On Linear Balancing Sets*

Arya Mazumdar
Department of ECE
University of Maryland
College Park, MD 20742, USA
arya@umd.edu

Ron M. Roth
Computer Science Department
Technion
Haifa 32000, Israel
ronny@cs.technion.ac.il

Pascal O. Vontobel
Hewlett–Packard Laboratories
Palo Alto, CA 94304, USA
pascal.vontobel@ieee.org

Abstract

Let n be an even positive integer and $\mathbb{F}$ be the field GF(2). A word in $\mathbb{F}^n$ is called balanced if its Hamming weight is $n/2$. A subset $\mathcal{C} \subseteq \mathbb{F}^n$ is called a balancing set if for every word $\boldsymbol{y} \in \mathbb{F}^n$ there is a word $\boldsymbol{x} \in \mathcal{C}$ such that $\boldsymbol{y} + \boldsymbol{x}$ is balanced. It is shown that most linear subspaces of $\mathbb{F}^n$ of dimension slightly larger than $\frac{3}{2} \log_2 n$ are balancing sets. A generalization of this result to linear subspaces that are "almost balancing" is also presented. On the other hand, it is shown that the problem of deciding whether a given set of vectors in $\mathbb{F}^n$ spans a balancing set, is NP-hard. An application of linear balancing sets is presented for designing efficient error-correcting coding schemes in which the codewords are balanced.

1 Introduction

Let $\mathbb{F}$ denote the finite field GF(2) and assume hereafter that n is an even positive integer. For words (vectors) $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{F}^n$, denote by $\mathsf{w}(\boldsymbol{x})$ the Hamming weight of $\boldsymbol{x}$ and by $\mathsf{d}(\boldsymbol{x}, \boldsymbol{y})$ the Hamming distance between $\boldsymbol{x}$ and $\boldsymbol{y}$.

* The work of A. Mazumdar and R.M. Roth was done in part while visiting Hewlett–Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA. The work of R.M. Roth was supported in part by Grant No. 1280/08 from the Israel Science Foundation. Part of this work was presented at the *IEEE International Symposium of Information Theory*, Seoul, Korea, June 28–July 3, 2009.

We say that a word $\boldsymbol{z} \in \mathbb{F}^n$ is *balanced* if $\mathsf{w}(\boldsymbol{z}) = n/2$. For a word $\boldsymbol{x} \in \mathbb{F}^n$, define the set

$$\begin{aligned} \mathcal{B}(\boldsymbol{x}) &= \{\boldsymbol{x} + \boldsymbol{z} \,:\, \boldsymbol{z} \text{ is balanced}\} \\ &= \{\boldsymbol{y} \in \mathbb{F}^n \,:\, \mathsf{d}(\boldsymbol{y}, \boldsymbol{x}) = n/2\} \ . \end{aligned}$$

In particular, if $\mathbf{0}$ denotes the all-zero word in $\mathbb{F}^n$, then $\mathcal{B}(\mathbf{0})$ is the set of all balanced words in $\mathbb{F}^n$. It is known that

$$\frac{2^n}{\sqrt{2n}} \leq \binom{n}{n/2} = |\mathcal{B}(\boldsymbol{x})| \leq \frac{2^n}{\sqrt{\pi n/2}} \tag{1}$$

(see, for example, [12, p. 309]). We extend the notation $\mathcal{B}(\cdot)$ to subsets $\mathcal{C} \subseteq \mathbb{F}^n$ by

$$\mathcal{B}(\mathcal{C}) = \bigcup_{\boldsymbol{x} \in \mathcal{C}} \mathcal{B}(\boldsymbol{x}) \ .$$

A subset $\mathcal{C} \subseteq \mathbb{F}^n$ is called a *balancing set* if $\mathcal{B}(\mathcal{C}) = \mathbb{F}^n$; equivalently, $\mathcal{C}$ is a balancing set if for every $\boldsymbol{y} \in \mathbb{F}^n$ there exists an $\boldsymbol{x} \in \mathcal{C}$ such that $\mathsf{d}(\boldsymbol{y}, \boldsymbol{x}) = \mathsf{w}(\boldsymbol{y} + \boldsymbol{x}) = n/2$ (which is also the same as saying that for every $\boldsymbol{y} \in \mathbb{F}^n$ one has $\mathcal{B}(\boldsymbol{y}) \cap \mathcal{C} \neq \emptyset$). Using the terminology of Cohen *et al.* in [6, §13.1], a balancing set can also be referred to as an $\{n/2\}$-*covering code*.

An example of a balancing set of size n was presented by Knuth in [10]: his set consists of the words $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, where

$$\boldsymbol{x}_i = \underbrace{11\ldots1}_{i}\underbrace{00\ldots0}_{n-i} \ .$$

It was shown by Alon *et al.* in [1] that every balancing set must contain at least n words; hence, Knuth's balancing set has the smallest possible size.

As proposed by Knuth, balancing sets can be used to efficiently encode unconstrained binary words into balanced words as follows: given an information word $\boldsymbol{u} \in \mathbb{F}^n$, a word $\boldsymbol{x}$ in a balancing set $\mathcal{C}$ is found so that $\boldsymbol{u} + \boldsymbol{x}$ is balanced. The transmitted codeword then consists of $\boldsymbol{u} + \boldsymbol{x}$, appended by a recursive encoding of the index (of length $\lceil \log_2 |\mathcal{C}| \rceil$) of $\boldsymbol{x}$ within $\mathcal{C}$. Thus, when $|\mathcal{C}| = n$, the redundancy of the transmission is $(\log_2 n) + O(\log \log n)$. By (1), we can get a smaller redundancy of $\frac{1}{2}(\log_2 n) + O(1)$ using any one-to-one mapping into $\mathcal{B}(\mathbf{0})$. Such a mapping, in turn, can be implemented using enumerative coding, but the overall time complexity will be higher than Knuth's encoder.

The requirement that the transmitted codewords be balanced is found in many applications—especially in magnetic and optical storage systems [9, Ch. 1], [11, Ch. 5], [13, §1]. Moreover, in most of these applications, the transmitted codewords are also required to have some Hamming distance properties so as to provide error-correction capabilities [11, Ch. 6], [13, §7]. Placing an error-correcting encoder before applying any of the two balancing encoders mentioned earlier, will generally not work, since the balancing encoder may destroy any distance properties of its input. One possible solution would then be to encode the raw

information word directly into a codeword of a constant-weight error-correcting code, specifically, a code in which all codewords are in $\mathcal{B}(\mathbf{0})$. By a simple averaging argument one gets that for every code $\mathcal{C} \subseteq \mathbb{F}^n$ there is at least one word $\boldsymbol{x} \in \mathbb{F}^n$ for which the shifted set

$$\mathcal{C} + \boldsymbol{x} = \{\boldsymbol{y} \in \mathbb{F}^n \,:\, \boldsymbol{y} - \boldsymbol{x} \in \mathcal{C}\}$$

contains at least $(\binom{n}{n/2}/2^n)|\mathcal{C}| \geq |\mathcal{C}|/\sqrt{2n}$ balanced words. Yet, for most known constant-weight codes, the implementation of an encoder for such codes is typically quite complex compared to the encoding of linear codes or to the above-mentioned balancing methods [15].

In this work, we will be interested in *linear balancing sets*, namely, balancing sets that are linear subspaces of $\mathbb{F}^n$. Our main result, to be presented in Section 3, states that most linear subspaces of $\mathbb{F}^n$ of dimension which is at a (small) margin above $\frac{3}{2}\log_2 n$ are linear balancing sets. A generalization of this result to sets which are "almost balancing" (in a sense to be formally defined) will be presented in Section 4. On the other hand, we will prove (in Appendix B) that the problem of deciding whether a given set of vectors in $\mathbb{F}^n$ spans a balancing set, is NP-hard.

Our study of balancing sets was motivated by the potential application of these sets in obtaining efficient coding schemes that combine balancing and error correction, as we outline in Section 5. However, we feel that linear balancing sets could be interesting also on their own right, from a purely combinatorial point of view.

2 Existence result

From the result in [1] we readily get the following lower bound on the dimension of any linear balancing set.

Theorem 2.1. [1] *The dimension of every linear balancing set $\mathcal{C} \subseteq \mathbb{F}^n$ is at least $\lceil \log_2 n \rceil$.*

As mentioned earlier, we will show that most linear subspaces of $\mathbb{F}^n$ of dimension slightly above $\frac{3}{2}\log_2 n$ are in fact balancing sets. We start with the following simpler existence result, as some components of its proof (in particular, Lemma 2.3 below) will be useful also for our random-coding result.

Theorem 2.2. *There exists a linear balancing set in $\mathbb{F}^n$ of dimension $\lceil \frac{3}{2}\log_2 n \rceil$.*

Theorem 2.2 can be seen as the balancing-set counterpart of the result of Goblick [8] regarding the existence of good linear covering codes (see also Berger [2, pp. 201–202], Cohen [5], Cohen *et al.* [6, §12.3], and Delsarte and Piret [7]); in fact, our proof is strongly based on their technique. In what follows, we will adopt the formulation of [7].

