VII

Information Theoretic Models in Language Evolution

R. Ahlswede, E. Arikan, L. Bäumer, and C. Deppe*

Abstract. We study a model for language evolution which was introduced by Nowak and Krakauer ([12]). We analyze discrete distance spaces and prove a conjecture of Nowak for all metrics with a positive semidefinite associated matrix. This natural class of metrics includes all metrics studied by different authors in this connection. In particular it includes all ultra-metric spaces.

Furthermore, the role of feedback is explored and multi-user scenarios are studied. In all models we give lower and upper bounds for the fitness.

1 Introduction

The human language is used to store and transmit information. Therefore there is a significant interest in the mathematical models of language development. These models aim to explain how natural selection can lead to the gradual emergence of human language. Nowak and coworkers ([12], [13]) created a mathematical model, in which they introduced the *fitness of a language* as a measure for the communicative performance of a signalling system. In this model the signals can be misinterpreted with certain probabilities. In this case it was shown that the performance of such systems is intrinsically limited, meaning that the fitness can not be increased over a certain threshold by adding more and more signals to the repertoire of the communicators. This limitation can be overcome by concatenating signals or phonemes to form words, which increases significantly the fitness.

In the model the signals are elements of a given distance space. The fitness of the distance space is then defined as the supremum of the fitness values taken over all languages. In [13] and [5] the fitnesses of different metric spaces were investigated. Nowak conjectures that the fitness of a product-space is equal to the product of the fitnesses of the individual spaces. In the following we will refer to this conjecture as *product conjecture*.

In this paper we analyze discrete distance spaces. We prove the product conjecture for this model under assumptions which are sufficiently general so that the result includes all the models of metric spaces considered before in [12], [13] and [5].

We also show in this model that Hamming codes asymptotically achieve the maximal possible fitness.

* Supported in part by INTAS-00-738.

R. Ahlswede et al. (Eds.): Information Transfer and Combinatorics, LNCS 4123, pp. 769–787, 2006.

This model for simple signalling systems and their fitness suggests the investigations of other classical information theoretical problems in this context. We will start this direction of research by considering feedback problems and transmission problems for multiway channels. In the feedback model that we introduced we show that feedback-fitness can be bigger than the fitness without feedback.

In [14] a relation between Shannon's noisy coding theorem and the fitness of a language is shown. They show that Shannon's error probability is inversely proportional to the fitness function.

2 Definitions, Notations and Known Results

We consider a special case of a model which was introduced in [13]. In this model a group of individuals can communicate about a given number of objects. We denote this set of objects by

$$\mathcal{O} = \{o_1, \ldots, o_N\}.$$

These are objects of the environment, other individuals, concepts or actions. Each object is mapped to a sequence of signals by the function

$$r : \mathcal{O} \to \mathcal{X}^n.$$

We represent each signal-sequence by a sequence of length n, where $\mathcal{X}$ is the set of all possible signals in the language. We call a signal-sequence, which describes an object, a word of the language. It is possible, that several objects are mapped to the same word. We assume that we have a distance function

$$d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$$

and $(\mathcal{X}, d)$ forms a distance space. We always write $\mathcal{X}$ for the distance space, if it is clear which distance function we use. If d satisfies, in addition, the following triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in \mathcal{X}$ and $d(x, y) = 0$ holds only for $x = y$, then $(\mathcal{X}, d)$ is called a metric space.

We denote by x_t for $1 \leq t \leq n$ the t-th letter of a word x^n, thus $x^n = (x_1, \ldots, x_n)$. The distance between two words is defined by $d^n(x^n, y^n) = \sum_{t=1}^{n} d(x_t, y_t)$, where $x^n, y^n \in \mathcal{X}^n$.

As in [13] we define the similarity of two words by $s : \mathcal{X}^n \times \mathcal{X}^n \to \mathbb{R}_+$, where

$$s(x^n, y^n) = exp(-d^n(x^n, y^n)).$$

We call a family

$$\mathcal{L} = \{x^n(i) \ : i = 1, \ldots, N\}$$

with $x^n(i) = r(o_i)$ a language for N objects in $\mathcal{X}^n$. Note that in this way it is allowed to use the same word in order to describe different objects.

The probability of understanding y^n when x^n was signalled is given by

$$p(x^n, y^n) = \frac{s(x^n, y^n)}{\sum_{i=1}^{N} s(x^n, x^n(i))}.$$

We assume that successful communication is of benefit to speaker and listener. Thus for each correct transmitted word for the i-th object both get a payoff a_i, which defines the value of this object. We assume here that $a_i = 1$ for all i.

With this restriction we define the fitness of a language $\mathcal{L}$ of length N in $\mathcal{X}^n$ by

$$F(\mathcal{L}, \mathcal{X}^n) = \sum_{i=1}^{N} p(x^n(i), x^n(i)).$$

The fitness of the distance space $\mathcal{X}^n$ is then defined as the maximal possible value of the fitness of all languages in $\mathcal{X}^n$. Thus

$$F(\mathcal{X}^n) = \sup\{F(\mathcal{L}, \mathcal{X}^n) : \mathcal{L} \textit{ language in } \mathcal{X}^n\}.$$

If we restrict the languages to be for a fixed number N of objects we define correspondingly:

$$F(\mathcal{X}^n, N) = \sup\{F(\mathcal{L}, \mathcal{X}^n) : \mathcal{L} \textit{ language in } \mathcal{X}^n \textit{for } N \textit{ objects}\}.$$

The next statement shows how the fitness values behave if we form languages of product type.

Let $\mathcal{L}$ be a language in the space $\mathcal{X}$ then the product language $\mathcal{L}^n$ is defined as the n-fold Cartesian product of $\mathcal{L}$, i.e., $\mathcal{L}^n = \times_{k=1}^{n} \mathcal{L}_k$, with $\mathcal{L}_k = \mathcal{L}$ for all k and the elements of the family $\mathcal{L}^n$ consist of all possible concatenations of n words from $\mathcal{L}$.

Proposition 1. *Let $\mathcal{L}$ be a language in the space $\mathcal{X}$. Then*

$$F(\mathcal{L}^n, \mathcal{X}^n) = F(\mathcal{L}, \mathcal{X})^n$$

and therefore

$$F(\mathcal{X}^n) \geq F(\mathcal{X})^n.$$

In [13] the authors considered three models for $\mathcal{X}$.

