THE ASYMPTOTIC BEHAVIOUR OF DIAMETERS IN THE AVERAGE

Rudolf Ahlswede and Ingo Althöfer

Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
33501 Bielefeld
Germany

Abstract

In 1975 Ahlswede and Katona posed the following average distance problem ([1], page 10): For every cardinality $a \in \{1, \dots, 2^n\}$ determine subsets A of $\{0,1\}^n$ with $\#A = a$, which have **minimal** average **inner** Hamming distance. Recently Althöfer and Sillke gave an exact solution of this problem for the central value $a = 2^{n-1}$.

Here we present nearly optimal solutions for $a = 2^{\lambda n}$ with $0 < \lambda < 1$: Asymptotically it is not possible to do better than choosing

$$A_n = \big\{(x_1, \dots, x_n) \mid \sum_{t=1}^{n} x_t = \lfloor \alpha n \rfloor \big\},$$

where $\lambda = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$.

Next we investigate the following more general problem, which occurs for instance in the construction of good "Write–Efficient–Memories [WEMs]":

Given any finite set M with an arbitrary cost function $d : M \times M \to \mathbb{R}$, the corresponding sum type cost function $d_n : M^n \times M^n \to \mathbb{R}$ is defined by $d_n\big((x_1, \dots, x_n)$, $(y_1, \dots, y_n)\big) = \sum_{i=1}^{n} d(x_t, y_t)$. The task is to find sets A_n of a given cardinality, which minimize the average inner cost $\frac{1}{(\#A_n)^2} \sum_{a \in A_n} \sum_{a' \in A_n} d_n(a, a')$. We prove that asymptotically optimal sets can be constructed by using "mixed typical sequences" with at most two different local configurations. As a non–trivial example we look at the Hamming distance for $M = \{1, \dots, m\}$ with $m \geq 3$.

1. $\{0,1\}^n$ and the Hamming Distance

For two elements $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in $\{0,1\}^n$ the Hamming distance is defined by

$$d(x,y) = \#\{t \mid x_t \neq y_t\} .$$

For a set $A \subset \{0,1\}^n$ the average inner distance is defined by

$$\overline{d}(A) = \frac{1}{(\#A)^2} \sum_{x \in A} \sum_{y \in A} d(x,y) .$$

Let $\qquad \overline{d}_n(a) = \min_{\substack{A \subset \{0,1\}^n : \\ \#A = a}} \overline{d}(A) \qquad$ for all $a \in \{0, 1, \ldots, 2^n\}$.[1]

We derive asymptotically tight bounds for $\overline{d}_n(a)$, when $a \approx \binom{n}{\alpha n}$ with a constant $\alpha \in (0, \frac{1}{2})$.

We show in this section

Theorem 1.1:

Let $(a_n)_{n=1}^\infty$ be a sequence of natural numbers with $0 \leq a_n \leq 2^n$ for all n and $\lim_{n\to\infty} \inf \frac{a_n}{\binom{n}{\lfloor \alpha n \rfloor}} > 0$ for some constant $\alpha \in (0, \frac{1}{2})$. Then

$$\lim_{n\to\infty} \inf \frac{\overline{d}_n(a_n)}{n} \geq 2\alpha(1-\alpha) .$$

The optimality of this bound is readily demonstrated.

The weight of $x \in \{0,1\}^n$ is defined by $w(x) = \#\{t | x_t = 1\}$. The level sets $A_n = \{x \in \{0,1\}^n \mid w(x) = \lfloor \alpha n \rfloor\}$ fulfill the cardinality conditions and yield average distances as desired. Notice also that for $a = 2^{n-1}$ the subcube — and not the sphere — is the best configuration ([2]).

Proof of Theorem 1.1:

The key idea in the proof is to generalize the problem by studying probability distributions on $\{0,1\}^n$ with a given entropy instead of sets with a given cardinality.