Before proving Theorem 2.2, we introduce some notation. We denote the union $\mathcal{C} \cup (\mathcal{C} + \boldsymbol{x})$ by $\mathcal{C} + \mathbb{F}\boldsymbol{x}$. (When $\mathcal{C}$ is a linear subspace of $\mathbb{F}^n$ then so is $\mathcal{C} + \mathbb{F}\boldsymbol{x}$, and $\mathcal{C} + \boldsymbol{x}$ is a coset of $\mathcal{C}$ within $\mathbb{F}^n$.)

We also define

$$Q(\mathcal{C}) = 2^{-n} \left| \mathbb{F}^n \setminus \mathcal{B}(\mathcal{C}) \right| = 1 - \frac{|\mathcal{B}(\mathcal{C})|}{2^n} \ .$$

Namely, $Q(\mathcal{C})$ is the probability that $\mathcal{B}(\boldsymbol{x}) \cap \mathcal{C} = \emptyset$, for a randomly and uniformly selected word $\boldsymbol{x} \in \mathbb{F}^n$.

The proof of Theorem 2.2 makes use of the following lemma.

Lemma 2.3. *For every subset $\mathcal{C} \subseteq \mathbb{F}^n$,*

$$2^{-n} \sum_{\boldsymbol{x} \in \mathbb{F}^n} Q(\mathcal{C} + \mathbb{F}\boldsymbol{x}) = \left(Q(\mathcal{C})\right)^2 \ .$$

Proof. The proof is essentially the first part of the proof of Theorem 3 in [7], except that we replace the Hamming sphere by $\mathcal{B}(\cdot)$. For the sake of completeness, we include the proof in Appendix A. □

Proof of Theorem 2.2. Again, we follow the steps of the proof of Theorem 3 in [7]. Write $\ell = \lceil \frac{3}{2} \log_2 n \rceil$. We construct iteratively linear subspaces $\mathcal{C}_0 \subset \mathcal{C}_1 \subset \cdots \subset \mathcal{C}_\ell$ as follows. The subspace $\mathcal{C}_0$ is simply $\{\mathbf{0}\}$. Given now the subspace $\mathcal{C}_{i-1}$, we let

$$\mathcal{C}_i = \mathcal{C}_{i-1} + \mathbb{F}\boldsymbol{x}_i \ ,$$

where $\boldsymbol{x}_i$ is a word in $\mathbb{F}^n$ such that

$$Q(\mathcal{C}_{i-1} + \mathbb{F}\boldsymbol{x}_i) \leq \left(Q(\mathcal{C}_{i-1})\right)^2 \ ;$$

by Lemma 2.3, such a word indeed exists. Now,

$$Q(\mathcal{C}_0) = 1 - \frac{|\mathcal{B}(\mathbf{0})|}{2^n} = 1 - 2^{-n} \binom{n}{n/2} \leq 1 - \frac{1}{\sqrt{2n}} \ , \tag{2}$$

where the last step follows from the lower bound in (1). Hence,

$$Q(\mathcal{C}_\ell) \leq \left(Q(\mathcal{C}_0)\right)^{2^\ell} \leq \left(1 - \frac{1}{\sqrt{2n}}\right)^{n^{3/2}} \leq \mathrm{e}^{-n/\sqrt{2}} < 2^{-n} \ .$$

As $2^n Q(\mathcal{C}_\ell)$ is an integer, we conclude that $Q(\mathcal{C}_\ell)$ is necessarily zero, namely, $\mathcal{B}(\mathcal{C}_\ell) = \mathbb{F}^n$. □

3 Most linear subspaces are balancing sets

The next theorem is our main result. Hereafter, $\mathbb{N}$ stands for the set of natural numbers, and the notation $\mathsf{exp}(z)$ stands for an expression of the form $a \cdot 2^{bz}$, for some positive constants a and b.

Theorem 3.1. *Given a function $\rho : (2\mathbb{N}) \to \mathbb{N}$, let $\mathcal{C}$ be a random linear subspace of $\mathbb{F}^n$ which is spanned by $\lceil \frac{3}{2} \log_2 n \rceil + \rho(n)$ words that are selected independently and uniformly from $\mathbb{F}^n$. Then,*

$$\mathsf{Prob}\left\{\mathcal{C} \text{ is a balancing set}\right\} \geq 1 - \mathsf{exp}(-\rho(n)) \; .$$

(Thus, as long as $\rho(n)$ goes to infinity with n, all but a vanishing fraction of the ensemble of linear subspaces of $\mathbb{F}^n$ of dimension $\lceil \frac{3}{2} \log_2 n \rceil + \rho(n)$ are balancing sets.)

Theorem 3.1 is the balancing-set counterpart of a result originally obtained by Blinovskii [3], showing that most linear codes attain the sphere-covering bound. An alternate proof for his result (with slightly different convergence rates as $n \to \infty$) was then presented by Cohen *et al.* in [6, §12.3]. The proof that we provide for Theorem 3.1 can be seen as an adaptation (and refinement) of the proof of Cohen *et al.* to the balancing-set setting.

We break the proof of Theorem 3.1 into three lemmas. To maintain the flow of the exposition, we will defer the proofs of the lemmas until after the proof of Theorem 3.1.

Lemma 3.2. *Let $\mathcal{C}_0$ be a random linear subspace of $\mathbb{F}^n$ which is spanned by $\lceil \frac{1}{2} \log_2 n \rceil$ random words that are selected independently and uniformly from $\mathbb{F}^n$. There exists an absolute constant $\beta \in [0, 1)$ independent of n (e.g., $\beta = \frac{3}{4}$) such that*

$$\mathsf{Prob}\left\{Q(\mathcal{C}_0) > \beta\right\} \leq \mathsf{exp}(-n) \; .$$

Lemma 3.3. *Let $\mathcal{C}_0$ be a linear subspace of $\mathbb{F}^n$. Fix a positive integer r, and let $\mathcal{C}_1$ be a random linear subspace of $\mathbb{F}^n$ which is spanned by $\mathcal{C}_0$ and r random words from $\mathbb{F}^n$ that are selected uniformly and independently. Then*

$$\mathsf{Prob}\left\{Q(\mathcal{C}_1) > (Q(\mathcal{C}_0))^{(r/2)+1}\right\} < (Q(\mathcal{C}_0))^{r/2} \; .$$

Lemma 3.4. *Let $\mathcal{C}_1$ be a linear subspace of $\mathbb{F}^n$ and let $\mathcal{C}_2$ be a random linear subspace of $\mathbb{F}^n$ which is spanned by $\mathcal{C}_1$ and $\lceil \log_2 n \rceil$ random words from $\mathbb{F}^n$ that are selected uniformly and independently. Then*

$$\mathsf{Prob}\left\{Q(\mathcal{C}_2) > 0\right\} \leq 8Q(\mathcal{C}_1) \; .$$

Proof of Theorem 3.1. It is known (e.g., from [12, p. 444, Theorem 9]) that

$$\mathsf{Prob}\left\{\mathcal{C} \neq \mathbb{F}^n\right\} \leq \mathsf{exp}(n - \rho(n)) \; .$$

Hence, we can assume hereafter in the proof that $\rho(n)$ is at most linear in n.

Let $\mathcal{U}$ be the list of $|\mathcal{U}| = \lceil \frac{3}{2}\log_2 n \rceil + \rho(n)$ random words from $\mathbb{F}^n$ that span $\mathcal{C}$, and write $\ell = \lceil \frac{1}{2}\log_2 n \rceil$, $t = \lceil \log_2 n \rceil$, and $r = |\mathcal{U}| - \ell - t$. We partition the words in $\mathcal{U}$ into three sub-lists, $\mathcal{U}_0$, $\mathcal{U}_1$, and $\mathcal{U}_2$, of sizes ℓ, r, and t, respectively. We denote by $\mathcal{C}_0$, $\mathcal{C}_1$, and $\mathcal{C}_2$ the linear spans of $\mathcal{U}_0$, $\mathcal{U}_0 \cup \mathcal{U}_1$, and $\mathcal{U}_0 \cup \mathcal{U}_1 \cup \mathcal{U}_2$, respectively.

Take $\beta = \frac{3}{4}$ (say). By Lemma 3.2 we get that

$$\mathsf{Prob}\left\{Q(\mathcal{C}_0) > \beta\right\} \leq \mathsf{exp}(-n) \; . \tag{3}$$

By Lemma 3.3 we have

$$\mathsf{Prob}\left\{ Q(\mathcal{C}_1) > \beta^{(r/2)+1} \;\middle|\; Q(\mathcal{C}_0) \leq \beta \right\} < \beta^{r/2} \; . \tag{4}$$

Finally, by Lemma 3.4 we get

$$\mathsf{Prob}\left\{ Q(\mathcal{C}_2) > 0 \;\middle|\; Q(\mathcal{C}_1) \leq \beta^{(r/2)+1} \right\} \leq (8\beta) \cdot \beta^{r/2} \; . \tag{5}$$

The result is now obtained by combining (3)–(5) and noting that $\beta^{r/2} = \mathsf{exp}(-\rho(n))$. □

Next, we turn to the proofs of the lemmas.

