21-238, Math Studies Algebra 2, Department of Mathematical Sciences, Carnegie Mellon University
Spring 2012: Monday, Wednesday, Friday, 10:30 am, Doherty Hall 1211.
Luc TARTAR, University Professor of Mathematics, Wean Hall 6212, tartar@cmu.edu

24- Monday March 19, 2012.

Remark 24.1: Finite fields are used in some applications like coding or encrypting messages, and it then is useful to know a little about what algebraic questions are necessary for understanding such applications.

The distinction between *coding* and *encrypting* was probably made during World War II, and the first mathematician to study coding in a theoretical way may have been SHANNON.[1] Encrypting was already used by Caesar, but since coding is about detecting and correcting errors which occur during the transmission of a message, encrypted or not, it could only become important once messages were transmitted by electric or electronic means.

Th dstnctn btwn cdng nd ncrptng ws prbbl md drng Wrld Wr II, nd th frst mthmtcn t std cdng n thrtcl w m hv bn SHNNN. ncrptng ws lrd sd b Csr, bt snc cdng s bt dtctng nd crrctng rrrs whch ccr drng th trnsmssn f mssg, ncrptd r nt, t cld nl bcm mprtnt nc mssgs wr trnsmttd b lctrc r lctrnc mns.

The preceding paragraph is just the same than the first paragraph without the footnotes but with vowels deleted, and most of the words can easily be completed by English speaking readers, although the context must be used for deciding which is the correct reading when different unrelated words have the same skeleton of consonants.[2] Writing only consonants is usual for semitic languages, and the vowels in the Tanakh (Hebrew Bible) were only introduced at the time of the Masoretes,[3] and the (short) vowels as well as the diacritical dots for distinguishing various consonants were only introduced in the Quran (i.e. in the official version made by the Caliph 'UTHMAN) in the beginning of the 8th century, by AL HAJJAJ.[4]

Coding is not about deciding as in the previous example of a possible reading when some ambiguity occurs, since the purpose is to transmit sequences of numbers which have no meaning for most people, sometimes because the message is encrypted and can only be deciphered by people who possess the *key*. Errors occur in transmission lines because of some electric/electronic "noise" which one does not control,[5] and the purpose of coding is to transmit "words" of a fixed length, by coding each word in such a way that errors on a few symbols can be detected, and then corrected if they are not too numerous.

Remark 24.2: A first step is to transform a (long) message into a succession of words of fixed length, and for this one recalls that a *bit* is a binary digit, i.e. 0 or 1, and a *byte* is a string of 8 bits, so that there are $2^8 = 256$ bytes, but one also uses a 4-bit string called a *nibble*, for which one uses the hexadecimal system, i.e. 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F, so that a byte corresponds to two nibbles. For transforming letters and punctuation into bytes, one either uses the *ASCII* system (American Standard Code for Information Interchange) or the *EBCDIC* system (Extended Binary Coded Decimal Interchange Code): the ASCII system uses $2^7 = 128$ characters, i.e. bytes with last digit 0, but it only has 95 printable characters

[1] Claude Elwood SHANNON, American mathematician and electronic engineer, 1916–2001. He worked at MIT (Massachusetts Institute of Technology) in Cambridge, MA, and at Bell Laboratories in Murray Hill, NJ. He is considered the father of information theory.

[2] In the preceding paragraph, lrd could mean already, lord, and lured, sd could mean sad, said, and used, bt could mean about, bait, bat, bet, bit, bite, bout, but, and byte, for example.

[3] The Masoretes (whose name is derived from a Hebrew word) were groups of Jewish scribes and scholars working between the 7th and the 10th century, who compiled a system for fixing the pronunciation, paragraph and verse divisions, and cantillation of the Jewish Bible (Tanakh).

[4] AL HAJJAJ ibn Youcef, –714. Governor of Mesopotomia and the Iranian provinces (694–714) under the Umayyad Caliphs, he introduced the diacritical dots for distinguishing various consonants in Arabic, as well as the signs for the short vowels.