1. $\mathcal{X} = [0, a] \subset \mathbb{R}$ and $d(x, y) = |x - y|$,
2. $\mathcal{X} = [0, 1) \subset \mathbb{R}$ and $d(x, y) = \min\{|x - y|, 1 - |x - y|\}$,
3. $\mathcal{X} = \{0, d\}$ and $d(x, y) = \begin{cases} 0 \text{ , if x=y} \\ d \text{ , else} \end{cases}$

For the model 1 they obtained the

Theorem (NKD, [5])

1. $F([0, a]) = 1 + \frac{a}{2}$.
2. $F([0, a] \times [0, b]) = F([0, a])F([0, b])$.
3. $F([0, a]^n) = \left(1 + \frac{a}{2}\right)^n$.

Motivated by some experiments and this result Nowak formulated the following

Conjecture 1 (Product conjecture). *Let $(\mathcal{X}, d)$ be a metric space, then*

$$F(\mathcal{X}^n) = (F(\mathcal{X}))^n.$$

3 The Product Conjecture

Let $(\mathcal{X}, d)$ be a finite distance space. For a language $\mathcal{L}$ with N words (of length 1, that is letters) from $\mathcal{X}$ we introduce a language vector $\lambda = (\lambda_x)_{x\in\mathcal{X}}$, with

$$\lambda_x = \frac{Number\ of\ occurences\ of\ the\ word\ x}{N},$$

so that λ is a probability distribution (PD) on $\mathcal{X}$. With these definitions we can denote

$$F(\mathcal{L}, \mathcal{X}) = F(\mathcal{X}, \lambda) = \sum_x \frac{\lambda_x}{\sum_y \lambda_y e^{-d(x,y)}}.$$

For the fitness of the space $\mathcal{X}$ we can write

$$F(\mathcal{X}) = \max_\lambda F(\mathcal{X}, \lambda).$$

For a PD λ on $\mathcal{X}$, let λ^n denote the product-form distribution on $\mathcal{X}^n$ with marginals λ.

Property 1 now takes the form $F(\mathcal{X}^n, \lambda^n) = F(\mathcal{X}, \lambda)^n$ and $F(\mathcal{X}^n) \geq F(\mathcal{X})^n$. The product conjecture states that equality holds here for any metric space.

Supposition. In the following we shall assume, unless stated otherwise, that

(i) the diameter $D(\mathcal{X})$ of the set $\mathcal{X}$, defined as the maximum of $d(x, y)$ over all pairs (x, y) in $\mathcal{X}$, is finite, and
(ii) the matrix $[e^{-d(x,y)}]_{x,y\in\mathcal{X}}$ is positive semi-definite (psd.),

that is a self-adjoint square matrix with $A = A^T$ (Hermitian matrix) and all of whose eigenvalues are nonnegative. In our case all matrices are Hermitian because they are symmetric. We shall prove the product conjecture for such spaces. Recall that $d^n(x^n, y^n) = \sum_{t=1}^n d(x_t, y_t)$ is of sum-type.

We note that if $[e^{-d(x,y)}]_{x,y\in\mathcal{X}}$, is psd., then $[e^{-d^n(x^n,y^n)}]_{x^n,y^n\in\mathcal{X}^n}$ is psd.

This follows from the fact that $[e^{-d^n(x^n,y^n)}]$ is the nth tensor power of $[e^{-d(x,y)}]$.

3.1 A Lower Bound on $F(\mathcal{X})$

Since $F(\mathcal{X}) \geq F(\mathcal{X}, \lambda)$ for all PDs λ on $\mathcal{X}$, we obtain a lower bound on $F(\mathcal{X})$ for any choice of λ. Let λ^* be a PD that achieves the minimum in

$$\min_\lambda \sum_x \sum_y \lambda_x \lambda_y e^{-d(x,y)}. \tag{1}$$

Since by assumption the matrix $[e^{-d(x,y)}]$ is psd., the necessary and sufficient conditions for λ^* to achieve this minimum are given by the Karush-Kuhn-Tucker conditions, namely,

$$\sum_y \lambda_y^* e^{-d(x,y)} \geq c, \quad \text{for all } x \text{ with equality if } \lambda_x^* > 0, \tag{2}$$

where c is a constant whose value can be found by multiplying the two sides of the inequality by λ_x^* and summing over x,

$$c = \sum_x \lambda_x^* \sum_y \lambda_y^* e^{-d(x,y)}. \tag{3}$$

It turns out that the parameter $R_0(\mathcal{X})$ defined by

$$R_0(\mathcal{X}) = -\log c \tag{4}$$

plays a crucial role here. In terms of this parameter, we notice that

$$F(\mathcal{X}, \lambda^*) = \sum_x \lambda_x^* \frac{1}{e^{-R_0(\mathcal{X})}} = e^{R_0(\mathcal{X})}. \tag{5}$$

This gives us the following lower bound.

Proposition 2. *Under our Supposition for a space* $\mathcal{X}$*,*

$$F(\mathcal{X}) \geq e^{R_0(\mathcal{X})}, \tag{6}$$

where

$$R_0(\mathcal{X}) = -\log \min_\lambda \sum_x \sum_y \lambda_x \lambda_y e^{-d(x,y)}. \tag{7}$$

As an example we note that for $\mathcal{X}$ as Hamming space, $\mathcal{X} = \{0, 1\}$ with

$$d(x, y) = \begin{cases} 0 \text{ , if } x{=}y \\ 1 \text{ , else} \end{cases},$$

$R_0(\mathcal{X}) = \log[2/(1 + e^{-1})]$ and the lower bound is

$$F(\mathcal{X}) \geq \frac{2}{1 + e^{-1}}. \tag{8}$$

For use in the next section we note that $R_0(\mathcal{X}^n) = nR_0(\mathcal{X})$. This follows by observing that the optimality conditions (2) written for the space $\mathcal{X}^n$ are satisfied by a product-form distribution with marginals equal to λ^*. (We note the similarity of this result to the "parallel channel theorem" in [7], Chapter 5).

3.2 An Upper Bound

The following upper bound combined with the above lower bound establishes the product conjecture.

Proposition 3. *For all* $n \geq 1$

$$F(\mathcal{X}^n) \leq e^{nR_0(\mathcal{X}) + o(n)}.$$

Before proving this proposition, let us show that the product conjecture follows as a consequence.

