21st International Symposium on
Mathematical Theory of Networks and Systems
July 7-11, 2014. Groningen, The Netherlands

Editing in Manhattan

Lin Sok[1], Patrick Solé[1,2], Aslan Tchamkerten[1]
[1]Telecom ParisTech, CNRS LTCI, 46 rue Barrault 75 634 Paris, France.
[2] King Abdulaziz University, Department of Mathematics, Jeddah, Saudi Arabia.
{lin.sok;patrick.sole;aslan.tchamkerten}@telecom-paristech.fr

***Abstract*— The problem of constructing Insertion/Deletion codes for the editing distance is reduced to constructing codes over the integers for the Manhattan distance by run length coding. These latter codes are constructed by truncation of translates of lattices. These lattices in turn are obtained from Construction A applied to binary codes and $\mathbb{Z}_4-$codes. Complete weight enumerators of these codes allow to compute the generalized theta series of the corresponding lattices. Asymptotic Gilbert and Hamming type of bounds are derived in large dimensions. Constructive bounds better than the Gilbert type bound are derived by use of geometric codes.**

Keywords: Insertion/ Deletion codes, lattice, Lee metric, Construction A, weight enumerator, ν-series

I. Introduction

Coding for the insertion/deletion channel remains a major challenge for coding theorists. Part of the reason for this is that the use of standard block algebraic coding techniques (parity-checks, cosets, syndromes) is precluded due to the specificity of the channel which produces output vectors of variable lengths. A variation of this channel is the so-called segmented insertion/deletion channel where at most a fixed number of $r-1$ errors can occur within segments of given size [13], [12]. By looking at the input-output runlengths of symbols, the channel becomes a standard memoryless channel for which algebraic coding techniques can be used. Specifically, we construct lattice-based codes, which, in principle, can be decoded when obtained via Construction A from Lee metric codes with known decoding algorithms [6].

The proposed code constructions are analoguous to the so-called $(d,k)-$codes in magnetic recording where each codeword contains runs of zeros of length at least d and at most k while each run of one has unit length [8], [11]. Given d,k and assuming a constant number of runs of zeros, label the runs by integers modulo m and consider block codes over the ring of integers modulo m—the smallest possible m depends on d and k.

Our approach differs from the one in [8], [11] in two ways. First, we relax the unit length runlength of the ones. Second, we consider lattices rather than codes over the integers modulo m to allow a wider choice of parameters. Indeed our codes are obtained as sets of vectors in a lattice with certain metric properties. A code determines a lattice by Construction A but not conversely. We extend some results of [1], [17] on generalized theta series, called there $\nu-$series, to enumerate effectively these special sets of vectors in the lattice. In particular, if the lattice is obtained via Construction A from a code, the generalized $\nu-$series allows to enumerate these sets from the weight enumerators of the code.

This work was supported in part by an Excellence Chair Grant from the French National Research Agency (ACE project).

The paper is organized as follows. In Section 2, we formulate the problem. In Section 3, we state the main results on ν-series for Construction A lattices and provide some numerical results. In Section 4, we derive the analogue of the Gilbert and Hamming code size bounds for the Manhattan metric space. In section 5 the asymptotic versions of the analogues of respectively the Gilbert and Hamming bounds are derived. In Section 6, we improve constructively on these asymptotic bounds. In Section 7, we provide a few concluding remarks.

II. Background and Statement of the Problem

Consider a binary sequence of length N, that starts with a zero and ends with a one, and that contains n' runs of zeros and n' runs of ones. (Similar considerations occur for different choices of starting/ending symbol).

Example: The sequence 0011100011 contains $n'=2$ runs of each symbol for a length of $N=10$.

Working Hypothesis: In the whole paper, we assume that n' is the same for all the vectors in any given code so that we can work in constant length. Vectors are restricted to have runs of length at least r so that deletions of at most $r-1$ bits do not destroy the pattern of runs at the receiving end.

