R^*}$	PPM_R	$\frac{1}{R_{PPM}}$	AC_R	$\frac{1}{R_{AC}}$
4	0.65	0.65	–	–	–	0.29	0.53	–	–	0.578	0.239	0.247	0.97	0.69	3.32	0.64	3.1	0.65	3.12
4	0.72	0.72	–	–	–	0.301	0.608	–	–	0.49	0.253	0.266	0.95	0.61	3.3	0.55	3.001	0.58	3.13
4	0.8	0.8	–	–	–	0.318	0.708	–	–	0.379	0.273	0.294	0.93	0.5	3.23	0.43	2.76	0.48	3.1
4	0.86	0.86	–	–	–	0.372	0.81	–	–	0.286	0.293	0.321	0.91	0.41	3.16	0.32	2.52	0.38	2.9
8	−0.65	0.6	−0.55	–	–	−0.31	0.23	−0.23	–	0.595	0.236	0.244	0.97	0.71	3.34	0.66	3.14	0.67	3.16
16	0.6	0.6	0.58	0.6	–	0.104	0.17	.164	0.327	0.57	0.229	0.249	0.92	0.7	3.12	0.66	2.95	0.67	2.98
32	0.62	0.7	0.55	0.55	0.599	0.324	0.821	−0.438	−0.219	0.49	0.242	0.266	0.91	0.62	3.06	0.56	2.77	0.6	2.964

TABLE I. Results for $q = 4,\ 8,\ 16,\ 32$ and selected sets of autocorrelations $\{C_k\}$.

Table I indicates the following main results: (a) For $q = 4$ (the upper part of Table I) a degradation in the performance is observed as the correlations are enhanced and as a result the entropy decreases. The degradation seems to be significant as the entropy is below ~ 0.3 (or for the test case $R = 1/3$, $f_c \geq 0.3$). [22] A similar degradation was observed as the entropy decreases for larger values of q. (b) The efficiency of our joint source-channel coding technique is superior to the alternative standard gzip compression in the source channel separation technique. For high entropy the gain of the MN-DBP is about $5-10\%$. This gain disappears as the entropy and the performance of the DBP algorithm are decreased. (c) In comparison to the standard gzip, the compression rate is improved by $2-5\%$ using the AC method. A further improvement of a few percents is achieved by the PPM compression. This later improvement seems to be significant in the event of low entropy. (d) Our DBP joint source-channel coding seems to be superior (by $\sim 3\%$)to the separation method based on the PPM compression for high entropy. However for ensemble of sequences characterized by low entropy this gain disappears. (e) With respect to the computational time of the source channel coding, our limited experience indicates that the DBP joint-source channel coding is faster than the AC separation method and the PPM separation method is substantially slower.

For a given set of autocorrelations where C_{k_0} is the maximal one taken, the MN-DBP algorithm can be implemented with any field $q \geq 2^{k_0}$. If one wishes to optimize the complexity of the decoder it is clear that one has to work with the minimal allowed field, $q = 2^{k_0}$. However, when the goal is to optimize the performance of the code and to maximize the threshold, the selection of the optimal field, q, is in question. In order to answer this question we present in Fig. 2 results for $k_0 = 2$ ($C_1 = C_2 = 0.86$) and $q = 4,\ 16,\ 64$. It is clear that the threshold, f_c, increases as a function of q as was previously found for the case of unbiased signals. [12] More precisely, the estimated thresholds for $q = 4,\ 16,\ 64$ are $\sim 0.293,\ 0.3,\ 0.309$, respectively, and the corresponding Ratios ($\equiv f_c/f_{Sh}$) are $0.913,\ 0.934, 0.962$. where $f_{Sh} = 0.321$. Note that the extrapolation of f_c for large q seems asymptotically to be consistent with $f_c(q) \sim 0.316 - 0.18/q$.

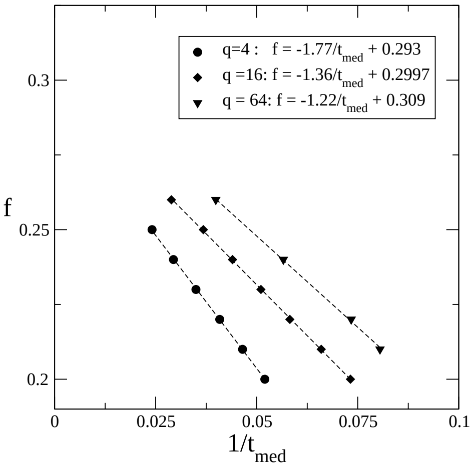

FIG. 2. The scaling behavior, f as a function of $1/t_{med}$, for $C_1 = C_2 = 0.86$ and $q = 4,\ 16,\ 64$. The lines are a result of a linear regression fit. The estimated thresholds for $q = 4,\ 16,\ 64$ are $0.293,\ 0.3,\ 0.309$, and the corresponding *Ratio* $\equiv f_c/f_{Sh} = 0.913,\ 0.934, 0.962$, where $f_{Sh} = 0.321$.