[5] One usually assumes that "noise" results from some random effect, because one does not know what really happens, but in a situation of conflict an opponent may try to disturb one's electric/electronic transmissions on purpose.

(and 33 non printable characters), as shown in the following table,

$x=$	0	1	2	3	4	5	6	7
$00x$	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL
$01x$	BS	HT	LF	VT	FF	CR	SO	SI
$02x$	DLE	$DC1$	$DC2$	$DC3$	$DC4$	NAK	SYN	ETB
$03x$	CAN	EM	SUB	ESC	FS	GS	RS	US
$04x$	SP	!	”	#	$	%	&	′
$05x$	(	)	∗	+	,	−	.	/
$06x$	0	1	2	3	4	5	6	7
$07x$	8	9	:	;	<	=	>	?
$10x$	@	A	B	C	D	E	F	G
$11x$	H	I	J	K	L	M	N	O
$12x$	P	Q	R	S	T	U	V	W
$13x$	X	Y	Z	[	\	]	ˆ	_
$14x$	`	a	b	c	d	e	f	g
$15x$	h	i	j	k	l	m	n	o
$16x$	p	q	r	s	t	u	v	w
$17x$	x	y	z	{	|	}		DEL

in which the 128 characters correspond to numbers written in octal (i.e. in base 8), so that the letter “a” has number 141, for example. The meanings of the 33 non printable characters are : NUL = null, SOH = start of heading, STX = start of text, ETX = end of text, EOT = end of transmission, ENQ = enquiry, ACK = acknowledge, BEL = bell, BS = backspace, HT = horizontal tab, LF = line feed, NL = new line, VT = vertical tab, FF = form feed, NP = new page, CR = carriage return, SO = shift out, SI = shift in, DLE = data link escape, DCj = device control j, NAK = negative acknowledge, SYN = synchronous idle, ETB = end of trans. block, CAN = cancel, EM = end of medium, SUB = substitute, ESC = escape, FS = file separator, GS = group separator, RS = record separator, US = unit separator, SP = space.

The *EBCDIC* (Extended Binary Coded Decimal Interchange Code) standard uses $2^8 = 256$ characters, i.e. bytes, but it only has 184 printable characters (and 72 non printable characters).[6]

Remark 24.3: One way to encode is to repeat: if one replaces 0 by 000, and 1 by 111, then there are only two codewords 000 and 111 among eight possible words (which may be received because of errors during the transmission); in such a code, two errors can be detected and one can be corrected, so that receiving 100, 010, or 001 means that 0 was transmitted with one error of transmission, and receiving 011, 101, or 110 means that 1 was transmitted with one error of transmission (and not 0 with two errors of transmission, in which case the error would not be corrected).

Definition 24.4: For words of length n expressed in an *alphabet* A of q symbols, one defines the *Hamming distance* between two words:[7] if $x = x_1 \cdots x_n, y = y_1 \cdots y_n \in A^n$, the Hamming distance $d(x, y)$ is the number of $i \in \{1, \ldots, n\}$ with $x_i \neq y_i$.[8]

A *q-ary code* C of length n is a set W (of words to transmit) together with an injective mapping of W into A^n for an alphabet A with q characters; a *codeword* is an element of the image. The *minimum distance* $d(C)$ of a code C is the smallest Hamming distance between two distinct codewords. An *(n, M, d)-code* is a code of length n, consisting of M codewords, and with minimum distance d.

[6] I am not sure if the change from ASCII to EBCDIC served to add characters from other alphabets, and accents: in French, the only new “letter” is œ, but c may receive a cedilla (ç), and there are eleven cases of accents: e may receive four different accents (é, è, ê, ë), while a, i, and u may receive two different accents each (à, â, î, ï, ù, û), and o may receive one accent (ô).