There is a natural correspondence between such a sequence and a sequence defined over the natural integers. Let x_i and y_i denote the ith run length of zeros and ones, respectively. Then we can form a sequence of length $n=2n'$ over the integers defined by

$$(x_1,y_1,\ldots,x_i,y_i,\ldots,x_{n'},y_{n'}).$$

Example: The binary sequence 0011100011 corresponds to $(2,3,3,2)$.

This approach is a natural generalization of [11] which considers the case where the y_i's are all equal to one.

Note that the integer sequence so constructed satisfies the constraint

$$N=\sum_{i=1}^{n'}(x_i+y_i).$$

Denote by ϕ the above correspondence from $\mathbb{F}_2^N$ to $\mathbb{Z}^n$. The **Levenshtein distance** between two binary vectors is the least number of insertions/deletions to go from one to the other. Alternatively it is the complement to the total length of both vectors of the length of the largest common

ISBN: 978-90-367-6321-9

subsequence. The **Manhattan distance** between two vectors $\mathbf{w}, \mathbf{z} \in \mathbb{Z}^n$ is given by the expression

$$|\mathbf{w} - \mathbf{z}| = \sum_{i=1}^{n} |w_i - z_i|.$$

The following observation is trivial but crucial.

Proposition 2.1: Under the above working hypothesis, the map ϕ is an isometry between the Levenshtein distance and the Manhattan distance. *Proof:* Let

$$\mathbf{z} = (x_1, y_1, \cdots, x_n, y_n)$$

denote the sequence of runs. Let j be an integer $\leq r-1$. Any insertion /deletion of j zeros (resp. ones) into run number i will result into a change of x_i (resp. y_i) into $x_i \pm j$ (resp. $y_i \pm j$) yielding a sequence $\mathbf{z}'$ at Manhattan distance j away from $\mathbf{z}$.

■

The problem we consider is to characterize $A(n, d, N, r)$, the largest number of length n vectors of nonnegative integers at Manhattan distance at least d apart and with coordinates summing up to N. Any set of length n vectors with integral entries $\geq r$, at Manhattan distance at least d apart, and coordinates summing up to N, we refer to as an $(n, d, N, r)-$set.

III. Enumeration for construction A lattices

By a **code** C of $\mathbb{Z}_m^n$, we shall mean a $\mathbb{Z}_m-$ submodule of $\mathbb{Z}_m^n$. Define the **complete weight enumerator** (cwe) of C as the homogeneous polynomial in the variables x_i given by the formula

$$cwe_C(x_1, x_2, \ldots, x_m) = \sum_{c \in C} \prod_{i=0}^{m-1} x_i^{n_i(C)},$$

where $n_i(c)$ is the number of entries equal to i in the vector c. For $m = 2$, we let $W_C(x, y) = cwe_C(x, y)$, the classical **weight enumerator** of a binary code.

By a **lattice** of $\mathbb{R}^n$, we shall mean a discrete additive subgroup of $\mathbb{R}^n$. A lattice L is said to be obtained by **Construction A** from a code C of $\mathbb{Z}_m^n$ if it is the inverse image of C in $\mathbb{Z}^n$ by reduction modulo m componentwise. This will be denoted by $L = A(C)$. An important parameter of a lattice is its minimum distance (norm) which is given by the following proposition. Recall that the **Lee weight** of a symbol $x \in \mathbb{Z}_m = \{0, 1, \cdots, m-1\}$, is $\min(x, m-x)$. The weight of a vector is the sum of the weight of its components, and the Lee distance of two vectors is the Lee weight of their difference. The **Lee distance** of a linear code $C \subseteq \mathbb{Z}_m^n$ is then the minimum nonzero weight of one its element.

Proposition 3.1 ([15]): Let L be a lattice constructed from a code C with reduction modulo m. Then its minimum distance

$$d = \min(d', m),$$

where d' is the minimum Lee distance of C.

