Generalized witness sets

Gérard Cohen
Ecole Nationale Supérieure des Télécommunications
Telecom-Paristech, UMR 5141, CNRS,
46, rue Barrault, 75634 PARIS cedex 13, France.
Email: cohen@telecom-paristech.fr

Sihem Mesnager
LAGA, UMR 7539, CNRS,
Department of Mathematics,
University of Paris VIII and University of Paris XIII,
2 rue de la liberté, 93526 Saint-Denis Cedex, France.
Email: smesnager@univ-paris8.fr

***Abstract*—Given a set C of q-ary n-tuples and $c \in C$, how many symbols of c suffice to distinguish it from the other elements in C ? This is a generalization of an old combinatorial problem, on which we present (asymptotically tight) bounds and variations.**

I. Introduction

Set $[q] = \{0, 1, ...q-1\}, [n] = \{1, 2, ...n\}$. A code C of length n is a subset of $[q]^n$. Coding theory asks for large codes such that every codeword is "different" (has a large Hamming distance to all other codewords). The notion of difference adopted here is that there should exist a small subset $W \subset [n]$ of coordinates such that c differs from every other codeword in W, so that c can be singled out by examining a small subset of coordinates. Equivalently, c can be losslessly compressed to its projection on a small subset. More precisely, for $x \in [q]^n$, and $W \subset [n]$ let us define the projection π_W

$$\begin{array}{rccc} \pi_W & : [q]^{[n]} & \rightarrow & [q]^W \\ & x & \mapsto & (x_i)_{i \in W} \end{array}$$

and let us say that $W = W(c)$ is a *witness set* (or a witness for short) for $c \in C$ if $\pi_W(c) \neq \pi_W(c')$ for every $c' \in C$, $c \neq c'$. Codes with small witnesses arise in particular in machine learning theory [1], [5] see [6, Ch. 12] for a short survey of known results, [2] and references therein for a more recent discussion, and [4] for a study of the binary case. Finally, let us mention [7] for numerical constructions and upper bounds based on semidefinite programming.

Let us now say that a code is a w-*witness code*, if every one of its codewords has a witness of size w. Denote by $f(q, n, w)$ the maximum cardinality of a w-witness code of length n.

The paper is organized as follows. Section II presents some easy facts. Section III is devoted to asymptotics: we obtain tight bounds on the exponent of $f(q, n, w)$. Section IV deals with constant weight w-witness codes. Uniform witnesses and the linear case are considered in Section V. Finally, Section VI concludes with some open problems.

II. Warming up

First, two easy facts

- If C is a w-witness code, so is any translate $C + x$,
- $f(q, n, w)$ is an increasing function of n and w.

Example 1: Let C be the set of the $n(q-1)$ vectors of length n and weight (number of non-zero coordinates) equal to 1. Then every codeword of C has a witness of size 1, namely its support (set of non-zero coordinates). Note that for the slightly larger code $C \cup \{\mathbf{0}\}$, the all-zero vector $\mathbf{0}$ has no witness of size less than n.

A simple upper bound is :

$$f(q, n, w) \leq q^w \binom{n}{w}. \tag{1}$$

Indeed, a w-subset of $[n]$ can be a witness for at most q^w codewords and there are at most $\binom{n}{w}$ such sets.

We also have the following lower bound on $f(q, n, w)$.

Proposition 1: $f(q, n, w) \geq (q-1)^w \binom{n}{w}$.

Proof: Let C be the set of all vectors of weight w. Notice that every $c \in C$ has its support as witness. ■

Theorem 1: Let $g(q, n, w) = f(q, n, w)/\binom{n}{w}$. Then, for fixed q and w, $g(q, n, w)$ is a decreasing function of n.

Proof: Call $i \in [n]$ *indispensable* for c if $i \in \cap_{W \in \binom{[n]}{w}} W(c)$, and define

- $I(c)$ the set of indispensable i's for a given c;
- $C(i)$ the set of codewords for which i is indispensable.

