Hill verses Hamming

It's easy to imagine the 19th century Philadelpia wool dealer Frank J. Primrose as a happy man. I envision him shearing sheep during the day, while in the evening he brings his wife flowers and plays games with his little children until bedtime. However, in 1887 Frank J. Primrose was not a happy man. This is because in June of that year, he had telegraphed his agent in Kansas instructions to buy a certain amount of wool. However, the telegraph operator made a single mistake in transmitting his message and Primrose unintentionally bought far more wool than he could possibly sell. Ordinarily, such a small error has little consequence, because errors can often be detected from the context of the message. However, this was an unusual case and the mistake cost him about a half-million dollars in today's money. He promptly sued and his case eventually made its way to the Supreme Court. The famous 1894 United States Supreme Court case Primrose v. Western Union Telegraph Company decided that the telegraph company was not liable for the error in transmission of a message.

Thus was born the need for error-correcting codes.

Introduction

Lester Hill is most famously known for the Hill cipher, frequently taught in linear algebra courses today. We describe this cryptosystem in more detail in one of the sections below, but here is the rough idea. In this system, developed and published in the 1920's, we take an $k \times k$ matrix K, composed of integers between 0 and 25, and encipher plaintext $\mathbf{p}$ by

$$\mathbf{p} \mapsto \mathbf{c} = K\mathbf{p}, \tag{1}$$

where the arithmetical operations are performed $\pmod{26}$. Here K is the key, which should be known only to the sender and the intended receiver, and $\mathbf{c}$ is the ciphertext transmitted to the receiver.

On the other hand, Richard Hamming is known for the Hamming codes, also frequently taught in a linear algebra course. This will be describes in more detail in one of the sections below, be here is the basic idea. In this scheme, developed in the 1940's, we take an $k \times n$ matrix G over a finite field, constructed in a very particular way, and encode a message $\mathbf{m}$ by

$$\mathbf{m} \mapsto \mathbf{c} = \mathbf{m}G, \tag{2}$$

where the arithmetical operations are performed in $\mathbb{F}$. The matrix G is called the generator matrix and $\mathbf{c}$ is the codeword transmitted to the receiver.

Here, in a nutshell, is the mystery at the heart of this article. These schemes (1) and (2), while algebraically very similar, have quite different aims. One is intended for secure communication, the other for reliable communication. However, in an unpublished paper [**7**] Hill developed a hybrid encryption/error-detection scheme, what we shall call "Hill codes" (described in more detail below). Why wasn't Hill's result published and therefore Hill, more than Hamming, known as a pioneer of error-correcting codes?

Perhaps Hill himself hinted at the answer. In an overy optimistic statement, Hill wrote (italics mine):

> Further problems connected with checking operations in finite fields will be treated *in another paper*. Machines may be devised to render almost quite automatic the evaluation of checking elements $c_1, c_2, \dots, c_q$ according to any proposed reference matrix of the general type described in Section 7, whatever the finite field in which the operations are effected. Such machines would enable us to dispense entirely with tables of any sort, and checks could be determined with great speed. But before checking machines could be seriously planned, the following problem – which is one, incidentally, of considerable interest from the standpoint of pure number theory – *would require solution*.
>
> - Lester Hill, [**7**]

By my interpretation, this suggests Hill wanted to answer the question below before moving on. As simple looking as it is, this problem is still, as far as I know, unsolved at the time of this writing.

Question 1. *(*Hill's Problem*) Given k and q, find the largest r such that there exists a $k \times r$ matrix of the form*

$$\begin{pmatrix} a_1 & a_2 & \dots & a_r \\ a_1^2 & a_2^2 & \dots & a_r^2 \\ \vdots & & & \\ a_1^k & a_2^k & \dots & a_r^k \end{pmatrix}$$

with the property that every square submatrix is non-singular.

Indeed, this is closely related to the following related question from MacWilliams-Sloane [**8**], also still unsolved at this time. (Since Cauchy matrices do give a large family of matrices with the desired property, I'm guessing Hill was not aware of them.)