For an integer $r \geq 0$ denote by $\nu_L(r; q)$ the shifted $\nu-$series in the indeterminate q of the lattice L

$$\nu_L(r; q) = \sum \{q^{|\mathbf{x}|} : \mathbf{x} \in L, \& \min_i x_i \geq r\}.$$

This definition extends trivially to any discrete subset L of $\mathbb{R}^n$. The motivation for this generating function, whose case $r = 0$ is the $\nu-$series of [1], [16], is Propositionn3.2, whose trivial proof is omitted. We use the Waterloo notation for coefficients of generating series. **Notation:** If the $q-$series $f = \sum_i f_i q^i$, we denote by $[q^i]f(q)$ the coefficient f_i.

Proposition 3.2: Keep the above notation. If L is a lattice of $\mathbb{R}^n$, of minimum Manhattan distance d then the set of vectors of L with coordinate entries bounded below by r and Manhattan norm N form a $(n, d, N, r)-$set of size $[q^N]\nu_L(r; q) \leq A(n, d, N, r)$.

We now show how to compute (shifted) $\nu-$series of lattices from (complete) weight enumerators of codes.

Theorem 3.3: If $L = A(C)$ and $m = 2$, then

$$\nu_L(r; q) = W_C(\frac{q^a}{1-q^2}, \frac{q^b}{1-q^2}),$$

where a (resp. b) is the first even (resp. odd) integer $\geq r$. If $L = A(C)$ and $m = 4$, then

$$\nu_L(r; q) = cwe_C(\frac{q^a}{1-q^4}, \frac{q^b}{1-q^4}, \frac{q^c}{1-q^4}, \frac{q^d}{1-q^4}),$$

where a, b, c, d are the first integers $\geq r$, congruent to $0, 1, 2, 3$ modulo 4 respectively.

Proof: By the same argument as in [1], [17], writing $A(C)$ as a disjoint union of cosets of $m\mathbb{Z}^n$, we have

$$\nu_L(r; q) = W_C(\nu_{2\mathbb{Z}}(r; q), \nu_{2\mathbb{Z}+1}(r; q))$$

for $m = 2$, and

$$\nu_L(r; q) =$$
$$cwe_C(\nu_{4\mathbb{Z}}(r; q), \nu_{4\mathbb{Z}+1}(r; q), \nu_{4\mathbb{Z}+2}(r; q), \nu_{4\mathbb{Z}+3}(r; q))$$

respectively, for $m = 4$. The result follows by observing that

$$\nu_{4\mathbb{Z}}(r; q) = \frac{q^a}{1-q^4}$$

and so on by summing appropriate geometric series of reason q^2 or q^4. ■

IV. Bounds on $A(n, d, N, r)$

We will use the enumerative results of the previous section. First we recall a well-known identity of formal power series.

Lemma 4.1: For any integer $n \geq 1$, we have

$$\frac{1}{(1-q)^n} = \sum_{i=0}^{\infty} \binom{i+n-1}{n-1} q^i.$$

Proof: Differentiate the geometric series

$$\frac{1}{(1-q)} = \sum_{i=0}^{\infty} q^i$$

with respect to q and use induction on n. ■

Using generating functions, we compute the volume $V(n, e)$ of the Manhattan ball of radius e in $\mathbb{Z}^n$.

Lemma 4.2: For any integers $n \geq e \geq 1$, we have

$$V(n, e) = [q^e]\frac{(1+q)^n}{(1-q)^{n+1}} = \sum_{i=0}^{\min(n,e)} 2^i \binom{n}{i}\binom{e}{i}.$$

Proof:

$$V(n,e) = \sum_{i=0}^{e}[q^i]\nu_{\mathbb{Z}^n}(-\infty,q) = \sum_{i=0}^{e}[q^i](\tfrac{1+q}{1-q})^n = [q^e]\tfrac{(1+q)^n}{(1-q)^{n+1}}.$$

The second expression is from [10]. It can be derived from the above generating series by expanding

$$(1+\frac{2q}{1-q})^{n+1} = \sum_{i=0}^{n}\binom{n}{i}2^i\frac{q^i}{(1-q)^{i+1}}$$

by Lemma 4.1. ∎

By the same techniques, we can compute the volume of the ambient space $A(n,1,N,r)$.