Theorem 1. *For spaces satisfying our Supposition, the fitness function is given by* $F(\mathcal{X}^n) = e^{nR_0(\mathcal{X})}$.

Proof: Suppose to the contrary that for some m, $F(\mathcal{X}^m) \geq e^{mR_0(\mathcal{X})+\epsilon}$ for some $\epsilon > 0$. Then, by the fact that $F((\mathcal{X}^m)^k) \geq (F(\mathcal{X}^m))^k$, we have $F(\mathcal{X}^{mk}) \geq e^{km(R_0(\mathcal{X})+\epsilon/m)}$. Since ϵ/m is not an $o(m)$ term, this contradicts *Proposition 2.* Hence, we must conclude that for all $m \geq 1$, $F(\mathcal{X}^m) \leq e^{mR_0(\mathcal{X})}$. Since the reverse inequality $F(\mathcal{X}^m) \geq e^{mR_0}$ has already been established, the conclusion follows. □

Proof of Proposition 2: Fix $n \geq 1$ arbitrarily. Let λ be any PD on $\mathcal{X}^n$. Let S be the support set of λ. For each $x \in S$, define

$$A_x = \sum_y \lambda_y e^{-d(x,y)}$$

Note that for all $x \in S$

$$e^{-nD} \leq A_x \leq 1$$

where $D = D(\mathcal{X})$ is the diameter of $\mathcal{X}$ which is finite by assumption. Fix $\delta > 0$ arbitrarily and put $K = \lceil nD/\delta \rceil$. For $k = 1, \ldots, K$ define

$$S_k = \{x \in S : e^{-k\delta} < A_x \leq e^{-(k-1)\delta}\}$$

Note that these sets form a partition of S. So, we may write and justify afterwards

$$F(\mathcal{X}^n, \lambda) = \sum_{k=1}^{K} \sum_{x \in S_k} \lambda_x \frac{1}{A_x} \tag{9}$$

$$= \sum_k \lambda(S_k) \sum_{x \in S_k} \frac{\lambda_x}{\lambda(S_k)} \frac{1}{A_x} \tag{10}$$

$$\leq \sum_k \lambda(S_k) \frac{e^{\delta}}{\sum_{x \in S_k} \frac{\lambda_x}{\lambda(S_k)} A_x} \tag{11}$$

$$= \sum_k \lambda(S_k) \frac{e^{\delta}}{\sum_{x \in S_k} \frac{\lambda_x}{\lambda(S_k)} \sum_{y \in S} \lambda_y e^{-d(x,y)}} \tag{12}$$

$$\leq \sum_k \lambda(S_k) \frac{e^{\delta}}{\sum_{x \in S_k} \frac{\lambda_x}{\lambda(S_k)} \sum_{y \in S_k} \lambda_y e^{-d(x,y)}} \tag{13}$$

$$= \sum_k \frac{e^{\delta}}{\sum_{x \in S_k} \frac{\lambda_x}{\lambda(S_k)} \sum_{y \in S_k} \frac{\lambda_y}{\lambda(S_k)} e^{-d(x,y)}} \tag{14}$$

$$\leq \sum_k \frac{e^{\delta}}{e^{-nR_0(\mathcal{X})}} \tag{15}$$

$$= K e^{\delta} e^{nR_0(\mathcal{X})} \tag{16}$$

In (10) we have used $\lambda(S_k) = \sum_{x \in S_k} \lambda_x$. Inequality (11) follows by the following argument. For shorthand put $p_x = \lambda_x / \lambda(S_x)$ and recall that, for all $x \in S_k$,

$e^{-k\delta} < A_x \leq e^{-(k-1)\delta}$. Then,

$$\sum_{x \in S_k} p_x \frac{1}{A_x} \leq \sum_{x \in S_k} p_x \frac{1}{e^{-k\delta}} \tag{17}$$

$$= \frac{1}{\sum_{x \in S_k} p_x e^{-k\delta}} \tag{18}$$

$$\leq \frac{e^{\delta}}{\sum_{x \in S_k} p_x A_x} \tag{19}$$

In line (15), we used the assumption (ii) that the distance matrix is psd., hence $R_0(\mathcal{X}^n) = nR_0(\mathcal{X})$. The remaining inequalities are self-explanatory. Now, we may choose $\delta = \sqrt{n}$, say, then $K \approx \sqrt{n}$, and we have

$$F(\mathcal{X}^n, \lambda) \leq e^{nR_0(\mathcal{X}) + o(n)}.$$

Since the upper bound holds uniformly for all *PDs* λ, the fitness of the space is also upper-bounded by $e^{nR_0(\mathcal{X})+o(n)}$. This completes the proof. □

Remark 1

1. *It does not follow from the above results that $F(\mathcal{X}, \lambda)$ is a concave function of λ.*
2. *The proof can possibly be extended to any distance space with a bounded distance function but generalization to arbitrary distance spaces is not at all obvious.*
3. *The assumption about the positive semidefiniteness of the distance matrix appears to be essential. The Hamming metric, the metrics $|x-y|$ and $(x-y)^2$ defined on real spaces satisfy this constraint, as we show in the next section.*

3.3 A Connection Between Fitness and Parameters of Communication Channels

It is noteworthy that the Nowak fitness has an interesting relationship to pairwise error probabilities in noisy channels. Given a discrete memoryless channel $W : \mathcal{A} \to \mathcal{B}$, the Bhattacharyya distance (B-distance) between two input letters $a, a' \in \mathcal{A}$ is defined as

$$d_B(a, a') = -\log \sum_{b \in \mathcal{B}} \sqrt{W(b|a)W(b|a')}.$$

The cutoff rate parameter of the channel is defined as

$$R_0(W) = -\log \min_{\lambda} \left[\sum_{a \in \mathcal{A}} \sum_{a' \in \mathcal{A}} \lambda_a \lambda_{a'} e^{-d(a,a')} \right],$$

where the minimum is over all *PDs* $\lambda = \{\lambda_a : a \in \mathcal{A}\}$.