Proof of Lemma 3.2. Write $\ell = \lceil \frac{1}{2}\log_2 n \rceil$, and let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell$ denote the random words that span $\mathcal{C}_0$. The proof is based on the fact that, with high probability, the Hamming weight of each nonzero word in $\mathcal{C}_0$ is close to $n/2$. Indeed, fix some nonzero vector $(a_i)_{i=1}^{\ell}$ in $\mathbb{F}^\ell$. Then the sum $\boldsymbol{x} = \sum_{i=1}^{\ell} a_i \boldsymbol{x}_i$ is uniformly distributed over $\mathbb{F}^n$ and, so, by the Chernoff bound, for every $\delta > 0$ there exists an $\eta = \eta(\delta) > 0$ such that

$$\mathsf{Prob}\left\{ |\mathsf{w}(\boldsymbol{x}) - \frac{n}{2}| > \delta n \right\} \leq 2^{-\eta n} \; .$$

Given some $\delta \in [0, \frac{1}{2})$, let $\mathcal{E}$ denote the event that $\mathcal{C}_0$ has dimension (exactly) ℓ and each nonzero word in $\mathcal{C}_0$ has Hamming weight within $\left(\frac{1}{2} \pm \delta\right) n$; namely,

$$\mathcal{E} = \left\{ |\mathsf{w}(\boldsymbol{x}) - \frac{n}{2}| \leq \delta n \text{ for every } \boldsymbol{x} = \textstyle\sum_{i=1}^{\ell} a_i \boldsymbol{x}_i \text{ where } (a_i)_{i=1}^{\ell} \in \mathbb{F}^\ell \setminus \{\boldsymbol{0}\} \right\} .$$

By the union bound we readily get that

$$\mathsf{Prob}\left\{\mathcal{E}\right\} > 1 - 2^{\ell} \cdot 2^{-\eta n} = 1 - \mathsf{exp}(-n) \; .$$

Let $\boldsymbol{x}$ and $\boldsymbol{x}'$ be two distinct words in $\mathcal{C}_0$, write $\mathsf{d}(\boldsymbol{x}, \boldsymbol{x}') = \tau n$, and suppose that $\frac{1}{2} - \delta \leq \tau \leq \frac{1}{2} + \delta$. If there exists a word $\boldsymbol{y}$ that is at equal distance from $\boldsymbol{x}$ and $\boldsymbol{x}'$ then $\mathsf{d}(\boldsymbol{x}, \boldsymbol{x}')$ must be even. Therefore, if τn is odd then $|\mathcal{B}(\boldsymbol{x}) \cap \mathcal{B}(\boldsymbol{x}')| = 0$. Otherwise,

$$|\mathcal{B}(\boldsymbol{x}) \cap \mathcal{B}(\boldsymbol{x}')| \;\; = \;\; \binom{\tau n}{\tau n/2} \binom{(1-\tau)n}{(1-\tau)n/2}$$

$$\begin{aligned} &\leq \frac{2^{\tau n}}{\sqrt{\pi\tau n/2}} \cdot \frac{2^{(1-\tau)n}}{\sqrt{\pi(1-\tau)n/2}} \\ &= \frac{2^{n+1}}{\pi n\sqrt{\tau(1-\tau)}} \\ &\leq \frac{2^{n+2}}{\pi n\sqrt{1-4\delta^2}}\,, \end{aligned} \tag{6}$$

where the second step follows from the upper bound in (1).

Conditioning on the event $\mathcal{E}$, we get the next chain of inequalities, where the first inequality is a direct application of de Caen's lower bound [4], and the third inequality follows from the lower bound in (1) and the fact that for any $a > 0$, the real function $x \mapsto x^2/(a+x)$ is increasing whenever $x > 0$:

$$\begin{aligned} |\mathcal{B}(\mathcal{C}_0)| &\geq \sum_{\boldsymbol{x}\in\mathcal{C}_0} \frac{|\mathcal{B}(\boldsymbol{x})|^2}{\sum_{\boldsymbol{x}'\in\mathcal{C}_0} |\mathcal{B}(\boldsymbol{x})\cap\mathcal{B}(\boldsymbol{x}')|} \\ &> 2^\ell \binom{n}{n/2}^2 \Big/ \left(\frac{2^\ell\cdot 2^{n+2}}{\pi n\sqrt{1-4\delta^2}} + \binom{n}{n/2}\right) \\ &\geq 2^n \Big/ \left(\frac{8}{\pi\sqrt{1-4\delta^2}} + \frac{\sqrt{2n}}{2^\ell}\right). \end{aligned} \tag{7}$$

On the other hand, we also have $2^\ell \geq \sqrt{n}$ and, so, writing

$$\beta(\delta) = 1 - \left(\frac{8}{\pi\sqrt{1-4\delta^2}} + \sqrt{2}\right)^{-1},$$

we get that, conditioned on the event $\mathcal{E}$,

$$Q(\mathcal{C}_0) = 1 - \frac{|\mathcal{B}(\mathcal{C}_0)|}{2^n} \leq \beta(\delta)\,. \tag{8}$$

The result follows by recalling that $\mathsf{Prob}\{\mathcal{E}\} \geq 1 - \mathsf{exp}(-n)$ and observing that $\beta(\delta) < 1$ for every $\delta \in [0, \frac{1}{2})$ (in particular, there is some δ for which $\beta(\delta) = \frac{3}{4} > \beta(0)$). □

Remark 3.1. Suppose that $\mathcal{C}_0(m,\ell)$ is an ℓ-dimensional linear subspace of the linear $[n{=}2^m, m, 2^{m-1}]$ code over $\mathbb{F}$ obtained by appending a fixed zero coordinate to every codeword of the binary $[2^m{-}1, m, 2^{m-1}]$ simplex code. In this case, we can substitute $\delta = 0$ in (8) and obtain that $Q(\mathcal{C}_0(m,\ell)) \leq \beta(0) \approx 0.748$, for every ℓ in the range $m/2 \leq \ell \leq m$. Thus, $\mathcal{C}_0(m,\ell)$ can replace the random code $\mathcal{C}_0$ in Lemma 3.2. If ℓ grows sufficiently fast with m so that $\ell{-}(m/2)$ tends to infinity, then from (7) it follows that

$$\lim_{m,\,\ell-(m/2)\to\infty} Q(\mathcal{C}_0(m,\ell)) \leq 1 - \frac{\pi}{8} \approx 0.607\,.$$

Let $\mathcal{C}_0' = \mathcal{C}_0'(m,\ell)$ be given by $\mathcal{C}_0(m,\ell) + \mathbb{F}\boldsymbol{x}$, where $\boldsymbol{x}$ is an odd-weight word in $\mathbb{F}^n$. For $m > 1$ we have $|\mathcal{B}(\mathcal{C}_0')| = 2|\mathcal{B}(\mathcal{C}_0(m,\ell))|$. Therefore, when m, $\ell{-}(m/2) \to \infty$, we can bound $Q(\mathcal{C}_0')$ from above by $1 - (\pi/4) \approx 0.215$. □

Proof of Lemma 3.3. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r$ be the random words that, together with $\mathcal{C}_0$, span (the random code) $\mathcal{C}_1$. Obviously, $\mathcal{B}(\mathcal{C}_0 + \boldsymbol{x}_i) \subseteq \mathcal{B}(\mathcal{C}_1)$ and $Q(\mathcal{C}_0 + \boldsymbol{x}_i) = Q(\mathcal{C}_0)$ for every $i = 1, 2, \ldots, r$. Hence, the expected value of $Q(\mathcal{C}_1)$ (taken over all the independently and uniformly distributed words $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r \in \mathbb{F}^n$) satisfies

$$\begin{aligned}
\mathsf{E}\left\{Q(\mathcal{C}_1)\right\} &= 2^{-n} \sum_{\boldsymbol{y} \in \mathbb{F}^n} \mathsf{Prob}\left\{\boldsymbol{y} \notin \mathcal{B}(\mathcal{C}_1)\right\} \\
&\leq 2^{-n} \sum_{\boldsymbol{y} \in \mathbb{F}^n \setminus \mathcal{B}(\mathcal{C}_0)} \prod_{i=1}^{r} \mathsf{Prob}\left\{\boldsymbol{y} \notin \mathcal{B}(\mathcal{C}_0 + \boldsymbol{x}_i)\right\} \\
&= \left(Q(\mathcal{C}_0)\right)^{r+1} .
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathsf{Prob}\left\{Q(\mathcal{C}_1) > \left(Q(\mathcal{C}_0)\right)^{(r/2)+1}\right\} \\
&\quad \leq \mathsf{Prob}\left\{Q(\mathcal{C}_1) > \left(Q(\mathcal{C}_0)\right)^{-r/2} \mathsf{E}\left\{Q(\mathcal{C}_1)\right\}\right\} \\
&\quad < \left(Q(\mathcal{C}_0)\right)^{r/2} ,
\end{aligned}$$

