Open problems in coding and cryptography

Gérard Cohen

May 2, 2012

Outline

Notation and packings

$\{0,1\}^n = F^n$: binary Hamming hypercube.
$x = (x_i), i = 1, ...n$, $y = (y_i)$ ... vectors
$d(x,y) = |\{i : x_i \neq y_i\}|$: Hamming distance
A code: $C \subset F^n$
Linear code: $C[n, k.d]$, $C < F^n$, $dim\ C = k$
$d = 2r + 1$: minimum distance between codewords

A code is a packing by spheres of radius r

$\mathbf{H}$ $(n-k) \times n$: parity-check matrix
Syndrome: $\sigma(x) = \mathbf{H}^t x$
$\sigma(c) = 0$ ssi $c \in C$.

W*M

Binary storage medium of n cells
to store and update information.
Operations performed under some constraints,
dictated by technology, cost, efficiency, speed, fashion ...
The latest: **Flash memories**.

EXAMPLES OF W*M:

- write-unidirectional memory (WUM)
- write-isolated memory (WIM)
- reluctant memories (WRM)
- defective memories (WDM)

Constrained memories

Memory is in *state* $y \in F^n$
Due to the constraints, only a subset $A(y)$ of F^n is reachable from y.
The (directed) *constraint graph* (F^n, A):
digraph with vertex set F^n
an arc from y to y' if and only if y' is reachable from y.
The state y can be updated to $v(y)$ states, where $v(y)$ is the *outdegree* of y.
To store one among M messages, the following must clearly hold:

Theorem

$M \leq \max_{y \in F^n} v(y)$.

Simple bound tight in some cases.
Here *symmetric* constraints (A is symmetric).

Asymptotically maximum achievable rate κ of the W*M

$$\kappa = (1/n) \log_2 M \ ?$$

Translation-invariant constraints

$$A(y) = y + A(0) = \{y + x : x \in A(0)\}$$

Set $A(0) = A, |A| = a_n$
$A(x)$: *A-set centred at* x
Translation-invariance is stronger than symmetry
Implies that the constraint graph is regular:
for all $y \in F^n$, $|A(y)| = a_n$.
Wlog assume we are in the state 0.
By the theorem:

$$M \leq a_n$$

Cloud encoding — packing by coverings

A coding strategy based on A-coverings

A subset $B = \{b_i\}$ of F^n is a A-*covering* or *cloud* if

$$\bigcup_{b_i \in B} A(b_i) = F^n.$$

That is, F^n is covered by the A-sets centred at the elements of B.

If a cloud B is an A-covering, so is any translate $B + x$, $x \in F^n$.
To write on a W*M, use the following encoding function:
to a message m_i associate an A-covering C_i of F^n

$$m_i \leftrightarrow C_i = \{c_{i,1}, c_{i,2}, \ldots\},$$

where, for all i

$$\bigcup_{c_{i,j} \in C_i} A(c_{i,j}) = F^n.$$

In that way, whatever the state y of the memory is, y can be updated to one of the $c_{i,j}$'s encoding m_i, while satisfying the constraints.

Packing many coverings

Theorem

If $B_1, B_2, \ldots, B_M$ are pairwise disjoint A-coverings, they yield a W*M-code of size M.

What is the maximum number of A-coverings of packable in F^n,
i.e., having void pairwise intersection?

Group coverings

The upper bound in the theorem is asymptotically tight.
1. Existence of small A-*group coverings* of F^n (i.e., clouds which are groups).
2. Finding pairwise disjoint clouds, becomes simple:
if G is a group A-covering with $|G| = 2^k$,
then there are 2^{n-k} pairwise disjoint A-coverings,
namely the cosets of G.
To that end, we use a greedy algorithm in a group version.

Theorem

There exists a group covering G of F^n of size 2^k, with

$$k = n - \log_2 a_n + \log_2 n + O(1).$$

Example. Balancing sets (application to magnetic and optical storage systems)
$A(0) = B_{n/2}(0)$.
$k = (3/2)\log_2 n + O(1)$.

Capacity

This scheme gives

$$M = 2^{n-k} = \Omega(a_n/n),$$

and the following result.

Theorem

$$\kappa = \lim_{n\to\infty} n^{-1} \log_2 a_n.$$

More graph notation

$B_r(v)$ the *ball* (resp. $S_r(v)$ the *sphere*) of radius r centred at v
the set of vertices within (resp. at) distance r from v.
Two vertices v_1 and v_2 such that $v_1 \in B_r(v_2)$ (resp. $v_1 \in S_r(v_2)$)
r-*cover* (resp. *exactly* r-*cover*) each other.
A set (exactly) $X \subseteq V$ r-covers a set $Y \subseteq V$ if every vertex in Y is (exactly)
r-covered by at least one vertex in X.
$K_{C,r}(v) = C \cap B_r(v)$ (resp.$X_{C,r}(v) = C \cap S_r(v)$) is the set of codewords
r-covering (resp. exactly r-covering) v.