To illustrate the connection between fitness and channel coding, let $\mathcal{X} = \{0, 1\}$ with d the Hamming metric. The Hamming distance $d(x, y)$ for any $x, y \in \mathcal{X}$ equals the B-distance $d_B(x, y)$ of a binary symmetric channel $W : \mathcal{X} \to \mathcal{X}$ with crossover probability ϵ chosen so that $d_B(0, 1) = 1$, i.e., $\sqrt{4\epsilon(1-\epsilon)} = e^{-1}$. For W chosen this way, the cutoff rate of the BSC equals $R_0(W) = \log[2/(1 + e^{-1})]$. Thus, the $R_0(\mathcal{X})$ that appears as the exponent in the fitness growth rate for space $(\mathcal{X}, d)$ can be identified as the cutoff rate $R_0(W)$ of the associated BSC W.

This type of association between the metrics considered by Nowak et al. and B-distances of DMC's can be established in certain other cases as well. E.g., the metric $|x-y|$ is the B-distance for an exponential noise channel $W : X \to X+N$, where $X \geq 0$ is the channel input and $X + N$ is the channel output with N equal to an independent exponentially distributed random variable with intensity $\mu = 2$ (mean 1/2). Likewise, the metric $(x - y)^2$ can be interpreted as the B-distance for a Gaussian noise channel. Whenever a distance d can be associated with the B-distance of a channel, the matrix $[e^{-d(x,y)}]$ is a Gramm matrix and hence psd. Thus, the product conjecture holds for such distances on finite spaces.

This association between the fitness model and noisy communication channels is significant in that it explains the confoundability of phonemes as the result of the phonemes being sent across a noisy channel. This association also helps interpret Nowak's formula in terms of well-studied concepts in information theory, such as pairwise error probabilities and average list sizes in list-decoding.

3.4 Embedding of Distance Spaces

Let (X, d) and (X', d') be two distance spaces. Then (X, d) is said isometrically embeddable into (X', d') if there exists a mapping Φ (the isometric embedding) from X to X' such that $d(x, y) = d'(\Phi(x), \Phi(y))$ for all $x, y \in X$. For any $p \geq 1$, the vector space $\mathcal{R}^m$ can be endowed with the l_p-norm defined by $||x||_p = (\sum_{k=1}^{m} |x_k|^p)^{\frac{1}{p}}$ for $x \in \mathcal{R}^m$. The associated metric is denoted by d_{l_p}. The metric space (R^m, d_{l_p}) is abbreviated as l_p^m. A distance space is said to be l_p-embeddable, if (X, d) is isometrically embeddable into the space l_p^m for some integer $m \geq 1$. We call a distance space psd., if the corresponding matrix $[e^{-d(x,y)}]$ is psd.

Lemma 1. *If a distance space is psd., then all distance subspaces are also psd. Furthermore all distance spaces, which can be isometrically embedded in a subspace of a psd. distance space are psd.*

Proof: If $[e^{-d(x,y)}]$ is psd., then for all non-zero vectors x in $\mathcal{R}^n$ we have

$$x^T [e^{-d(x,y)}] x \geq 0.$$

This property remains if we delete a finite number of columns and rows of $[e^{-d(x,y)}]$. Therefore the remaining space is still psd. □

With the help of this lemma it is possible for us to establish the product conjecture for an arbitrary distance space whenever it is possible to embed it in a larger distance space which is psd. The following theorems of Vestfried and Fichet are very useful.

Theorem 1 (V, [16]). *Any separable ultrametric space is l_2-embeddable.*

Theorem 2 (F, [6]). *Any metric space with 4 points is l_1-embeddable.*

We describe now in a proposition situations where this technique applies. Recall that in an ultra-metric space for any three points a, b, c holds $d(a, b) \leq \max(d(a, c), d(c, b))$.

Proposition 4. *1. All ultra-metric spaces are psd.*
2. All finite metric spaces with up to 4 elements are psd.
3. There exist some metric spaces with 5 elements which are not psd.
4. For every distance space there exists a scaling, such that the space becomes psd.

Proof: 1. follows from the theorem of Vestfried and Lemma 1.
2. follows from the theorem of Fichet. To show 3. consider the following metric space on five points: Let for $i \neq j$ $d(i, j) = a$ if $i, j \in \{1, 2, 3\}$ and $d(i, j) = \frac{a}{2}$ otherwise. Then if $0 < a < 7.07 \cdot 10^{-6}$ the corresponding matrix is not psd. 4. follows, because the matrix $[e^{-\alpha d(x,y)}]$ converges for $\alpha \to \infty$ to the identity matrix. □

4 A Hamming Code Is a Good Language

In the previous section we have shown that the product conjecture is true in particular for the Hamming model. The optimal fitness is attained at $\lambda = (\frac{1}{2^n}, \ldots, \frac{1}{2^n})$. But this means, that one has to use all possible words in the language to achieve the optimal fitness. In general the memory of the individuals is restricted. For this reason we look for languages, which use only a fraction of all possible words, but have large fitness.

We consider simple and perfect codes: The Hamming codes ([8]). A q-ary block-code of length n is a map c from a finite set $\mathcal{O}$ to $\{0, 1, \ldots, q-1\}^n$. $c(o)$ with $o \in \mathcal{O}$ is called a codeword and $\mathcal{C} = \{c(o) : o \in \mathcal{O}\}$ is called the code. Thus we can view each code as a language. There exists a lot of work about codes (see [11]). A special class of codes are the t-error correcting block-codes. These codes have the property that for two different codewords the Hamming-distance is larger than $2t+1$. For a block-code of length n the weight-distribution $(A_0, A_1, \ldots, A_n)$ and the distance distribution $(B_0, \ldots, B_n)$ are defined. A_i denotes the number of codewords of weight i and B_i is the number of ordered pairs of codewords (u, v) such that $d(u, v) = i$ divided by the number of messages. We summarize the properties of the single-error-correcting Hamming-codes.

Proposition 5. *1. Hamming codes exist for the lengths $2^k - 1$.*
2. Their number of codewords is $N = 2^{2^k-1-k}$. The minimal distance is 3.
3. The weight distribution is the same for each word.

In [9] and [10] it is shown that the weight distribution is very easy to calculate. Let $(A_0, A_1, \ldots, A_n)$ be the distance-distribution of the Hamming-code $\mathcal{C}$, then we define the Hamming weight enumerator by

$$W_{\mathcal{C}}(x) = \sum_{i=0}^{n} A_i x^i.$$

Theorem 3 (McW, [9],[10]). *Let $(A_0, A_1, \ldots, A_n)$ be the distance-distribution of the Hamming-code $\mathcal{C}$, then the Hamming weight enumerator of this code is given by*

$$W(x) = \frac{1}{n+1}\left((1+x)^n + n(1-x)(1-x^2)^{\frac{n-1}{2}}\right).$$

With $F_H(n)$ we denote the fitness of a Hamming Code of length n.