[7] Richard Wesley Hamming, American mathematician, 1915–1998. He worked at University of Louisville, KY, in the Manhattan Project (Los Alamos, NM), at Bell Telephone Laboratories (Murray Hill, NJ), at City College of New York (New York, NY), and at the NPS (Naval Postgraduate School, Monterey, CA). The Hamming distance and the Hamming codes are named after him.

[8] A distance is the same as a metric, i.e. it satisfies $d(x, y) = d(y, x) \geq 0$ for all x, y, $d(x, y) = 0$ if and only if $y = x$, and the triangle inequality holds: $d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, z.

Remark 24.5: If d is the minimum distance of a code C, and $t = \lfloor \frac{d-1}{2} \rfloor$,[9] then C can detect $d-1 = 2t$ errors and can correct up to t errors in any transmitted codeword.

Definition 24.6: A code C is called *perfect* if it has minimum distance $d(C) = 2t+1$ and for every $y \in A^n$ there exists a codeword $x \in C$ with $d(x,y) \leq t$. $A_q(n,d)$, the largest value of M such that there exists a q-ary (n,M,d)-code, is the *sphere-packing bound*, which satisfies $A_q(n,d) \leq \frac{q^n}{\sum_{m=0}^{t} \binom{n}{m} (q-1)^m}$, and a q-ary (n,M,d)-code is perfect if and only if $M = \frac{q^n}{\sum_{m=0}^{t} \binom{n}{m} (q-1)^m}$.

Remark 24.7: If u and v are distinct codewords, the (closed) balls $\overline{B}_t(u)$ and $\overline{B}_t(v)$ of radius t centered respectively at u and at v do not intersect. That a code is perfect means that A^n is the union of all such (closed) balls centered at all the codewords. Since the number of elements at distance m from any $u \in A^n$ is $\binom{n}{m}(q-1)^m$, the number of elements in a (closed) ball of radius r centered u is $|\overline{B}_r(u)| = \sum_{m=0}^{r} \binom{n}{m} (q-1)^m$, so that expressing that the various (closed) balls of radius t do not intersect gives the sphere-packing bound $M \sum_{m=0}^{t} \binom{n}{m} (q-1)^m \leq q^n$, hence the upper bound for $A_q(n,d)$.

Remark 24.8: Linear algebra enters the picture once one considers the family of *linear codes*, where A is a finite field F with q elements (so that q is a power of the characteristic p of F) and codewords form a subspace; encoding a word from $W = F^k$ into a codeword in $V = F^n$ is done in a linear way.

It is usual in coding theory to consider that $x \in F^n$ means that x is a row vector, and x^T denotes the corresponding column vector.

Since the *inner product* of two vectors $x, y \in F^n$ is $x \cdot y = \sum_{i=1}^{n} x_i y_i = x\, y^T = y\, x^T$, and two vectors $x, y \in F^n$ are said to be *orthogonal* if $x \cdot y = 0$, one should pay attention to the fact that a non-zero vector $x \in F^n$ may be orthogonal to itself, if $\sum_i x_i^2$ is a multiple of the characteristic p of the field F.

Definition 24.9: An $[n,k]$-*code* C over a finite field F is a subspace of dimension k in the vector space F^n; $n-k$ is called the *redundancy* of the code C, and $\frac{k}{n}$ is called the *transmission rate* of the code C; if F has q elements, a q-ary $[n,k]$-code then has q^k codewords.

If the $[n,k]$-code has minimum distance d, one calls it a $[n,k,d]$-*code*.

Two $[n,k]$-codes C and C' over F are said to be *equivalent* if there exists a bijective mapping f from C onto C', non-zero scalars $\alpha_1, \ldots, \alpha_n \in F^*$, and a permutation σ of $\{1, \ldots, n\}$, such that $f(x_1, \ldots, x_n) = (\alpha_1 x_{\sigma(1)}, \ldots, \alpha_n x_{\sigma(n)})$ for all $x_1, \ldots, x_n \in F$.

