Codes in the Space of Multisets—Coding for Permutation Channels with Impairments

Mladen Kovačević, *Member, IEEE*, and Vincent Y. F. Tan, *Senior Member, IEEE*

***Abstract*—Motivated by communication channels in which the transmitted sequences are subject to random permutations, as well as by DNA storage systems, we study the error control problem in settings where the information is stored/transmitted in the form of multisets of symbols from a given finite alphabet. A general channel model is assumed in which the transmitted multisets are potentially impaired by insertions, deletions, substitutions, and erasures of symbols. Several constructions of error-correcting codes for this channel are described, and bounds on the size of optimal codes correcting any given number of errors derived. The construction based on the notion of Sidon sets in finite Abelian groups is shown to be optimal, in the sense of minimal asymptotic code redundancy, for any "error radius" and any alphabet size. It is also shown to be optimal in the sense of maximal code cardinality in various cases.**

***Index Terms*—Error correction, multiset code, lattice packing, diameter-perfect code, Sidon set, difference set, permutation channel, insertion, deletion, DNA storage.**

I. Introduction

INFORMATION storage systems using pools of DNA molecules as storage media have recently been proposed in the literature [19]. A distinctive feature of those models is that the data is "written" in the pools in an unordered fashion. This kind of storage is rather different from the traditional ones, and requires information to be encoded in the form of *multisets*[1] of symbols—objects which are unordered by definition. A similar situation arises in communication channels in which the input sequences are subject to random permutations. In such channels, the order of the symbols belonging to the input sequence cannot be inferred by the receiver with a reasonable degree of confidence, and the only carrier of information is again the multiset of the transmitted symbols.

One of the necessary ingredients of both kinds of systems described above are codes capable of protecting the stored/transmitted multisets from various types of noise. Motivated by this observation, we study error-correcting codes in the space of multisets over an arbitrary finite alphabet. The error model that we adopt is "worst-case" (as opposed to probabilistic); in other words, we focus on constructions and bounds on the cardinality of optimal codes capable of correcting a given number of errors.

Date: July 10, 2017.

This work was supported by the Singapore Ministry of Education (MoE) Tier 2 grant "Network Communication with Synchronization Errors: Fundamental Limits and Codes" (Grant number R-263-000-B61-112). Part of the work was presented at the 2017 IEEE International Symposium on Information Theory (ISIT) [27].

M. Kovačević is with the Department of Electrical & Computer Engineering, National University of Singapore (email: mladen.kovacevic@nus.edu.sg).

V. Y. F. Tan is with the Department of Electrical & Computer Engineering and the Department of Mathematics, National University of Singapore (email: vtan@nus.edu.sg).

[1]Informally, a multiset is a set with repetitions of elements allowed.

A. The channel model

We next describe in abstract terms the channel model that will be referred to throughout the paper. More concrete examples of communication channels that motivated introducing this model are mentioned in the following subsection.

The channel inputs are multisets of symbols from a given alphabet $\mathbb{A}$. Let $U = \{u_1, \ldots, u_n\}$ denote the generic input, where $u_i \in \mathbb{A}$ and the u_i's are not necessarily all distinct. The channel acts on the transmitted multiset U by removing some of its elements (deletions), by adding some elements to it (insertions), by replacing some of its elements with other symbols from $\mathbb{A}$ (substitutions), and by replacing some of its elements with the symbol $? \notin \mathbb{A}$ (erasures). What is obtained at the channel output is another multiset $\tilde{U} = \{\tilde{u}_1, \ldots, \tilde{u}_{\tilde{n}}\}$, where $\tilde{u}_i \in \mathbb{A} \cup \{?\}$ and $\tilde{n} = |\tilde{U}|$ in general need not equal $n = |U|$. The goal of the receiver is to reconstruct U from $\tilde{U}$.

As pointed out above, our main object of study are codes enabling the receiver to uniquely recover the transmitted multiset U after it has been impaired by a given number of insertions, deletions, substitutions, and erasures.

B. Motivation

1) Permutation Channels: Consider a communication channel that acts on the transmitted sequences by permuting their symbols in a random fashion. In symbolic notation:

$$u_1\ u_2\ \cdots\ u_n \quad \rightsquigarrow \quad u_{\pi(1)}\ u_{\pi(2)}\ \cdots\ u_{\pi(n)},$$

where $(u_1, \ldots, u_n) \in \mathbb{A}^n$ is a *sequence* of symbols from the input alphabet $\mathbb{A}$, and π is drawn uniformly at random from the set of all permutations over $\{1, \ldots, n\}$. Apart from shuffling their symbols, the channel is assumed to impose other kinds of impairments on the transmitted sequences as well, namely insertions, deletions, substitutions, and erasures of symbols. Let us refer to this model as the *permutation channel* [25].

As discussed in [25], the appropriate space for defining error-correcting codes for the permutation channel[2] is the set of all *multisets* over the channel alphabet $\mathbb{A}$, and therefore the results of this paper are relevant precisely for such models. The

[2]To our knowledge, apart from [25], there is only a handful of papers discussing coding and related problems for channels with random reordering of symbols, e.g., [49], [37], [14], [24]. Channel models with restricted reordering errors have also been studied in the literature, e.g., [2], [28], [31], [40]; in these and similar works it is assumed that only certain permutations are admissible during a given transmission and, consequently, they require quite different methods of analysis.

reasoning behind this observation is very simple: in the permutation channel, no information can be transferred in the *order* of the transmitted symbols, because this information cannot be inferred by the receiver with any probability greater than what could be obtained by random guessing. Consequently, the only carrier of information is the multiset of symbols $U = \{u_1, \ldots, u_n\}$.

The communication scenario that motivated introducing the above model are packet networks employing multipath routing as a means for end-to-end packet transfer [29]. In such networks, packets belonging to the same "generation" usually traverse paths of different lengths, bandwidths, congestion levels, etc., on their way to the receiver, which causes their delays to be different and unpredictable. Consequently, the packets may arrive at the destination in a different order than the one they were transmitted in. Furthermore, packets can be deleted in the network due to buffer overflows in network routers, link failures, etc., and they can also experience other types of errors for various reasons. Therefore, this can be seen as an instance of the permutation channel whose alphabet is the set of all possible packets.

Models related to the permutation channel are also relevant for diffusion-based molecular communications [38], where reordering errors, as well as deletions, are frequent due to the nature of the channel.

2) DNA Storage Systems: A class of data storage systems that uses synthesized DNA molecules as information carriers was recently proposed[3] and studied in [19]. In this model, information is written onto n DNA molecules of length ℓ each, which are then stored in an unordered way. In other words, information is stored in the form of multisets of cardinality n over an alphabet of size 4^ℓ.

Furthermore, it is reasonable to incorporate noise in the model of any such data storage system. For example, one might consider deletions and substitutions of DNA molecules as errors that can happen during the reading process. While [19] does not consider any such type of noise and assumes that molecules are stored and read in an error-free way, these impairments form an integral part of the present work. Another two salient differences vis-à-vis [19] are: (i) the molecules from the stored multiset are sampled with replacement by the receiver, and (ii) the fundamental limits of the system, in terms of rate and under a vanishing error probability formalism, are studied. Here, we are mostly concerned with the fundamental limits of the code sizes in the space of multisets when the aforementioned impairments are present.

C. Contributions and paper organization

In Section II we introduce multiset codes formally and demonstrate some of their basic properties. We prove that all four types of impairments considered here are in a sense equivalent, so one can focus on analyzing only one of them, e.g., deletions. We also introduce a metric that is appropriate for the problem at hand and that characterizes error correction capability of multiset codes.

[3]A related model is also discussed in an unpublished manuscript [33], albeit from a different perspective.

A geometric restatement of the problem, given in Section II-A, reveals a close connection between multiset codes and codes in the so-called A_q lattices [9], which prompted us to investigate the latter in their own right. These results, presented in Section III, will be used subsequently to derive some properties of multiset codes, but are also of independent interest. In particular, we demonstrate that linear codes in A_q lattices are geometric analogs of the so-called Sidon sets, a notion well-known in additive combinatorics [39]. Perfect and diameter-perfect codes in A_q lattices are also studied here, and several (non-)existence results in this regard provided.

In Section IV we describe our main construction of multiset codes, based on Sidon sets, and derive bounds on the size of optimal codes correcting a given number of errors. This construction is shown to be optimal, in the sense of minimal asymptotic redundancy, for any "error radius" and any alphabet size. It is also shown to be optimal in the (stricter) sense of maximal code cardinality in various cases. It turns out that codes in the space of multisets are closely related to codes for classical binary insertion/deletion channels with restrictions either on the noise model, or on the channel inputs. We discuss this fact in Section IV-D and show that our results improve upon the existing results for those channels.

In Section V we describe two additional code constructions that are provably suboptimal, but are of interest nonetheless. One of them is based on indexing—a method that essentially turns a sequence into a set by adding a sequence number prefix to each of its symbols. This approach is frequently used to protect the packets from possible reorderings in networking applications [29]. We prove that this construction is strictly suboptimal, and quantify this fact by comparing the rates achievable by optimal codes obtained in this way to those achievable by optimal multiset codes. The other construction provided here has the algebraic flavor usually encountered in coding theory, and is based on encoding information in the roots of a suitably defined polynomial.

A brief conclusion and several pointers for further work are stated in Section VI.

II. General properties of multiset codes

Throughout the paper we assume that the channel input alphabet is $\mathbb{A} = [0{:}q] \coloneqq \{0, 1, \ldots, q\}$, for some $q \geq 1$.

A. Basic definitions and geometric representation

For a multiset $U = \{u_1, \ldots, u_n\}$ over $[0{:}q]$, denote by $\mathbf{x}^U = \left(x_0^U, x_1^U, \ldots, x_q^U\right) \in \mathbb{Z}^{q+1}$ the corresponding vector of multiplicities of its elements, meaning that x_i^U is the number of occurrences of the symbol $i \in [0{:}q]$ in U. The vector $\mathbf{x}^U$ satisfies $x_i^U \geq 0$ and $\sum_{i=0}^q x_i^U = |U| = n$. Multisets over a given alphabet are uniquely specified by their multiplicity vectors. For that reason we shall mostly use the vector notation and terminology in the sequel, occasionally referring to multisets; it should be clear that these are just two different ways of expressing the same notion.

We emphasize that, even though the codewords will be described by integer vectors, these vectors are not actually sent through the channel described in Section I-A. Namely,

if $\mathbf{x} = (x_0, x_1, \ldots, x_q) \in \mathbb{Z}^{q+1}$ is a codeword, then what is being transmitted is a multiset containing x_0 copies of the symbol 0, x_1 copies of the symbol 1, etc. Therefore, $n = \sum_{i=0}^{q} x_i$—the cardinality of the multiset in question—will be referred to as the *length* of the codeword $\mathbf{x}$ in this setting because this is the number of symbols that are actually being transmitted. Also, a deletion in the channel is understood not as a deletion of an element of $\mathbf{x}$, but rather a deletion of an element of the multiset represented by $\mathbf{x}$, and similarly for the other types of noise.

If we impose the usual requirement that all codewords are of the same length n, we end up with the following code space:

$$\triangle_n^q = \left\{ \mathbf{x} \in \mathbb{Z}^{q+1} : x_i \geq 0, \sum_{i=0}^{q} x_i = n \right\}. \tag{1}$$

The set $\triangle_n^q$ is a discrete simplex of "sidelength" n, dimension q, and cardinality $|\triangle_n^q| = \binom{n+q}{q}$.

Definition 1. *A multiset code of length n over the alphabet $[0{:}q]$ is a subset of $\triangle_n^q$ having at least two elements.* ▲

The requirement that the code has at least two codewords is clear from the practical perspective. It is made explicit here to avoid discussing trivial cases in the sequel.

The metric on $\triangle_n^q$ that is appropriate for our purposes is essentially the ℓ_1 distance:

$$d(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \sum_{i=0}^{q} |x_i - y_i|. \tag{2}$$

The metric space $(\triangle_n^q, d)$ can be visualized as a graph with $|\triangle_n^q| = \binom{n+q}{q}$ vertices and with edges joining vertices at distance one, see Figure 1. The quantity $d(\mathbf{x}, \mathbf{y})$ is precisely the graph distance between the vertices corresponding to $\mathbf{x}$ and $\mathbf{y}$, i.e., the length of the shortest path between them. The minimum distance of a code $\mathcal{C} \subseteq \triangle_n^q$ with respect to $d(\cdot, \cdot)$ is denoted $d(\mathcal{C})$.