Theorem 2. *The fitness of the Hamming code approaches asymptotically the optimal fitness. Not only* $\lim_{n\to\infty} \frac{1}{n} F_H(n) = \lim_{n\to\infty} \frac{1}{n} F(\mathcal{X}^n)$ *and* $\lim_{n\to\infty} \frac{F_H(n)}{F(\mathcal{X}^n)} = 1$, *but even the stronger condition*

$$\lim_{n\to\infty} F_H(n) - F(\mathcal{X}^n) = 0$$

holds.

Proof

The fitness of the Hamming code can be expressed using the weight enumerator W.

$$F_H(n) = \frac{2^{2^k-1-k}}{W(exp(-1))} = \frac{2^{n-\log_2(n+1)}}{W(exp(-1))}.$$

We now show that the difference $F(\mathcal{X}^n) - F_H(n)$ goes to zero.

$$F(\mathcal{X}^n) - F_H(n) = \left(\frac{2}{1+exp(-1)}\right)^n - \frac{2^{n-\log_2(n+1)}}{W(exp(-1))}$$

$$= \left(\frac{2}{1+e^{-1}}\right)^n - \frac{2^{n-\log_2(n+1)}}{\frac{1}{n+1}(1+e^{-1})^n + n(1-e^{-1})(1-e^{-2})^{\frac{n-1}{2}}}$$

$$= \left(\frac{2}{1+e^{-1}}\right)^n - \frac{2^n}{(1+e^{-1})^n + n(n+1)(1-e^{-1})(1-e^{-2})^{\frac{n-1}{2}}}$$

$$= \frac{2^n n(n+1)(1-e^{-1})(1-e^{-2})^{\frac{n-1}{2}}}{(1+e^{-1})^{2n} + (1+e^{-1})^n n(n+1)(1-e^{-1})(1-e^{-2})^{\frac{n-1}{2}}}$$

$$\leq \frac{2^n n(n+1)(1-e^{-1})(1-e^{-2})^{\frac{n-1}{2}}}{(1+e^{-1})^{2n}},$$

$$= \frac{\left(2\sqrt{(1-e^{-2})}\right)^n n(n+1)(1-e^{-1})}{((1+e^{-1})^2)^n \sqrt{(1-e^{-2})}}.$$

The last term goes to zero if n goes to infinity, because

$$\frac{2\sqrt{1-e^{-2}}}{(1+e^{-1})^2} < 1,$$

($\frac{2\sqrt{1-e^{-2}}}{(1+e^{-1})^2} < 0.995$). Since the difference is always positive the proof is complete. □

Next we show that *ratewise* the fitness of the Hamming space is attained if we choose the middle level as a language.

Suppose that n is even and let the language $\mathcal{L}$ consist of all words x^n with exactly $\frac{n}{2}$ ones, i.e. $w(x^n) = \frac{n}{2}$. If we fix any word from this language then there are $\binom{\frac{n}{2}}{j}^2$ words in $\mathcal{L}$ at a distance of $2j$, $(j = 0, \ldots, \frac{n}{2})$. Therefore the fitness of $\mathcal{L}$ is

$$F(\mathcal{L}, \mathcal{X}^n) = \frac{\binom{n}{\frac{n}{2}}}{\sum_{j=0}^{\frac{n}{2}} \binom{\frac{n}{2}}{j}^2 e^{-2j}}.$$

Let $j^\star(n)$ denote the j for which the summand in the denominator is maximal and let $\tau^\star(n) = \frac{j^\star(n)}{n}$. Then we can estimate the rate of the fitness of $\mathcal{L}$ as follows. Let $\epsilon > 0$.

$$\frac{1}{n} \log F(\mathcal{L}, \mathcal{X}^n) = \frac{1}{n} \log \binom{n}{\frac{n}{2}} - \frac{1}{n} \log \sum_{j=0}^{\frac{n}{2}} \binom{\frac{n}{2}}{j}^2 e^{-2j}$$

$$\geq \frac{1}{n} \log \binom{n}{\frac{n}{2}} - \frac{1}{n} \log \left((\frac{n}{2} + 1) \binom{\frac{n}{2}}{j^\star(n)}^2 e^{-2j^\star(n)} \right)$$

$$= \frac{1}{n} \log \binom{n}{\frac{n}{2}} - \frac{1}{n} \log \left(\frac{n}{2} + 1 \right) - \frac{1}{\frac{n}{2}} \log \binom{\frac{n}{2}}{2\tau^\star(n) \cdot \frac{n}{2}} + 2\tau^\star(n) \log(e),$$

which we can bound further for sufficiently large n by

$$\geq 1 - 0 + \min_{\tau} \{-h(2\tau) + 2\tau \log(e)\} - \epsilon, \tag{20}$$

where h is the binary entropy function, $h(\tau) = -\tau \log \tau - (1 - \tau) \log(1 - \tau)$.

We can find the minimum of the convex function $-h(2\tau) + 2\tau \log(e)$ by looking at the root of the first derivative. The first derivative is $2 \log(2\tau) - 2 \log(1 - 2\tau) + 2 \log(e)$, which is zero for $\tau = \frac{1}{2(1+e)}$. Substituting this in (20) we can conclude that for sufficiently large n

$$\frac{1}{n} \log F(\mathcal{L}, \mathcal{X}^n) \geq 1 - \log(1 + e^{-1}) - \epsilon.$$

The opposite inequality $\frac{1}{n} \log F(\mathcal{L}, \mathcal{X}^n) \leq 1 - \log(1 + e^{-1})$ is also true because we know from Theorem 1 that for the Hamming space $F(\mathcal{X}^n) = \left(\frac{2}{1+e^{-1}} \right)^n$. Therefore we can summarize our result in the following theorem.