where the last step follows from Markov's inequality. □

Proof of Lemma 3.4. The result is obvious when $Q(\mathcal{C}_1) \notin (0, \frac{1}{8})$; so we assume hereafter in the proof that $Q(\mathcal{C}_1)$ is within that interval. Write $t = \lceil \log_2 n \rceil$, and let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t$ be the random words that, together with $\mathcal{C}_1$, span $\mathcal{C}_2$. For $i = 0, 1, 2, \ldots, t$, define the linear space $\mathcal{L}_i$ iteratively by $\mathcal{L}_0 = \mathcal{C}_1$ and

$$\mathcal{L}_i = \mathcal{L}_{i-1} + \mathbb{F}\boldsymbol{x}_i .$$

Letting $\boldsymbol{Q}_i$ stand for (the random variable) $Q(\mathcal{L}_i)$ and ω_i for $2^i/(8Q(\mathcal{C}_1))$, by Lemma 2.3 and Markov's inequality we get for every $i = 1, 2, \ldots, t$ that, conditioned on an instance of $\mathcal{L}_{i-1}$,

$$\begin{aligned}
&\mathsf{Prob}\left\{\boldsymbol{Q}_i > \boldsymbol{Q}_{i-1}^2 \omega_i \;\middle|\; \mathcal{L}_{i-1}\right\} \\
&\quad = \mathsf{Prob}\left\{Q(\mathcal{L}_{i-1} + \mathbb{F}\boldsymbol{x}_i) > \boldsymbol{Q}_{i-1}^2 \omega_i \;\middle|\; \mathcal{L}_{i-1}\right\} \\
&\quad \leq \frac{1}{\omega_i} = (8Q(\mathcal{C}_1)) \cdot 2^{-i} .
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathsf{Prob}\left\{\boldsymbol{Q}_t > \boldsymbol{Q}_0^{2^t} \prod_{i=1}^{t} \omega_i^{2^{t-i}}\right\} \\
&\quad \leq \mathsf{Prob}\left\{\bigcup_{i=1}^{t} (\boldsymbol{Q}_i > \boldsymbol{Q}_{i-1}^2 \omega_i)\right\}
\end{aligned}$$

$$\leq \sum_{i=1}^{t} \mathsf{Prob}\left\{\boldsymbol{Q}_i > \boldsymbol{Q}_{i-1}^2 \omega_i\right\}$$
$$\leq \sum_{i=1}^{t} \frac{1}{\omega_i} < 8Q(\mathcal{C}_1) \ .$$

Substituting $\boldsymbol{Q}_0 = Q(\mathcal{C}_1)$ and $\boldsymbol{Q}_t = Q(\mathcal{C}_2)$, we conclude that

$$\mathsf{Prob}\Big\{Q(\mathcal{C}_2) > (Q(\mathcal{C}_1))^{2^t} \prod_{i=1}^{t} \omega_i^{2^{t-i}}\Big\} < 8Q(\mathcal{C}_1) \ ,$$

where

$$\begin{aligned}(Q(\mathcal{C}_1))^{2^t} \prod_{i=1}^{t} \omega_i^{2^{t-i}} &< (Q(\mathcal{C}_1))^{2^t} \Big(\prod_{i=1}^{\infty} \omega_i^{2^{-i}}\Big)^{2^t} \\ &= (Q(\mathcal{C}_1))^{2^t} \left(\frac{2^{\sum_{i=1}^{\infty} i2^{-i}}}{(8Q(\mathcal{C}_1))^{\sum_{i=1}^{\infty} 2^{-i}}}\right)^{2^t} \\ &= 2^{-2^t} \leq 2^{-n} \ .\end{aligned}$$

The result follows by recalling that the events "$Q(\mathcal{C}_2) \geq 2^{-n}$" and "$Q(\mathcal{C}_2) > 0$" are identical. □

Figure 1 lists the generator matrices of linear $[n, k, d]$ codes over $\mathbb{F}$ that form linear balancing sets, for several values of n that are divisible by 4. These matrices were found using a greedy algorithm and they do not necessarily generate the smallest sets for a given code length n, except for $n = 12$ and $n = 20$, where the sets attain the lower bound of Theorem 2.1 (in addition, for the case $n = 20$, the set attains the Griesmer bound [12, §17.5]).

Remark 3.2. In view of Remark 3.1, when $n = 2^m$ (or, more generally, when n is "close" to 2^m), Theorem 3.1 holds also for the smaller ensemble where we fix $\lceil m/2 \rceil$ basis elements of the random code $\mathcal{C}$ to be linearly independent codewords of the code $\mathcal{C}_0(m, \lceil m/2 \rceil)$ defined in Remark 3.1. Furthermore, if these $\lceil m/2 \rceil$ rows are replaced by ℓ basis elements of the code $\mathcal{C}_0'(m, \ell)$ (as defined in that remark), then the value β in the proof of Theorem 3.1 can be taken as $1 - (\pi/4)$ (≈ 0.215) whenever $\ell - (m/2)$ goes to infinity (yet more slowly than $\rho(n)$). □

We leave it open to find an explicit construction of linear balancing sets in $\mathbb{F}^n$ of dimension $O(\log n)$. We also mention the following intractability result.

Theorem 3.5. *Given as input a basis of a linear subspace $\mathcal{C}$ of $\mathbb{F}^n$, the problem of deciding whether $\mathcal{C}$ is a balancing set, is NP-hard.*

The proof of Theorem 3.5 is obtained by some modification of the reduction in [14] from Three-Dimensional Matching. We include the proof in Appendix B.

$$
\begin{array}{ll}
[8,3,3]: & \begin{pmatrix} 00001111 \\ 01110010 \\ 10001100 \end{pmatrix} \\
[12,4,5]: & \begin{pmatrix} 000000111111 \\ 000111001110 \\ 101001011100 \\ 111100001000 \end{pmatrix} \\
[16,5,7]: & \begin{pmatrix} 0000000011111111 \\ 0001111100001110 \\ 0110011101111100 \\ 1101011011001000 \\ 1111111100010000 \end{pmatrix} \\
[20,5,9]: & \begin{pmatrix} 00000000001111111111 \\ 00000111110000111110 \\ 01111001110111001100 \\ 11100101101100101000 \\ 11010111010010010000 \end{pmatrix} \\
[24,6,9]: & \begin{pmatrix} 000000000000111111111111 \\ 000000011111000000111110 \\ 000111100111001111011100 \\ 001001111011110011111000 \\ 111111110100000000010000 \\ 110101011010100000100000 \end{pmatrix} \\
[28,6,11]: & \begin{pmatrix} 0000000000000001111111111111111 \\ 0000000111111100000011111110 \\ 0001111000111100111100111100 \\ 0010011011001111001111111000 \\ 1111111011011000000000010000 \\ 1011011101100110000000100000 \end{pmatrix} \\
[32,7,13]: & \begin{pmatrix} 00000000000000000111111111111111 \\ 00000000011111110000000011111110 \\ 00000111100000110000011101111100 \\ 01110011001001010101110110101000 \\ 01101110100101011111011010110000 \\ 10110111000111000110111001100000 \\ 10100100100000011101111101000000 \end{pmatrix}
\end{array}
$$

Figure 1: Bases of linear balancing sets for $n = 8, 12, 16, \ldots, 32$.

4 Linear almost-balancing sets

While the code $\mathcal{C}_0(m, \ell{=}m)$ in Remark 3.1 is such that $Q(\mathcal{C}_0(m,m))$ is bounded away from zero, this code can be seen as "almost balancing" in the following sense: for every word $\boldsymbol{y} \in \mathbb{F}^n$ (where $n = 2^m$) there exists a codeword $\boldsymbol{x} \in \mathcal{C}_0(m,m)$ such that $|\mathsf{d}(\boldsymbol{y},\boldsymbol{x}) - (n/2)| \le \sqrt{n}/2$. The proof of this fact is similar to the one showing that the covering radius of the first-order Reed–Muller code is at most $(n - \sqrt{n})/2$ [6, pp. 241–242] (specifically, in the line following Eq. (9.2.4) therein, simply reverse the inequality in "$|\langle \cdot, \cdot \rangle| \ge \sqrt{n}$"; see also (11) below).