Identification

Definition

A code $C \subseteq V$ is called r-*identifying* if all the sets $K_{C,r}(v)$, $v \in V$, are nonempty and distinct.

- every vertex is r-covered by at least one codeword
- every pair of vertices is r-separated by at least one codeword.

Application to fault diagnosis in multiprocessor computer systems.

Theorem

Consider $M \geq 1$ vertices $c^1, c^2, \cdots, c^M$ (non necessarily distinct) of F^n and M non-negative radii $r_1, r_2, \cdots, r_M$ such that

$$F^n = \bigcup_{j=1}^{M} S_{r_i}(c^j).$$

Then $M \geq n$ if n is even, and $M \geq n+1$ if n is odd.

Tightness

Bounds given by the theorem are tight :
for any vertex x we have

$$F^n = \bigcup_{i=0}^{n} S_i(x).$$

If n is even, then

$$F^n = \bigcup_{i=1}^{n-1} S_i(x) \cup S_{n/2}(y)$$

where y is any vertex satisfying $d(x, y) = n/2$.

Corollary

Let $C = \{c^i, L_i\}$ be a covering of the binary n-cube by shells, then $\Sigma_i |L_i| \geq n$.

Frequency allocation

In order to provide mobile telephone service using a limited band in the radio spectrum, the strategy is to dispatch users into cells.
A call is allocated a radio frequency.
The same frequency may be used simultaneously by another user, provided the distance between the cells they originate from exceeds some threshold, say r, to avoid interferences.
Let $\Gamma = (V, E)$ be the graph where vertices are cells and edges connect neighbouring cells with the usual metric.
$f(\mathbf{x})$ is the *call* function, number of (active) users in cell $\mathbf{x}$.

Covering by packings

The *call colouring problem* on Γ consists in assigning $f(\mathbf{x})$ colours (frequencies) to each vertex $\mathbf{x}$ in V with the constraint that, within every ball of a given radius r centred at $\mathbf{x}$, no other point has a colour in common with $\mathbf{x}$.
The cells of a given colour clearly make for a code of minimum distance $r+1$ (i.e., a *packing*).
In the case when $f=1$, i.e., when exactly one user per cell is active, these packings are disjoint.
The problem is then to find a minimum covering by packings.

Witness

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.

Motivation

Coding theory asks for maximal codes such that every codeword is different (has a large Hamming distance to all other codewords).

The notion of difference here is:
there should exist a small subset of coordinates on which
a codeword differs from every other,
so that it can be singled out by a small witness.

Context

Equivalently, every codeword can be losslessly compressed to its projection on a small subset.
Such codes arise in a variety of contexts,

in particular in machine learning theory,
where a witness is also called a specifying set or a discriminant.

Definitions

A subset $W(= W(c)) \in \binom{[n]}{w}$ is a (minimal) **Witness** for $c \in C$ if:

$$\forall c' \in C, c' \neq c : \pi_W(c') \neq \pi_W(c)$$

where π_W is the projection on W.

Pattern: $\pi_W(c) = \pi_{W(c)}(c)$.

$f(q, n, w)$:
Maximal size of a code with minimal witnesses of size at most w.

Previous work (binary case)

The average size of a witness is considered by Kushilevitz et al.
For a survey, see Jukna, where the following upper bound is given:

$$f(2, n, w) \leq \binom{n}{w} 2^w$$

Proof. Pigeon-hole principle:
there are at most this number of available patterns.

Immediate generalization to the q-ary case:

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

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

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

Proof. Pick $C = S_w(\mathbf{0})$.
Notice that $W(c) = support(c)$ for all c:

Every codeword has a **unique** pattern, namely its support.

An improved upper bound

(See [C.,Randriam, Zémor] for the binary ; [C., Mesnager] for the q-ary case).
For an optimal code (realizing $|C| = f(q,n,w)$), set

$g(q,n,w) := f(q,n,w)/\binom{n}{w}$.

Theorem

For q, w fixed, $g(q,n,w)$ is decreasing with n.

Corollary

For fixed q, w,
$lim_{n\to\infty} g(q,n,w) = f(q,n,w)/\binom{n}{w}$ *exists.*

Asymptotics

Set $w = \omega n$,
$h_q(x)$ the entropy function
$h_q(x) := -x \log_q x - (1-x)\log_q(1-x) + x \log_q(q-1)$:
$lim_{n\to\infty} n^{-1} log_q f(q, n, \omega n) = h_q(\omega), 0 \leq \omega \leq (q-1)/q.$

Witness with distance

$f(q, n, w, \geq d) :=$
maximal size of a w-witness code with minimum distance at least d.