Question 2. *(*Research Problem (11.1d) in [**8**]*) Given k and q, find the largest r such that there exists a $k \times r$ matrix having entries in $GF(q)$ with the property that every square submatrix is non-singular.*

In this paper, after brief biographies, an even more brief description of the Hill cipher and Hamming codes is given, with examples. Finally, the above-mentioned unpublished paper in which Hill discovered error-correcting codes is discussed in more detail.

Short biographies

Who is Hill? Recent short biographies have been published by C. Christensen and his co-authors. Modified slightly from [**1**] and [**2**] is the following information.

Lester Sanders Hill was born on January 19, 1890 in New York. He graduated from Columbia University in 1911 with a B. A. in Mathematics and earned his Masters Degree in 1913. He taught mathematics for a few years at Montana University, then at Princeton University. He served in the United States Navy Reserves during World War I. After the WWI, he taught at the University of Maine and then at Yale, from which he earned his Ph.D. in mathematics in 1926. His Ph.D. advisor is not definitely known at this writing but I think a reasonable guess is Wallace Alvin Wilson.

In 1927, he accepted a position with the faculty of Hunter College in New York City, and he remained there, with one exception, until his resignation in 1960 due to illness. The one exception was for teaching at the G.I. University in Biarritz in 1946, during which time he may have been reactivated as a Naval Reserves officer. Hill died January 9, 1961.

Thanks to an interview that David Kahn had with Hill's widow reported in [**1**], we know that Hill loved to read detective stories, to tell jokes and, while not shy, enjoyed small gatherings as opposed to large parties.

Who is Hamming? His life is much better known and details can be readily found in several sources.

Richard Wesley Hamming was born on February 11, 1915, in Chicago. Hamming earned a B.S. in mathematics from the University of Chicago in 1937, a masters from the University of Nebraska in 1939, and a PhD in mathematics (with a thesis on differential equations) from the University of Illinois at Urbana-Champaign in 1942. In April 1945 he joined the Manhattan Project at the Los Alamos Laboratory, then left to join the Bell Telephone Laboratories in 1946. In 1976, he retired from Bell Labs and moved to the Naval Postgraduate School in Monterey, California, where he worked as an Adjunct Professor and senior lecturer in computer science until his death on January 7, 1998.

Hill's cipher

The Hill cipher is a polygraphic cipher invented by Lester S. Hill in 1920's. Hill and his colleague Wisner from Hunter College filed a patent for a telegraphic device encryption and error-detection device which was roughly based on ideas arising from the Hill cipher. It appears nothing concrete became of their efforts to market the device to the military, banks or the telegraph company (see Christensen, Joyner and Torres [**2**] for more details). Incidently, Standage's excellent book [**10**] tells the amusing story of the telegraph company's attempt to add a relatively simplistic error-detection to telegraph codes during that time period. `#epicfail`, in the twitter-verse of today.

Some books state that the Hill cipher never saw any practical use in the real world. However, research by historians F. L. Bauer and David Kahn uncovered the fact that the Hill cipher saw some use during World War II encrypting three-letter groups of radio call signs [**1**]. Perhaps insignificant, at least compared to the practical value of Hamming codes, none-the-less, it was a real-world use.

The following discussion assumes an elementary knowledge of matrices. First, each letter is first encoded as a number, namely $A \leftrightarrow 0$, $B \leftrightarrow 1$, ..., $Z \leftrightarrow 25$. The integers $\{0, 1, \dots, 25\}$ will be denoted by by $\mathbb{Z}/26\mathbb{Z}$. This is closed under addition and multiplication $\pmod{26}$, and sums and products $\pmod{26}$ satisfy the usual associative and distributive properties. For $R = \mathbb{Z}/26\mathbb{Z}$, let $GL(m, R)$ denote the set of invertible matrix transformations from $R^m \to R^m$ (that is, one-to-one and onto linear functions).