Next, we formalize the notion of almost balancing sets and present generalizations for Theorems 2.2 and 3.1. In what follows, we fix some function $\lambda : 2\mathbb{N} \to \mathbb{N}$ such that $\lambda(n) < n/2$, and write $\lambda = \lambda(n)$ for simplicity. For a word $\boldsymbol{x} \in \mathbb{F}^n$ define the set

$$\mathcal{B}_\lambda(\boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{F}^n \,:\, |\mathsf{d}(\boldsymbol{y},\boldsymbol{x}) - n/2| \le \lambda\} \ .$$

As was the case for $\lambda = 0$, the notation $\mathcal{B}_\lambda(\cdot)$ can be extended to subsets $\mathcal{C} \subseteq \mathbb{F}^n$ by

$$\mathcal{B}_\lambda(\mathcal{C}) = \bigcup_{\boldsymbol{x} \in \mathcal{C}} \mathcal{B}_\lambda(\boldsymbol{x}) \ .$$

A subset $\mathcal{C} \subseteq \mathbb{F}^n$ is called a λ–*almost-balancing set* if $\mathcal{B}_\lambda(\mathcal{C}) = \mathbb{F}^n$; equivalently, $\mathcal{C}$ is a λ–almost-balancing set if for every $\boldsymbol{y} \in \mathbb{F}^n$ there exists an $\boldsymbol{x} \in \mathcal{C}$ such that $|\mathsf{d}(\boldsymbol{y}, \boldsymbol{x}) - n/2| \leq \lambda$. (In the terminology of Cohen *et al.* in [6, §19.1], a λ–almost-balancing set is an L-covering code, where $L = \{i \in \mathbb{N} : n/2 - \lambda \leq i \leq n/2 + \lambda\}$.)

The following theorem can be seen as a generalization of Theorem 2.2.

Theorem 4.1. *Suppose that $\lambda = \lambda(n) = O(\sqrt{n})$. There exists a linear λ–almost-balancing set of dimension $\left\lceil \frac{3}{2}\log_2 n - \log_2(2\lambda+1) + O(\lambda^2/n) \right\rceil$.*

Proof. We follow the steps of the proof of Theorem 2.2, with $Q(\mathcal{C}_i)$ replaced by a term $Q_\lambda(\mathcal{C}_i)$ which equals $1 - 2^{-n}\mathcal{B}_\lambda(\mathcal{C}_i)$, and with (2) replaced by an upper bound on $Q_\lambda(\mathcal{C}_0) = Q_\lambda(\{\mathbf{0}\})$ which we shall now derive.

Let $\mathsf{H} : [0,1] \to [0,1]$ be the binary entropy function $\mathsf{H}(z) = -(z\log_2 z) - (1-z)\log_2(1-z)$. Then,

$$\begin{aligned} |\mathcal{B}_\lambda(\mathbf{0})| &= \sum_{i=(n/2)-\lambda}^{(n/2)+\lambda} \binom{n}{i} \\ &\geq (2\lambda+1)\binom{n}{n/2-\lambda} \\ &\geq \frac{2\lambda+1}{\sqrt{2n(1-4(\lambda/n)^2)}} \cdot 2^{n\mathsf{H}(\frac{1}{2}-\frac{\lambda}{n})} \\ &\geq \frac{2\lambda+1}{\sqrt{2n}} \cdot 2^{n\mathsf{H}(\frac{1}{2}-\frac{\lambda}{n})}\,, \end{aligned} \tag{9}$$

where the penultimate step follows from a well known lower bound on binomial coefficients [12, p. 309]. From (9) we have,

$$Q_\lambda(\mathcal{C}_0) \leq 1 - \frac{2\lambda+1}{\sqrt{2n}} \cdot 2^{-n(1-\mathsf{H}(\frac{1}{2}-\frac{\lambda}{n}))}\,,$$

thereby obtaining the counterpart of (2). Proceeding as in the proof Theorem 2.2, we see that $\left\lceil \frac{3}{2}\log_2 n - \log_2(2\lambda+1) + n\left(1 - \mathsf{H}\left(\frac{1}{2} - \frac{\lambda}{n}\right)\right)\right\rceil$ basis elements are sufficient to span a linear λ–almost-balancing set.

Finally, using the Taylor series expansion for $\mathsf{H}(\frac{1}{2} - z)$ and recalling that $\lambda = O(\sqrt{n})$, we obtain

$$\begin{aligned} n\Big(1 - \mathsf{H}\Big(\frac{1}{2} - \frac{\lambda}{n}\Big)\Big) &= \frac{1}{\ln 2}\Big(2\frac{\lambda^2}{n} + \frac{4}{3}\frac{\lambda^4}{n^3} + \ldots\Big) \\ &= \frac{\lambda^2}{n}\Big(\frac{2}{\ln 2} + o(1)\Big) = O\Big(\frac{\lambda^2}{n}\Big)\,, \end{aligned} \tag{10}$$

thereby completing the proof. □

Observe that for $n = 2^m$ and $\lambda = \lfloor \sqrt{n}/2 \rfloor$, the code $\mathcal{C}_0(m,m)$ realizes the dimension guaranteed in Theorem 4.1.

The following theorem is a generalization of Theorem 3.1.

Theorem 4.2. *Suppose that $\lambda = \lambda(n) = O(\sqrt{n})$. Given a function $\rho : 2\mathbb{N} \to \mathbb{N}$, let $\mathcal{C}$ be a random linear subspace of $\mathbb{F}^n$ that is spanned by $\left\lceil \frac{3}{2}\log_2 n - \log_2(2\lambda+1) \right\rceil + \rho(n)$ words selected independently and uniformly from $\mathbb{F}^n$. Then,*

$$\mathsf{Prob}\left\{\mathcal{C} \text{ is a } \lambda\text{–almost-balancing set}\right\} \geq 1 - \mathsf{exp}(-\rho(n)) \ .$$

Proof. The proof is the same as that of Theorem 3.1, except that $Q(\cdot)$ is replaced by $Q_\lambda(\cdot)$ in Lemmas 3.3 and 3.4 (and in their proofs), and Lemma 3.2 is replaced by the following lemma. □

Lemma 4.3. *Suppose that $\lambda = O(\sqrt{n})$, and let $\mathcal{C}_0$ be a random linear subspace of $\mathbb{F}^n$ which is spanned by $\lceil \frac{1}{2}\log_2 n - \log_2(2\lambda+1) \rceil$ random words that are selected independently and uniformly from $\mathbb{F}^n$. There exists an absolute constant $\beta \in [0,1)$ such that*

$$\mathsf{Prob}\left\{Q_\lambda(\mathcal{C}_0) > \beta\right\} \leq \mathsf{exp}(-n) \ .$$

The proof of Lemma 4.3 can be found in Appendix C.

While Theorems 4.1 and 4.2 only cover the case where $\lambda = O(\sqrt{n})$, we next show that when $\lambda = \Omega(\sqrt{n})$, it is fairly easy to obtain an explicit construction for linear λ–almost-balancing sets with relatively small dimensions. Specifically, let s and m be any two positive integers, and set $n = s \cdot 2^m$ and $\lambda = \lfloor \sqrt{sn}/2 \rfloor$. The construction described below yields a linear λ–almost-balancing set of dimension at most $2(\log_2 n - \log_2(2\lambda))$.

Given m and s, let $\mathcal{C}_0 = \mathcal{C}_0(m,m)$ be the linear $[M{=}2^m, m, 2^{m-1}]$ code over $\mathbb{F}$ as in Remark 3.1, and let $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_M$ denote the codewords of $\mathcal{C}_0$. It is shown in [6, p. 242] that for every word $\boldsymbol{y} \in \mathbb{F}^M$,

$$\sum_{i=1}^{M} \left(M - 2\mathsf{d}(\boldsymbol{y}, \boldsymbol{c}_i)\right)^2 = M^2 \tag{11}$$

(from which one gets that there exists at least one codeword $\boldsymbol{c}_i \in \mathcal{C}_0$ such that $|(M/2) - \mathsf{d}(\boldsymbol{y}, \boldsymbol{c}_i)| \leq \sqrt{M}/2$; see the discussion at the beginning of this section).