For a given q, there are many sets of autocorrelations (or a finite fraction of $\{C_k\}$ in $k_0 - 1$ dimensions) obeying the same entropy. An interesting question is whether the performance of our DBP algorithm measured by the Ratio ($\equiv f_c/f_{Sh}$) is a function of the entropy only. Our numerical simulations indicate that the entropy is not the only parameter which controls the performance of the DBP algorithm. For the same entropy and q the Ratio can fluctuate widely among different sets of correlations. For illustration, in Table II results for two sets of autocorrelations with *the same entropy* are summarized for each $q = 4,\ 8,\ 16$ and 32. It is clear that as the Ratio ($\equiv f_c/f_{Sh}$) is much degradated the gzip performance is superior (the second example with $q = 8$ and 32 in Table II where the Ratio is 0.8 and 0.72, respectively). The crucial question is to find the criterion to classify the performance of the DBP algorithm among all sets of autocorrelations obeying the same entropy. Our generic criterion is *the decay of the correlation function over distances beyond two successive blocks.* However, before the examination of this criterion, we would like to turn back to some aspects of statistical physics.

The entropy of sequences with given first k_0 correlations are determined via the effective Hamiltonian consisting of k_0 interactions, eqs. 2-3. As a result the entropy of these sequences is *the same* as the entropy of the effective Hamiltonian, $H\{y_k\}$, at the inverse temperature $\beta = 1$, eq. 5. As for the usual scenario of the transfer matrix method, the leading order of quantities such as the free energy and the entropy are a function of the *largest eigenvalue* of the transfer matrix only. On the other hand the decay of the correlation function is a function of the whole spectrum of the 2^{k_0} eigenvalues. Asymptotically, the decay of the correlation function is determined from the ratio between the second largest eigenvalue and the largest eigenvalue, λ_2/λ_{max}. From the statistical mechanical point of view one may wonder, why the first k_0 correlations can be determined using the information of λ_{max} only. The answer to this question is that once the transfer matrix is defined as a function of $\{y_k\}$, eqs. 3-5, *all eigenvalues* are determined as well as λ_{max}. There is no way to determine λ_{max} independently of all other eigenvalues.

In Table II results of the DBP-MN algorithm for $q = 4,\ 8,\ 16,\ 32$ are presented. For each q, two different sets of autocorrelations characterized by the *same entropy* and threshold f_{Sh} are examined. The practical method we used to generate different sets of autocorrelations with the same entropy was a simple Monte Carlo over the space of $\{C_k\}$. [23] The additional column in Table II (in comparison with Table I) is the ratio between λ_2/λ_{max}, which characterizes the decay of the correlation function over large distances. Independent of q, it is clear that for a given entropy as λ_2/λ_{max} increases/decreases the performance of the DBP algorithm measured by the Ratio f_c/f_{Sh} is degradated/enhanced. The new criterion to classify the performance of the DBP algorithm among all sets of autocorrelations obeying the same entropy is the decay of the correlation function. This criterion is consistent with the tendency that as the first k_0 autocorrelations are increased/decreased a degradation/enhancement in the performance is observed (see Table I). The physical intuition is that as the correlation length increases, the relaxation time increases and flips on larger scales than nearest neighbor blocks are required.

Note that the decay of the correlation function in the intermediate region of a small number of blocks is a function of all the 2^{k_0} eigenvalues. Hence, in order to enhance the effect of the fast decay of the correlation function in the case of small λ_2/λ_{max}, we also try to enforce in our Monte Carlo search that all other $2^{k_0} - 2$ eigenvalues be less than $A\lambda_{max}$ with the minimal possible constant A. This additional constraint was easily fulfilled for $q = 4$ with $A = 0.1$, but for $q = 32$ the minimal A was around 0.5.

Finally we raise the question of whether for a given entropy a degradation in the *typical* performance of the DBP algorithm is expected as q increases. This is crucial since the superiority, if any, of the DBP joint source-channel coding method over advanced compression methods is in question. As explained above, our Monte Carlo simulations indicate that for a given entropy the suppression of the correlation function is more difficult as q increases. [23] This is a strong indication that as q increases a degradation in the typical performance of the DBP decoder is expected, but its nature and significance have still to be examined in further research.