The construction

Suppose your message m consists of n capital letters, with no spaces. This may be regarded an n-tuple M with elements in $\mathbb{Z}/26\mathbb{Z}$. Identify the message M as a sequence of column vectors $\mathbf{p} \in (\mathbb{Z}/26\mathbb{Z})^k$. A *key* in the Hill cipher is a $k \times k$ matrix K, all of whose entries are in $\mathbb{Z}/26\mathbb{Z}$, such that the matrix K is invertible. It is important to keep k and K secret.

The encryption is performed by computing

$$\mathbf{c} = K\mathbf{p}, \tag{3}$$

and rewriting the resulting vector as a string over the same alphabet. Decryption is performed similarly by computing

$$\mathbf{p} = K^{-1}\mathbf{c}. \tag{4}$$

Example 1. Suppose m is the message“BWGN” (“Beat Westpoint, Go Navy”) . Transcoding into numbers, the plaintext is rewritten $p_0 = 1, p_1 = 22, p_2 = 6, p_3 = 13$. Suppose the key is $K = \left(\begin{array}{cc} 1 & 3 \\ 5 & 12 \end{array}\right)$. Using (3), note that $c_0 = 7, c_1 = 3, c_2 = 24, c_3 = 3$. (Verification is left to the reader as an exercise.)

Security concerns This cipher is linear and can be broken by a known plaintext attack, described below.

In the 2×2 case, the plaintext is broken into blocks of size 2 as follows:

$$p = ((p_0, p_1), (p_2, p_3), \ldots, (p_{2\ell}, p_{2\ell+1})),$$

where each $p_i \in \mathbb{Z}/26\mathbb{Z}$. In this case, the key K is a 2×2 matrix of the form

$$K = \left(\begin{array}{cc} k_{11} & k_{12} \\ k_{21} & k_{22} \end{array}\right).$$

Write the ciphertext as follows:

$$c = ((c_0, c_1), (c_2, c_3), \ldots, (c_{2\ell}, c_{2\ell+1})),$$

where each $c_i \in \mathbb{Z}/26\mathbb{Z}$ is determined by

$$\left(\begin{array}{c} c_0 \\ c_1 \end{array}\right) = \left(\begin{array}{cc} k_{11} & k_{12} \\ k_{21} & k_{22} \end{array}\right)\left(\begin{array}{c} p_0 \\ p_1 \end{array}\right), \left(\begin{array}{c} c_2 \\ c_3 \end{array}\right) = \left(\begin{array}{cc} k_{11} & k_{12} \\ k_{21} & k_{22} \end{array}\right)\left(\begin{array}{c} p_2 \\ p_3 \end{array}\right), \ldots .$$

Here is a plaintext attack: assume c_0, c_1, c_2, c_3 and p_0, p_1, p_2, p_3 are both known. The first two matrix equations above can be rewritten as

$$p_0 \cdot k_{11} + p_1 \cdot k_{12} + 0 \cdot k_{21} + 0 \cdot k_{22} = c_0,$$

$$0 \cdot k_{11} + 0 \cdot k_{12} + p_0 \cdot k_{21} + p_1 \cdot k_{22} = c_1,$$

$$p_2 \cdot k_{11} + p_3 \cdot k_{12} + 0 \cdot k_{21} + 0 \cdot k_{22} = c_2,$$

$$0 \cdot k_{11} + 0 \cdot k_{12} + p_2 \cdot k_{21} + p_3 \cdot k_{22} = c_3,$$

in other words, as

$$\begin{pmatrix} p_0 & p_1 & 0 & 0 \\ 0 & 0 & p_0 & p_1 \\ p_2 & p_3 & 0 & 0 \\ 0 & 0 & p_2 & p_3 \end{pmatrix} \begin{pmatrix} k_{11} \\ k_{12} \\ k_{21} \\ k_{22} \end{pmatrix} = \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix}. \qquad (5)$$

Observe that this is invertible if any only if $P^* = \begin{pmatrix} p_0 & p_1 \\ p_2 & p_3 \end{pmatrix}$ is invertible if and only if $p_0p_3 \not\equiv p_1p_2 \pmod{26}$. Therefore, under mild conditions on the plaintext, the (secret) key K can be computed using matrix theory.