Theorem 3. *Let $\mathcal{L}$ be the language in the Hamming space $\mathcal{X}^n$ that consists of all words of weight $\frac{n}{2}$. Then the fitness of the language $\mathcal{L}$ is ratewise optimal, i.e.,*

$$\lim_{n \to \infty} \frac{1}{n} \log F(\mathcal{L}, \mathcal{X}^n) - \frac{1}{n} \log F(\mathcal{X}^n) = 0.$$

Theorem 4. *Let c be a fixed integer and $\mathcal{L}$ be the language in the Hamming space $\mathcal{X}^n$ that consists of all words of weight $\frac{n}{2}$ with $\lceil \frac{c}{2} \rceil$ fixed position with 0's and $\lfloor \frac{c}{2} \rfloor$ fixed positions with 1's. Then the fitness of the language $\mathcal{L}$ is also ratewise optimal, i.e.,*

$$\lim_{n\to\infty} \frac{1}{n} \log F(\mathcal{L}, \mathcal{X}^n) - \frac{1}{n} \log F(\mathcal{X}^n) = 0.$$

Proof: We assume that n and c are even. Following the same idea as in Theorem 3 we get for $\epsilon > 0$.

$$\frac{1}{n} \log F(\mathcal{L}, \mathcal{X}^n) = \frac{1}{n} \log \left(\frac{1}{2^c} \binom{n}{\frac{n}{2}} \right) - \frac{1}{n} \log \sum_{j=0}^{\frac{n-c}{2}} \binom{\frac{n-c}{2}}{j}^2 e^{-2j}$$

$$\geq \frac{1}{n} \log \left(\frac{1}{2^c} \binom{n}{\frac{n}{2}} \right) - \frac{1}{n} \log \left((\frac{n-c}{2} + 1) \binom{\frac{n-c}{2}}{j^\star(n)}^2 e^{-2j^\star(n)} \right)$$

$$= \frac{1}{n} \log \left(\frac{1}{2^c} \binom{n}{\frac{n}{2}} \right) - \frac{1}{n} \log \left(\frac{n-c}{2} + 1 \right) - \frac{1}{\frac{n}{2}} \log \binom{\frac{n-c}{2}}{2\tau^\star(n) \cdot \frac{n}{2}} + 2\tau^\star(n) \log(e),$$

which we can bound further for sufficiently large n by

$$\geq 1 - 0 + \min_{\tau} \{-h(2\tau) + 2\tau \log(e)\} - \epsilon. \tag{21}$$

□

5 A Language with Noiseless Feedback

In this section we consider a language with noiseless feedback. The channel model is well known in Information Theory ([3], [2]). It can be described in our language model as follows. Individual A signalled a letter (word of length 1) and is informed which letter individual B understood (because of some reaction of B). Individual A has a special strategy for each object. After n repetitions of this procedure B notices some object with a certain probability. We denote the set of objects like before by $\mathcal{O} = \{o_1, \ldots, o_N\}$.

The functions

$$st_j(o_i, y^{j-1})$$

for $j = 1, \ldots, n$ define the next signal given by the speaker if he wants to speak about object i and the listener understands $y^n \in \{0,1\}^n$. Thus

$$st_j : \mathcal{O} \times \{0,1\}^{j-1} \to \{0,1\}.$$

We define the set of error vectors by

$$\mathcal{E} = \{0,1\}^n.$$

This is the set of all possible error vectors. Let $e^n = (e^n(1), \ldots, e^n(n)) \in \mathcal{E}$, if $e^n(t) = 1$ then an error happened at the t-th position, otherwise $e^n(t) = 0$.

We set $0^n = (0, \ldots, 0)$ a vector of length n. The error vector and the strategy determine what the the speaker says. Thus we have a function

$$st : \mathcal{O} \times \mathcal{E} \to \{0,1\}^n$$

where $st(o_i, e^n)$ is defined by

$$(st_1(o_i), st_2(o_i, st_1(o_1) + e_1 = y^1), \ldots, st_n(o_i, st_{n-1}(o_{n-1}, y^{n-2}) + e_{n-1} = y^{n-1})).$$

We define the feedback-language as $\mathcal{L}^{st} = (st(o_t, 0^n))_{t=1}^N$. We need a distance-function to define the fitness in this case. We define the similarity for two words as follows. $s(x^n, y^n) = e^{-t}$, where

$$t = \begin{cases} \min\{w(e^n) : st(o_i, e^n) \oplus e^n = y^n\} \; if \; \exists y^n : st(o_i, e^n) \oplus e^n = y^n \\ 0 \qquad\qquad\qquad\qquad otherwise \end{cases}.$$

The feedback fitness of a strategy is defined as

$$F^f(st, \mathcal{X}^n) = \sum_{t=1}^{|\mathcal{L}^{st}|} \sum_{e^n : st(o_t, e^n) \in \mathcal{L}} \frac{1}{\sum_{e^n : st(o_t, e^n) \in \mathcal{L}} s(st(o_t, 0^n), st(o_t, e^n))}$$

and the fitness is defined as the maximal possible value of the fitness of all strategies in $\mathcal{X}^n$. Thus

$$F^f(\mathcal{X}^n) = \sup_{st}\{F^f(st, \mathcal{X}^n)\}.$$

This is a generalization of the model without feedback. If the speaker just ignores the feedback, we get the same model like before. We write F^f for all fitness definitions, if we consider the fitness with feedback.

Proposition 6

$$F(\mathcal{X}^n) = F(\mathcal{X}^n, \mathcal{X}^n) = F^f(\mathcal{X}^n, \mathcal{X}^n).$$

Proof: The property holds because, if we use all possible words of a language, all similarities between the words occur in the fitness formula in the summands just in another order. □

Now we will give an example where the feedback-fitness is bigger than the usual fitness. For the case $n = 3$ we know that $F(\{0,1\}^3) = \left(\frac{2}{1+exp(-1)}\right)^3$. We will show now that the fitness can be increased with feedback. We give an example for a feedback-language with seven objects and a bigger fitness.

Example: Strategy f: Map the i-th object to the binary representation of i. If a 1 is understood as a 0 start saying 0.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7
$t=0$	001	010	011	100	101	110	111
	000	000	000	000	000	000	000
$t=1$	011	110	111	101	100	111	110
	101	011	010	110	111	100	100
	100	001	001	001	001	001	001
$t=2$	010	111	110	010	010	010	010
	111	100	100	111	110	101	101
$t=3$	110	101	101	011	011	011	011

Obviously $\mathcal{L}^{st} = \{0,1\}^n \backslash 0^n$.