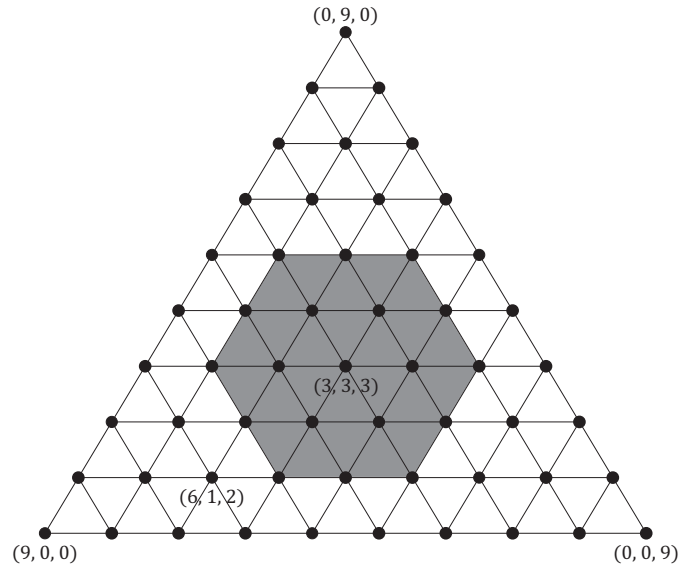


Fig. 1. The graph of the simplex $\triangle_9^2$ representing the set of all multisets of cardinality 9 over the ternary alphabet $\{0, 1, 2\}$, and an illustration of a ball of radius 2 in this graph.

The following definition is motivated by the structure of the code space $\triangle_n^q$. Namely, this space can be seen as the translated A_q lattice restricted to the non-negative orthant, where

$$A_q = \left\{ \mathbf{x} \in \mathbb{Z}^{q+1} : \sum_{i=0}^{q} x_i = 0 \right\}. \tag{3}$$

Definition 2. *We say that a multiset code $\mathcal{C} \subseteq \triangle_n^q$ is linear if $\mathcal{C} = (\mathcal{L} + \mathbf{t}) \cap \triangle_n^q$ for some* lattice *$\mathcal{L} \subseteq A_q$ and some vector $\mathbf{t} \in \mathbb{Z}^{q+1}$ with $\sum_{i=0}^{q} t_i = n$.* ▲

In other words, $\mathcal{C}$ is linear if it is obtained by translating a linear code in A_q (a sublattice of A_q) and keeping only the codewords with non-negative coordinates.

B. Error correction capability of multiset codes

Let $\mathbf{e}_i \in \mathbb{Z}^{q+1}$, $i \in [0{:}q]$, be the unit vector having a 1 at the i'th coordinate and 0's elsewhere. Let U be the transmitted multiset. If the received multiset $\tilde{U}$ is produced by inserting a symbol i to U, then $\mathbf{x}^{\tilde{U}} = \mathbf{x}^U + \mathbf{e}_i$. Similarly, deletion of i from U means that $\mathbf{x}^{\tilde{U}} = \mathbf{x}^U - \mathbf{e}_i$, and a substitution of $i \in U$ by j that $\mathbf{x}^{\tilde{U}} = \mathbf{x}^U - \mathbf{e}_i + \mathbf{e}_j$.

We say that a code can correct h_{ins} insertions, h_{del} deletions, and h_{sub} substitutions if no two distinct codewords can produce the same channel output after being impaired by *arbitrary patterns* of $\leq h_{\text{ins}}$ insertions, $\leq h_{\text{del}}$ deletions, and $\leq h_{\text{sub}}$ substitutions. In other words, every codeword can be uniquely recovered after being impaired by such an error pattern.

Theorem 1. *Let $\mathcal{C} \subseteq \triangle_n^q$ be a multiset code, h_{ins}, h_{del}, h_{sub} non-negative integers, and $h = h_{\text{ins}} + h_{\text{del}} + 2h_{\text{sub}}$. The following statements are equivalent:*

(a) $\mathcal{C}$ can correct h_{ins} insertions, h_{del} deletions, and h_{sub} substitutions,
(b) $\mathcal{C}$ can correct h insertions,
(c) $\mathcal{C}$ can correct h deletions.

Proof: Since any substitution can be thought of as a combination of a deletion and an insertion, and vice versa, the statement *(a)* is equivalent to the following:

(a') $\mathcal{C}$ can correct $h_{\text{ins}} + h_{\text{sub}}$ insertions and $h_{\text{del}} + h_{\text{sub}}$ deletions.

We show next that *(a')* $\Rightarrow$ *(c)*; the remaining implications can be proven in a similar way. We shall assume that $n \geq h$, the statement otherwise being vacuously true.

Suppose that *(c)* does not hold. This means that there are two different codewords (multisets) that can produce the same channel output after h elements have been deleted from both of them[4]. In other words, there exist $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\mathbf{x} \neq \mathbf{y}$, such that $\mathbf{x} - \mathbf{f} = \mathbf{y} - \mathbf{g}$ for some vectors $\mathbf{f}, \mathbf{g}$ with $f_i, g_i \geq 0$, $\sum_{i=0}^{q} f_i = \sum_{i=0}^{q} g_i = h$ ($\mathbf{f}$ and $\mathbf{g}$ represent patterns of h deletions from $\mathbf{x}$ and $\mathbf{y}$, respectively). Write $\mathbf{f} = \mathbf{f}^{\text{del}} + \mathbf{f}^{\text{ins}}$ and $\mathbf{g} = \mathbf{g}^{\text{del}} + \mathbf{g}^{\text{ins}}$, where $\mathbf{f}^{\text{del}}, \mathbf{f}^{\text{ins}}, \mathbf{g}^{\text{del}}, \mathbf{g}^{\text{ins}}$ are arbitrary vectors satisfying $f_i^{\text{del}}, f_i^{\text{ins}}, g_i^{\text{del}}, g_i^{\text{ins}} \geq 0$, $\sum_{i=0}^{q} f_i^{\text{del}} = \sum_{i=0}^{q} g_i^{\text{del}} = h_{\text{del}} + h_{\text{sub}}$, $\sum_{i=0}^{q} f_i^{\text{ins}} = \sum_{i=0}^{q} g_i^{\text{ins}} = h_{\text{ins}} + h_{\text{sub}}$. Then

[4]A code can correct up to h deletions if and only if it can correct *exactly* $\min\{h, n\}$ deletions (meaning that exactly $\min\{h, n\}$ symbols of the transmitted multiset are being deleted in the channel).

$\mathbf{x} - \mathbf{f}^{\text{del}} + \mathbf{g}^{\text{ins}} = \mathbf{y} - \mathbf{g}^{\text{del}} + \mathbf{f}^{\text{ins}}$, which means that $\mathcal{C}$ cannot correct $h_{\text{ins}} + h_{\text{sub}}$ insertions and $h_{\text{del}} + h_{\text{sub}}$ deletions. Hence, *(a')* does not hold. ■

Remark 1 (Erasures)**.** It is easy to include erasures in the model too, but we have chosen not to do so here because it would slightly complicate notation (due to the additional symbol '?' in the output alphabet). Namely, in the same way as in the above proof one can show that erasures are as damaging as deletions: A code $\mathcal{C} \subseteq \triangle_n^q$ can correct h erasures if and only if it can correct h deletions. We emphasize that this is only true for codes whose codewords are all of the same length, i.e., codes in $\triangle_n^q$. In the case of variable-length codes, which we do not analyze here, erased symbols can reveal some information about the cardinality of the transmitted multiset to the receiver, unlike deleted symbols which do not appear at the channel output. ▲

In light of Theorem 1 and Remark 1, we can assume that deletions are the only type of noise in the channel.

The following statement gives a metric characterization of the error correction capability of a multiset code $\mathcal{C}$.

Theorem 2. *A multiset code $\mathcal{C} \subseteq \triangle_n^q$ can correct h deletions if and only if its minimum distance is $d(\mathcal{C}) > h$.*

Proof: Let $\mathbf{x}, \mathbf{y}$ be two codewords at distance $d(\mathcal{C})$. Then $\mathbf{f} = \mathbf{x} - \mathbf{y}$ satisfies $\sum_{i=0}^q f_i = 0$ and $\sum_{i=0}^q |f_i| = 2d(\mathcal{C})$; see (2). Let $\mathbf{f}^+ = \max\{\mathbf{f}, \mathbf{0}\}$ and $\mathbf{f}^- = \max\{-\mathbf{f}, \mathbf{0}\}$ be the positive and negative part of $\mathbf{f}$, respectively, so that $\mathbf{f} = \mathbf{f}^+ - \mathbf{f}^-$ (here max is the coordinate-wise maximum). Then $\mathbf{x} - \mathbf{f}^+ = \mathbf{y} - \mathbf{f}^-$. Since $f_i^+, f_i^- \geq 0$ and $\sum_{i=0}^q f_i^+ = \sum_{i=0}^q f_i^- = d(\mathcal{C})$, both $\mathbf{f}^+$ and $\mathbf{f}^-$ can be thought of as noise vectors describing patterns of $d(\mathcal{C})$ deletions from $\mathbf{x}$ and $\mathbf{y}$, respectively. This means that $\mathcal{C}$ cannot correct $d(\mathcal{C})$ deletions. Reversing the argument, one sees that $\mathcal{C}$ can always correct $< d(\mathcal{C})$ deletions because assuming otherwise would imply that there exist two codewords at distance $< d(\mathcal{C})$, which is a contradiction. ■

C. Error detection capability of multiset codes

We now briefly discuss the error *detection* problem for the studied channel and show that it admits a metric characterization similar to the one obtained for error correction.

We say that a code can detect h_{ins} insertions, h_{del} deletions, h_{sub} substitutions, and h_{ers} erasures if no codeword $\mathbf{x}$ can produce another codeword $\mathbf{y} \neq \mathbf{x}$ at the channel output after being impaired by an *arbitrary pattern* of $\leq h_{\text{ins}}$ insertions, $\leq h_{\text{del}}$ deletions, $\leq h_{\text{sub}}$ substitutions, and $\leq h_{\text{ers}}$ erasures. In other words, such error patterns result in the receiver obtaining either the transmitted codeword $\mathbf{x}$, or something which is not a codeword at all, meaning that it can decide with certainty whether an error has happened during transmission or not.

Erasures are trivial to detect. Also, if the number of insertions that occur in the channel is different than the number of deletions, the received multiset will have a different cardinality than the transmitted one and the detection is easy. If the number of insertions and deletions is the same, say s, then this can be thought of as s substitutions, as discussed before. Therefore, for the purpose of analyzing error detection, it is not a loss of generality to consider substitutions as the only type of noise in the channel.

Theorem 3. *A multiset code $\mathcal{C} \subseteq \triangle_n^q$ can detect h substitutions if and only if its minimum distance is $d(\mathcal{C}) > h$.*

In other words, a code $\mathcal{C} \subseteq \triangle_n^q$ can detect h substitutions if and only if it can correct h deletions.

Proof: That $\mathcal{C}$ *cannot* detect h substitutions means that there are two different codewords $\mathbf{x}, \mathbf{y}$, and a vector $\mathbf{f}$ with $\sum_{i=0}^q f_i = 0$, $\sum_{i=0}^q |f_i| \leq 2h$, such that $\mathbf{y} = \mathbf{x} + \mathbf{f}$ ($\mathbf{f}$ represents a pattern of h substitutions). If this is the case, then $d(\mathbf{x}, \mathbf{y}) \leq h$, and hence $d(\mathcal{C}) \leq h$. The other direction is similar. ■

III. Codes in A_q lattices

As we observed in Section II-A, the space in which multiset codes are defined is a translated A_q lattice, restricted to the non-negative orthant. This restriction is the reason why the space $\triangle_n^q$ lacks some properties that are usually exploited when studying bounds on codes, packing problems, and the like. In order to analyze the underlying geometric problem, we shall disregard these constraints in this section, and investigate the corresponding problems in the metric space (A_q, d). In particular, we shall discuss constructions of codes in A_q lattices having a given minimum distance, bounds on optimal codes, and (non-)existence of perfect and diameter-perfect codes in (A_q, d). These results will later on be used to study the corresponding questions for multiset codes (see Section IV), but are also of independent interest.