We have the following:

$|C|w \geq \Sigma_{c \in C}|I(c)| = \Sigma_{i \in [n]}|C(i)| := nE_{i \in [n]}(|C(i)|)$,

first inequality coming from the obvious $|I(c)| \leq w$; first equality from double counting the pairs $\{c, i\}$ with $i \in I(c)$, and $E_{i \in [n]}(|C(i)|)$ denoting the mean value of $|C(i)|$.

Suppose coordinate n, say, achieves $Min_i\{|C(i)|\}$, then every $c \in C \setminus C(n)$ has a witness in $[1, n-1]$. Thus $f(q, n-1, w) \geq |C \setminus C(n)| \geq (n-w)|C|/n$.

Taking C maximal with the w-witness property:

$f(q, n, w) \leq (n/(n-w))f(q, n-1, w)$,

and the result follows. ■

III. Asymptotics

Theorem 1 has the following immediate consequence:

Corollary 2: For fixed w, $\lim_{n \to \infty} g(q, n, w)$ exists.

From now on, set $\mu := (q-1)/q$. When dealing with asymptotics, we assume q fixed, n growing and omit floor and ceiling signs since they are not crucial here. Denote for $0 < x \leq 1$ by $h_q(x)$ the entropy function

$$h_q(x) := -x \log_q x - (1-x)\log_q(1-x) + x\log_q(q-1),$$

978-0-7695-4528-8/11 \$26.00 © 2011 IEEE
DOI 10.1109/CCP.2011.12

with $h_q(0) := 0$.
The function $h_q(x)$ increases to 1 for $0 < x \leq \mu$ and decreases after.

Standard estimates give for $0 < \lambda \leq \mu$:

$$n^{-1}q^{nh_q(\lambda)} \leq (q-1)^{\lambda n}\binom{n}{\lambda n} \leq \sum_{0\leq i\leq \lambda n}(q-1)^i\binom{n}{i} \leq q^{nh_q(\lambda)}, \tag{2}$$

Note that the problem of computing $f(q,n,w)$ is essentially solved for $w \geq \mu n$: since $f(q,n,w)$ is increasing with w, we then have:
$q^n \geq f(q,n,w) \geq f(q,n,\mu n) \geq (q-1)^{\mu n}\binom{n}{\mu n} \geq q^n/n$.
Thus

$$\lim_{n\to\infty} n^{-1}\log_q f(q,n,\mu n) = 1.$$

We shall therefore focus in the sequel on the case $w \leq \mu n$.

Although the gap between (1) and Proposition 1 is pretty small (at least for q large), we now narrow it even more by improving on (1).

By Theorem 1, for $n \geq v \geq w$, $g(q,n,w) \leq g(q,v,w)$.

We use the trivial $g(q,v,w) \leq q^v/\binom{v}{w}$ and minimize the right-hand side over the choice of v.

Set $w := \sigma v$. Applying the left-most inequality of (2) for $q = 2$, we get:
$2^{vh_2(\sigma)} \leq v\binom{v}{w}$ and thus
$q^v/\binom{v}{w} \leq vq^{w/\sigma}/2^{wh_2(\sigma)/\sigma} := vq^{wz(\sigma)}$,
where we have set $z(\sigma) = (1 - h_2(\sigma)\log_q 2)/\sigma$.

The minimum of $z(\sigma)$ can be seen to be reached for $\sigma = \mu$ and equals $\log_q(q-1)$, yielding
$g(q, v = w/\mu, w) \leq (w/\mu)(q-1)^w \leq n(q-1)^w$ and finally
$(q-1)^w \leq g(q,n,w) \leq n(q-1)^w$.

Corollary 3:

$$\lim_{n\to\infty} n^{-1}\log_q f(q,n,\omega n) = h_q(\omega) \qquad \text{for } 0 \leq \omega \leq \mu.$$

IV. CONSTANT-WEIGHT CODES

Denote now by $f(q,n,w,k)$ the maximal size of a w-witness code with codewords of weight k.

Proposition 2 (A la Bassalygo-Elias): We have:

$$\max_k f(q,n,w,k) \leq f(q,n,w) \leq \min_k \frac{f(q,n,w,k)q^n}{(q-1)^k\binom{n}{k}}.$$

Proof: The lower bound is trivial.