Consider now the code $\mathcal{C}_0^{(s)}$ which consists of the words $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M$, where

$$\boldsymbol{x}_i = \underbrace{(\boldsymbol{c}_i \,|\, \boldsymbol{c}_i \,|\, \ldots \,|\, \boldsymbol{c}_i)}_{s \text{ times}} , \quad i = 1, 2, \ldots, M \ .$$

Clearly, $\mathcal{C}_0^{(s)}$ is a linear $[n{=}sM, m]$ code over $\mathbb{F}$. Given a word $\boldsymbol{y} \in \mathbb{F}^n$, we write it as $(\boldsymbol{y}_1 \,|\, \boldsymbol{y}_2 \,|\, \ldots \,|\, \boldsymbol{y}_s)$ where each block $\boldsymbol{y}_j$ is in $\mathbb{F}^M$, and define

$$z_{i,j} = M - 2\mathsf{d}(\boldsymbol{y}_j, \boldsymbol{c}_i) \ , \quad i = 1, 2, \ldots, M \ , \quad j = 1, 2, \ldots, s \ .$$

Obviously,

$$n - 2\mathsf{d}(\boldsymbol{y}, \boldsymbol{x}_i) = \sum_{j=1}^{s} z_{i,j} \ , \quad i = 1, 2, \ldots, M \ ,$$

and, so,

$$\begin{aligned} \sum_{i=1}^{M} \Big(n - 2\mathsf{d}(\boldsymbol{y}, \boldsymbol{x}_i) \Big)^2 &= \sum_{i=1}^{M} \Big(\sum_{j=1}^{s} z_{i,j} \Big)^2 \\ &\leq s \sum_{i=1}^{M} \sum_{j=1}^{s} z_{i,j}^2 \\ &= s \sum_{j=1}^{s} \sum_{i=1}^{M} z_{i,j}^2 \overset{(11)}{=} s^2 M^2 \ , \end{aligned}$$

where the inequality follows from the convexity of $z \mapsto z^2$. Hence, there is at least one index $i \in \{1, 2, \ldots, M\}$ for which

$$|n - 2\mathsf{d}(\boldsymbol{y}, \boldsymbol{x}_i)| \leq s\sqrt{M} = \sqrt{sn} \ .$$

We conclude that $\mathcal{C}_0^{(s)}$ is a linear λ–almost-balancing set with $\lambda = \lfloor \sqrt{sn}/2 \rfloor$, and its dimension is $m = \log_2 (n/s) \leq 2(\log_2 n - \log_2 (2\lambda))$.

We end this section by comparing our results to the following generalization of Theorem 2.1.

Theorem 4.4. [1] *The dimension of every linear λ–almost-balancing set $\mathcal{C} \subseteq \mathbb{F}^n$ is at least $\lceil \log_2 n - \log_2(2\lambda + 1) \rceil$.*

For $\lambda = O(\sqrt{n})$, there is still an additive gap of approximately $\frac{1}{2} \log_2 n$ between the lower bound and the upper bound guaranteed by Theorem 4.1, and for $\lambda = \Omega(\sqrt{n})$, the dimension of $\mathcal{C}_0^{(s)}$ is approximately twice the lower bound.

5 Balanced error-correcting codes

In this section, we consider a potential application of linear balancing sets in designing an efficient coding scheme that maps information words into balanced words that belong to a linear error-correcting code; as such, the scheme combines error-correction capabilities with the balancing property.

The underlying idea is as follows. Let $\mathcal{C}$ be a linear $[n, k, d]$ code over $\mathbb{F}$ with the length n and minimum distance d chosen so as to satisfy the required correction capabilities. Suppose,

in addition, that we can write $\mathcal{C}$ as a direct sum of two linear subspaces $\mathcal{C}'$ and $\mathcal{C}''$ of dimensions k' and k'', respectively,

$$\mathcal{C} = \mathcal{C}' \oplus \mathcal{C}'' = \{\boldsymbol{c} + \boldsymbol{x} \,:\, \boldsymbol{c} \in \mathcal{C}', \boldsymbol{x} \in \mathcal{C}''\} \ , \tag{12}$$

where $\mathcal{C}''$ is a balancing set[1]. Now, if k'' is "small" (which means that k' is close to k), we can encode by first mapping a k'-bit information word $\boldsymbol{u}$ into a codeword $\boldsymbol{c} \in \mathcal{C}'$, and then finding a word $\boldsymbol{x} \in \mathcal{C}''$ so that $\boldsymbol{c} + \boldsymbol{x}$ is balanced. The transmitted codeword is then the (balanced) sum $\boldsymbol{c} + \boldsymbol{x}$. The mapping $\boldsymbol{u} \mapsto \boldsymbol{c}$ can be implemented simply as a linear transformation, whereas the balancing word $\boldsymbol{x}$ can be found by exhaustively searching over the $2^{k''}$ elements of $\mathcal{C}''$. At the receiving end, we apply a decoder for $\mathcal{C}$ (for correcting up to $(d-1)/2$) errors) to a (possibly noisy) received word $\boldsymbol{c} + \boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{e}$ is the error word. Clearly, if $\mathsf{w}(\boldsymbol{e}) \leq (d-1)/2$, we will be able to recover $\boldsymbol{c}+\boldsymbol{x}$ successfully, thereby retrieving $\boldsymbol{u}$.

Obviously, such as scheme is useful only when k'' is indeed small: first, k'' affects the effective rate (given by $k'/n = (k-k'')/n$) and, secondly, the encoding process—as described—is exponential in k''. Yet, not always is there a decomposition of $\mathcal{C}$ as in (12) that results in a small dimension k'' of $\mathcal{C}''$ (in fact, for some codes $\mathcal{C}$, such a composition does not exist at all).

A possible solution would then be to reverse the design process and start by first selecting the code $\mathcal{C}'$ so that it has the desired rate $R = k'/n$ and a "slightly" higher minimum distance d' than the desired value d. In addition, we assume that there is an efficient (i.e., polynomial-time) decoding algorithm $\mathcal{D}'$ for $\mathcal{C}'$ that corrects any pattern of up to $(d-1)/2$ errors.

Next, we select $\mathcal{C}''$ to be a random linear code spanned by $k'' = \lceil \frac{3}{2} \log_2 n \rceil + \rho(n)$ words that are chosen independently and uniformly from $\mathbb{F}^n$, for some function $\rho(n) = o(\log n)$ that grows to infinity. By Theorem 3.1, the code $\mathcal{C}''$ will be a balancing set with probability $1 - \mathsf{exp}(-\rho(n)) = 1 - o(1)$, and the choice of k'' guarantees that an exhaustive search for the balancing word $\boldsymbol{x}$ during encoding will take $O(n^{3/2+\epsilon})$ iterations, for an arbitrarily small $\epsilon > 0$ (if the search fails—an event that may occur with probability $o(1)$—we can simply replace the code $\mathcal{C}''$). The receiving end can be informed of the choice of the code $\mathcal{C}''$ by, say, using pseudo-randomness instead of randomness (and flagging a skip when failing to find a balancing word $\boldsymbol{x}$).

It remains to consider the distance properties of the direct sum $\mathcal{C} = \mathcal{C}' \oplus \mathcal{C}''$; specifically, we need the subset of balanced words in $\mathcal{C}$ to have minimum distance at least d; in particular, every balanced word in $\mathcal{C}$ should have a unique decomposition of the form $\boldsymbol{c}+\boldsymbol{x}$ where $\boldsymbol{c} \in \mathcal{C}'$ and $\boldsymbol{x} \in \mathcal{C}''$. When this condition holds, the decoding can proceed as follows. Given a received word $\boldsymbol{y} \in \mathbb{F}^n$, we enumerate over all words $\boldsymbol{x} \in \mathcal{C}''$ and then apply the decoder $\mathcal{D}'$ to each difference $\boldsymbol{y} - \boldsymbol{x}$. Decoding will be successful if the number of errors did not exceed $(d-1)/2$, and the decoding complexity will be $O(n^{3/2+\epsilon})$ times the complexity of $\mathcal{D}'$.

[1]For the scheme to work, it actually suffices that words in $\mathcal{C}''$ balance only the elements of $\mathcal{C}'$, rather than all the words in $\mathbb{F}^n$.

The next lemma considers the case where the code $\mathcal{C}'$ lies below the Gilbert–Varshamov bound. Hereafter, $V(n,t)$ stands for $\sum_{i=0}^{t}\binom{n}{i}$.

Lemma 5.1. *Suppose that $\mathcal{C}'$ is a linear $[n,k',d']$ code over $\mathbb{F}$ that satisfies $2^{k'}\cdot V(n,d'-1)\le 2^n$. For every $d\le d'$, the minimum distance $\mathsf{d}(\cdot)$ of (the random code) $\mathcal{C}=\mathcal{C}'\oplus\mathcal{C}''$ satisfies*

$$\mathsf{Prob}\left\{\mathsf{d}(\mathcal{C})<d\right\}<2^{k''}\cdot\frac{V(n,d-1)}{V(n,d'-1)}\,.$$

Proof. The code $\mathcal{C}$ contains $|\mathcal{C}|-|\mathcal{C}'|$ random codewords, each being uniformly distributed over $\mathbb{F}^n$ and therefore each having probability $V(n,d-1)/2^n$ to be of Hamming weight less than d. The result follows from the union bound. □

It is well known (see [12, p. 310]) that for any integer $t=\theta n\le n/2$,

$$\frac{1}{\sqrt{2n}}\cdot 2^{n\mathsf{H}(\theta)}\le V(n,t)\le 2^{n\mathsf{H}(\theta)}\,,$$

where $\mathsf{H}:[0,1]\to[0,1]$ is the binary entropy function defined earlier. Hence, taking $k''\le(\frac{3}{2}+\epsilon)\log_2 n$, we get from Lemma 5.1 and the concavity of $z\mapsto\mathsf{H}(z)$ that

$$\mathsf{Prob}\left\{\mathsf{d}(\mathcal{C})<d\right\}<\sqrt{2}\cdot n^{2+\epsilon}\Big(\frac{d'-1}{n-d'+1}\Big)^{d'-d}\,.$$

Thus, to achieve a vanishing probability, $\mathsf{Prob}\left\{\mathsf{d}(\mathcal{C})<d\right\}$, of ending up with a "bad" code $\mathcal{C}$ as n goes to infinity, it suffices to take $d'=d+O(\log n)$ when d/n is fixed and bounded away from zero, or $d'=d+O(1)$ when d is fixed.