A. A_q lattice under ℓ_1 metric

We first state some properties of A_q lattices under the metric $d(\cdot, \cdot)$ defined in (2). As in the case of multiset codes, we denote the minimum distance of a code $\mathcal{C} \subseteq A_q$ with respect to the metric $d(\cdot, \cdot)$ by $d(\mathcal{C})$. A code $\mathcal{C} \subseteq A_q$ is said to be linear if it is a sublattice of A_q. For $S, \mathcal{C} \subseteq A_q$, both nonempty, we say that $(S, \mathcal{C})$ is a *packing* in A_q if the translates $S + \mathbf{x}$ and $S + \mathbf{y}$ are disjoint for every $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, $\mathbf{x} \neq \mathbf{y}$. If $\mathcal{C}$ is a lattice, such a packing is called a lattice packing. The definitions for $\mathbb{Z}^q$ in place of A_q, and for an arbitrary metric in place of $d(\cdot, \cdot)$, are similar.

Another way of describing codes in the metric space (A_q, d) will be convenient for our purpose. For $\mathbf{x} = (x_1, \ldots, x_q)$, $\mathbf{y} = (y_1, \ldots, y_q) \in \mathbb{Z}^q$, define the metric

$$d_{\text{a}}(\mathbf{x}, \mathbf{y}) \coloneqq \max \left\{ \sum_{\substack{i=1 \\ x_i > y_i}}^{q} (x_i - y_i), \sum_{\substack{i=1 \\ x_i < y_i}}^{q} (y_i - x_i) \right\}. \quad (4)$$

This metric is used in the theory of codes for asymmetric channels (hence the subscript 'a'); see [23, Ch. 2.3 and 9.1].

Theorem 4. *(A_q, d) is isometric to $(\mathbb{Z}^q, d_{\text{a}})$.*

Proof: For $\mathbf{x} = (x_0, x_1, \ldots, x_q)$, denote $\mathbf{x}' = (x_1, \ldots, x_q)$. The mapping $\mathbf{x} \mapsto \mathbf{x}'$ is the desired isometry. Just observe that, for $\mathbf{x}, \mathbf{y} \in A_q$,

$$d(\mathbf{x}, \mathbf{y}) = \sum_{\substack{i=0 \\ x_i > y_i}}^{q} (x_i - y_i) = \sum_{\substack{i=0 \\ x_i < y_i}}^{q} (y_i - x_i) \tag{5}$$

because $\sum_{i=0}^{q} x_i = \sum_{i=0}^{q} y_i = 0$. Then, by examining the cases $x_0 \lesseqgtr y_0$, it follows that

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \max \left\{ \sum_{\substack{i=1 \\ x_i > y_i}}^{q} (x_i - y_i), \sum_{\substack{i=1 \\ x_i < y_i}}^{q} (y_i - x_i) \right\} \\ &= d_{\mathrm{a}}(\mathbf{x}', \mathbf{y}'). \end{aligned} \tag{6}$$

Furthermore, the mapping $\mathbf{x} \mapsto \mathbf{x}'$ is bijective. ∎

Therefore, packing and similar problems in (A_q, d) are equivalent to those in $(\mathbb{Z}^q, d_{\mathrm{a}})$, and hence we shall use these metric spaces interchangeably in the sequel. When discussing packings in $(\mathbb{Z}^q, d_{\mathrm{a}})$, the following sets naturally arise:

$$S_q(r^+, r^-) \coloneqq \left\{ \mathbf{x} \in \mathbb{Z}^q : \sum_{\substack{i=1 \\ x_i > 0}}^{q} x_i \leq r^+, \sum_{\substack{i=1 \\ x_i < 0}}^{q} |x_i| \leq r^- \right\}, \tag{7}$$

where $r^+, r^- \geq 0$. $S_q(r^+, r^-)$ is an anticode [1] of maximum distance $r^+ + r^-$ in $(\mathbb{Z}^q, d_{\mathrm{a}})$, i.e., a subset of $\mathbb{Z}^q$ of diameter $r^+ + r^-$. In particular, $S_q(r, r)$ is the ball of radius r around $\mathbf{0}$ in $(\mathbb{Z}^q, d_{\mathrm{a}})$, and $S_q(r, 0)$ is the simplex—the set of all non-negative vectors in $\mathbb{Z}^q$ with coordinates summing to $\leq r$.

Lemma 1. *The cardinality of the anticode $S_q(r^+, r^-)$ is*

$$|S_q(r^+, r^-)| = \sum_{j=0}^{\min\{q, r^+\}} \binom{q}{j} \binom{r^+}{j} \binom{r^- + q - j}{q - j}. \tag{8}$$

Proof: The j'th summand in (8) counts the vectors in $S_q(r^+, r^-)$ having j strictly positive coordinates. These coordinates can be chosen in $\binom{q}{j}$ ways. For each choice, the "mass" $\leq r^+$ can be distributed over them in $\sum_{t=j}^{r^+} \binom{t-1}{j-1} = \binom{r^+}{j}$ ways (think of placing $t \leq r^+$ balls into j bins, where at least one ball is required in each bin). Similarly, the mass $\leq r^-$ can be distributed over the remaining coordinates in $\sum_{t=0}^{r^-} \binom{t+q-j-1}{q-j-1} = \binom{r^- + q - j}{q-j}$ ways. ∎

The following claim provides a characterization of codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$ in terms of the anticodes $S_q(r^+, r^-)$. We omit the proof as it is analogous to the proof of the corresponding statement for finite spaces represented by distance-regular graphs [11], [1].

Theorem 5. *Let $\mathcal{C} \subseteq \mathbb{Z}^q$ be a code, and r^+, r^- non-negative integers. Then $(S_q(r^+, r^-), \mathcal{C})$ is a packing if and only if $d_{\mathrm{a}}(\mathcal{C}) > r^+ + r^-$.* ∎

Hence, whether $(S_q(r^+, r^-), \mathcal{C})$ is a packing depends on the values r^+, r^- only through their sum.

B. Codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$: An upper bound

Since the space $(\mathbb{Z}^q, d_{\mathrm{a}})$ is infinite, we cannot use the cardinality of a code as a measure of "how well it fills the space". The infinite-space notion that captures this fact is the *density* of a code, defined as

$$\mu(\mathcal{C}) \coloneqq \lim_{k \to \infty} \frac{|\mathcal{C} \cap \{-k, \ldots, k\}^q|}{(2k+1)^q}. \tag{9}$$

In case the above limit does not exist, one can define the upper ($\overline{\mu}(\mathcal{C})$) and the lower ($\underline{\mu}(\mathcal{C})$) density by replacing $\lim$ with $\limsup$ and $\liminf$, respectively. For a linear code $\mathcal{C}$ the density $\mu(\mathcal{C})$ exists and is equal to $\mu(\mathcal{C}) = \frac{1}{|\mathbb{Z}^q/\mathcal{C}|}$, where $\mathbb{Z}^q/\mathcal{C}$ is the quotient group of the code/lattice $\mathcal{C}$. Clearly, the higher the required minimum distance, the lower the achievable density. The following theorem quantifies this fact.

Theorem 6. *Let $\mathcal{C}$ be a code in $(\mathbb{Z}^q, d_{\mathrm{a}})$ with minimum distance $d_{\mathrm{a}}(\mathcal{C}) = d$. Then, for all non-negative integers r^+, r^- with $r^+ + r^- < d$,*

$$\overline{\mu}(\mathcal{C}) \leq |S_q(r^+, r^-)|^{-1}. \tag{10}$$

In particular, for $2 \leq d \leq 2q+1$,

$$\overline{\mu}(\mathcal{C}) < \left\lceil \frac{d-1}{2} \right\rceil! \left\lfloor \frac{d-1}{2} \right\rfloor! \left(q + 1 - \left\lceil \frac{d-1}{2} \right\rceil \right)^{1-d}, \tag{11}$$

and, for $1 \leq q < d$,

$$\overline{\mu}(\mathcal{C}) < 2^q q!^3 (2q)!^{-1} (d-q)^{-q}. \tag{12}$$

In words, the theorem gives upper bounds on the density of codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$ having a given minimum distance d and dimension q. This result will be used to derive an upper bound on the cardinality of optimal multiset codes with specified minimum distance d and alphabet size $q+1$ (Theorem 14 in Section IV).

Proof: The bound in (10) follows from Theorem 5 and is a version of the code-anticode bound [11], [1] adapted to the space studied here. Namely, if $d_{\mathrm{a}}(\mathcal{C}) = d > r^+ + r^-$, then every translate of $S_q(r^+, r^-)$ in $\mathbb{Z}^q$ contains at most one codeword from $\mathcal{C}$, and hence $\overline{\mu}(\mathcal{C}) \cdot |S_q(r^+, r^-)| \leq 1$.
By Lemma 1 we then have, for $q \geq r^+$,

$$\begin{aligned} \overline{\mu}(\mathcal{C})^{-1} &\geq \sum_{j=0}^{r^+} \binom{q}{j} \binom{r^+}{j} \binom{r^- + q - j}{q - j} \\ &\geq \sum_{j=0}^{r^+} \frac{(q-j+1)^j}{j!} \binom{r^+}{j} \frac{(q-j+1)^{r^-}}{r^-!} \\ &> \frac{(q - r^+ + 1)^{r^+ + r^-}}{r^+!\, r^-!}, \end{aligned} \tag{13}$$

where we used $\binom{q}{j} \geq \frac{(q-j+1)^j}{j!}$, and the last inequality is obtained by keeping only the summand $j = r^+$. Taking $r^+ =$

$\lceil \frac{d-1}{2} \rceil$ and $r^- = \lfloor \frac{d-1}{2} \rfloor$, we get (11).
Similarly, for $0 \le r^+ - q \le r^-$, we have

$$\begin{aligned} \overline{\mu}(\mathcal{C})^{-1} &\ge \sum_{j=0}^{q} \binom{q}{j} \binom{r^+}{j} \binom{r^- + q - j}{q - j} \\ &\ge \sum_{j=0}^{q} \binom{q}{j} \frac{(r^+ - j + 1)^j}{j!} \frac{(r^- + 1)^{(q-j)}}{(q-j)!} \\ &= \frac{1}{q!} \sum_{j=0}^{q} \binom{q}{j}^2 (r^+ - j + 1)^j (r^- + 1)^{(q-j)} \\ &> \frac{(r^+ - q + 1)^q}{q!} \binom{2q}{q}. \end{aligned} \tag{14}$$

In the last step we used the assumption $r^- \ge r^+ - q$ and the identity $\sum_{j=0}^{q} \binom{q}{j}^2 = \binom{2q}{q}$. Letting $r^+ = \lfloor \frac{d-1+q}{2} \rfloor$ and $r^+ + r^- = d - 1$, we get (12). ∎

C. Codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$: Construction based on B_h sets

In this subsection, we describe a method of construction of codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$ having a given minimum distance. As we shall demonstrate in Section III-D, the construction is optimal for some sets of parameters, and in fact produces perfect or diameter-perfect codes in those instances.

Let G be an Abelian group of order v, written additively. A set $B = \{b_0, b_1, \ldots, b_q\} \subseteq G$ is said to be a B_h *set* (or B_h *sequence*, or *Sidon set* of order h) if the sums $b_{i_1} + \cdots + b_{i_h}$, $0 \le i_1 \le \cdots \le i_h \le q$, are all different. If B is a B_h set, then so is $B - b_0 \equiv \{0, b_1 - b_0, \ldots, b_q - b_0\}$, and vice versa; we shall therefore assume in the sequel that $b_0 = 0$. With this convention, the requirement for B to be a B_h set is that the sums $b_{i_1} + \cdots + b_{i_u}$ are different for all $u \in \{0, 1, \ldots, h\}$ and $1 \le i_1 \le \cdots \le i_u \le q$. Among the early works on these and related objects we mention Singer's construction [41] of optimal B_2 sets in $\mathbb{Z}_v := \mathbb{Z}/v\mathbb{Z}$, for q a prime power and $v = q^2 + q + 1$, and a construction by Bose and Chowla [7] of B_h sets in $\mathbb{Z}_v$ for arbitrary $h \ge 1$ when: 1) q is a prime power and $v = q^h + q^{h-1} + \cdots + 1$, and 2) $q + 1$ is a prime power and $v = (q+1)^h - 1$. Since these pioneering papers, research in the area has become quite extensive, see [39] for references, and has also found various applications in coding theory, e.g., [4], [12], [17], [23], [32], [48].