For the upper bound, fix k, pick an optimal w-witness code C and consider its q^n translates by all possible vectors. Every n-tuple, in particular those of weight k, occurs exactly $|C|$ times in the union of the translates; hence there exists a translate (also an optimal w-witness code of size $f(q,n,w)$ - see beginning of Section II) containing at least the average number $|C|(q-1)^k\binom{n}{k}q^{-n}$ of vectors of weight k. Since k was arbitrary, the result follows. ∎

We now deduce from the previous proposition the exact value of the function $f(q,n,w,k)$ in some cases.

Corollary 4: For constant-weight codes we have:
If $k \leq w \leq \mu n$ then $f(q,n,w,k) = (q-1)^k\binom{n}{k}$;
an optimal code is given by $S_k(\mathbf{0})$, the Hamming sphere of radius k centered on $\mathbf{0}$.

Proof: If $k \leq w \leq \mu n$, we have the following series of inequalities:
$(q-1)^k\binom{n}{k} \leq f(q,n,k,k) \leq f(q,n,w,k) \leq (q-1)^k\binom{n}{k}$. ∎

V. UNIFORM WITNESSES AND LINEAR CODES

Call C a *uniform* w-witness code if there exists a subset of $[n]$ of size w that is a witness for *all* codewords (a uniform witness). The upper bound $|C| \leq q^w$ clearly holds for uniform w-witness codes.

Assume now that q is a prime power and that C is a linear subspace of F_q^n, the n-dimensional vector space over the finite field F_q. It is easy to check that a linear w-witness code is necessarily uniform; indeed, if $\mathbf{0}$ has witness $W(\mathbf{0})$, no two distinct codewords c and c' can coincide on it (otherwise, the non-zero codeword $c - c'$ would be all-zero on $W(\mathbf{0})$, a contradiction). Thus $W(\mathbf{0})$ is a uniform witness for C. Denote by $f[q,n,w]$ the maximum cardinality of a linear w-witness code. We have just proved that

Proposition 3: $f[q,n,w] = q^w$.

In the binary case, Bondy ([3], [6]) shows

Proposition 4: If $|C| \leq n$, then C is a uniform w-witness code with $w \leq |C| - 1$.

Proof: We give a simple coding proof of this known result, generalized to the q-ary case. We may assume by translation that $\mathbf{0} \in C = \{\mathbf{0}, c^{(1)}...c^{(m-1)}\}$, with $m \leq n$. Thus, the rank s of $\{c^{(1)}...c^{(m-1)}\}$ is at most $m-1$ and the elements of C span a linear subspace $C*$ of dimension s of F_q^n. As such, $C*$ (and thus C) possesses a uniform s-witness (refered to in coding as an *information set*). ∎

VI. CONCLUSION AND OPEN PROBLEMS

We have determined the asymptotic size of optimal w-witness codes. A few issues remain open, among which:

- When is the sphere $S_w(\mathbf{0})$ an optimal w-witness code ?
- Denoting by $f(q,n,w,\geq \delta n)$ the maximal size of a w-witness code with minimum Hamming distance $d \geq \delta n$, can the asymptotics of Corollary 3 be improved to

$$n^{-1}\log_q f(q,n,\omega n, \geq \delta n) < h_q(\omega) \text{ ?}$$

REFERENCES

[1] M. Anthony, G. Brightwell, D. Cohen, J. Shawe-Taylor: On exact specification by examples, *5th Workshop on Computational learning theory* 311-318, 1992.

[2] M. Anthony and P. Hammer: A Boolean Measure of Similarity, *Discrete Applied Mathematics* Volume 154, Number 16, 2242 - 2246, 2006.

[3] J.A. Bondy: Induced subsets, *J. Combin. Theory* (B) 12, 201-202, 1972.

[4] G. Cohen, H. Randriam, G. Zémor: Witness sets, Springer LNCS 5228, 37-45 (2008).

[5] S.A. Goldman, M.J. Kearns: On the complexity of teaching, *4th Workshop on Computational learning theory* 303-315, 1991.

[6] S. Jukna: *Extremal Combinatorics*, Springer Texts in Theoretical Computer Science 2001.

[7] N. Makriyannis, B. Meyer: Some constructions of maximal witness codes, *IEEE-ISIT 2011*, to appear.