Lemma 4.3: For any integer $N > nr$ and $r > e \geq 1$, we have

$$A(n,1,N,r) = \binom{N-nr+n-1}{n-1}.$$

Proof:

$$A(n,1,N,r) = [q^N]\nu_{\mathbb{Z}^n}(r,q) = [q^N](q^r\tfrac{1}{1-q})^n = [q^{N-nr}]\tfrac{1}{(1-q)^n}.$$

The result follows from Lemma 4.1. ∎

We are now in a position to formulate the analogues of the Gilbert and Hamming bound in the present context.

Theorem 4.4: For any integers $N > nr, n \geq d$, and $r > e = \lfloor (d-1)/2 \rfloor \geq 1$, we have

$$\frac{\binom{N-nr+n-1}{n-1}}{V(n,d-1)} \leq A(n,d,N,r) \leq \frac{\binom{N-nr+n-1}{n-1}}{V(n,e)}.$$

Proof: Combine Lemma 4.2 and Lemma 4.3 with the standard arguments. ∎

Since all codewords have constant Manhattan distance, it is natural to use the Johnson bound in the Lee metric.

Theorem 4.5: If $d > N(1-1/2n)$, then we have

$$A(n,d,N,r) \leq \frac{d}{d-N(1-1/2n)}.$$

Proof: Reduce all vectors modulo $Q = 2N$. Use Lemma 13.62 of [5] with $\overline{D} = Q/4 = N/2$, and $x = 1/n$. ∎

V. Asymptotic bounds on $A(n,d,N,r)$

We assume that r is fixed, that $N \to \infty$, and that $n \sim \eta N/r$, $d \sim \delta N$ for some constants η, δ with $\eta \in (0,1)$, and $\delta \geq 0$. Because each codeword has weight N, the triangle inequality in the Manhattan metric shows that $\delta \in (0,2)$. Denote by R the asymptotic exponent of $A(n,d,N,r)$, that is

$$R = \limsup \frac{1}{N}\log A(n,d,N,r).$$

The asymptotic form of Theorem 4.5 shows that $\delta \in (0,1)$ whenever $R \neq 0$.

Let

$$L(x) = x\log_2 x + \log_2(x+\sqrt{x^2+1}) - x\log_2(\sqrt{x^2+1}-1).$$

It was proved in [9] that when $x \to \infty$ and $e \sim \epsilon n$, then

$$\lim \frac{1}{n}\log_2 V(n,e) = L(\epsilon).$$

For convenience, let $H(q) = -q\log q - (1-q)\log(1-q)$ denote the binary entropy function and let

$$f(x,y,z) = [1-y+y/x]H(\frac{y}{y+x(1-y)}) - (y/x)L(\frac{xz}{y}).$$

We now state and prove the asymptotic version of Theorem 4.4.

Theorem 5.1: With the above notation, we have

$$f(r,\eta,\delta) \leq R \leq f(r,\eta,\delta/2).$$

Proof: The result follows from Theorem 4.4 by standard entropic estimates for binomial coefficients for the numerator and the result on large alphabet Lee balls from [9] for the denominators. ∎

VI. Improvement of Gilbert type bound

Let L be a Construction A lattice in $\mathbb{Z}^{n-1}$ with L^1-distance d. Define

$$\widehat{L} = \{(x_1,x_2,\ldots,x_{n-1},N-\sum_{i=1}^{n-1}x_i) | (x_1,\ldots,x_{n-1}) \in L\}$$

and

$$E(n,N,r) = \{y \in \mathbb{R}^{n-1} | y_i \geq r, \sum_{i=1}^{n-1} y_i \leq N-r-\rho\},$$

where ρ is the covering radius for L.