Let $P = \big(P(x)\big)_{x \in \{0,1\}^n}$ be a probability distribution on $\{0,1\}^n$. The average inner distance of P is defined by

[1] In this paper $d, \overline{d}, \overline{d}_n$ etc. are functions related to distances or cost functions. When the same symbol is used for more than one function, their differences are made clear by the symbols used for the arguments. As a benefit for this loose notation the reader is not burdened with too many symbols.

$$\overline{d}(P) = \sum_{x \in \{0,1\}^n} \sum_{y \in \{0,1\}^n} P(x)P(y)d(x,y) \ .$$

The entropy of P is given by

$$H(P) = \sum_{x \in \{0,1\}^n} -P(x) \log P(x) \ .$$

(In this note we take the logarithm with base 2.)

We have $0 \leq H(P) \leq n$ for every distribution P on $\{0,1\}^n$.

Let

$$\hat{d}_n(H) = \min \overline{d}(P),$$

where the min is taken over all P with $H(P) \geq H$.

Lemma 1.2:

Let $(H_n)_{n=1}^{\infty}$ *be a sequence of real numbers with* $0 \leq H_n \leq n$ *for all* $n \in \mathbb{N}$ *and* $\lim_{n \to \infty} \inf \frac{H_n}{n} \geq \lambda$ *for some constant* $\lambda \in (0,1]$.

Then

$$\lim_{n \to \infty} \inf \frac{\hat{d}_n(H_n)}{n} \geq 2\alpha(1-\alpha),$$

where $\alpha \in (0, \frac{1}{2})$ with $\lambda = h(\alpha) := -\alpha \log \alpha - (1-\alpha) \log(1-\alpha)$.

The theorem can be derived from this lemma in the following way.

A set $A \subset \{0,1\}^n$ corresponds in a natural way to the probability distribution P_A , given by

$$P_A(x) = \begin{cases} \frac{1}{\#A}, & \text{if } x \in A \\ 0, & \text{if } x \notin A. \end{cases}$$

We have

$$\overline{d}(A) = \overline{d}(P_A)$$

and

$$H(P_A) = \log \#A \ .$$

Theorem 1.1 follows from Lemma 1.2, as by Stirling's formula

$$\lim_{n\to\infty} \inf \frac{a_n}{\binom{n}{\lfloor \alpha n \rfloor}} > 0$$

implies

$$\lim_{n\to\infty} \inf \frac{\log a_n}{n} \geq h(\alpha) \ .$$

It remains to prove the lemma.

Proof of Lemma 1.2:

For a probability distribution P on $\{0,1\}^n$ we define marginal 1–probabilities

$$p_t = \sum_{\substack{x\in\{0,1\}^n: \\ x_t=1}} P(x) \quad \text{for} \quad t = 1, \dots, n \ .$$

From the properties of the entropy function [3] it follows that

$$H(P) \leq \sum_{t=1}^{n} h(p_t) = \sum_{t=1}^{n} -p_t \log p_t - (1-p_t)\log(1-p_t), \tag{1.1}$$

where equality holds iff P is the product of n distributions $(1-p_t, p_t)$ on $\{0,1\}$.

For the average inner distance of P we have

$$\overline{d}(P) = \sum_{t=1}^{n} 2p_t(1-p_t), \tag{1.2}$$

hence it is completely determined by the p_t .

The problem of minimizing $\overline{d}(P)$ for a fixed entropy level $H(P)$ is equivalent to maximizing $H(P)$ for a fixed distance level $\overline{d}(P)$. Thus by (1.1) and (1.2) it is sufficient to solve the following analytical problem. For $f(p_1, p_2, \dots, p_n) = \sum_{t=1}^{n} h(p_t)$ find

$$\left.\begin{array}{l} \displaystyle\max_{0\leq p_t\leq 1 \text{ for } t=1,\dots,n} f(p_1,\dots,p_n) \\ \displaystyle\text{under the constraint} \sum_{t=1}^{n} 2p_t(1-p_t) = 2\alpha(1-\alpha)n \ . \end{array}\right\} \tag{1.3}$$

By the symmetry of $h(p)$ and $p(1-p)$ in p and $(1-p)$ we may assume without loss of generality that $0 \leq p_t \leq \frac{1}{2}$ for all t .