It holds $F^f(\mathcal{L}^*) = 3,19 > F(\mathcal{X}^n, \mathcal{X}^n)$. Our strategy can be generalized and gives a lower bound for the feedback fitness.

Proposition 7

$$F^f(\mathcal{X}^n) \geq \frac{2^n - 1}{\left(\sum_{j=0}^{n} \binom{n}{j} e^{-j}\right) - e^{-1}}.$$

Proof: Use the generalization of the strategy in the example and the result follows. □

It is also possible to give a trivial upper bound.

Proposition 8

$$F^f(\mathcal{X}^n, N) \leq \frac{N}{1 + (N-1)e^{-n}}.$$

Proof: The smallest possible similarity between two different words is e^{-n}. Thus we assume that all similarities of all possible words are as small as possible and get the upper bound for the fitness. □

6 List-Language

In a "list-language", we divide the words of a language $\mathcal{L}$ into lists (subfamilies). For example words about food, words about danger e.t.c.. The goal of the listener is just to find out about which list the speaker speaks. To simplify the situation we assume that all words of the language $\mathcal{L}$ belong to exactly one list, all lists are of the same size l and we look only at languages with $l|N$, $(N = |\mathcal{L}|)$. In general, if $|\mathcal{L}| = l \cdot k + r$ with $r < l$, we have r lists of size $l+1$ and $k-r$ lists of size l, i.e., here we assume that $r = 0$ and call such a language an l-list-language.

We denote the lists by $\mathcal{L}_i$ for $i = 1, \ldots, k$. We set $L(x^n) = \mathcal{L}_i$, if the word x^n belongs to the list $\mathcal{L}_i$.

In a list-language the individuals get some profit, if the listener understands the list of the speaker. Therefore we define

$$F^l(\mathcal{L}, \mathcal{X}^n) = \sum_{x^n \in \mathcal{L}} \sum_{y^n \in L(x^n)} p(x^n, y^n).$$

Then naturally the question of the best l-list-language arises:

$$F^l(\mathcal{X}^n) = \sup\{F^l(\mathcal{L}, \mathcal{X}^n) : \mathcal{L} \; is \; l - \text{list} - \text{language in } \mathcal{X}^n\}.$$

Next we calculate the fitness of list-languages in a special case, namely that of *constant similarity*. Let $C > 0$ be a constant and let d be the following metric on $\mathcal{X}$

$$d(x, y) = \begin{cases} 0 \text{ , if } x{=}y \\ C \text{ , else} \end{cases}.$$

In this case the following proposition holds.

Proposition 9. $F^l(\mathcal{L}, \mathcal{X}) \leq F(\mathcal{L}, \mathcal{X}) + l - 1$

Proof

$$F^l(\mathcal{L}, \mathcal{X}) = \sum_{x \in \mathcal{L}} \sum_{y \in L(x)} \frac{exp(-d(x, y))}{\sum_{z \in \mathcal{L}} exp(-d(x, z))}$$

$$= F(\mathcal{L}, \mathcal{X}) + \sum_{x \in \mathcal{L}} \sum_{y \in L(x), y \neq x} \frac{exp(-d(x, y))}{\sum_{z \in \mathcal{L}} exp(-d(x, z))}$$

$$= F(\mathcal{L}, \mathcal{X}) + \frac{Nexp(-C)(l-1)}{1 + (N-1)exp(-C)} \leq F(\mathcal{L}, \mathcal{X}) + l - 1.$$

□

7 Multi-access-Language

In this section we will consider the following situation. Two individuals speak simultaneously. There is some interference and one individual wants to understand both. We look at two models. In the first model the speakers use the same language, in the second model they use different languages. Such models are well known in Information Theory. They were introduced in [1].

7.1 Model I

In this model $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, 2\}$. The individuals can only speak words which contain the signals 0 and 1. The listener understands 0 if both use the signal 0. He understands 1, if one individual uses the signal 0 and the other the signal 1 and he understands 2 if both use the signal 1. The listener understands some word in $\mathcal{Y}^n$. We search now for a language with the biggest multi-access-fitness. This model is known in information theory as the binary adder channel.

We set $d((x^n, y^n), (v^n, w^n)) = d_H(x^n + y^n, v^n + w^n)$, where $x^n + y^n = (x_1 + y_1, \ldots, x_n + y_n)$ and define the fitness of a multi-access-adder-language as

$$F_A(\mathcal{L}, \mathcal{X}^n) = \sum_{i=1}^{N} \sum_{j=1}^{N} p((x(i)^n, y(j)^n), (x(i)^n, y(j)^n)).$$

The probability and the similarity are defined as before. We will consider an example for $n = 2$. The language contains all elements of $\{0, 1\}$ exactly once. Thus we get the following table:

	$r(o_1) = 00$	$r(o_2) = 01$	$r(o_3) = 10$	$r(o_4) = 11$
$r(o_1) = 00$	00	01	10	11
$r(o_2) = 01$	01	02	11	12
$r(o_3) = 10$	10	11	20	21
$r(o_4) = 11$	11	12	21	22

Now we have for example $d((01, 01), (10, 01)) = d(02, 11) = 2$. The fitness of this language is $F_A(\mathcal{L}, \mathcal{X}^n) \approx 2.83$.

Proposition 10

$$F_A(\mathcal{X}^n) \leq F(\{0, \ldots, 2|\mathcal{X}|\}.$$

We know consider a generalization of this model. The speaker uses two different languages over the same distance space. We search for two languages which have the biggest common multi-access-fitness.

$$F_A(\mathcal{L}, \mathcal{M}, \mathcal{X}^n) = \sum_{i=1}^{m} \sum_{j=1}^{k} p((x(i)^n, y(j)^n), (x(i)^n, y(j)^n)).$$

For example let $\mathcal{L} = (00, 01, 10, 11)$ and $\mathcal{M} = (00, 11)$. Then we get the following table.

	00	01	10	11
00	00	01	10	11
11	11	12	21	22

The fitness of this language is $F_A(\mathcal{L}, \mathcal{M}) = 2,71$.

7.2 Model II

In this model $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and $d((x^n, y^n), (x^n, y^n)) = d_H(x^n \oplus y^n, x^n \oplus y^n)$, where $\oplus$ is the sum modulo $|\mathcal{X}| = 2$ in all components. All other definition are the same. Let us look at our example:

	00	01	10	11
00	00	01	10	11
01	01	00	11	10
10	10	11	00	01
11	11	10	01	00

All words are contained four times in the table. Thus this language attains the maximum, because the product conjecture holds. This can be generalized.

