Locally Balanced Constraints

Ryan Gabrys,* **Han Mao Kiah**,† **Alexander Vardy**,* **Eitan Yaakobi**,‡ and **Yiwei Zhang**§

*University of California San Diego, La Jolla, CA 92093
†Nanyang Technological University, Singapore 637371
‡Technion — Israel Institute of Technology, Haifa, Israel 3200003
§School of Cyber Science and Technology, Shandong University, Qingdao, Shandong, China 266237
ryan.gabrys@gmail.com, hmkiah@ntu.edu.sg, avardy@ucsd.edu, yaakobi@cs.technion.ac.il, ywzhang@sdu.edu.cn

***Abstract*—Three new constraints are introduced in this paper. These constraints are characterized by limitations on the Hamming weight of every subword of some fixed even length ℓ. In the (ℓ,δ)-*locally-balanced constraint*, the Hamming weight of every length-ℓ subword is bounded between $\ell/2-\delta$ and $\ell/2+\delta$. The *strong*-(ℓ,δ)-*locally-balanced constraint* imposes the locally-balanced constraint for any subword whose length is at least ℓ. Lastly, the Hamming weight of every length-ℓ subword which satisfies the (ℓ,δ)-*locally-bounded constraint* is at most $\ell/2-\delta$. It is shown that the capacity of the strong-(ℓ,δ)-locally-balanced constraint does not depend on the value of ℓ and is identical to the capacity of the $(2\delta+1)$-*RDS constraint*. The latter constraint limits the difference between the number of zeros and ones in every prefix of the word to be at most $2\delta+1$. This value is also a lower bound on the capacity of the (ℓ,δ)-locally-balanced constraint, while a corresponding upper bound is given as well. Lastly, it is shown that if δ is not large enough, namely for $\delta<\sqrt{\ell}/2$, then the capacity of the (ℓ,δ)-locally-bounded constraint approaches 1 as ℓ increases.**

I. Introduction

This paper initiates the study of a few constraints which impose limitations on binary words in the form of the Hamming weight of their subwords. For an even positive integer ℓ and a nonnegative integer δ, a binary word $\boldsymbol{x}$ is said to satisfy the (ℓ,δ)-*locally-balanced constraint* if the Hamming weight of each of its length-ℓ subwords is between $\ell/2-\delta$ and $\ell/2+\delta$. Furthermore, $\boldsymbol{x}$ satisfies the *strong*-(ℓ,δ)-*locally-balanced constraint* if the Hamming weight of each of its length-ℓ' subwords is between $\ell'/2-\delta$ and $\ell'/2+\delta$, for all even $\ell'\geqslant\ell$. In the third constraint, $\boldsymbol{x}$ satisfies the (ℓ,δ)-*locally-bounded constraint* if the Hamming weight of every length-ℓ subword is at most $\ell/2-\delta$, while δ can be both negative and positive. The main goal of this paper is studying the capacity values of these three constraints for several values of the parameters of ℓ and δ. First notice that if $\delta=0$ then the capacity of the locally-balanced and strong-locally-balanced constraints is 0. Thus, in order to allow non-zero capacity values, strict balanced constraints cannot be achieved and one has to allow every subword to be *almost* balanced.

For every fixed values of ℓ and δ the locally-balanced and locally-bounded constraints can be represented by a graph of a constrained system and thus the capacity in these two cases can be solved by calculating the largest eigenvalue of the adjacency matrix of the graph. For example, for $\ell=4$ and $\delta=1$, in the locally-balanced constraint the forbidden patterns are 0000 and 1111, while for the locally-bounded constraint the only permitted length-4 subwords are $0000, 1000, 0100, 0010, 0001$. See Table I for the capacity values of the locally-balanced constraint when $4\leqslant\ell\leqslant 14$ and $\delta=1,2$. However, the strong-locally-balanced constraint cannot be solved by this method and furthermore we will be mostly interested in studying the capacity values of these constraints when ℓ increases. Hence, representing the constraints by a graph is no longer a feasible solution.