Then covering the volume of $E(n,N,r)$ by Voronoi domains for L yields the following bound

$$Vol(E(n,N,r)) \leq A(n,N,r,d)Vol(L),$$

where $Vol(L)$ is the volume of the Voronoï cell for a lattice L.

The volume $Vol(E(n,N,r)$ is equal to the volume of the unit sphere multiplied by $\frac{(N-r-\rho)^{n-1}}{2^{n-1}}$, that is

$$Vol(E(n,N,r)) = \frac{2^{n-1}}{(n-1)!}\frac{(N-r-\rho)^{n-1}}{2^{n-1}}.$$

Taking C as a linear code over $\mathbb{F}_p$ with parameter $[n-1,k]$ and $L = A(C)$, that is

$$L = \bigcup_{c\in C}(c+p\mathbb{Z}^{n-1}),$$

the volume of L is equal to p^{n-1-k}.

Thus

$$A(n,N,r,d) \geq \frac{(N-r-\rho)^{n-1}}{(n-1)!p^{n-1-k}}. \quad (1)$$

By applying C as the concatenation code constructed in Corollary 4.2 of [14], we get the following theorem.

Theorem 6.1: With the above normalization, there exists a deletion code over $\mathbb{Z}_p$ whose asymptotic rate satisfies

$$R \geq \frac{1}{\alpha}\left(\log_2(\alpha e) + (-1+\frac{1}{2}(1-\alpha\delta))\log_2 p\right),$$

where $\alpha = \frac{r}{\eta}$.

Proof: With $r, \rho \sim p$ fixed, $N \to \infty, n \sim \frac{\eta}{r}N, d \sim \delta N$, We have

$$\frac{\log_2(A(n,N,d,r))}{n} \geq \log_2 N - \frac{\log_2((n-1)!)}{n} - \frac{\delta r}{2\eta}\log_2 p.$$

Using Stirling's approximation for $n! \sim n^n e^{-n}\sqrt{2\pi n}$, we get

$$\log_2(n!) \sim \log_2 n - \log_2 e.$$

Then in the said asymptotic regime we obtain

$$\lim(\frac{\log_2(A(n,N,d,r))}{N}) \geq \frac{\eta}{r}\left(\log_2\left(\frac{re}{\eta}\right) - (1-\frac{k}{n})\log_2 p\right).$$

Taking $\frac{k}{n} = \frac{1}{2}(1 - \frac{r}{\eta}\delta)$, we get the theorem as claimed. ■

By applying the code C as in Corollary 4.8 of [14], we get:

Theorem 6.2: With the above normalization, there exists a deletion code whose asymptotic rate satisfies

$$R \geq \frac{1}{\alpha}\left(\log_2(\alpha e) + (-1 + R'(\delta))\log_2 p\right),$$

where $\alpha = \frac{r}{\eta}$ and
$R'(\delta) =$

$$\begin{cases} \left(\frac{-v-\sqrt{\Delta}}{2}\right)^{1/3} + \left(\frac{-v+\sqrt{\Delta}}{2}\right)^{1/3} + \frac{4}{3} - \frac{1}{p-1} & \text{if } \Delta \geq 0 \\ 2\sqrt{\frac{-u}{3}}\cos\left(\frac{1}{3}\cos^{-1}(-\sqrt{\frac{27v^2}{-4u^3}}) + \frac{2\pi}{3}\right) + \frac{4}{3} - \frac{1}{p-1} & \text{if } \Delta < 0 \end{cases}$$

with $\Delta = 6912\delta'^6 + 2112\delta'^4 - \frac{16\delta'^2}{3}$, $u = 36\delta'^2 - \frac{1}{3}$ and $v = (48\delta'^2 - \frac{2}{7})$ and $\delta' = \alpha\delta$.

Proof: Following the same idea as the proof above and taking $R'(\delta) = \frac{k}{n}$, we get the theorem. ■

Recall that in [14], the Victoria+descent construction is better than the concatenation for large alphabet p. The graphs of comparison show that for small alphabet p, for instance $p = 7$, the latter construction using the concatenation code [14] improves the former with larger relative distance δ while, for large alphabet, for instance $p = 17$, the one using the Victoria+descent code [14] does with smaller relative distance δ.