Theorem 5. *The optimal fitness for the adder model II is attained, if the language consists of all possible codewords.*

Another configuration with the same fitness as the previous example:

	00 01
00	00 01
10	10 11

If we allow two different languages for the two speakers, we find more configurations, which attain the optimal fitness.

Let us look at our example:

	00 01 10 11
00	00 01 10 11

8 Broadcast to Two Different Languages

In this section we will consider the following situation. We have two individuals with two different languages $\mathcal{L} = (x(1), \ldots, x(N))$ and $\mathcal{M} = (y(1), \ldots, y(N))$ on the same distance space $(\mathcal{X}, d)$, such that $x(i)$ describes the same object as $y(i)$ for all $i = 1, \ldots, N$. Our goal is to find a good language for a third individual, which wants to communicate with both of them simultaneously. In Information Theory this kind of models were introduced in [4].

We define the fitness between two languages as

$$F(\mathcal{L}, \mathcal{M}) = \sum_{i=1}^{N} \frac{exp(-d(x(i), y(i)))}{\sum_{j=1}^{N} exp(-d(x(i), y(j))}.$$

There exists also examples in human language, where both people can speak in their own language and understand each other. An example is a conversation between a Swede and a Dane, who both speak in their language.

We define the fitness of a broadcast-language $\mathcal{N}$ as

$$F_B\left(\mathcal{N}, (\mathcal{L}, \mathcal{M}), \mathcal{X}^n\right) = \frac{1}{2}\left(F\left(\mathcal{N}, \mathcal{L}\right) + F\left(\mathcal{N}, \mathcal{M}\right)\right).$$

Proposition 11

$$F_B\left(\mathcal{N}, (\mathcal{L}, \mathcal{M}), \mathcal{X}^n\right) \geq \frac{1}{2} \max\{F(\mathcal{L}, \mathcal{X}^n) + 1, F(\mathcal{M}, \mathcal{X}^n) + 1\}.$$

9 Language Without Multiplicity

In all previous sections we allowed multiplicity of words. That means the individuals were allowed to use one word for more than one object. We will show that in the case without multiplicity there are examples, where the fitness of a product space is bigger than the product of the fitnesses of the single spaces.

Again we consider the set of objects

$$\mathcal{O} = \{o_1, \ldots, o_N\}$$

and now each object is mapped to a sequence of signals by the injective function

$$r : \mathcal{O} \to \mathcal{X}^n.$$

We call the languages of this type *injective* and denote the corresponding fitness values by F_{in}.

We consider the metric space $(\mathcal{M} = \{a, b, c\}, d)$, where the distance is defined as follows:

d	a	b	c
a	0	0.01	3
b	0.01	0	3
c	3	3	0

In this case holds:

$$F_{in}(\mathcal{M}) = F_{in}(\{a, c\}, \mathcal{M}) = \frac{2}{1 + e^{-3}} > F_{in}(\{a, b, c\}, \mathcal{M}),$$

but for the product we have:

$$F_{in}(\mathcal{M}^2) = F_{in}(\{aa, ac, cb, cc\}, \mathcal{M}^2) > F_{in}(\{aa, ac, ca, cc\}, \mathcal{M}^2).$$

Thus $F_{in}(\mathcal{M})^2 < F_{in}(\mathcal{M}^2)$. This means the product conjecture does not hold for injective languages. The reason for this behavior is, that the distance between a and b is very small and the optimal fitness does not consist of all possible letters. In the product space we can use the unused letter to improve the fitness. This counterexample does not work in the original problem, because in the case of such a finite metric space it is always better to choose all elements with a certain multiplicity.

Acknowledgment. The authors would like to thank V. Blinovsky and E. Telatar for discussions on these problems and P. Harremoes for drawing their attention to the counter-example in the case without multiplicity.

References

1. R. Ahlswede, Multi-way communication channels, Proceedings of 2nd International Symposium on Information Theory, Thakadsor, Armenian SSR, Sept. 1971, Akademiai Kiado, Budapest, 23-52, 1973.
2. R. Ahlswede, I. Wegener, Suchprobleme, Teubner, 1979, English translation: Wiley, 1987, Russian translation: MIR, 1982.
3. E.R. Berlekamp, Block coding for the binary symmetric channel with noiseless, delayless feedback in H.B.Mann, Error Correcting Codes, Wiley, 61-85, 1968.
4. T.M. Cover, Broadcast channels, IEEE Trans. Inform. Theory, Vol. 18, 2-14, 1972.
5. A. Dress, The information storage capacity of a metric space, preprint.
6. B. Fichet, Dimensionality problems in l_1-norm representations, in Classification and Dissimilarity Analysis, Lecture Notes in Statistics, Vol. 93, 201-224, Springer-Verlag, Berlin, 1994.

7. R.G. Gallager, Information Theory and Reliable Communication, New York, Wiley, 1968.
8. R.V. Hamming, Error detecting and error correcting codes, Bell Sys. Tech. Journal, 29, 147-160, 1950.
9. F.J. MacWilliams, Combinatorial problems of elementary group theory, Ph.D. Thesis, Harvard University, 1962.
10. F.J. MacWilliams, A theorem on the distribution of weights in a systematic code, Bell syst. Tech. J., Vol. 42, 79-94, 1963.
11. F.J. MacWilliams and N.J.A. Sloane, The Theory of Error-Correcting Codes, Elsevier Science Publishers B.V., 1977.
12. M.A. Nowak and D.C. Krakauer, The evolution of language, PNAS 96, 14, 8028-8033, 1999.
13. M.A. Nowak, D.C. Krakauer, and A. Dress, An error limit for the evolution of language, Proceedings of the Royal Society Biological Sciences Series B, 266, 1433, 2131-2136, 1999.
14. J.B. Plotkin and M.A. Nowak, Language evolution and information theory, J. theor. Biol. 205, 147-159, 2000.
15. C.E. Shannon, The zero-error capacity of a noisy channel, IRE Trans. Inform. Theory E, 8–19, 1956.
16. A.F. Timan and I.A. Vestfried, Any seperable ultrametric space is isometrically embeddable in l_2, Funk. Anal. Pri. 17, 1, 85-86, 1983.