The statement of the lemma suggests that the solution of (1.3) is to choose $p_t = \alpha$ for all t . This will be proved below by a simple exchange argument between only two coordinates:

Find $\max_{0 \le p_1, p_2 \le \frac{1}{2}} f(p_1, p_2)$

under the constraint $g(p_1,p_2) = 2p_1(1-p_1) + 2p_2(1-p_2) = c$ for some constant $c \in [0,1]$. (1.4)

Claim: For every constant $c \in [0,1]$, (1.4) is solved by choosing $p_1 = p_2$.

Proof of the Claim:

A necessary condition for an inner point $(p_1, p_2) \in (0, \frac{1}{2})^2$ to be at least a local (maximum or minimum) solution of (1.4) is that

$$k_\kappa(p_t) := \log(1-p_t) - \log p_t - \kappa(1-2p_t) = 0 \quad \text{for} \quad t = 1, 2 \ ,$$

where $\kappa \in \mathbb{R}$ is a Lagrange multiplier.

For $\kappa \le 2$ $k_\kappa(.)$ is strictly positive for all $p \in (0, \frac{1}{2})$. For every $\kappa > 2$ there exists some $p^*(\kappa) \in (0, \frac{1}{2})$ such that

$$k_\kappa(p) \begin{cases} < 0, & \text{if } p^*(\kappa) < p < \frac{1}{2}, \\ = 0, & \text{if } p^*(\kappa) = p, \\ > 0, & \text{if } 0 \le p < p^*(\kappa). \end{cases}$$

Hence the only candidates for local solutions of (1.4) are inner points (p_1, p_2) with $p_1 = p_2$ or boundary points which are of the form $(0, p)$ for $c \le \frac{1}{2}$, or $(p, \frac{1}{2})$ for $c \ge \frac{1}{2}$.

$h' := \frac{dh}{dp}$ is continuous in p in the interval $(0, \frac{1}{2}]$. As $h'(0) = +\infty$ and $h'(p) < +\infty$ for all $p \in (0, \frac{1}{2}]$, (1.4) has a local minimum at $(0, p)$. Hence for $c \le \frac{1}{2}$ (1.4) is solved by the point (p_1, p_2) with $p_1 = p_2$.

For $c \in [\frac{1}{2}, 1]$ let $p_c, q_c \in [0, \frac{1}{2}]$ be the real numbers satisfying $g(p_c, p_c) = c = g(q_c, \frac{1}{2})$.

We define

$$\widetilde{f}(c) = f(p_c, p_c) - f(q_c, \frac{1}{2}) \ .$$

As (p_c, p_c) and $(q_c, \frac{1}{2})$ are the only candidates for a solution of (1.4), we are done if $\widetilde{f}(c)$ is non–negative for all $c \in [\frac{1}{2}, 1]$.

$\widetilde{f}(\frac{1}{2}) > 0$, $\widetilde{f}(1) = 0$, and $\widetilde{f}$ is continuous in c . If there were some $c \in (\frac{1}{2}, 1)$ with $\widetilde{f}(c) < 0$, there would have to be another parameter $c^* \in (\frac{1}{2}, c)$ with $\widetilde{f}(c^*) = 0$.

But it can not be that (p_{c^*}, p_{c^*}) and $(q_{c^*}, \frac{1}{2})$, $(\frac{1}{2}, q_{c^*})$ are the only candidates for min or max solutions of (1.4), if $f(p_{c^*}, p_{c^*}) = f(q_{c^*}, \frac{1}{2})$.

This completes the proof of both the claim and the lemma.

■

Next we extend the analytical method and generalize Theorem 1.1.