Example 2. Suppose the following are known: (a) the plaintext is $p_0 = 1, p_1 = 22, p_2 = 3, p_3 = 13$, (b) the ciphertext is $c_0 = 7, c_1 = 3, c_2 = 24, c_3 = 3$, and (c) key length is 2. In this case, you can compute $K = \begin{pmatrix} 17 & 22 \\ 7 & 1 \end{pmatrix}$ using (5) and matrix theory. (Verification is left to the reader as an exercise.)

The 3×3 and larger cases are similar. As in the 2×2 case, we can convert $K\mathbf{p} = \mathbf{c}$ into a system of linear equations of the form $P\mathbf{k} = \mathbf{c}$, which can be solved by inverting P. If K is a $k \times k$ matrix then the associated $k^2 \times k^2$ matrix P is invertible if and only if the corresponding matrix P^* is invertible, where P^* is the $k \times k$ matrix of the first k^2 plaintext digits (ordered in the array left-to-right, top-to-bottom).

Hamming codes

Richard Hamming is a pioneer of coding theory, introducing the binary Hamming codes in the late 1940's. In the days when an computer error could crash the computer and force the programmer to retype his punch cards, Hamming, out of frustration, designed a system whereby the computer could automatically correct certain errors. The family of codes named after him can easily correct one error, as we will see below.

Let $\mathbb{F}$ denote a finite field. A *(linear error-correcting) code C of length n over $\mathbb{F}$* is a vector subspace C of $\mathbb{F}^n$ (provided with a fixed basis) and its elements are called *codewords*. A codeword will be identified with its coordinate vector representation in the fixed basis of $\mathbb{F}^n$. The *Hamming weight* of a vector is simply its distance from the origin:

$$\mathbf{wt}\,(\mathbf{v}) = |\{i \mid v_i \neq 0\}|,$$

where $\mathbf{v} = (v_1, \ldots, v_n)$. The smallest Hamming weight achieved by the non-zero codewords in C is called the *minimum distance* $d = d_C$ of the code. The minimum distance is one quantity used to measure how "good" a code is, from the practical point of view. Note the minimum distance function is not invariant under a change of basis (and now you see why it is important that $\mathbb{F}^n$ be provided with a *fixed* basis). A code of length n, dimension k (as a vector space over $\mathbb{F}$) and minimum distance d is called an $[n, k, d]$-*code*.

The construction

Let $r > 2$, $n = 2^r - 1$, $k = 2^r - r - 1$, and let $\mathbb{F} = GF(2)$ denote the finite field having 2 elements. Let H denote the $r \times n$ matrix whose columns are all the non-zero vectors in $\mathbb{F}^r$ (arranged in some fixed order, say lexicographically). Define the *binary Hamming* $[n, k, 3]$-*code* C to be the subspace of $\mathbb{F}^n$ of dimension k, given by the kernel of the ("parity check") matrix H,

$$C = ker(H).$$

These codes are widely used in computer memory detection and correction.

Example 3. Consider the binary Hamming code C with parameters $[7, 4, 3]$. We may simply define C to be the subset of vectors of $\mathbb{F}^7$ of the form

$$E(\mathbf{m}) = (m_1, m_2, m_3, m_4, m_1 + m_3 + m_4, m_1 + m_2 + m_4, m_1 + m_2 + m_3 + m_4),$$

where $\mathbf{m} = (m_1, m_2, m_3, m_4)$ run over all possible elements in $\mathbb{F}^4$. Think of $\mathbf{m}$ as the "information" you want to transmit over a noisy channel and $E(\mathbf{m})$ as the message you send. The message contains the information plus some redundancy. There is enough redundancy for the receiver to recover the information if one error was made during transmission.

Let

$$\mathbf{b_1} = (1, 0, 0, 0, 1, 1, 1), \ \ \mathbf{b_2} = (0, 1, 0, 0, 0, 1, 1),$$

$$\mathbf{b_3} = (0, 0, 1, 0, 1, 0, 1), \ \ \mathbf{b_4} = (0, 0, 0, 1, 1, 1, 0).$$

Then, in the notation of (2), we can write $E(\mathbf{m})$ as

$$E(\mathbf{m}) = m_1\mathbf{b_1} + m_2\mathbf{b_2} + m_3\mathbf{b_3} + m_4\mathbf{b_4} = \mathbf{m}G,$$

where

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

In this case,

$$H = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

Decoding Hamming codes Let C be the Hamming code with parameter $r > 2$ over $\mathbb{F} = GF(2)$.