VII. Conclusion

In this work, we have approached a problem of binary coding for the Levenshtein distance by using lattices for the Manhattan metric. These lattices are obtained by Construction A applied to binary and quaternary codes. Finding the densest lattice for the L^1-metric in a given dimension is still an open problem. Therefore it is worth varying codes, alphabets and use other constructions to improve the constructions of $(n,d,N,r)-$sets.

References

[1] M. Barlaud, M. Antonini, P. Solé, P. Mathieu, T. Gaidon "A pyramidal scheme for lattice vector quantization of wavelet transform coefficients applied to image coding" IEEE Trans. on Image Processing. 3 (1994) 367-381.

[2] A. Bonnecaze, P. Solé, C. Bachoc, B. Mourrain "Type II Codes over$\mathbf{Z}_4$" , IEEE Trans. on Information Theory, IT-43 (1997) 969-976.

[3] A. Bonnecaze, P. Solé, R. Calderbank, "Quaternary Quadratic Residue Codes and Unimodular Lattices" IEEE Trans. on Information Theory IT-41 (1995) 366-377.

[4] W. Bosma and J. Cannon, *Handbook of Magma Functions*, Sydney, 1995.

[5] E. Berlekamp, *Algebraic Coding Theory*, Aegean Park Press (1984).

[6] Antonio Campello, Grasiele C. Jorge, Sueli I. R. Costa, Decoding q-ary lattices in the Lee metric,http://arxiv.org/abs/1105.5557

[7] J. H. Conway, N. J. A. Sloane, "Fast quantizing and decoding algorithms for lattice quantizers and codes" IEEE Trans. on Information Theory IT-28(2): 227-231 (1982).

[8] AJ. Han Vinck, H. Morita, Codes over the ring of integers modulo m, IEICE Fundamentals, (1998) 2013–2018. http://www.exp-math.uni-essen.de/vinck/reference-papers/ vinck-morita-integer.pdf

[9] D. Gardy, P. Solé,"Saddle Point Techniques in Asymptotic Coding Theory." Congrès Franco-Soviétique de codage algébrique, Paris (1991), Springer Lecture Notes in Computer Science 573 (1991) 75–81. ftp://ftp.cs.brown.edu/pub/.../91/cs91-29.pdf

[10] S.W. Golomb, L.R. Welch, Perfect codes in the Lee metric and the packing of polyominoes, SIAM J. on Applied Math, Vol. 18, No 2, (1970) 302–317.

[11] Vladimir I. Levenshtein, A. J. Han Vinck: Perfect $(d,k)-$codes capable of correcting single peak-shifts. IEEE Transactions on Information Theory 39(2): 656-662 (1993)

[12] H. Mirghasemi, A. Tchamkerten: On the capacity of the one-bit deletion and duplication channel, Allerton (2012).

[13] Z. Liu, M. Mitzenmacher, Codes for deletion and insertion channels with segmented errors, ISIT (2007) 846–850.

[14] H. Randriam, L. Sok, and P. Solé, " Lower bound on the minimum distance of long codes in the Lee metric," *Designs Codes and Cryptography,* DOI: 10.1007/s10623-013-9870-z

[15] Rush J. A. and Sloane N. J. A. An improvement to the Minkowski-Hlawka bound for packing superball, Mathematika, vol. 34 (1987), pp. 8-18

[16] N. J. A. Sloane, On Single-Deletion-Correcting Codes, Codes and Designs, Ohio State University, May 2000 (Ray-Chaudhuri Festschrift), K. T. Arasu and A. Seress (editors), Walter de Gruyter, Berlin, 2002, pp. 273-291. http://neilsloane.com/doc/dijen.pdf

[17] P. Solé, Counting lattice points in pyramids. Discrete Mathematics, Volume 139, Number 1, 24 May 1995 , pp. 381-392