Remark 5.1. Instead of a decoding process whereby we enumerate over the codewords of $\mathcal{C}''$ and then apply the decoder $\mathcal{D}'$, we could use a decoder for the whole direct sum $\mathcal{C}$, if techniques such as iterative decoding are applicable to $\mathcal{C}$: in such circumstances, the advantage of the linearity of $\mathcal{C}$ is apparent. Linearity certainly helps if we are interested only in error detection rather than full correction, in which case the decoding amounts to just computing a syndrome with respect to any parity-check matrix of $\mathcal{C}$. □

Appendices

A Proof of Lemma 2.3

Proof. We have,

$$\begin{aligned}|\mathcal{B}(\mathcal{C}+\mathbb{F}\boldsymbol{x})| &= |\mathcal{B}(\mathcal{C}\cup(\mathcal{C}+\boldsymbol{x}))|\\ &= |\mathcal{B}(\mathcal{C})|+|\mathcal{B}(\mathcal{C}+\boldsymbol{x})|-|\mathcal{B}(\mathcal{C})\cap\mathcal{B}(\mathcal{C}+\boldsymbol{x})|\\ &= 2|\mathcal{B}(\mathcal{C})|-|\mathcal{B}(\mathcal{C})\cap(\mathcal{B}(\mathcal{C})+\boldsymbol{x})|\,.\end{aligned}$$

Hence,

$$\sum_{\boldsymbol{x}\in\mathbb{F}^n} |\mathcal{B}(\mathcal{C}+\mathbb{F}\boldsymbol{x})| = 2^{n+1}|\mathcal{B}(\mathcal{C})| - \sum_{\boldsymbol{x}\in\mathbb{F}^n} |\mathcal{B}(\mathcal{C}) \cap (\mathcal{B}(\mathcal{C})+\boldsymbol{x})| \ .$$

Now,

$$\begin{aligned}
&\sum_{\boldsymbol{x}\in\mathbb{F}^n} |\mathcal{B}(\mathcal{C}) \cap (\mathcal{B}(\mathcal{C})+\boldsymbol{x})| \\
&\qquad = |\{(\boldsymbol{x},\boldsymbol{y}) : \boldsymbol{x}\in\mathbb{F}^n, \boldsymbol{y}\in\mathcal{B}(\mathcal{C}), \boldsymbol{y}\in\mathcal{B}(\mathcal{C})+\boldsymbol{x}\}| \\
&\qquad = |\{(\boldsymbol{x},\boldsymbol{y}) : \boldsymbol{x}\in\mathbb{F}^n, \boldsymbol{y}\in\mathcal{B}(\mathcal{C}), \boldsymbol{x}\in\mathcal{B}(\mathcal{C})+\boldsymbol{y}\}| \\
&\qquad = |\mathcal{B}(\mathcal{C})|^2 \ .
\end{aligned}$$

Therefore,

$$\sum_{\boldsymbol{x}\in\mathbb{F}^n} |\mathcal{B}(\mathcal{C}+\mathbb{F}\boldsymbol{x})| = 2^{n+1}|\mathcal{B}(\mathcal{C})| - |\mathcal{B}(\mathcal{C})|^2 \ .$$

Using the definition of $Q(\cdot)$ the lemma is proved. □

B Proof of Theorem 3.5

We prove Theorem 3.5 below, starting by recalling the reduction that is used in [14] to show the intractability of computing the covering radius of a linear code.

Let $\mathcal{G} = (V_1{:}V_2{:}V_3, E)$ be a tripartite hyper-graph with a vertex set which is the union of the disjoint sets V_1, V_2, and V_3 of the same size t, and a hyper-edge set $E = \{e_1, e_2, \ldots, e_m\} \subseteq V_1 \times V_2 \times V_3$.

The reduction in [14] maps $\mathcal{G}$ into a $3t \times 8m$ parity-check matrix $H = H_{\mathcal{G}} = (\, H_e \,)_{e\in E}$, where each block H_e is a $3t \times 8$ matrix over $\mathbb{F}$ whose rows and columns are indexed by $u \in V_1 \cup V_2 \cup V_3$ and $(a_1\, a_2\, a_3) \in \mathbb{F}^3$, respectively, and is computed from the hyper-edge $e = (v_{e,1}, v_{e,2}, v_{e,3})$ as follows:

$$(H_e)_{u,(a_1a_2a_3)} = \begin{cases} 0 & \text{if } u \neq v_{e,\ell} \text{ for } \ell = 1,2,3 \\ a_\ell & \text{if } u = v_{e,\ell} \end{cases} \ .$$

(Namely, the three nonzero rows in H_e are indexed by the vertices that are incident with the hyper-edge e, and these rows form a 3×8 matrix whose columns range over all the elements of $\mathbb{F}^3$.)

A *matching* in $\mathcal{G}$ is a subset $\mathcal{M} \subseteq E$ of size t such that no two hyper-edges in $\mathcal{M}$ are incident with the same vertex (thus, every vertex of $\mathcal{G}$ is incident with exactly one hyper-edge in $\mathcal{M}$).

For our purposes, we can assume that every vertex in $\mathcal{G}$ is incident with at least one hyper-edge (or else no matching exists). Under these conditions, $m \geq t$ and the matrix H has full rank (since it contains the identity matrix of order $3t$).

The proof in [14] is based on the following two facts:

(i) There is a matching $\mathcal{M}$ in $\mathcal{G}$ if and only if the all-one column vector $\mathbf{1}$ in $\mathbb{F}^{3t}$ can be written as a sum of (exactly) t columns of H (note that $\mathbf{1}$ cannot be the sum of less than t columns). Those columns then must be those that are indexed by $(1\,1\,1)$ in all blocks H_e such that $e \in \mathcal{M}$.

(ii) If $\mathcal{M}$ is a matching in $\mathcal{G}$ then every column vector in $\mathbb{F}^{3t}$ can be written as a sum $\sum_{e\in\mathcal{M}} \boldsymbol{h}_e$, where each $\boldsymbol{h}_e$ is a column in H_e.

Let $\mathcal{C} = \mathcal{C}_{\mathcal{G}}$ be the linear $[8m, 8m{-}3t]$ code over $\mathbb{F}$ with a parity-check matrix H. It readily follows from facts (i) and (ii) that $\mathcal{G}$ has a matching if and only if every coset of $\mathcal{C}$ within $\mathbb{F}^{8m}$ has a word of Hamming weight t.

From facts (i)–(ii) we get the following lemma.

Lemma B.1. *Suppose that $t > 1$ and that $\mathcal{G}$ contains a matching. Then every column vector in $\mathbb{F}^{3t}$ can be obtained as a sum of w distinct columns in H, for every w in the range $t \le w \le 8m{-}t$.*

Proof. Let $\mathcal{M}$ be a matching which is assumed to exist in $\mathcal{G}$. Given $w \in \{t, t{+}1, \ldots, 8m{-}t\}$, write

$$\sigma = \min\{8(m{-}t), w{-}t\} \ ,$$

and let $\boldsymbol{x}$ be a column vector in $\mathbb{F}^{3t}$ which is the sum of σ columns in H that do *not* belong to the t blocks H_e that correspond to $e \in \mathcal{M}$. Also, write

$$\tau = w - \sigma = \begin{cases} t & \text{if } w \le 8m{-}7t \\ w - 8(m{-}t) & \text{otherwise} \end{cases} \ ,$$

and note that $t \le \tau \le 7t$.

Given an arbitrary column vector $\boldsymbol{s} \in \mathbb{F}^{3t}$, we show that there are w distinct columns in H that sum to $\boldsymbol{s}$. By fact (ii), for every $e \in \mathcal{M}$ there is a column $\boldsymbol{h}_e$ in H_e such that

$$\boldsymbol{s} = \boldsymbol{x} + \sum_{e\in\mathcal{M}} \boldsymbol{h}_e \ . \tag{13}$$

Furthermore, it follows from the structure of each block H_e that when $\boldsymbol{h}_e \neq \mathbf{0}$, then for every integer r in the range $1 \le r \le 7$ there exist r distinct columns $\boldsymbol{h}_{e,1}, \boldsymbol{h}_{e,2}, \ldots, \boldsymbol{h}_{e,r}$ in H_e such that

$$\boldsymbol{h}_e = \sum_{j=1}^{r} \boldsymbol{h}_{e,j} \ .$$

The same holds also when $\boldsymbol{h}_e = \mathbf{0}$ for values of r in $\{0, 1, 3, 4, 5, 7, 8\}$.