Let's go asymptotics and set

$$\limsup_{n \to \infty} n^{-1} \log_q f(q, n, \omega n, \geq \delta n) := \phi(\omega, \delta).$$

From the previous proposition, we know that
$\phi(\omega, \delta) \leq h_q(\omega)$.

An open problem

The size of optimal w-witness codes is asymptotically known.

In the asymptotic case with minimum distance at least δn,

can we show

$$\phi(\omega, \delta) < h_q(\omega) \ ?$$

Non-malleable codes (NMC)

(Based on recent work with Chabanne, Flori and Patey).
Dziembowski et al. proposed a transposition of the cryptographic definition of non-malleability to the field of coding theory.

Informally, they define a NMC as a code such that, when a codeword is subject to modifications, its decoding procedure either
- corrects these errors and decodes to the original message or
- returns a value that is completely unrelated to the original message.

Bit-wise Independent Tampering

Bit-wise independent tampering is a special case of tampering where each bit of the codeword is tampered with independently.

Formally a function $f : F^n \mapsto F^n$ is bit-wise independent if we can find n independent functions $f_1, \ldots, f_n : F \mapsto F$ such that $\forall x \in F^n, f(x) = (f_1(x), \ldots, f_n(x))$.

There are four possibilities for each f_i : **keep**, **flip**, **0** and **1**, where **0** (resp. **1**) is the function that sets a bit to 0 (resp. 1), regardless of what it was before.

Theorem

Let $\mathcal{F} \subset F^{nF^n}$ be a family of bit-wise independent tampering functions such that: $\forall f = (f_1, \ldots, f_n) \in \mathcal{F}, |\{i | f_i = \mathbf{0} \text{ or } f_i = \mathbf{1}\}| \geq D$.
Let C be a $[n, k, d]$-linear code such that $D > n - d^{\perp}$, where $d^{\perp}$ is the minimal distance of its dual code $C^{\perp}$.
Then a linear coset-coding using C is non-malleable w.r.t. $\mathcal{F}$.

Generalized hashing

For a parameter $t \geq 2$ a code C is called t-*hashing* if
for any t distinct codewords $x^1, \ldots, x^t \in C$
there is a coordinate $1 \leq i \leq n$ such that all values x_i^j, $1 \leq j \leq t$ are distinct.

The concept of a hashing family is most central in Computer Science and Coding Theory.

Definition

Let $2 \leq t < u$ be integers.
A subset $C \subset Q^n$ is (t, u)-*hashing* if
for any two subsets T, U of C such that $T \subset U$, $|T| = t$, $|U| = u$,
there is some coordinate $i \in \{1, \ldots, n\}$ such that
for any $x \in T$ and any $y \in U, y \neq x$, we have $x_i \neq y_i$.

The concept of (t, u)-hashing generalizes the standard notion of hashing.

Indeed, when $u = t + 1$, a (t, u)-hashing family is $(t + 1)$-hashing.

Parent-identifying codes

Let C be an (n, M)-code. Suppose $X \subseteq C$. For any coordinate i define the *projection*

$$P_i(X) = \bigcup_{x \in X} \{x_i\}.$$

Define the *envelope* $e(X)$ of X by:

$$e(X) = \{x \in Q^n : \forall i, x_i \in P_i(X)\}.$$

Elements of the envelope $e(X)$ will be called *descendants* of X.

Observe that $X \subseteq e(X)$ for all X, and $e(X) = X$ if $|X| = 1$.

Given a word $s \in Q^n$ (a son) which is a descendant of X, we would like to identify without ambiguity at least one member of X (a parent).

Definition

For any $s \in Q^n$ let $\mathcal{H}_t(s)$ be the set of subsets $X \subset C$ of size at most t such that $s \in e(X)$. We shall say that C has the *identifiable parent property of order* t (or is a t-identifying code, or has the t-IPP, for short) if for any $s \in Q^n$, either $\mathcal{H}_t(s) = \emptyset$ or

$$\bigcap_{X \in \mathcal{H}_t(s)} X \neq \emptyset.$$

Barg et al. discovered a connection between (t, u)-hashing and t-IPP. Specifically, they proved the following:

Lemma

Let $u = \lfloor (t/2+1)^2) \rfloor$. If C is (t, u)-hashing then C is a t-identifying code.

The study of parent identifying codes is motivated by its connection to digital fingerprinting and schemes against software piracy.