The following statement demonstrates that linear codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$ (or, equivalently, in (A_q, d)) are in fact geometric analogs of B_h sets.

Theorem 7. *Let $h \ge 1$ be an integer.*

(a) Assume that $B = \{0, b_1, \ldots, b_q\}$ is a B_h set in an Abelian group G of order v, and that B generates G. Then the code

$$\mathcal{L} = \left\{ \mathbf{x} \in \mathbb{Z}^q : \sum_{i=1}^{q} x_i b_i = 0 \right\} \tag{15}$$

has minimum distance $d_{\mathrm{a}}(\mathcal{L}) > h$ and density $\mu(\mathcal{L}) = \frac{1}{v}$. (Here $x_i b_i$ denotes the sum in G of $|x_i|$ copies of b_i if $x_i > 0$, or $-b_i$ if $x_i < 0$.)

(b) Conversely, if $\mathcal{L}' \subseteq \mathbb{Z}^q$ is a linear code with minimum distance $d_{\mathrm{a}}(\mathcal{L}') > h$, then the group $G = \mathbb{Z}^q/\mathcal{L}'$ contains a B_h set of cardinality $q + 1$ that generates G.

Proof: Both statements follow from Theorem 5 and the familiar group-theoretic formulation of lattice packing problems [44], [18], [45], [15], [20], [46], [47], so we only sketch the proof of *(a)*.
If $B = \{0, b_1, \ldots, b_q\}$ is a B_h set, then $(S_q(h, 0), \mathcal{L})$ is a packing in $\mathbb{Z}^q$, or, in coding-theoretic terminology, the "error-vectors" from $S_q(h, 0)$ are correctable and have different syndromes. Namely, we see from (15) that the syndromes are of the form $b_{i_1} + \cdots + b_{i_u}$, where $u \in \{0, 1, \ldots, h\}$ and $1 \le i_1 \le \cdots \le i_u \le q$, and the condition that they are all different is identical to the condition that $\{0, b_1, \ldots, b_q\}$ is a B_h set. This implies $d_{\mathrm{a}}(\mathcal{L}) > h$ (see Theorem 5). Furthermore, if B generates G, then G is isomorphic to $\mathbb{Z}^q/\mathcal{L}$, meaning that $\mu(\mathcal{L}) = \frac{1}{|G|}$. ∎

Example 1. The set $\{(0,0), (1,1), (0,5)\}$ is a B_3 set in the group $\mathbb{Z}_2 \times \mathbb{Z}_6$. The corresponding lattice packing $(S_2(2,1), \mathcal{L})$ in $\mathbb{Z}^2$ is illustrated in Figure 2. Notice that this is in fact a perfect packing, i.e., a tiling of the grid $\mathbb{Z}^2$. This means that $\mathcal{L}$ is a diameter-perfect linear code of minimum distance $d_{\mathrm{a}}(\mathcal{L}) = 4$, see Section III-D3 ahead. ▲

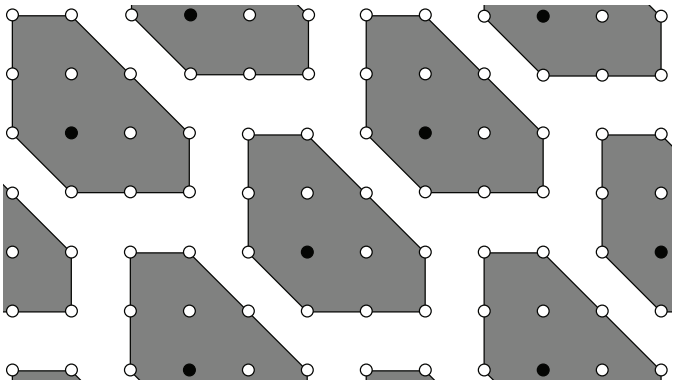


Fig. 2. Tiling of $\mathbb{Z}^2$ by the anticode $S_2(2,1)$.

The significance of Theorem 7 is twofold. First, using the known constructions of B_h sets one automatically obtains good codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$. This in particular gives a lower bound on the achievable density of codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$, i.e., the "achievability" counterpart of Theorem 6. For example, a construction of Bose and Chowla mentioned above asserts the existence of linear codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$ of minimum distance d and density $> (q+1)^{1-d}$, for any $d \ge 2$ and $q \ge 1$ with $q+1$ a prime power. Second, this geometric interpretation enables one to derive the best known bounds on the parameters of B_h sets. Namely, having in mind the correspondence between the density of the code and the size of the group containing a B_h set, between the minimum distance of the code and the parameter h, and between the dimension of the code and the cardinality of the B_h set, one can restate the inequalities from Theorem 6 in terms of the parameters of B_h sets. The resulting bounds are either equivalent to, or improve upon the known bounds[5]: The bound in (11) is equivalent to those in [22, Thm 2] and [8, Thm 2], but with an explicit error term, while the

[5]To our knowledge, the best known bounds for B_h sets in finite groups are stated in [22], [8].

bound in (12) improves upon that in [22, Thm 1(v)] by a factor of two. See [26] for an explicit statement of these bounds and their further improvements.

D. Perfect codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$

A code is said to be *r-perfect* if balls of radius r around the codewords are disjoint and cover the entire space. Perfect codes are the best possible codes having a given error-correction radius; it is therefore important to study their existence, and methods of construction when they do exist. Notice that, by Theorem 7, linear r-perfect codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$ correspond to B_{2r} sets of cardinality $q+1$ in Abelian groups of order $v = |S_q(r, r)|$.

1) 1-Perfect codes and planar difference sets: Linear 1-perfect codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$ correspond to B_2 sets of cardinality $q+1$ in Abelian groups of order $v = |S_q(1,1)| = q^2+q+1$. Such sets are better known in the literature as *planar* (or *simple*) *difference sets*. The condition that all the sums $b_i + b_j$ are different, up to the order of the summands, is equivalent to the condition that all the differences $b_i - b_j$, $i \neq j$, are different (hence the name), and the condition that the order of the group is $v = q^2+q+1$ means that *every* nonzero element of the group can be expressed as such a difference. If D is a planar difference set of size $q+1$, then q is referred to as the *order* of D. If G is Abelian, cyclic, etc., then D is also said to be Abelian, cyclic, etc., respectively.

Planar difference sets and their generalizations are extensively studied objects [6], and have also been applied in communications and coding theory in various settings, see for example [3], [30], [34], [13]. The following claim, which is a corollary to Theorem 7, states that these objects are essentially equivalent to linear 1-perfect codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$, and can be used to construct the latter via (15).

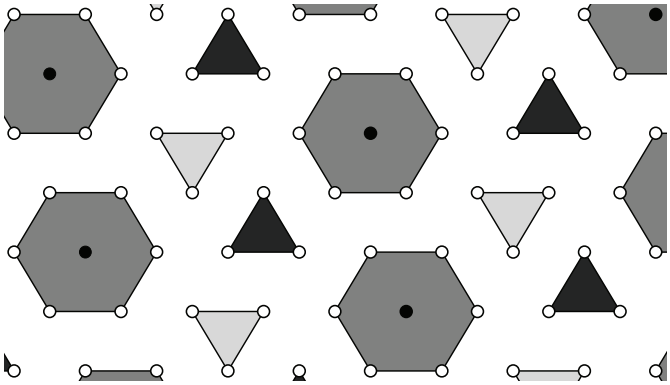
Theorem 8. *The space $(\mathbb{Z}^q, d_{\mathrm{a}})$ admits a linear 1-perfect code if and only if there exists an Abelian planar difference set of order q.* ■

Example 2. Consider a planar difference set $\{0, 1, 3, 9\}$ in the cyclic group $\mathbb{Z}_{13}$. The corresponding 1-perfect code in (A_3, d) is illustrated in Figure 3(a). The figure shows the intersection of A_3 with the plane $x_0 = 0$. Intersections of a ball of radius 1 in (A_3, d)—a cuboctahedron—with the planes $x_0 = \text{const}$ are shown in Figure 3(b) for clarification. ▲

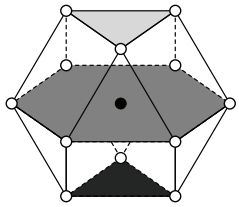
Abelian planar difference sets of prime power orders q were first constructed by Singer [41]. It is believed that this condition on q is in fact necessary, but this question—now known as the prime power conjecture [6, Conj. 7.5, p. 346]—remains open for nearly eight decades. By Theorem 8, the statement can be rephrased as follows:

Conjecture 1 (Prime power conjecture)**.** *There exists a linear 1-perfect code in $(\mathbb{Z}^q, d_{\mathrm{a}})$ if and only if the dimension q is a prime power.* ▲

2) r-perfect codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$: Theorem 8, together with a direct inspection of the one- and two-dimensional case (see also [10]), yields the following fact.



(a) The code viewed in the plane $x_0 = 0$.



(b) Intersections of a ball in (A_3, d) with the planes $x_0 = \text{const}$.

Fig. 3. 1-perfect code in (A_3, d).

Theorem 9. *There exists an r-perfect code in $(\mathbb{Z}^q, d_{\mathrm{a}})$ for:*

- *$q \in \{1, 2\}$, r arbitrary;*
- *$q \geq 3$ a prime power, $r = 1$.* ■

Proving (non-)existence of perfect codes for arbitrary pairs (q, r) seems to be a highly non-trivial problem[6]. We shall not be able to solve it here, but Theorem 10 below is a step in this direction.

For $S \subset \mathbb{Z}^q$, denote by S^{cub} the body in $\mathbb{R}^q$ defined as the union of unit cubes translated to the points of S, namely, $S^{\mathrm{cub}} = \bigcup_{\mathbf{y} \in S} (\mathbf{y} + [-1/2, 1/2]^q)$, and by S^{con} the convex hull in $\mathbb{R}^q$ of the points in S (see Figure 4).

Lemma 2. *Let $S_q(r) \equiv S_q(r, r)$ be the ball of radius r around $\mathbf{0}$ in $(\mathbb{Z}^q, d_{\mathrm{a}})$. The volumes of the bodies $S_q^{\mathrm{cub}}(r)$ and $S_q^{\mathrm{con}}(r)$ are given by*

$$\mathrm{Vol}(S_q^{\mathrm{cub}}(r)) = \sum_{j=0}^{\min\{q,r\}} \binom{q}{j} \binom{r}{j} \binom{r+q-j}{q-j} \tag{16}$$

$$\mathrm{Vol}(S_q^{\mathrm{con}}(r)) = \frac{r^q}{q!} \binom{2q}{q}. \tag{17}$$

Furthermore, $\lim_{r \to \infty} \mathrm{Vol}\left(S_q^{\mathrm{con}}(r)\right) / \mathrm{Vol}\left(S_q^{\mathrm{cub}}(r)\right) = 1$.

Proof: $S_q^{\mathrm{cub}}(r)$ consists of $|S_q(r)|$ unit cubes so (16) follows from Lemma 1.

To compute the volume of $S_q^{\mathrm{con}}(r)$, observe its intersection with the orthant $x_1, \ldots, x_j > 0$, $x_{j+1}, \ldots, x_q \leq 0$, where $j \in [0{:}q]$. The volume of this intersection is the product of the volumes of the j-simplex $\left\{(x_1, \ldots, x_j) : x_i > 0,\right.$

[6]It should also be contrasted with the well-known Golomb-Welch conjecture [16] (see also, e.g., [21]) stating that r-perfect codes in $\mathbb{Z}^q$ under the ℓ_1 metric exist only in the following cases: 1) $q \in \{1, 2\}$, r arbitrary, and 2) $r = 1$, q arbitrary.