We conclude that we can find t nonnegative integers $(r_e)_{e\in\mathcal{M}}$ such that the following two conditions hold:

- $\sum_{e\in\mathcal{M}} r_e = \tau$ ($\in \{t, t{+}1, \ldots, 7t\}$), and
- For each $e \in \mathcal{M}$, the column vector $\boldsymbol{h}_e$ can be written as a sum of (exactly) r_e distinct columns of H_e.

Thus, the right-hand side of (13) can be expressed as a sum of $\sigma + \sum_{e\in\mathcal{M}} r_e = \sigma + \tau = w$ distinct columns in H. □

Proof of Theorem 3.5. Given a hyper-graph $\mathcal{G}$, consider the linear $[16m{-}2t, 8m{-}3t]$ code $\mathcal{C}'_{\mathcal{G}}$ over $\mathbb{F}$ with an $(8m{+}t) \times (16m{-}2t)$ parity-check matrix

$$H' = H'_{\mathcal{G}} = \left(\begin{array}{c|c} 0 & I \\ \hline H & 0 \end{array}\right) ,$$

where $H = H_{\mathcal{G}}$ and I is the identity matrix of order $8m{-}2t$. Next, we show that there is a matching in $\mathcal{G}$ if and only if every coset of $\mathcal{C}'_{\mathcal{G}}$ contains a balanced word (i.e., a word of Hamming weight $8m{-}t$).

Suppose that $\mathcal{G}$ contains a matching $\mathcal{M}$. We show that every column vector $\boldsymbol{s} \in \mathbb{F}^{8m+t}$ can be expressed as a sum of (exactly) $8m{-}t$ distinct columns of H'. Write $\boldsymbol{s}^T = (\boldsymbol{s}_1^T \mid \boldsymbol{s}_2^T)$, where $\boldsymbol{s}_1$ consists of the first $8m{-}2t$ entries of $\boldsymbol{s}$ (and $\boldsymbol{s}_2$ consists of the remaining $3t$ entries). By Lemma B.1, there exist $w = 8m{-}t{-}\mathsf{w}(\boldsymbol{s}_1)$ distinct columns in H that sum to $\boldsymbol{s}_2$. Hence, by the structure of H' it follows that H' contains $w + \mathsf{w}(\boldsymbol{s}_1) = 8m{-}t$ columns that sum to $\boldsymbol{s}$.

Conversely, suppose that every coset of $\mathcal{C}'_{\mathcal{G}}$ contains a balanced word. In particular, this means that the all-one vector in $\mathbb{F}^{8m+t}$ can be expressed as a sum of $8m{-}t$ columns of H'. Now, the last $8m{-}2t$ columns of H' must be included in this sum; this, in turn, implies that the all-one vector $\mathbf{1}$ in $\mathbb{F}^{3t}$ can be written as a sum of t columns of H. The result follows from fact (i). □

C Proof of Lemma 4.3

Proof. We will follow along the steps of the proof of Lemma 3.2, except that (6) needs to be replaced by a different upper bound which we now derive. Given some $\delta \in [0, \frac{1}{2})$, let $\boldsymbol{x}$ and $\boldsymbol{x}'$ be two distinct words in $\mathcal{C}_0$ with $\mathsf{d}(\boldsymbol{x}, \boldsymbol{x}') = \tau n$ where $\frac{1}{2} - \delta \le \tau \le \frac{1}{2} + \delta$. The number of words $\boldsymbol{y} \in \mathbb{F}^n$ such that $\mathsf{d}(\boldsymbol{x}, \boldsymbol{y}) = i$ and $\mathsf{d}(\boldsymbol{x}', \boldsymbol{y}) = j$ is given by

$$p_{i,j}^{(\tau n)} = \binom{\tau n}{(j{-}i{+}\tau n)/2} \binom{(1{-}\tau)n}{(i{+}j{-}\tau n)/2}$$

(here we assume that the binomial coefficient $\binom{m}{k}$ is equal to 0 unless m and k are both nonnegative integers and $m \ge k$). Hence,

$$|\mathcal{B}_\lambda(\boldsymbol{x}) \cap \mathcal{B}_\lambda(\boldsymbol{x}')| \le \sum_{i=(n/2)-\lambda}^{(n/2)+\lambda} \sum_{j=(n/2)-\lambda}^{(n/2)+\lambda} p_{i,j}^{(\tau n)} .$$

It can be easily verified that when τn is even then

$$\max_{i,j} p_{i,j}^{(\tau n)} = \binom{\tau n}{\tau n/2}\binom{(1-\tau)n}{(1-\tau)n/2} \stackrel{(1)}{\leq} \frac{2^{\tau n}}{\sqrt{\pi\tau n/2}} \cdot \frac{2^{(1-\tau)n}}{\sqrt{\pi(1-\tau)n/2}} ,$$

and when τn is odd then

$$\begin{aligned} \max_{i,j} p_{i,j}^{(\tau n)} &= \binom{\tau n}{(\tau n+1)/2}\binom{(1-\tau)n}{((1-\tau)n+1)/2} \\ &= \frac{1}{4}\binom{\tau n+1}{(\tau n+1)/2}\binom{(1-\tau)n+1}{((1-\tau)n+1)/2} \\ &\stackrel{(1)}{\leq} \frac{2^{\tau n}}{\sqrt{\pi\tau n/2}} \cdot \frac{2^{(1-\tau)n}}{\sqrt{\pi(1-\tau)n/2}} . \end{aligned}$$

In either case we have:

$$\begin{aligned} |\mathcal{B}_\lambda(\boldsymbol{x}) \cap \mathcal{B}_\lambda(\boldsymbol{x}')| &\leq (2\lambda+1)^2 \max_{i,j} p_{i,j}^{(\tau n)} \\ &\leq (2\lambda+1)^2 \frac{2^{n+1}}{\pi n\sqrt{\tau(1-\tau)}} \\ &\leq (2\lambda+1)^2 \frac{2^{n+2}}{\pi n\sqrt{1-4\delta^2}} . \end{aligned} \tag{14}$$

In addition, from (9) and (10) we get:

$$|\mathcal{B}_\lambda(\boldsymbol{x})| \geq \frac{2\lambda+1}{\sqrt{2n}} \cdot 2^{n-O(1)} . \tag{15}$$

We now proceed as in the proof of Lemma 3.2, with (14) replacing (6) and with (15) replacing the lower bound in (1): by de Caen's lower bound [4] we get a bound which is similar to (7), in which we plug $\ell = \lceil \frac{1}{2}\log_2 n - \log_2(2\lambda+1) \rceil$. The result follows. □

References

[1] N. Alon, E.E. Bergmann, D. Coppersmith, and A.M. Odlyzko, *Balancing sets of vectors, IEEE Trans. Inform. Theory,* 34 (1988), 128–130.

[2] T. Berger, *Rate Distortion Theory,* Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[3] V.M. Blinovskii, *Covering the Hamming space with sets translated by linear code vectors, Probl. Inform. Transm.,* 26 (1990), 196–201.

[4] D. de Caen, *A lower bound on the probability of a union, Disc. Math.,* 169 (1997), 217–220.

[5] G. Cohen, *A nonconstructive upper bound on covering radius, IEEE Trans. Inform. Theory,* 29 (1983), 352–353.

[6] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering Codes,* North-Holland, Amsterdam, 1997.

[7] P. Delsarte and P. Piret, *Do most binary linear codes achieve the Goblick bound on the covering radius?, IEEE Trans. Inform. Theory,* 32 (1986), 826–828.

[8] T.J. Goblick, Jr., *Coding for a discrete information source with a distortion measure,* Ph.D. dissertation, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1962.

[9] K.A.S. Immink, *Codes for Mass Data Storage Systems,* Second Edition, Shannon Foundation Publishers, Eindhoven, The Netherlands, 2004.

[10] D.E. Knuth, *Efficient balanced codes, IEEE Trans. Inform. Theory,* 32 (1986), 51–53.

[11] E.M. Kurtas and B. Vasic (Editors), *Advanced Error Control Techniques for Data Storage Systems,* CRC Press, Boca Raton, Florida, 2006.

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

[13] B.H. Marcus, R.M. Roth, and P.H. Siegel, *Constrained systems and coding for recording channels,* in *Handbook of Coding Theory,* V.S. Pless and W.C. Huffman (Editors), Elsevier Scientific Publishers, Amsterdam, 1998.

[14] A.M. McLoughlin, *The complexity of computing the covering radius of a code, IEEE Trans. Inform. Theory,* 30 (1984), 800–804.

[15] N. Sendrier, *Encoding information into constant weight words, Proc. 2005 Int'l Symp. Inform. Theory (ISIT 2005),* Adelaide, Australia (2005), 435–438.