For decoding, assume that for each message sent by the sender over the noisy channel, the transmission received by the receiver contains at most one error. Mathematically, this means that if the sender transmits the codeword $\mathbf{c} \in C$ then the receiver either received $\mathbf{c}$ or $\mathbf{c} + \mathbf{e_i}$, for some i. Here $\mathbf{e_i}$ is the i-th standard basis element, with a 1 in the ith position and a 0 elsewhere.

> **Decoding algorithm**: Assume that for each message sent by the sender over the noisy channel, the transmission received by the receiver contains exactly one error.
>
> INPUT: The received vector $\mathbf{v} \in \mathbb{F}^n$.
> OUTPUT: The codeword $\mathbf{c} \in C$ closest to $\mathbf{v}$ in the Hamming metric.
> ALGORITHM:
>
> - Order the columns of the check matrix H of C in some fixed way.
> - Compute $\mathbf{s} = H\mathbf{v}$ (this is called the *syndrome* of $\mathbf{v}$).
> - If $\mathbf{s} = \mathbf{0}$ then $\mathbf{v}$ is a codeword. Let $\mathbf{c} = \mathbf{v}$.
>
>   If $\mathbf{s} \neq \mathbf{0}$ then $\mathbf{v} = \mathbf{c} + \mathbf{e_i}$ for some codeword $\mathbf{c}$ and some $\mathbf{e_i}$. In this case,
>
> $$\mathbf{s} = H\mathbf{v} = H(\mathbf{c} + \mathbf{e_i}) = H\mathbf{c} + H\mathbf{e_i} = \mathbf{0} + H\mathbf{e_i} = H\mathbf{e_i}$$
>
>   is the i-th column of H. This tells us what i is. Also, this means that there was an error in the i-th coordinate of $\mathbf{c}$. Let $\mathbf{c} = \mathbf{v} + \mathbf{e_i}$.
> - Return $\mathbf{c}$.

Hill's unpublished paper

While he was a student at Yale, Hill published three papers in *Telegraph and Telephone Age* [**3**], [**4**], [**5**]. In these papers Hill described a mathematical method for checking the accuracy of telegraph communications. There is some overlap with these papers and [**7**], so it seems likely to me that Hill's unpublished paper [**7**] dates from this time (that is, during his later years at Yale or early years at Hunter). In [**7**], Hill describes a family of linear block codes over a finite field and an algorithm for error-detection (which can be easily extended to error-correction). In it, he states the construction of what I'll call the "Hill codes," (defined below), gives numerous computational examples, and concludes by recording Hill's Problem (stated above as Question 1). It is quite possibly Hill's best work.

Here is how Hill describes his set-up.

> Our problem is to provide convenient and practical accuracy checks upon a sequence of n elements $f_1, f_2, \ldots, f_r$ in a finite algebraic field F. We send, in place of the simple sequence $f_1, f_2, \ldots, f_r$, the amplified sequence

$f_1, f_2, \ldots, f_r, c_1, c_2, \ldots, c_k$ consisting of the "operand" sequence and the "checking" sequence.

- Lester Hill, [7]

Then Hill continues as follows. Let $F = GF(p)$ denote the finite field having p elements, where $p > 2$ is a prime number. The checking sequence contains k elements of F as follows:

$$c_j = \sum_{i=1}^{r} a_i^j f_i,$$

for $j = 1, 2, \ldots, k$. The checks are to be determined by means of a fixed matrix

$$A = \begin{pmatrix} a_1 & a_2 & \ldots & a_r \\ a_1^2 & a_2^2 & \ldots & a_r^2 \\ \vdots & & & \vdots \\ a_1^k & a_2^k & \ldots & a_r^k \end{pmatrix}$$

of elements of F, the matrix having been constructed according to the criteria in Hill's Problem above. In other words, if the operand sequence (i.e., the message) is the vector $\mathbf{f} = (f_1, f_2, \ldots, f_r)$, then the amplified sequence (or codeword in the Hill code) to be transmitted is

$$\mathbf{c} = \mathbf{f}G,$$

where $G = (I_r, A)$ and where I_r denotes the $r \times r$ identity matrix. The *Hill code* is the row space of G.