We thank Shlomo Shamai and David MacKay for valuable discussions.

q	c_1	c_2	c_3	c_4	c_5	y_1	y_2	y_3	y_4	y_5	λ_2/λ_{max}	H	f_c	f_{Sh}	Ratio	Z_R	$\frac{1}{R^*}$
4	0.785	0.636	–	–	–	0.88	0.11	–	–	–	0.65	0.49	0.256	0.266	0.97	0.61	3.48
4	0.499	0.769	–	–	–	0.091	0.921	–	–	–	0.83	0.49	0.242	0.266	0.91	0.61	3.01
8	0.77	0.69	0.561	–	–	0.544	0.598	−0.167	–	–	0.59	0.48	0.259	0.269	0.96	0.6	3.42
8	0.557	0.577	0.75	–	–	−0.04	0.218	0.774	–	–	0.86	0.48	0.215	0.269	0.80	0.62	2.47
16	0.721	0.489	0.353	0.303	–	1.247	−0.212	−0.149	0.169	–	0.4	0.57	0.241	0.249	0.97	0.68	3.32
16	0.6	0.6	0.58	0.6	–	0.104	0.17	.164	0.327	–	0.82	0.57	0.229	0.249	0.92	0.7	3.12
32	0.62	0.7	0.55	0.55	0.599	0.324	0.821	−0.438	−0.219	0.51	0.78	0.49	0.242	0.266	0.91	0.62	3.06
32	0.5	0.62	0.5	0.6	0.669	−0.159	0.351	−0.266	0.366	0.66	0.89	0.49	0.191	0.266	0.72	0.64	2.26

TABLE II. Results for $q = 4,\ 8,\ 16,\ 32$ and different sets of autocorrelations. For each q, two different sets of autocorrelations characterized by the same entropy and threshold f_{Sh} are examined. As λ_2/λ_{max} increases/decreases the performance of the DBP algorithm measured by the Ratio f_c/f_{Sh} is degradated/enhanced.

We thank Shlomo Shamai and David MacKay for valuable discussions.

[1] J. Kliewer and R. Thobaben, "Combining FEC and Optimal Soft-Input Source Decoding for Reliable Transmission of Correlated Variable-Length Encoded Signals" (to appear in DCC2002)

[2] S. Shamai and S. Verdu, *European Transactions on Telecommunications*, vol. 6, no. 5, pp. 587-600, Sep.-Oct. 1995

[3] A.D. Liveris, Z. Xiong and C.N. Georghiades, "Compression of Binary Sources with Side Information Using Low-Density Parity-Check Codes", paper CTS-08-7, GLOBECOM2002, 17-21, November, Taipei, Taiwan

[4] J. Garcia-Frias and J.D. Villasenor, *IEEE J. Selec. Areas Commun.*, Vol. 19, No. 9, pp. 1671-1679, Sept. 2001

[5] C.-C. Zhu and F. Alajaji, *IEEE Commun. Letters*, Vol. 6, No. 2, pp. 64-66, February 2002

[6] J. Garcia-Frias and Z. Ying, *IEEE Commun. Letters*, Vol. 6, No. 9, pp. 394-396, September 2002

[7] C. E. Shannon. A mathematical theory of communication. *Bell System Technical J.*, **27**, 379-423, 623-656, 1948.

[8] I. Kanter and H. Rosemarin, cond-mat-0301005

[9] N. Sourlas, *Nature* **339**, 6227, 1989

[10] N. Sourlas, *Physica A* **302**, 14 (2001)

[11] R. J. Baxter. Exactly Solved Models In Statistical Mechanics. *Academic Press*, 1982

[12] M. C. Davey and D. J. C. Mackay, *IEEE Communications Letters*, Vol 2, No. 6, June 1998

[13] We thank David MacKay for bringing to our attention the reduced complexity of the decoder from $O(q^2)$ to $O(qlogq)$

[14] D. J. C. MacKay and M. C. Davey, "Gallager Codes for Short Block Length and High Rate Applications", Codes, Systems and Graphical Models", IMA Volumes in Mathematics and its Applications, Springer-Verlag (2000).

[15] T. Richardson and R. Urbanke, *IEEE Transactions on Information Theory*, **47**, 599 (2001).

[16] I. Kanter and D. Saad. *Phys. Rev. Lett.* **83**, 2660, 1999

[17] I. Kanter and D. Saad. *J. Phys. A* **33**, 1675 (2000)

[18] In practice we define t_{med} to be the average convergence time of all samples with $t \leq$ the median time.

[19] A. Priel, M. Blatt, T. Grossman, E. Domany and I. Kanter *Phys. Rev. E* **50,** 577 (1994)

[20] The used PPMZ software can be downloaded from www.cbloom.com/src/ppmz.html

[21] The used AC software can be downloaded from www.cs.mu.oz.au/ alistair/arith_coder

[22] A similar degradation in the performance was observed for $q = 2$ and biased binary messages (each source bit is equal 0/1 and is chosen with probability $p/1-p$). As $|p-0.5|$ increases the entropy decreases and a degradation in the performance of the MN algorithm was observed.

[23] H. Kfir and I. Kanter (in preparation)