A lower bound

Theorem

Let $u \geq t+1$, $q = t+1$ and $\varepsilon > 0$. Infinite sequences of (t,u)-hashing codes exist for all rates R such that

$$R + \varepsilon \leq \frac{t!(u-t)^{u-t}}{u^u(u-1)\ln(t+1)} \ .$$

Conclusion

Abstraction: Maximum packings of *different* objects

Classical: Diff= Distant
More general: c diff $\{c^1, c^2, ...\}$
Examples
$(1,t)$-separation: For every $\{c, c^1, ...c^t\} \in C$,
there exists $i \in [1,n]$ s.t. $c_i \notin \{c_i^1, ...c_i^t\}$.
Hashing $= (1,1,...1)$-separation
Applications to tracing traitors, broadcast encryption,...

(w,t)-witness:
For every $\{c, c^1, ...c^t\} \in C$, there exists $W \subset [1,n], |W| = w$ s.t.
$c/W \notin \{c^1/W, ...c^t/W\}$.
Application to computational learning theory.

Different ambient spaces: $[0, q-1]^n$, S_n (the symmetric group),...

Bibliography

- N. Alon, E. Bergmann, D. Coppersmith, A. Odlyzko: Balancing sets of vectors, *IEEE Transactions on Information Theory*, Vol. 34(1), pp. 128–130, 1988.
- D. Auger, G. Cohen: Sphere coverings and identifying codes, *Des. Codes Crypto* online 22 March 2012.
- H. Chabanne, G. Cohen, J.P. Flori, A. Patey: Non-Malleable codes from the wire-tap Channel, ITW 2011.
- G. Cohen, I. Honkala, S. Litsyn, A. Lobstein: *Covering Codes*. Amsterdam: Elsevier, 1997.
- G. Karpovsky, K. Chakrabarty, L.B. Levitin: On a new class of codes for identifying vertices in graphs, *IEEE Transactions on Information Theory*, Vol. 44(2), pp. 599–611, 1998.
- A. Mazumbar, R. Roth, P. Vontobel: On linear balancing sets, *Advances in mathematics of Communications*, Vol. 4 (3), 2010,345-361.

Bibliography for witnesses

M. Anthony, G. Brightwell, D. Cohen, J. Shawe-Taylor: On exact specification by examples, *5th Workshop on Computational learning theory* 311-318, 1992.
M. Anthony and P. Hammer: A Boolean Measure of Similarity, *Discrete Applied Mathematics* Volume 154, Number 16, 2242 - 2246, 2006.
J.A. Bondy: Induced subsets, *J. Combin. Theory* (B) 12, 201-202, 1972.
G. Cohen, S. Mesnager: Generalized witness sets, *2011 CCP* 255-256.
G. Cohen, H. Randriam and G. Zémor, "Witness sets", *Springer-Verlag LNCS* 5228 (2008) 37-45.
S. Jukna, *Extremal Combinatorics* Springer Texts in Theoretical Computer Science 2001.
E. Kushilevitz, N. Linial, Y. Rabinovitch and M. Saks: Witness sets for families of binary vectors, *J. Combin. Theory* (A) 73, 376-380, 1996.
N. Makriyannis, B. Meyer: Some constructions of maximal witness codes, *IEEE-ISIT 2011*.

Bibliography for generalized hashing

N. Alon, J. Bruck, J. Naor, M. Naor and R. Roth, "Construction of asymptotically good, low-rate error-correcting codes through pseudo-random graphs", IEEE Transactions on Information Theory, 38 (1992), 509-516.

N. Alon, E. Fischer and M. Szegedy, "Parent-identifying codes", *J. Combin. Theory Ser. A* **95** 2001, pp. 349–359.

A. Barg, G. Cohen, S. Encheva, G. Kabatiansky and G. Zémor, "A hypergraph approach to the identifying parent property", *SIAM J. Disc. Math.*, **14** 2001, pp. 423-432.

D. Boneh and M. Franklin, "An efficient public-key traitor-tracing scheme", *Crypto'99*, LNCS 1666 (1999), pp. 338–353.

Bibliography for generalized hashing II

B. Chor, A. Fiat and M. Naor, "Tracing traitors", *Crypto'94* LNCS 839 (1994), pp. 257–270.

M. Fredman and J. Komlós, "On the size of separating systems and perfect hash functions", *SIAM J. Algebraic and Disc. Meth*, **5** (1983), pp. 61–68.

H. D. L. Hollmann, J. H. van Lint, J.-P. Linnartz and L. M. G. M. Tolhuizen, "On codes with the identifiable parent property", *J. Combin. Theory Ser. A*, **82** 1998, pp. 121–133.

J. Körner, "Fredman-Komlós bounds and information theory", *SIAM J. Algebraic and Disc. Methods*, **7** 1986, pp. 560–570.

J. Körner and K. Marton, "New bounds for perfect hashing via information theory", *Europ. J. Combinatorics*, **9** 1988, pp. 523–530.

A. Nilli, "Perfect hashing and probability", *Combinatorics, Probability and Computing*, **3** 1994, pp. 407–409.