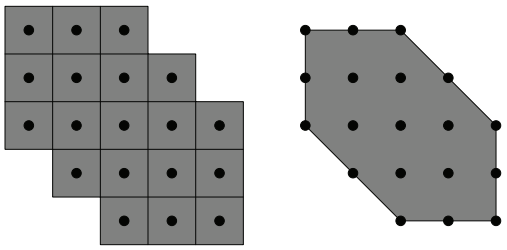


Fig. 4. Bodies in $\mathbb{R}^2$ corresponding to a ball of radius 2 in $(\mathbb{Z}^2, d_{\mathrm{a}})$: The cubical cluster $S_2^{\mathrm{cub}}(2)$ (left) and the convex hull $S_2^{\mathrm{con}}(2)$ (right).

$\sum_{i=1}^{j} x_i \leq r\}$, which is known to be $r^j/j!$, and of the $(q-j)$-simplex $\{(x_{j+1}, \ldots, x_q) : x_i \leq 0, \sum_{i=j+1}^{q} x_i \geq -r\}$, which is $r^{q-j}/(q-j)!$. This implies that $\mathrm{Vol}\left(S_q^{\mathrm{con}}(r)\right) = \sum_{j=0}^{q} \binom{q}{j} \frac{r^j}{j!} \frac{r^{q-j}}{(q-j)!}$, which reduces to (17) by using the identity $\sum_{j=0}^{q} \binom{q}{j}^2 = \binom{2q}{q}$.

Finally, when $r \to \infty$ we have $\binom{r}{j} \sim \frac{r^j}{j!}$ and so $\mathrm{Vol}\left(S_q^{\mathrm{cub}}(r)\right) \sim \frac{r^q}{q!}\binom{2q}{q} = \mathrm{Vol}\left(S_q^{\mathrm{con}}(r)\right)$. (Here $f(r) \sim g(r)$ stands for $\lim_{r\to\infty} f(r)/g(r) = 1$.) ∎

Theorem 10. *There are no r-perfect codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$, $q \geq 3$, for large enough r, i.e., for $r \geq r_0(q)$.*

Proof: The proof relies on the idea used to prove the corresponding statement for r-perfect codes in $\mathbb{Z}^q$ under ℓ_1 distance [16]. First observe that an r-perfect code in $(\mathbb{Z}^q, d_{\mathrm{a}})$ would induce a tiling of $\mathbb{R}^q$ by $S_q^{\mathrm{cub}}(r)$, and a *packing* by $S_q^{\mathrm{con}}(r)$. The relative efficiency of the latter with respect to the former is defined as the ratio of the volumes of these bodies, $\mathrm{Vol}\left(S_q^{\mathrm{con}}(r)\right) / \mathrm{Vol}\left(S_q^{\mathrm{cub}}(r)\right)$, which by Lemma 2 converges to 1 as r tends to infinity. This has the following consequence: If an r-perfect code exists in $(\mathbb{Z}^q, d_{\mathrm{a}})$ for infinitely many r, then there exists a tiling of $\mathbb{R}^q$ by translates of $S_q^{\mathrm{cub}}(r)$ for infinitely many r, which further implies that a packing of $\mathbb{R}^q$ by translates of $S_q^{\mathrm{con}}(r)$ exists which has efficiency arbitrarily close to 1. But then there would also be a packing by $S_q^{\mathrm{con}}(r)$ of efficiency 1, i.e., a tiling (in [16, Appendix] it is shown that there exists a packing whose density is the supremum of the densities of all possible packings with a given body). This is a contradiction. Namely, it is known [36, Thm 1] that a necessary condition for a convex body to be able to tile space is that it be a polytope with centrally symmetric[7] facets, which $S_q^{\mathrm{con}}(r)$ fails to satisfy for $q \geq 3$. For example, the facet which is the intersection of $S_q^{\mathrm{con}}(r)$ with the hyperplane $x_1 = -r$ is the simplex $\{(x_2, \ldots, x_q) : x_i \geq 0, \sum_{i=2}^{q} x_i \leq r\}$, a non-centrally-symmetric body. ∎

3) Diameter-perfect codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$: The following generalization of a notion of perfect code, adjusted to our setting, was introduced in [1]. We say that a code $\mathcal{C} \subseteq \mathbb{Z}^q$ of minimum distance $d_{\mathrm{a}}(\mathcal{C})$ is *diameter-perfect* if there exists an anticode $S \subset \mathbb{Z}^q$ of diameter $d_{\mathrm{a}}(\mathcal{C}) - 1$ such that $\mu(\mathcal{C}) \cdot |S| = 1$. Namely, by the arguments from [11], [1], we know that for any such code-anticode pair, we must have $\mu(\mathcal{C}) \cdot |S| \leq 1$ (see also Theorems 5 and 6). Therefore, a code is said to be diameter-perfect if it achieves this bound. This notion is especially interesting when the minimum distance of a code is even, which can never be the case for perfect codes.

[7] A polytope $P \subset \mathbb{R}^q$ is centrally symmetric if its translation $\tilde{P} = P - \mathbf{c}$ satisfies $\tilde{P} = -\tilde{P}$ for some $\mathbf{c} \in \mathbb{R}^q$.

Theorem 11. *There exists a diameter-perfect code of minimum distance $2r$ in $(\mathbb{Z}^q, d_{\mathrm{a}})$ for:*

- *$q \in \{1, 2\}$, r arbitrary;*
- *$q \geq 3$, $r = 1$.*

Proof. We show that, in all the stated cases, there exists a lattice tiling $(S_q(r, r-1), \mathcal{L})$, $\mathcal{L} \subseteq \mathbb{Z}^q$. This will prove the claim because $S_q(r, r-1)$ is an anticode of diameter $2r-1$. Dimension $q = 1$ is trivial. In dimension $q = 2$, one can check directly that the lattice generated by the vectors (r, r) and $(0, 3r)$ defines a tiling by $S_2(r, r-1)$ for any $r \geq 1$, see Figure 2. It can also be shown that this lattice is unique—there are no other diameter-perfect codes of minimum distance $2r$ in $(\mathbb{Z}^2, d_{\mathrm{a}})$. The statement for $r = 1$ is left. By Theorem 7, a lattice tiling $(S_q(1, 0), \mathcal{L})$, $\mathcal{L} \subseteq \mathbb{Z}^q$, exists if and only if a B_1 set exists in some Abelian group G of order $|S_q(1, 0)| = q+1$. Notice that any G is itself such a set. □

In dimensions $q \geq 3$, one can show in a way analogous to Theorem 10 that tilings of $\mathbb{Z}^q$ by the anticodes $S_q(r, r-1)$ do not exist for r large enough.

IV. Multiset codes: Construction and bounds

In this section we describe a construction of multiset codes, i.e., codes in the simplex $\triangle_n^q$, inspired by the corresponding construction of codes in the A_q lattice (Section III-C). We then derive bounds on the cardinalities of optimal multiset codes and examine their asymptotic behavior in several regimes of interest. These bounds will, in particular, demonstrate optimality of the presented construction for some sets of parameters.

A. Construction based on Sidon sets

The construction given next is inspired by the observation in Theorem 7, which states that linear codes in A_q lattices are essentially equivalent to Sidon sets in Abelian groups.

Let $B = \{b_0, b_1, \ldots, b_q\} \subseteq G$ and $b \in G$, where G is an Abelian group. Define

$$\mathcal{C}_n^{(G,B,b)} = \left\{ \mathbf{x} \in \triangle_n^q : \sum_{i=0}^{q} x_i b_i = b \right\}. \tag{18}$$

Theorem 12. *If B is a B_h set, then the code $\mathcal{C}_n^{(G,B,b)}$ can correct h deletions.*

In other words, if B is a B_h set and $\left|\mathcal{C}_n^{(G,B,b)}\right| \geq 2$, then $d\left(\mathcal{C}_n^{(G,B,b)}\right) > h$ (see Theorem 2).

Proof: Suppose that $\mathcal{C}_n^{(G,B,b)}$ cannot correct h deletions, i.e., there exist two different codewords $\mathbf{x}, \mathbf{y}$ and two different vectors $\mathbf{f}, \mathbf{g}$ such that $f_i, g_i \geq 0$, $\sum_{i=0}^{q} f_i = \sum_{i=0}^{q} g_i = h$, and $\mathbf{x} - \mathbf{f} = \mathbf{y} - \mathbf{g}$. This implies that $\sum_{i=0}^{q}(x_i - f_i)b_i = \sum_{i=0}^{q}(y_i - g_i)b_i$ and, since $\sum_{i=0}^{q} x_i b_i = \sum_{i=0}^{q} y_i b_i = b$, we get $\sum_{i=0}^{q} f_i b_i = \sum_{i=0}^{q} g_i b_i$. This means that $\{b_0, b_1, \ldots, b_q\}$ is not a B_h set. ∎

Theorem 13. *Let $\mathcal{C} = (\mathcal{L}+\mathbf{t}) \cap \triangle_n^q$ be a linear multiset code of minimum distance $d(\mathcal{C}) > h$, where $\mathbf{t}$ satisfies $t_i \geq h$ for all $i \in [0{:}q]$ (and hence $n \geq h(q+1)$). Then $\mathcal{C}$ is necessarily of the form* (18) *for some B_h set $\{b_0, b_1, \ldots, b_q\}$.*

Proof: Let $\mathcal{C} = (\mathcal{L}+\mathbf{t}) \cap \triangle_n^q$ be a linear code of minimum distance $d(\mathcal{C}) > h$. Notice that $\mathbf{t} \in \mathcal{C}$ since $\mathbf{0} \in \mathcal{L}$. If $\mathbf{t}$ satisfies the condition $t_i \geq h$ then the entire ball of radius h around $\mathbf{t}$ (regarded in $A_q + \mathbf{t}$) belongs to $\triangle_n^q$, i.e., all the points in this ball have non-negative coordinates. Moreover, since the code $\mathcal{C}$ has distance $> h$, this ball does not contain another codeword of $\mathcal{C}$. These two facts imply that $\mathcal{L}$ is a linear code of minimum distance $d(\mathcal{L}) > h$ in A_q. The claim then follows by invoking Theorem 7 which states that any such code is of the form $\left\{\mathbf{x} \in A_q : \sum_{i=0}^{q} x_i b_i = 0\right\}$, where $\{b_0, b_1, \ldots, b_q\}$ is a B_h set in the quotient group $A_q/\mathcal{L}$. ∎

Remark 2. Suppose $\{\mathcal{C}'_n\}_n$ is a family of linear multiset codes obtained from a family of lattices $\{\mathcal{L}'_n\}_n$, $\mathcal{L}'_n \subseteq A_q$, where n denotes the code block-length. If the density of $\mathcal{L}'_n$ is bounded, meaning that $\mu(\mathcal{L}'_n) = \mathcal{O}(1)$ when $n \to \infty$, then every code $\mathcal{C}'_n$ of sufficiently large block-length ($n \geq n_0$) will necessarily contain a codeword $\mathbf{t}$ satisfying $t_i \geq d(\mathcal{C}'_n) - 1$, and will by Theorem 13 be of the form (18). ▲

B. Bounds and asymptotics: Fixed alphabet case

Let $M_{q+1}(n,h)$ denote the cardinality of the largest multiset code of block-length n over the alphabet $[0{:}q]$ which can correct h deletions (or, equivalently, which has minimum distance $h+1$), and $M^{\text{L}}_{q+1}(n,h)$ the cardinality of the largest *linear* multiset code with the same parameters. We shall assume in the following that $n > h$; this condition is necessary and sufficient for the existence of nontrivial codes of distance $> h$, i.e., codes with at least two codewords.

When discussing asymptotics, the following conventions will be used: $f(n) \sim g(n)$ means $\lim_{n\to\infty} f(n)/g(n) = 1$, and $f(n) \gtrsim g(n)$ means $\limsup_{n\to\infty} f(n)/g(n) \geq 1$.

For notational convenience, denote by $\beta(h,q)$ the size of the anticode $S_q\left(\lceil \frac{h}{2} \rceil, \lfloor \frac{h}{2} \rfloor\right) \subset \mathbb{Z}^q$ of diameter h (see (8)). That is,

$$\beta(h,q) = \sum_{j=0}^{\min\{q, \lceil \frac{h}{2} \rceil\}} \binom{q}{j} \binom{\lceil \frac{h}{2} \rceil}{j} \binom{\lfloor \frac{h}{2} \rfloor + q - j}{q-j}. \tag{19}$$

Let also $\phi(h,q)$ denote the order of the smallest Abelian group containing a B_h set of cardinality $q+1$. The lower bounds that follow will be expressed in terms of this quantity; more explicit lower bounds stated in terms of the parameters h, q can be obtained from the known upper bounds on $\phi(h,q)$ [41], [7], [22], [26].