2. Arbitrary Sets M and Sum Type Cost Functions

In Section 1 we have investigated the problem of minimizing the average inner distance of subsets of $\{0,1\}^n$ of a given cardinality. This is only a special case of the following more general problem:

Let $M = \{1, \ldots, m\}$ be a finite set, and $d : M \times M \to \mathbb{R}$ an arbitrary real–valued cost function. For every $n \in \mathbb{N}$ the corresponding sum type cost function $d_n : M^n \times M^n \to \mathbb{R}$ is defined by

$$d_n(x, y) = \sum_{t=1}^{n} d(x_t, y_t)$$

for all $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in M^n$.

For a set $A \subset M^n$ the average inner cost are defined by

$$\overline{d}_n(A) = \frac{1}{(\#A)^2} \sum_{x \in A} \sum_{y \in A} d_n(x, y),$$

and for every $a \in \{1, \ldots, m^n\}$ we define

$$\overline{d}_n(a) = \min_{\substack{A \subset M^n: \\ \#A = a}} \overline{d}_n(A) \ .$$

We are interested in good bounds for the function $\overline{d}_n$.

These average inner cost play an important role for instance in the design of good WEM–codes [4].

For the presentation of the general result on asymptotically optimal configurations of cardinality $\approx 2^{\lambda n}$, $0 < \lambda < \log m$, we need a notation of **typical sequences**. Let $P = (P(1), \ldots, P(m))$ be a probability distribution on M . A tuple $(x_1, \ldots, x_n) \in M^n$ is of **type P** , if $\#\{t \mid x_t = i\} = P(i)n$ for all $i \in \{1, \ldots, m\}$. Let $T_n(P) = \{x \in M^n \mid x \text{ has type } P\}$. $\#T_n(P) \approx 2^{H(P)n}$, if $T_n(P) \neq \varnothing$.

See for instance the book [3] for a more detailed introduction and discussion of typical sequences.

Consider a constant ν , $0 \le \nu \le 1$, and two probability distributions P and P' on M . $(x_1, \ldots, x_n) \in M^n$ is said to be of the **mixed type** $(\nu P, (1-\nu)P')$, if

$$(x_1, \ldots, x_{\lfloor \nu n \rfloor}) \text{ is of type } P$$

and

$$(x_{\lfloor \nu n \rfloor + 1}, \ldots, x_n) \text{ is of type } P' .$$

Let $T_n(\nu, P, P') = \{x \in M^n \mid x \text{ is of the mixed type } (\nu P, (1-\nu)P')\}$. $\#T_n(\nu, P, P') \approx 2^{H(P)\nu n + H(P')(1-\nu)n}$.

Theorem 2.1:

Fix some finite set M *and a cost function* $d : M \times M \to \mathbb{R}$. *For every* λ , $0 < \lambda \le \log m$, *there exists a mixed type* (ν, P, P') *with* $\nu H(P) + (1-\nu)H(P') = \lambda$, *such that*

$$\limsup_{n\to\infty} \left[\overline{d}_n(T_n(\nu, P, P')) - \overline{d}_n(2^{\lambda n})\right] < +\infty .$$

In case of $\lim_{n\to\infty} \overline{d}_n(2^{\lambda n}) \in \{\pm\infty\}$ *this means*

$$\lim_{n\to\infty} \frac{\overline{d}_n(T_n(\nu, P, P'))}{\overline{d}_n(2^{\lambda n})} = 1 .$$

In other words, the sets $T_n(\nu, P, P')$ *have asymptotically minimal average inner cost.*

Proof:

As in the proofs of Section 1 we start by generalizing the problem to probability distributions Q on M^n , defining average inner cost $\overline{d}(Q)$ and substituting the cardinality condition by a lower bound on the entropy $H(Q)$. Given $q_t(k) = \sum_{x : x_t = k} Q(x)$ for all $t \in \{1, \ldots, n\}$, $k \in M$, we have

$$\overline{d}(Q) = \sum_{t=1}^{n} \left[\sum_{k=1}^{m} \sum_{\ell=1}^{m} q_t(k) q_t(\ell)\ d(k, \ell)\right]$$

and

$$H(Q) \le \sum_{t=1}^{n} H(q_t(1), \ldots, q_t(m)) .$$

In the last line equality holds iff Q is the product of its n 1–dimensional marginal distributions. For a fixed $\lambda \in [0, \log m]$, we want to solve the following analytical optimization problem: Find

$$\left.\begin{array}{l} \min \ \overline{d}(Q) \\ \text{under the constraint } \displaystyle\sum_{t=1}^{n} H(q_t(1), \ldots, q_t(m)) \geq \lambda n, \end{array}\right\} \tag{2.1}$$

where each tuple $(q_t(1), \ldots, q_t(m))$ is a probability distribution on M .

Our goal is to show that (2.1) is solved approximately by a combination of at most two different distributions P and P' on M , taking P for the first $\lfloor \nu n \rfloor$ coordinates and P' for the other $n - \lfloor \nu n \rfloor$ coordinates. Of course P, P' , and ν will depend on λ .

We start with

Lemma 2.2:

Consider real numbers $x_1, \ldots, x_n$, $y_1, \ldots, y_n$ *, and a probability distribution* $(\lambda_1, \ldots, \lambda_n)$ *on* $N = \{1, \ldots, n\}$ *. Then there exist two elements* $j, k \in \{1, \ldots, n\}$ *and some* $\mu \in [0, 1]$ *, such that*

$$\mu x_j + (1 - \mu) x_k \leq \overline{x} := \sum_{t=1}^{n} \lambda_t x_t \tag{2.2}$$

and

$$\mu y_j + (1 - \mu) y_k \geq \overline{y} := \sum_{t=1}^{n} \lambda_t y_t \ .$$

Proof of Lemma 2.2:

We proceed by induction in n . For $n = 1$ or 2 nothing has to be shown. Now assume $n \geq 3$ and $\lambda_t > 0$ for all $t \in N$.

If there is some t with $x_t \leq \overline{x}$ and $y_t \geq \overline{y}$, we are done by setting $j = t$ and $\mu = 1$. If there is some m with $x_m \geq \overline{x}$ and $y_m \leq \overline{y}$, we restrict ourselves to the reduced set $N - \{m\}$ (with normalized probabilities $\frac{\lambda_t}{1-\lambda_m}$ instead of λ_t and $\sum_{t \neq m} \frac{\lambda_t}{1-\lambda_m} x_t \leq \overline{x}$, $\sum_{t \neq m} \frac{\lambda_t}{1-\lambda_m} y_t \geq \overline{y}$) and apply induction hypothesis.

It remains to solve the case where for each t either $x_t \leq \overline{x}$, $y_t \leq \overline{y}$ or $x_t \geq \overline{x}$, $y_t \geq \overline{y}$. Let
$N_{\text{low}} = \left\{t \mid x_t \leq \overline{x}, y_t \leq \overline{y}\right\}$, $N_{\text{high}} = \left\{t \mid x_t \geq \overline{x}, y_t \geq \overline{y}\right\}$,

and assume without loss of generality $\#N_{\text{low}} \leq \#N_{\text{high}}$. This means $\#N_{\text{high}} \geq 2$, as $n \geq 3$. Let an arbitrary element $m \in N_{\text{high}}$ be selected. Now we reduce the problem to the basic case $n = 3$ by defining

$$\widetilde{\lambda}_1 = \sum_{t \in N_{\text{low}}} \lambda_t,\ \widetilde{\lambda}_2 = \lambda_m,\ \widetilde{\lambda}_3 = \sum_{\substack{t \in N_{\text{high}},\\ t \neq m}} \lambda_t,$$