Definition 24.10: A *generator matrix* G of the $[n,k]$-code C (over F) is a $k \times n$ matrix (with entries in F) whose rows form a basis of C, so that $C = \{u\, G \mid u \in F^k\}$ (and G has rank k); elements of $W = F^k$ are called *message words*, and the scheme for *encoding* from F^k to C is $u \mapsto u\, G$ (message words and codewords are row vectors).

Let C be an $[n,k]$-code (over F), then the *dual code* $C^\perp$ of C is $C^\perp = \{y \in F^n \mid x \cdot y = 0 \text{ for all } x \in C\}$; C is called *self-orthogonal* if $C \subset C^\perp$.

Remark 24.11: C is self-orthogonal if and only if any two codewords are orthogonal, and in particular $\sum_i x_i^2 = 0 \pmod{p}$ for every codeword x, where p is the characteristic of the field F.

$C^\perp$ is an $[n, n-k]$ code, and $(C^\perp)^\perp = C$. If G is a generator matrix of C, then $C^\perp$ is the *null space* $N = \{X \in F^n \mid G\, X = 0\}$ of G, so that two generator matrices of the same code have the same null space.

Definition 24.12: Let C be an $[n,k]$-code (over F), then a generator matrix H of the dual code $C^\perp$ is called a *parity-check matrix* of the code C, so that $C = \{x \in F^n \mid x\, H^T = 0 = H\, x^T\}$ (the equations $H\, x^T = 0$ are called the *parity-check equations*), and one has $G\, H^T = H\, G^T = 0$.[10]

A *canonical generator matrix* of C has the form $G_* = [I_k \mid A]$, where I_k is the identity matrix of size k, and it is associated with the *canonical parity-check matrix* $H_* = [A^T \mid I_{n-k}]$ of C.

Remark 24.13: The *weight* $w(x)$ of a vector $x \in F^n$ is the number of non-zero components of x, i.e. the Hamming distance $d(x,0)$, so that the minimum distance of a linear code C is $d(C) = \min\{w(x) \mid x \in C, x \neq$

[9] $\lfloor x \rfloor$ is the greatest integer $\leq x$.

[10] Conversely, if G is a $k \times n$ matrix of rank k and H is an $(n-k) \times n$ matrix of rank $n-k$ such that $G\, H^T = 0$, then H is a parity-check matrix of the code C if and only if G is a generator matrix of C.

0}: with M codewords, one needs only check the weights of $M-1$ vectors (instead of comparing the distances of $\frac{M\,(M-1)}{2}$ pairs). However, a code may have a very large M, or the codewords may be described implicitly by giving a parity-check matrix H, so that it is useful to deduce from H what the minimum distance of the code is: since $H\,x^T = 0$ means that a particular combination of columns of H is 0, one deduces that if an $[n,k]$-code C has parity-check matrix H, then $d(C)$ is the minimal number of linearly dependent columns of H (hence $d(C) \leq n-k+1$).

Example 24.14: The former *ISBN code* (International Standard Book Numbers) was a $[10,9]$-code over F_{11} ($\simeq \mathbb{Z}_{11}$) defined by the parity check $H = [\,1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10\,]$. The first part of an ISBN codeword was the *group identifier*, which identified a country or a language area, the second part was the *publisher identifier*, which identified a specific publisher in a specific group, the third part was the *title identifier*, which identified a specific publication of a specific publisher; the length of the three parts varied, but the total length was 9, and the *check-digit* x_{10} was written X if it was 10.

The revised ISBN code uses a 13-digit number:[11] the verification of the code $c_1 c_2 \cdots c_{13}$ is that one must have $c_1 + 3c_2 + c_3 + 3c_4 + \ldots + c_{11} + 3c_{12} + c_{13} = 0 \pmod{10}$.