Theorem 14. *For every $q \geq 1$ and $n > h \geq 1$,*

$$M_{q+1}(n,h) \geq \frac{\binom{n+q}{q}}{\phi(h,q)}, \tag{20}$$

$$M_{q+1}(n,h) \leq \frac{\binom{n+q}{q}}{\beta(h,q)} + \sum_{j=1}^{(q+1)\lceil \frac{h}{2} \rceil} \binom{n+q-j}{q-1}. \tag{21}$$

Proof: The lower bound in (20) follows from the construction described in the previous subsection. Fix n, an Abelian group G, and a B_h set $B \subseteq G$ with $|B| = q+1$. Then by Theorem 12 the codes $\mathcal{C}_n^{(G,B,b)}$ can correct h deletions. Furthermore, since they form a partition of $\triangle_n^q$, and since there are $|G|$ of them (one for each $b \in G$), at least one has cardinality $\geq |\triangle_n^q|/|G|$. To get the tightest bound take G to be the smallest group containing a B_h set with $q+1$ elements, i.e., $|G| = \phi(h,q)$.

The upper bound in (21) follows from an argument similar to the one that led to the code-anticode bound (10) in Theorem 6. The difficulty in directly applying that argument to codes in $\triangle_n^q$ is that, if a codeword $\mathbf{x}$ is too close to the "boundary" of $\triangle_n^q$, then the corresponding anticode around $\mathbf{x}$ will be "clipped" (see Figure 1) and will have cardinality smaller than $\beta(h,q)$. The vectors $\mathbf{x} \in \triangle_n^q$ for which this is not true, i.e., anticodes around which have cardinality $\beta(h,q)$, are those satisfying $x_i \geq \lceil \frac{h}{2} \rceil$ for all $i \in [0{:}q]$. The set of such sequences can be written as $\left(\lceil \frac{h}{2} \rceil, \ldots, \lceil \frac{h}{2} \rceil\right) + \triangle_{n'}^q$, where $n' = n - (q+1)\lceil \frac{h}{2} \rceil$, and is of cardinality $|\triangle_{n'}^q|$. Now, write $M_{q+1}(n,h) = M' + M''$, where M' is the number of codewords of an optimal code that belong to $\left(\lceil \frac{h}{2} \rceil, \ldots, \lceil \frac{h}{2} \rceil\right) + \triangle_{n'}^q$, and M'' the number of remaining codewords. By the code-anticode argument leading to (10), we have $M' \cdot \beta(h,q) \leq |\triangle_n^q|$, which gives the first summand in the upper bound (21). The second summand is the size of the remaining part of the simplex, $|\triangle_n^q| - |\triangle_{n'}^q| = \sum_{j=1}^{n-n'} \binom{n+q-j}{q-1}$, which is certainly an upper bound on M''. ∎

In the asymptotic case, as the block-length grows to infinity, we get the following bounds.

Theorem 15. *For every fixed $q \geq 1$ and $h \geq 1$, as $n \to \infty$,*

$$\frac{n^q}{q!\ \beta(h,q)} \gtrsim M_{q+1}(n,h) \gtrsim \frac{n^q}{q!\ \phi(h,q)} \tag{22}$$

$$M^{\text{L}}_{q+1}(n,h) \sim \frac{n^q}{q!\ \phi(h,q)}. \tag{23}$$

Proof: (22) follows from (20) and (21) by noting that $\binom{n+q}{q} \sim \frac{n^q}{q!}$, and that the second summand in (21) is of the order $\mathcal{O}(n^{q-1})$.

Theorem 7 implies that the largest density a sublattice of A_q with minimum distance $> h$ can have is $1/\phi(h,q)$. Since the dimension q and the minimum distance $h+1$ are fixed, and the size of the simplex grows indefinitely (as $n \to \infty$), it follows that $M^{\text{L}}_{q+1}(n,h) \sim |\triangle_n^q|/\phi(h,q)$, which gives (23). ∎

We see from Theorem 15 that the cardinality of optimal multiset codes over a fixed alphabet scales as $\Theta(n^q)$. Therefore, at most polynomially many codewords are available to the transmitter in this setting, as opposed to exponentially many codewords available in standard models where the transmitter sends *sequences* of symbols. This is the price paid for storing/transmitting information in an unordered way. Our main contribution in Theorem 15, however, is to establish bounds on the implied constant in the $\Theta(n^q)$ term for general multiset codes and the exact implied constant in the same term for linear multiset codes. These constants depend explicitly on the number of deletions h and the alphabet size $q+1$.

Remark 3. The bounds stated in Theorem 15 remain valid in the regime $n \to \infty$, $h \to \infty$, as long as h grows slower than n, i.e., $h = o(n)$. ▲

The following claim states that the construction described in Section IV-A produces asymptotically optimal multiset codes over binary and ternary alphabets for arbitrary minimum distance, and over arbitrary alphabets for small distances. These codes are in fact *asymptotically (diameter-) perfect*, meaning that they asymptotically achieve the upper bound (21) (they are also *perfect* over a binary alphabet, and in some special cases over a ternary alphabet [25]).

Corollary 16. *The following statements hold for multiset codes over the alphabet $[0{:}q]$, in the limit $n \to \infty$.*

- *For a binary alphabet ($q = 1$) and any $h \geq 1$,*

$$M_2(n,h) \sim M_2^{\text{L}}(n,h) \sim \frac{n}{h+1}. \tag{24}$$

- *For a ternary alphabet ($q = 2$) and any $r \geq 1$,*

$$M_3(n,2r) \sim M_3^{\text{L}}(n,2r) \sim \frac{n^2}{2(3r^2+3r+1)}, \tag{25}$$

$$M_3(n,2r-1) \sim M_3^{\text{L}}(n,2r-1) \sim \frac{n^2}{6r^2}. \tag{26}$$

- *For an arbitrary alphabet ($q \geq 1$) and $h = 1$,*

$$M_{q+1}(n,1) \sim M_{q+1}^{\text{L}}(n,1) \sim \frac{n^q}{(q+1)!}. \tag{27}$$

- *For a prime power $q \geq 1$, and $h = 2$,*

$$M_{q+1}(n,2) \sim M_{q+1}^{\text{L}}(n,2) \sim \frac{n^q}{q!\ (q^2+q+1)}. \tag{28}$$

Proof: In all the stated cases there exist lattice tilings $\left(S_q\left(\lceil \frac{h}{2} \rceil, \lfloor \frac{h}{2} \rfloor\right), \mathcal{L}\right)$ of $\mathbb{Z}^q$, implying that $\phi(h,q) = \beta(h,q)$ (these are the diameter-perfect codes of Theorems 9 and 11). The claim then follows from Theorem 15 after plugging in the expressions for $\beta(h,q)$ in these particular cases. ∎

C. Bounds and asymptotics: Growing alphabet case

We now discuss the case when the input alphabet is not necessarily fixed. In particular, we consider the regime when the size of the alphabet is a linear function of the block-length n, namely $q+1 = \lfloor \tilde{q}n \rfloor$ for an arbitrary positive real constant $\tilde{q}$. As we shall point out in Section V-A, this regime is well-motivated by the standard way of dealing with symbol reordering in networking applications.

The upper bound in (21) is useless in this regime, so we derive in Theorem 17 another bound appropriate for this case. The method we use is similar to [32], though the setting is quite different. Before stating the theorem, we give two auxiliary facts that will be needed in the proof.

Lemma 3. *Let $\mathbf{x} \in \triangle_n^q$ be a vector with i non-zero coordinates. The set of all vectors that can be obtained after $\mathbf{x}$ is impaired with r deletions and $h-r$ insertions*[8] *has at least $\binom{i}{r}\binom{q+h-2r}{h-r}$ elements.*

Proof: Subtract 1 from r of the positive coordinates of $\mathbf{x}$, and distribute a mass of $h-r$ over the remaining $q+1-r$ coordinates. The former can be done in $\binom{i}{r}$ ways, and the latter in $\binom{q+h-2r}{h-r}$. ∎

[8]Recall that we are always referring to deletions and insertions on the *multisets* represented by vectors from $\triangle_n^q$, not on the vectors themselves.

There are $\binom{q+1}{i}\binom{n-1}{i-1}$ vectors in $\triangle_n^q$ with exactly i non-zero coordinates. Consequently,

$$\sum_{i=1}^{\min\{q+1,n\}} \binom{q+1}{i}\binom{n-1}{i-1} = |\triangle_q^n| = \binom{n+q}{q}. \tag{29}$$

Theorem 17. *Fix $q \geq 1$ and $n > h \geq 1$. The following inequality is valid for all integers $r \in \{0,1,\ldots,h\}$ and $l \in \{r,\ldots,q+1\}$:*

$$M_{q+1}(n,h) \leq \frac{\binom{n+h-2r+q}{q}}{\binom{l}{r}\binom{q+h-2r}{h-r}} + \sum_{i=1}^{l-1}\binom{q+1}{i}\binom{n-1}{i-1}. \tag{30}$$

In particular, for $r = 0$, $l = 1$, this simplifies to

$$M_{q+1}(n,h) \leq \frac{\binom{n+h+q}{q}}{\binom{q+h}{h}}. \tag{31}$$

Proof: Let $\mathcal{C} \subseteq \triangle_n^q$ be an optimal multiset code correcting h deletions, i.e., $|\mathcal{C}| = M_{q+1}(n,h)$. Write $M_{q+1}(n,h) = M' + M''$, where M' is the number of codewords of $\mathcal{C}$ having at least l non-zero coordinates, and M'' the number of remaining codewords. Recall from Theorem 1 that $\mathcal{C}$ corrects h deletions if and only if it corrects r deletions and $h-r$ insertions. This implies that sets of outputs obtained by deleting r and inserting $h-r$ symbols to each of the codewords, are disjoint. Note that these outputs live in $\triangle_{n+h-2r}^q$, which, together with Lemma 3, implies that $M' \cdot \binom{l}{r}\binom{q+h-2r}{h-r} \leq \left|\triangle_{n+h-2r}^q\right| = \binom{n+h-2r+q}{q}$. This gives the first summand in the upper bound (30). The second summand is simply the cardinality of the set of all vectors in $\triangle_n^q$ having less than l non-zero coordinates, which is certainly an upper bound on M''. ∎

For a real $\tilde{q} > 0$ and an integer $h \geq 1$, define

$$c(h,\tilde{q}) = \min_{0 \leq r \leq h} (h-r)!r!(1+\tilde{q})^r. \tag{32}$$

To simplify the following expressions, we shall write $\tilde{q}n$ instead of $\lfloor \tilde{q}n \rfloor$ for the alphabet size, ignoring the fact that the former need not be an integer.

Theorem 18. *For any real $\tilde{q} > 0$ and any integer $h \geq 1$, as $n \to \infty$,*

$$\frac{\binom{n+\tilde{q}n}{\tilde{q}n}}{\tilde{q}^h n^h} \lesssim M_{\tilde{q}n}(n,h) \lesssim c(h,\tilde{q})\frac{\binom{n+\tilde{q}n}{\tilde{q}n}}{\tilde{q}^h n^h}, \tag{33}$$

where

$$\binom{n+\tilde{q}n}{\tilde{q}n} \sim \frac{2^{n(1+\tilde{q})H\left(\frac{1}{1+\tilde{q}}\right)}}{\sqrt{2\pi\frac{\tilde{q}}{1+\tilde{q}}n}}, \tag{34}$$

and $H(\cdot)$ denotes the binary entropy function.

Proof: The expression in (34) follows from Stirling's approximation for the factorial.

It was proven in [7] that $\phi(h,q) < (q+1)^h$ when $q+1$ is a prime power, implying that $\phi(h,\tilde{q}n) \lesssim \tilde{q}^h n^h$. This, together with (20), gives the lower bound in (33).