$$\widetilde{x}_1 = \sum_{t \in N_{\text{low}}} \frac{\lambda_t}{\widetilde{\lambda}_1} x_t,\ \widetilde{x}_2 = x_m,\ \widetilde{x}_3 = \sum_{\substack{t \in N_{\text{high}},\\ t \neq m}} \frac{\lambda_t}{\widetilde{\lambda}_3} x_t,$$

and

$$\widetilde{y}_1 = \sum_{t \in N_{\text{low}}} \frac{\lambda_t}{\widetilde{\lambda}_1} y_t,\ \widetilde{y}_t = y_m,\ \widetilde{y}_3 = \sum_{\substack{t \in N_{\text{high}},\\ t \neq m}} \frac{\lambda_t}{\widetilde{\lambda}_3} y_t\ .$$

Obviously

$$\sum_{t=1}^{3} \widetilde{\lambda}_t \widetilde{x}_1 = \overline{x},\ \sum_{t=1}^{3} \widetilde{\lambda}_t \widetilde{y}_1 = \overline{y},$$

and

$$\widetilde{x}_1 \leq \overline{x}, \widetilde{y}_1 \leq \overline{y},\ \widetilde{x}_2 \geq \overline{x}, \widetilde{y}_2 \geq \overline{y},\ \widetilde{x}_3 \geq \overline{x}, \widetilde{y}_3 \geq \overline{y}\ .$$

There is some $\lambda^* \in [0, \widetilde{\lambda}_1]$ such that

$$\lambda^* \widetilde{y}_1 + \widetilde{\lambda}_2 \widetilde{y}_2 = (\lambda^* + \widetilde{\lambda}_2)\overline{y}$$

and

$$(\widetilde{\lambda}_1 - \lambda^*)\widetilde{y}_1 + \widetilde{\lambda}_3 \widetilde{y}_3 = (\widetilde{\lambda}_1 - \lambda^* + \widetilde{\lambda}_3)\overline{y}\ .$$

By the equality on the right side of (2.2) at least one of the following inequalities holds:

$$\lambda^* \widetilde{x}_1 + \widetilde{\lambda}_2 \widetilde{x}_2 \leq (\lambda^* + \widetilde{\lambda}_2)\overline{x},$$
$$(\widetilde{\lambda}_1 - \lambda^*)\widetilde{x}_1 + \widetilde{\lambda}_3 \widetilde{x}_3 \leq (\widetilde{\lambda}_1 - \lambda^* + \widetilde{\lambda}_3)\overline{x}\ .$$

So either $j = 1, k = 2$, $\mu = \frac{\lambda^*}{\lambda^* + \widetilde{\lambda}_2}$ or $j = 1, k = 3, \mu = \frac{\widetilde{\lambda}_1 - \lambda^*}{\widetilde{\lambda}_1 - \lambda^* + \widetilde{\lambda}_3}$ is an appropriate choice.

If originally $n \geq 4$, then we have just reduced the problem to one of the cases with sets $N_{\text{low}} \cup \{m\}$ or $N - \{m\}$ instead of N . For these we apply induction hypothesis. This completes the proof of Lemma 2.2 .

■

For the next step consider a compact set $K \subset \mathbb{R}^m$, continuous functions $f, g : K \to \mathbb{R}$, and for all $n \in \mathbb{N}$ the optimization problem

$$\left.\begin{array}{l} \displaystyle\min_{(z_1,\dots,z_n)\in K^n} \sum_{t=1}^{n} f(z_t) \\ \text{under the constraint } \displaystyle\sum_{t=1}^{n} g(z_t) \geq cn, \end{array}\right\} \tag{2.3}$$

where $c \in \mathbb{R}$ is some fixed constant.