Example 24.15: A *binary Hamming code* $Ham(r,2)$ is defined for an integer $r > 1$ and for an $r \times (2^r - 1)$ parity-check matrix H whose columns are the distinct non-zero vectors in $(F_2)^r$ (and $F_2 \simeq \mathbb{Z}_2$), so that there are $(2^r-1)!$ equivalent codes). The length of the code is $n = 2^r - 1$, and its dimension is $k = n - r$, so that r is the redundancy of the code, and $Ham(r,2)$ is a $[2^r - 1, 2^r - r - 1]$-code; since a column may be the sum of two others, but no two columns are multiple, the minimum distance is 3.

$Ham(2,2)$ is a $[3,1]$-code with parity-check matrix $H = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ (not in canonical form), with $2^1 = 2$ codewords: it is the binary repetition code $\{000, 111\}$. $Ham(3,2)$ is a $[7,4]$-code with parity-check matrix $H = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$ (not in canonical form), with $2^4 = 16$ codewords. $Ham(r,2)$ is a perfect code with minimum distance 3, since $2^{n-r}\left[\binom{n}{0} + \binom{n}{1}\right] = 2^n$, because $1 + n = 2^r$.

Example 24.16: For q a power of a prime p, one considers the field F_q, and for an integer $r > 1$ one defines $n = \frac{q^r-1}{q-1}$, and a *q-ary Hamming code* $Ham(r,q)$ is defined by an $r \times n$ parity-check matrix H whose columns are non-zero vectors in $(F_q)^r$ such that no column is a scalar multiple of another (there are $n!\,(q-1)^n$ equivalent codes). The length of the code is n, and its dimension is $k = n - r$, so that r is the redundancy of the code, and $Ham(r,q)$ is a $[n, n-r]$-code; since a column may be the sum of two others, but no two columns are multiple, the minimum distance is 3. $Ham(r,q)$ is a perfect code with minimum distance 3, since $2^{n-r}\left[\binom{n}{0} + \binom{n}{1}(q-1)\right] = 2^n$, because $1 + n\,(q-1) = q^r$.

Remark 24.20: A decoding procedure for a (general) linear code is to prepare a look-up table which permits for each vector $y \in F^n$ to find one codeword $x \in C$ which is nearest to y (and choose one if there are many): it consists in finding a vector of least weight (called a *coset leader*) in the *coset* $y + C$. If C is an $[n,k]$-code over $F = F_q$, each coset has $M = q^k$ elements, and there are $N = q^{n-k}$ cosets: one denotes $e_1, \ldots, e_N$ the coset leaders, numbered in ascending order of weights, and $0 = c_1, c_2, \ldots, c_M$ the elements of C. One then forms an $N \times M$ matrix called a *standard array* for the code C by putting $e_i + c_j$ as entry (i,j): given a vector $y \in F^n$, one looks for it in the matrix, and if it is $e_i + c_j$ one decodes y as c_j, i.e. the entry at the top of the column containing y. This decoding procedure is only useful if the code length n is small.

Another procedure for decoding, called *syndrome decoding*, uses a parity-check matrix H for the $[n,k]$-code over $F = F_q$, and for each $y \in F^n$ defines the *syndrome* $S(y)$ of $y \in F^n$ as $S(y) = y\,H^T$. Since $S(y) = S(z)$ is equivalent to $z \in y + C$, one constructs a *syndrome table* with two columns, with the coset leaders $e_1, \ldots, e_N$ (with $N = q^{n-k}$) in the first column, and their syndromes $S(e_1), \ldots, S(e_N)$ in the second column: given a vector $y \in F^n$, one computes $S(y)$, which one looks for in the second column, so that it is $S(e_i)$, which gives e_i in the first column, and one decodes y as $y - e_i$.

[11] The ISBN numbers of my first three books, are 978-3-540-35743-8, 978-3-540-71482-8, and 978-3-540-77561-4, so that 978-3-540 seems to identify Springer (Berlin Heidelberg New York), but the ISBN number of my fourth book published by the same publisher is 978-3-642-05194-4.