The upper bound in (33) is obtained from (30) after choosing l appropriately. The idea is to set l large enough so that the first summand on the right-hand side of (30) is minimized, but small enough so that the second summand is still negligible compared to the first one. Observe the relation (29) and the

second summand on the right-hand side of (30). From (29) we see that, as $n \to \infty$ and $q = \tilde{q}n$, the sum $\sum_i \binom{q+1}{i}\binom{n-1}{i-1}$ grows exponentially in n with exponent $(1+\tilde{q})H\big(\frac{1}{1+\tilde{q}}\big)$ (see (34)), and since it has linearly many summands, there must exist $\lambda \in (0,1)$ such that $\binom{q+1}{\lambda n}\binom{n-1}{\lambda n}$ grows exponentially in n with the same exponent. By using Stirling's approximation, one can find the exponent of $\binom{q+1}{\lambda n}\binom{n-1}{\lambda n}$ as a function of λ, and check by differentiation that it is maximized for unique $\lambda = \lambda^* = \frac{\tilde{q}}{1+\tilde{q}}$. Informally speaking, the term $\binom{q+1}{\lambda^* n}\binom{n-1}{\lambda^* n}$ in the sum (29), and the terms "immediately around it", account for most of the space $\triangle_n^q$, and the remaining terms are negligible. More precisely, there exists a sublinear function $f(n) = o(n)$ such that $\sum_{i=1}^{\lambda^* n - f(n)} \binom{q+1}{i}\binom{n-1}{i-1} = o\big(n^{-h}\binom{n+q}{q}\big)$ [43]. Therefore, if we set $l = \lambda^* n - f(n)$, the second summand on the right-hand side of (30) will be asymptotically negligible compared to the first one, and the first summand will give precisely the upper bound in (33). ∎

In particular, the asymptotic expression for the cardinality of optimal single-deletion-correcting multiset codes has the following form.

Corollary 19. *For every $\tilde{q} > 0$ and $h = 1$, as $n \to \infty$,*

$$M_{\tilde{q}n}(n,1) \sim \frac{2^{n(1+\tilde{q})H\left(\frac{1}{1+\tilde{q}}\right)}}{\sqrt{2\pi\frac{\tilde{q}^3}{1+\tilde{q}}}\; n^{\frac{3}{2}}}. \tag{35}$$

Proof: Follows from the bounds in (33), and the fact that $c(1,\tilde{q}) = 1$. ∎

On the logarithmic scale, we obtain from Theorem 18:

$$\frac{1}{n}\log_2 M_{\tilde{q}n}(n,h) = (1+\tilde{q})H\Big(\frac{1}{1+\tilde{q}}\Big) - \Big(h+\frac{1}{2}\Big)\frac{\log_2 n}{n} + \mathcal{O}\Big(\frac{1}{n}\Big). \tag{36}$$

Hence, the largest rate achievable asymptotically by multiset codes of block-length n over an alphabet of size $\tilde{q}n$ is $(1+\tilde{q})H\big(\frac{1}{1+\tilde{q}}\big)$ bits per symbol. This is the capacity of the *noiseless* multiset channel described in Section I-A. The back-off from capacity at finite block-lengths scales as $\frac{1}{2}\frac{\log_2 n}{n} + \mathcal{O}\big(\frac{1}{n}\big)$ in the noiseless case, while an additional redundancy of $h\frac{\log_2 n}{n}$ is necessary if the code is required to have the capability of correcting h deletions. Note that the cardinality of codes produced by the construction based on Sidon sets has the same asymptotic expansion (36). In other words, the constructed codes are optimal in the sense of minimal asymptotic code redundancy, for any h and any $\tilde{q}$.

D. Connections to classical binary insertion/deletion channels

Codes in the simplex $\triangle_n^q$ are relevant not only for permutation channels and unordered data storage, but also for classical binary channels. We shall omit the detailed description of this connection, as it can be found in relevant references; instead, we only provide a brief comparison between the results in these references and the results obtained here.

1) Deletion and repetition channel with constrained inputs: The classical binary deletion channel with inputs constrained in such a way that all of them have the same number of runs of identical consecutive symbols, and that each run is of length at least r, can be reduced to the multiset channel treated here via run-length coding [42], [5]. More precisely, codes correcting deletions and repetitions of binary symbols can be equivalently described in the metric space $(\triangle_n^q, d)$, for appropriately defined n and q (these parameters have an entirely different meaning in this setting from that in ours). In [42], [5], the authors provide constructions and derive bounds on the cardinality of optimal codes in $(\triangle_n^q, d)$ having a specified minimum distance. Furthermore, the asymptotic regimes studied there correspond exactly to those we discussed in the previous two subsections.

We wish to point out that the bounds obtained in this paper, Theorems 15 and 18 in particular, are strictly better that the ones in [42], [5]. The main reason for this is that the authors in [42], [5] derive bounds on codes in $\triangle_n^q$ under ℓ_1 distance via packings in $\mathbb{Z}^q$ under ℓ_1 distance. However, as we have shown in Theorem 4, the ℓ_1 distance on the simplex corresponds to a different metric on $\mathbb{Z}^q$, namely d_{a}. The latter observation, together with the fact that optimal linear codes in $(\mathbb{Z}^q, d_{\mathrm{a}})$ can be constructed via Sidon sets (Theorem 7), enables one to derive much better bounds.

2) Channel with deletions of zeros: Restricting the inputs of a binary channel to sequences of the same Hamming weight, and describing such sequences by their *runs of zeros*, one again obtains $\triangle_n^q$ as the relevant code space. This representation of binary sequences is appropriate for the binary deletion channel in which only 0's can be deleted; see Levenshtein's work [32]. While [32] does not discuss the constant-weight case, but rather the binary channel with no constraints on inputs, the methods of analysis are similar, at least in some asymptotic regimes. For example, the construction via Sidon sets was given also in [32], and, when appropriately modified for the constant-weight case, implies the lower bound in Theorem 18. (Levenshtein was unaware of [7] and the construction of Sidon sets therein. Consequently, he stated in [32] a worse lower bound for the 0-deletion-correcting codes than his construction actually implies.)

It is worth noting that the upper bound we have derived in the previous subsection improves on that in [32]. In particular, there is no need to distinguish between the cases of odd and even h, as was done in [32].

V. Other constructions of multiset codes

In this section we describe two additional constructions of multiset codes. Both of these constructions are asymptotically suboptimal and result in codes of smaller cardinality compared with the construction based on Sidon sets, but are of interest nonetheless.

A. Construction based on sequence number prefixes

In networking applications, particularly those employing multipath routing, the problem of packet reordering is usually solved by supplying each packet with a sequence number in its header [29]. If up to n packets are being sent in one "generation", a sequence number will take up $\lceil \log_2 n \rceil$ bits. Therefore, if the original packets are of length m bits each, i.e., the cardinality of the channel alphabet is $\tilde{q} = 2^m$, then the "new"

packets with prepended sequence numbers will be of length $m + \lceil \log_2 n \rceil$ bits. Notice that by adding sequence number prefixes, we are actually changing the channel alphabet—the new alphabet is the set of all packets of length $m + \lceil \log_2 n \rceil$, and its cardinality is $2^{m+\lceil \log_2 n \rceil} \approx \tilde{q}n$.

Furthermore, in order to protect the packets from other types of noise, a classical code of length n in the $\tilde{q}$-ary Hamming space may be used [24]. To clarify what is meant here, we are assuming that: 1) a sequence of information packets to be transmitted, $(s_1, \ldots, s_n)$, is a codeword of a code $\mathcal{C}_{\text{H}}$ of length n over a $\tilde{q}$-ary alphabet having minimum Hamming distance $> h$, and 2) to each symbol/packet of this codeword we then prepend a sequence number indicating its position in the codeword, i.e., the sequence actually transmitted is $(u_1, \ldots, u_n)$, where $u_i = i \circ s_i$ ('$\circ$' denotes concatenation). Notice that the order in which u_i's are transmitted is irrelevant because it can easily be recovered from the sequence numbers. In other words, each codeword $(u_1, \ldots, u_n)$ obtained in the above-described way can be thought of as a multiset $\{u_1, \ldots, u_n\}$. Therefore, the resulting code $\mathcal{C}$ can in fact be seen as a multiset code over an alphabet of size $\tilde{q}n$, but a special case thereof in which no codeword contains two identical packets (as each of the n packets has a different sequence number prefix). Furthermore, the fact that $\mathcal{C}_{\text{H}}$ has minimum Hamming distance $> h$ implies that $\mathcal{C}$ can correct h packet deletions.

To compare this construction with the one given in Section IV-A, note that the size of the optimal code that can be obtained in this way, denoted $M^{\text{seq}}_{\tilde{q}n}(n, h)$, cannot exceed the sphere packing bound in the $\tilde{q}$-ary Hamming space:

$$M^{\text{seq}}_{\tilde{q}n}(n, h) \leq \frac{\tilde{q}^n}{\sum_{j=0}^{\lfloor \frac{h}{2} \rfloor} \binom{n}{j} (\tilde{q}-1)^j} \sim \frac{\lfloor \frac{h}{2} \rfloor! \ \tilde{q}^n}{(\tilde{q}-1)^{\lfloor \frac{h}{2} \rfloor} \ n^{\lfloor \frac{h}{2} \rfloor}}. \quad (37)$$

From Theorem 18 we then get, for any $\tilde{q} \geq 2$ and $h \geq 1$, as $n \to \infty$,

$$\begin{aligned} \frac{M_{\tilde{q}n}(n, h)}{M^{\text{seq}}_{\tilde{q}n}(n, h)} &\gtrsim \frac{(\tilde{q}-1)^{\lfloor \frac{h}{2} \rfloor} \ 2^{n(1+\tilde{q}) \log_2(1+\tilde{q}^{-1})}}{\lfloor \frac{h}{2} \rfloor! \sqrt{2\pi \frac{\tilde{q}}{1+\tilde{q}}} \ \tilde{q}^h \ n^{\lceil \frac{h}{2} \rceil + \frac{1}{2}}} \quad (38) \\ &= 2^{n(1+\tilde{q}) \log_2(1+\tilde{q}^{-1}) + o(n)} \end{aligned}$$

and hence

$$\begin{aligned} \frac{1}{n} \log_2 M_{\tilde{q}n}(n, h) - \frac{1}{n} \log_2 M^{\text{seq}}_{\tilde{q}n}(n, h) \gtrsim \\ (1+\tilde{q}) \log_2(1+\tilde{q}^{-1}). \end{aligned} \quad (39)$$

In words, the asymptotic rate achievable by multiset codes based on sequence number prefixes is strictly smaller than the corresponding rate achievable by general multiset codes. The lower bound on the difference, $(1+\tilde{q}) \log_2(1+\tilde{q}^{-1})$, is a monotonically decreasing function of $\tilde{q}$, and hence also of the length of information packets $m = \log_2 \tilde{q}$. Thus, the savings (in terms of rate) obtained by using optimal multiset codes instead of the ones based on sequence numbers are greater for large block-lengths and small alphabets.

B. Construction based on polynomial roots

Consider a polynomial

$$s(x) = x^n + s_{n-1} x^{n-1} + \cdots + s_1 x + s_0, \quad (40)$$

with coefficients s_i drawn from a finite field $\mathbb{F}_{p^m}$ ($p \geq 2$ is a prime and $m \geq 1$ an integer). Each such polynomial has n (not necessarily distinct) roots, $u_1, \ldots, u_n$, which are elements of the extended field $\mathbb{F}_{p^{mn}}$. The coefficients $(s_0, s_1, \ldots, s_{n-1})$ can always be recovered from the roots, e.g., by using Vieta's formulas:

$$s_{n-k} = (-1)^k \sum_{1 \leq i_1 < \cdots < i_k \leq n} u_{i_1} \cdots u_{i_k}, \quad (41)$$

for $k = 1, \ldots, n$. Therefore, *multisets* of roots $\{u_1, \ldots, u_n\}$ are in one-to-one correspondence with sequences of coefficients $(s_0, s_1, \ldots, s_{n-1})$, and the mapping:

$$(s_0, s_1, \ldots, s_{n-1}) \mapsto \{u_1, \ldots, u_n\}$$

defines a multiset code of length n over an alphabet of size p^{mn}. In channel coding parlance, $(s_0, s_1, \ldots, s_{n-1})$ is an information sequence, and $\{u_1, \ldots, u_n\}$ is the corresponding codeword to be transmitted.