Lemma 2.3:

There exist $\widetilde{z}_1, \widetilde{z}_2 \in K$ *and* $\nu \in [0,1]$ *, all depending on* c *, such that*

$$\lfloor \nu n \rfloor g(\widetilde{z}_1) + (n - \lfloor \nu n \rfloor) g(\widetilde{z}_2) \geq cn$$

and

$$\lfloor \nu n \rfloor f(\widetilde{z}_1) + (n - \lfloor \nu n \rfloor) f(\widetilde{z}_2) - \sum_{t=1}^{n} f(z_{t,n}^*) \leq \mid f(\widetilde{z}_1) - f(\widetilde{z}_2) \mid$$

for all $n \in \mathbb{N}$ *, where* $(z_{1,n}^*, \ldots, z_{n,n}^*)$ *is an optimal solution of (2.3).*

Proof:

The optimization problem

$$\left.\begin{array}{l} \min\limits_{(z_1,z_2)\in K^2, \nu\in[0,1]} \left[\nu f(z_1) + (1-\nu) f(z_2)\right] \\ \text{under the constraint } \ \nu g(z_1) + (1-\nu) g(z_2) \geq c \end{array}\right\} \tag{2.4}$$

has a solution, say $(\widetilde{z}_1, \widetilde{z}_2, \nu)$.

(2.5)

For the existence of this solution the continuity of f and g is needed. Without loss of generality assume $g(z_1) \leq g(z_2)$.

Now fix $n \in \mathbb{N}$.

Putting $x_t = f(z_{t,n}^*), y_t = g(z_{t,n}^*)$, and $\lambda_t = \frac{1}{n}$ for $t = 1, \ldots, n$, we can apply Lemma 2.2 and see that there are $j, k \in \{1, \ldots, n\}$ and $\mu \in [0,1]$, such that

$$\mu n \ f(z_{j,n}^*) + (1-\mu) n \ f(z_{k,n}^*) \leq \sum_{t=1}^{n} f(z_{t,n}^*)$$

and

$$\mu n \ g(z_{j,n}^*) + (1-\mu) n \ g(z_{k,n}^*) \geq cn \ .$$

Thus by (2.4) and (2.5) we also have

$$\nu n \ f(\widetilde{z}_1) + (1-\nu) n \ f(\widetilde{z}_2) \leq \sum_{t=1}^{n} f(z_{t,n}^*)$$

and

$$\nu n\; g(\widetilde{z}_1) + (1-\nu)n\; g(\widetilde{z}_2) \geq cn\ .$$

This completes the proof of Lemma 2.3 .

∎

Let $\hat{d}_n(H) = \min \overline{d}(Q)$, where the min is taken over all probability distributions Q on M^n with $H(Q) \geq nH$.

For a mixed type (ν, P, P') we define the corresponding product distribution Q_n on M^n by the marginal probabilities

$$q_t(k) = \begin{cases} P(k), & \text{if } 1 \leq t \leq \lfloor \nu n \rfloor, \\ P'(k), & \text{if } \lfloor \nu n \rfloor < t \leq n, \end{cases}$$

for all $k \in M$.

Lemma 2.4:

For every $\lambda \in [0, \log m]$ *there exists a mixed type* (ν, P, P') *, such that*

$$\limsup_{n\to\infty} \left[\overline{d}(Q_n) - \hat{d}_n(\lambda)\right] \leq \max_{j,k\in M} d(j,k) - \min_{j,k\in M} d(j,k) < \infty$$

and

$$\lfloor \nu n \rfloor H(P) + (n - \lfloor \nu n \rfloor) H(P') \geq \lambda n\ .$$

Proof of Lemma 2.4:

For probability distributions P on M we define two functions

$$f(P) = \sum_{j=1}^{m} \sum_{k=1}^{m} P(j)P(k)d(j,k)$$

and

$$g(P) = H(P)\ .$$

and apply Lemma 2.3.

Obviously $\mid f(P) - f(P') \mid \leq \max d(j,k) - \min d(j,k)$ for all P, P' .

This completes the proof of Lemma 2.4.

■

Theorem 2.1 follows immediately, as

$-c \leq \overline{d}_n(T_n(\nu, P, P')) - \overline{d}(Q_n) \leq c$ for all $n \in \mathbb{N}$, where the finite bound c depends only on m and $d : M \times M \to \mathbb{R}$.