Example 4. This example is not from Hill's paper, but is constructed using more recent research in the theory of error-correcting codes. According to Shokrollahi [9], the matrix over $GF(211)$,

$$A_0 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & \ldots & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & \ldots & 210 \\ 1 & 4 & 9 & 16 & 25 & 36 & \ldots & 1 \\ 1 & 8 & 27 & 64 & 125 & 5 & \ldots & 210 \\ 1 & 16 & 81 & 45 & 203 & 30 & \ldots & 1 \end{pmatrix},$$

has the property that every square submatrix is non-singular. Therefore,

$$A = \begin{pmatrix} 2 & 3 & 4 & 5 & 6 \\ 4 & 9 & 16 & 25 & 36 \\ 8 & 27 & 64 & 125 & 5 \\ 16 & 81 & 128 & 203 & 30 \\ 32 & 32 & 90 & 171 & 180 \end{pmatrix},$$

satisfies the condition in Hill's Problem. The matrix $G = (I_5, A)$ is a generator matrix of a Hill code. Clearly, this is a linear $[n, k, d]$ code over $GF(211)$ with length $n = 10$, dimension $k = 5$ and minimum distance $d \leq 6$. However, computer computations show that in fact, $d \leq 5$.

We conclude with one more open question.

Question 3. *What is the minimum distance of a Hill code?*

The minimum distance of any Hamming code is $d = 3$. Do all sufficiently long Hill codes have minimum distance greater than 3?

Summary

Most books today (for example, the excellent MAA publication written by Thompson [**11**]) date the origins of the theory of error-correcting codes to the late 1940s, due to Richard Hamming. However, this paper argues that the actual birth is in the 1920s due to Lester Hill. Topics discussed include why Hill's discoveries weren't publicly known until relatively recently, what Hill actually did that trumps Hamming, and some open (mathematical) questions connected with Hill's work.

Acknowledgements: Many thanks to Chris Christensen and Alexander Barg for helpful and encouraging conversations. I'd like to explicitly credit Chris Christensen, as well as historian David Kahn, for the original discoveries of the source material.

References

1. C. Christensen, *Lester Hill revisited*, Cryptologia 38(2014)293-332.
2. ——, D. Joyner and J. Torres, *Lester Hill's error-detecting codes*, Cryptologia 36(2012)88-103.
3. L. Hill, *A novel checking method for telegraphic sequences,* Telegraph and Telephone Age (October 1, 1926), 456 - 460.
4. ——, *The role of prime numbers in the checking of telegraphic communications, I* Telegraph and Telephone Age (April 1, 1927), 151 - 154.
5. ——, *The role of prime numbers in the checking of telegraphic communications, II* Telegraph and Telephone Age (July, 16, 1927), 323 - 324.
6. ——, Lester S. Hill to Lloyd B. Wilson, November 21, 1925. Letter.
7. ——, *Checking the accuracy of transmittal of telegraphic communications by means of operations in finite algebraic fields*, undated and unpublished notes, 40 pages.
   Available in latex at `http://wdjoyner.org/papers/hill/`
8. F. MacWilliams and N. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, 1977.
9. A. Shokrollahi, *On cyclic MDS codes*, in *Coding Theory and Cryptography: From Enigma and Geheimschreiber to Quantum Theory* (ed. D. Joyner), Springer-Verlag, 2000.
10. T. Standage, *The Victorian Internet*, Walker & Company, 1998.
11. T. Thompson, *From Error-Correcting Codes Through Sphere Packings to Simple Groups*, Mathematical Association of America, 1983.