Furthermore, we can extend this construction to obtain a code capable of correcting $h \geq 1$ deletions, which in the present terminology means that the coefficients of the "information polynomial" (40) can be recovered from any $n-h$ of the n transmitted roots. To do this, fix h coefficients beforehand, say $s_{n-1} = \cdots = s_{n-h} = 0$. We see from (41) that the remaining coefficients can indeed be recovered uniquely from any $n-h$ roots. In other words, the information sequence is now $(s_0, s_1, \ldots, s_{n-h-1})$, and the mapping:

$$(s_0, s_1, \ldots, s_{n-h-1}, 0, \ldots, 0) \mapsto \{u_1, \ldots, u_n\}, \quad s_i \in \mathbb{F}_{p^m},$$

defines a multiset code with the following properties: length n, minimum distance $> h$, and cardinality $p^{m(n-h)}$. The code is defined over an alphabet of size p^{mn} ($u_i \in \mathbb{F}_{p^{mn}}$).

VI. Concluding remarks and further work

We have described a coding-theoretic framework for a communication setting where information is being transmitted in the form of multisets over a given finite alphabet. General statements about the error correction capability of multiset codes have been obtained, constructions of such codes described, and bounds on the size of optimal codes derived. Furthermore, the *exact* asymptotic behavior of the cardinality of optimal codes has been obtained in various cases.

As we have shown, the study of multiset codes over a fixed alphabet reduces to the study of codes in A_q lattices, at least in the large block-length limit. In connection to this, there are several natural directions of further work on this topic:

- Improve the bounds on the density of optimal codes in (A_q, d) having a given minimum distance.
- Investigate whether the construction of codes in (A_q, d) via Sidon sets optimal, or it is possible to achieve larger densities with non-linear codes.
- Demonstrate (non-)existence of diameter-perfect codes in (A_q, d) with parameters different from those stated in Theorems 9 and 11.

We have also argued that the case where the alphabet size is a growing function of the block-length is a meaningful

asymptotic regime for multiset codes. Several problems worth investigating in this context are:

- Improve the bounds on $M_{\tilde{q}n}(n;h)$ for $h \geq 2$.
- Investigate whether the construction of multiset codes via Sidon sets optimal in this regime.
- Derive bounds on $M_{\tilde{q}n}\big(n;\tilde{h}n\big)$, for an arbitrary constant $\tilde{h} \in (0,1)$. Namely, the number of deletions growing linearly with the block-length is another natural asymptotic regime. Note that the upper bound in (31) is valid in this regime and can be expressed in the corresponding asymptotic form via (34). As for the lower bound, one can apply the familiar Gilbert-Varshamov bound. However, due to the structure of the space $\triangle_n^q$ and, in particular, the fact that balls in this space do not have uniform sizes, we do not expect this bound to be tight.

ACKNOWLEDGMENT

We thank Dr. Aslan Tchamkerten and Dr. Jossy Sayir for sending us the preprints [5] and [33], respectively, and Prof. David Tse and Dr. Netanel Raviv for helpful discussions on the subject of DNA storage.

REFERENCES

[1] R. Ahlswede, H. K. Aydinian, and L. H. Khachatrian, "On Perfect Codes and Related Concepts," *Des. Codes Cryptogr.*, vol. 22, no. 3, pp. 221–237, Jan. 2001.

[2] R. Ahlswede and A. H. Kaspi, "Optimal Coding Strategies for Certain Permuting Channels," *IEEE Trans. Inform. Theory*, vol. 33, no. 3, pp. 310–314, May 1987.

[3] M. D. Atkinson, N. Santoro, and J. Urrutia, "Integer Sets with Distinct Sums and Differences and Carrier Frequency Assignments for Nonlinear Repeaters," *IEEE Trans. Commun.*, vol. 34, no. 6, pp. 614–617, 1986.

[4] A. Barg and A. Mazumdar, "Codes in Permutations and Error Correction for Rank Modulation," *IEEE Trans. Inform. Theory*, vol. 56, no. 7, pp. 3158–3165, Jul. 2010.

[5] J.-C. Belfiore, L. Sok, P. Solé, and A. Tchamkerten, "Lattice Codes for Deletion and Repetition Channels," *IEEE Trans. Inform. Theory*, submitted for publication.

[6] T. Beth, D. Jungnickel, and H. Lenz, *Design Theory*, 2nd ed., Cambridge University Press, 1999.

[7] R. C. Bose and S. Chowla, "Theorems in the Additive Theory of Numbers," *Comment. Math. Helv.*, vol. 37, no. 1, pp. 141–147, Dec. 1962.

[8] S. Chen, "On the Size of Finite Sidon Sequences," *Proc. Amer. Math. Soc.*, vol. 121, no. 2, pp. 353–356, Jun. 1994.

[9] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, 3rd ed., Springer, 1999.

[10] S. I. R. Costa, M. Muniz, E. Agustini, and R. Palazzo, "Graphs, Tessellations, and Perfect Codes on Flat Tori," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2363–2377, Oct. 2004.

[11] P. Delsarte, "An Algebraic Approach to Association Schemes of Coding Theory," *Philips J. Res.*, vol. 10, pp. 1–97, 1973.

[12] H. Derksen, "Error-Correcting Codes and B_h-Sequences," *IEEE Trans. Inform. Theory*, vol. 50, no. 3, pp. 476–485, Mar. 2004.

[13] C. Ding, *Codes from Difference Sets*, World Scientific, 2015.

[14] M. Gadouleau and A. Goupil, "Binary Codes for Packet Error and Packet Loss Correction in Store and Forward," in *Proc. Int. ITG Conf. on Source and Channel Coding*, Siegen, Germany, Jan. 2010.

[15] S. Galovich and S. Stein, "Splittings of Abelian Groups by Integers," *Aequationes Math.*, vol. 22, no. 1, pp. 249–267, 1981.

[16] S. W. Golomb and L. R. Welch, "Perfect Codes in the Lee Metric and the Packing of Polyominoes," *SIAM J. Appl. Math.*, vol. 18, no. 2, pp. 302–317, Mar. 1970.

[17] R. L. Graham and N. J. A. Sloane, "Lower Bounds for Constant Weight Codes," *IEEE Trans. Inform. Theory*, vol. 26, no. 1, pp. 37–43, 1980.

[18] W. Hamaker, "Factoring Groups and Tiling Space," *Aequationes Math.*, vol. 9, no. 2-3, pp. 145–149, 1973.

[19] R. Heckel, I. Shomorony, K. Ramchandran, and D. N. C. Tse, "Fundamental Limits of DNA Storage Systems," in *Proc. 2017 IEEE Int. Symp. Inform. Theory (ISIT)*, Aachen, Germany, Jun. 2017. Available online: arXiv:1705.04732v1 [cs.IT].

[20] D. Hickerson, "Splittings of Finite Groups," *Pacific J. Math.*, vol. 107, no. 1, pp. 141-171, 1983.

[21] P. Horak, "On Perfect Lee Codes," *Discrete Math.*, vol. 309, no. 18, pp. 5551–5561, Sep. 2009.

[22] X.-D. Jia, "On Finite Sidon Sequences," *J. Number Theory*, vol. 44, no. 1, pp. 84–92, May 1993.

[23] T. Kløve, "Error Correcting Codes for the Asymmetric Channel," Technical Report, Dept. of Informatics, University of Bergen, 1981. (Updated in 1995.)

[24] M. Kovačević and D. Vukobratović, "Subset Codes for Packet Networks," *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 729–732, Apr. 2013.

[25] M. Kovačević and D. Vukobratović, "Perfect Codes in the Discrete Simplex," *Des. Codes Cryptogr.*, vol. 75, no. 1, pp. 81–95, Apr. 2015.

[26] M. Kovačević and V. Y. F. Tan, "Improved Bounds on Sidon Sets via Lattice Packings of Simplices," *SIAM J. Discrete Math.*, to appear. Available online: arXiv:1610.01341 [math.CO].

[27] M. Kovačević and V. Y. F. Tan, "Coding for the Permutation Channel with Insertions, Deletions, Substitutions, and Erasures," in *Proc. 2017 IEEE Int. Symp. Inform. Theory (ISIT)*, Aachen, Germany, Jun. 2017. Available online: arXiv:1612.08837 [cs.IT].

[28] V. Yu. Krachkovsky, "Bounds on the Zero-Error Capacity of the Input-Constrained Bit-Shift Channel," *IEEE Trans. Inform. Theory*, vol. 40, no. 4, pp. 1240–1244, Jul. 1994.

[29] J. F. Kurose and K. W. Ross, *Computer Networking*, 5th ed., Addison-Wesley, 2010.

[30] A. W. Lam and D. V. Sarwate, "On Optimum Time-Hopping Patterns," *IEEE Trans. Commun.*, vol. 36, no. 3, pp, 380–382, Mar. 1988.

[31] M. Langberg, M. Schwartz, and E. Yaakobi, "Coding for the ℓ_∞-Limited Permutation Channel," in *Proc. 2015 IEEE Int. Symp. Inform. Theory (ISIT)*, pp. 1936–1940, Hong Kong, Jun. 2015.

[32] V. I. Levenshtein, "Binary Codes with Correction for Deletions and Insertions of the Symbol 1" (In Russian), *Probl. Peredachi Inf.*, vol. 1, no. 1, pp. 12–25, 1965.

[33] D. J. C. MacKay, J. Sayir, and N. Goldman, "Near-Capacity Codes for Fountain Channels with Insertions, Deletions, and Substitutions, with Applications to DNA Archives," Unpublished manuscript, 2015.

[34] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, North-Holland Publishing Company, 1977.

[35] C. Martinez, R. Beivide, and E. Gabidulin, "Perfect Codes for Metrics Induced by Circulant Graphs," *IEEE Trans. Inform. Theory*, vol. 53, no. 9, pp. 3042–3052, Sep. 2007.

[36] P. McMullen, "Convex Bodies Which Tile Space by Translation," *Mathematika*, vol. 27, no. 1, pp. 113–121, Jun. 1980.

[37] J. J. Metzner, "Simplification of Packet-Symbol Decoding With Errors, Deletions, Misordering of Packets, and No Sequence Numbers," *IEEE Trans. Inform. Theory*, vol. 55, no. 6, pp. 2626–2639, Jun. 2009.

[38] T. Nakano, A. W. Eckford, and T. Haraguchi, *Molecular Communication*, Cambridge University Press, 2013.

[39] K. O'Bryant, "A Complete Annotated Bibliography of Work Related to Sidon Sequences," *Electron. J. Combin.*, #DS11, 39 pp, 2004.

[40] L. J. Schulman and D. Zuckerman, "Asymptotically Good Codes Correcting Insertions, Deletions, and Transpositions," *IEEE Trans. Inform. Theory*, vol. 45, no. 7, pp. 2552–2557, Nov. 1999.

[41] J. Singer, "A Theorem in Finite Projective Geometry and Some Applications to Number Theory," *Trans. Amer. Math. Soc.*, vol. 43, pp. 377–385, 1938.

[42] L. Sok, P. Solé, and A. Tchamkerten, "Lattice Based Codes for Insertion and Deletion Channels," in *Proc. 2013 IEEE Int. Symp. Inform. Theory (ISIT)*, pp. 684–688, Istanbul, Turkey, Jul. 2013.

[43] J. Spencer, *Asymptopia*, American Mathematical Society, 2014.

[44] S. Stein, "Factoring by Subsets," *Pacific J. Math.*, vol. 22, no. 3, pp. 523–541, 1967.

[45] S. Stein, "Algebraic Tiling," *Amer. Math. Monthly*, vol. 81, pp. 445–462, May 1974.

[46] S. Stein, "Packings of R^n by Certain Error Spheres," *IEEE Trans. Inform. Theory*, vol. 30, no. 2, pp. 356–363, Mar. 1984.

[47] S. Stein and S. Szabó, *Algebra and Tiling: Homomorphisms in the Service of Geometry*, The Mathematical Association of America, 1994.

[48] R. R. Varshamov, "A Class of Codes for Asymmetric Channels and a Problem from the Additive Theory of Numbers," *IEEE Trans. Inform. Theory*, vol. 19, no. 1, pp. 92–95, Jan. 1973.

[49] J. M. Walsh, S. Weber, and C. wa Maina, "Optimal Rate–Delay Tradeoffs and Delay Mitigating Codes for Multipath Routed and Network Coded Networks," *IEEE Trans. Inform. Theory*, vol. 55, no. 12, pp. 5491–5510, Dec. 2009.