■

Theorem 1.1 shows that the special case of a degenerated optimal mixed type (ν, P, P') with $\nu = 1$ occasionally occurs.

Let us now apply Theorem 2.1 to a non–trivial example.

Choose $M = \{1, 2, 3\}$ and

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

Hence d_n is the Hamming distance again, but now for alphabet size 3 instead of 2.

The results mentioned below have been found by computer runs. We omit the theoretical proofs. In the first step we have to understand the case $n = 1$.

Fact 2.5:

Fix some $\lambda \in [0, \log 3]$. Among all distributions P on M with $H(P) = \lambda$ the distribution with minimal average inner cost is of the form $(q, \frac{1-q}{2}, \frac{1-q}{2})$ with $q \geq \frac{1-q}{2}$.

Minimizing $\overline{d}_n$ for a given cardinality $2^{\lambda n}$ is equivalent to maximizing the cardinality under the condition $\overline{d}_n \leq cn$.

The computer results give

Fact 2.6:

Among all subsets of $\{1, 2, 3\}^n$ with average inner cost $\leq cn$ the following ones have asymptotically maximal cardinality:

(i) $T_n(P)$, where $P = (q, \frac{1-q}{2}, \frac{1-q}{2})$ with $q \geq \frac{1-q}{2}$ and $\overline{d}(P) = c$, if $0 \leq c \leq \frac{1}{2}$.

(ii) $T_n(\nu, P, P')$, where $P = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$, $P' = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, and $\nu \overline{d}(P) + (1 - \nu)\overline{d}(P') = c$, if $\frac{1}{2} \leq c \leq \frac{2}{3}$.

In the more general case with $M = \{1, \ldots, m\}$, $m \geq 3$, and

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

our computer results indicate that the optimal solutions have the following structure.

Observation 2.7:

Fix some $\lambda \in [0, \log m]$.

Among all distributions P on M with $H(P) = \lambda$ the one with minimal average inner cost is of the form $(q, \frac{1-q}{m-1}, \dots, \frac{1-q}{m-1})$ with $q \geq \frac{1-q}{m-1}$.

Observation 2.8:

For every $m \geq 3$ there is some threshold $c_m^* \in (0, \frac{m-1}{m})$ such that among all subsets of M^n with average inner cost $\leq cn$ the following ones have asymptotically maximal cardinality:

(i) $T_n(P)$, where $P = (q, \frac{1-q}{m-1}, \dots, \frac{1-q}{m-1})$ with $q \geq \frac{1-q}{m-1}$ and $\overline{d}(P) = c$, if $0 \leq c \leq c_m^*$.

(ii) $T_n(\nu, P, P')$, where $P = (q_m^*, \frac{1-q_m^*}{m-1}, \dots, \frac{1-q_m^*}{m-1})$ with $\overline{d}(P) = c_m^*$, $P' = (\frac{1}{m}, \dots, \frac{1}{m})$, and $\nu\overline{d}(P) + (1-\nu)\overline{d}(P') = c$, if $c_m^* \leq c \leq \frac{m-1}{m}$.

Observation 2.9:

m	3	4	5	6	7	10
c_m^*	0.5	0.4166..	0.35	0.3	0.26191..	$\frac{2}{11} = 0.1\overline{8}$

References

[1] R. Ahlswede and G. Katona, Contributions to the geometry of Hamming spaces, Discrete Mathematics 17 (1977), 1–22.

[2] I. Althöfer and T. Sillke, An “average distance” inequality for large subsets of the cube, to appear in Journal of Combinatorial Theory B.

[3] I. Csiszár and J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems, Acad. Press, New York, 1982.

[4] R. Ahlswede and Z. Zhang, On write–efficient memories, Information and Computation 83 (1989), 80–97.

[5] R. Ahlswede, N. Cai, and Z. Zhang, Diametric theorems in sequence spaces, Preprint 89–004 SFB 343 Bielefeld, to appear in Combinatorica.