In studying these constraints we will draw a close connection between the capacity of the first two constraints and the one of the well-studied *running digital sum* (*RDS*) constraint. The RDS of a binary word $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)$ is given by the word $\boldsymbol{s}=(s_0,s_1,\ldots,s_n)$, where $s_0=0$ and $s_i=\sum_{j=1}^{i}(-1)^{1-x_j}$. Given some $\delta>0$, a word is said to satisfy the δ-*RDS constraint* if the difference between the values of $\max_{0\leqslant i\leqslant n}\{s_i\}$ and $\min_{0\leqslant i\leqslant n}\{s_i\}$ is at most δ.

In this work, we completely solve the capacity of the strong-locally-balanced constraint for all ℓ and δ. It is shown that for a given value of δ, the capacity does not depend on the value ℓ and equals the capacity of the $(2\delta+1)$-RDS constraint. This result clearly provides also a lower bound on the capacity of the locally-balanced constraint. On the other hand, it is shown that $\frac{\log(2+2^{\mathbb{C}^{RDS}(2\delta)})}{2}$, when $\mathbb{C}^{RDS}(2\delta)$ is the capacity of the 2δ-RDS constraint, is an upper bound on the capacity of the (ℓ,δ)-locally-balanced constraint for ℓ large enough. In the locally bounded constraint, it is shown that for $\delta<\sqrt{\ell}/2$, the capacity of the (ℓ,δ)-locally-bounded constraint approaches 1 when ℓ increases.

One of the prominent motivations to study these constraints originates from DNA storage, an area which attracted significant interest lately due to the extreme density and durability of DNA [13]. The DNA storage channel can be divided into three main phases: *synthesis*, *storage*, and *sequencing*. In the synthesis phase, data is converted into a set of DNA strands, which are sequences over the nucleotide alphabet A, T, C, G. In the storage phase, the strands are kept in a container with compartments; notably, the arrangement or order of the strands in the set is *not* preserved. Finally, in the sequencing phase, the strands are collected, and the original data is (hopefully) recovered.

During the storage phase, media degradation, and in par-

The work of E. Yaakobi was partially supported by the United States-Israel BSF grant 2018048. Y. Zhang is also with Key Laboratory of Cryptologic Technology and Information Security, Ministry of Education, Shandong University.

ticular breaks, can arise in DNA due to factors that include radiation, humidity, and high temperatures. In [7], the authors proposed to encapsulate the stored DNA in a silica substrate and then to employ custom error-correcting codes to mitigate the effects of these errors. Another approach to dealing with media degradation is to generate strands of DNA that have approximately balanced GC-content[1], and this approach has been leveraged in several existing works such as [5], [17], [18]. However, in some cases it may be desirable to have more stringent GC-balancing requirements in order to prolong the lifetime of the DNA strands. In fact, naturally occurring DNA strands have approximately balanced GC-content for short k-mers, i.e., subwords, of length at most 10, and it has been postulated that this balancing is necessary to ensure the stability of the DNA structure over time [2]–[4], [9], [12], [14]. This motivates the need to design DNA strands that are *locally GC-balanced*.

We note that designing locally balanced DNA strands is related to constructing codes with an equal number of zeros and ones, which is one of the more well-studied problems in coding theory [8], [10]. The key feature of our problem is that the balancing constraint is enforced *locally*. In another context, codes satisfying the locally-bounded constraints were studied to facilitate the simultaneous energy and information transfer in low power devices [6], [15], [16]. These codes were referred to as *sliding window codes* and of significance to this work is [16] where the authors provided lower bounds on the capacity of the locally-bounded constraint when $\delta = \epsilon\ell$.

The rest of the paper is organized as follows. In Section II the three constraints studied in the paper are formally defined together with several preliminary results and observations. In Section III we calculate the capacity of the strong-locally-balanced constraint for all ℓ and δ. In Section IV an upper bound on the capacity of the locally-balanced constraint is presented. Section V presents our results on the locally-bounded constraint. Lastly, Section VI concludes the paper and lists several open problems.

II. Definitions and Preliminaries

For a positive integer n, the set $\{1, 2, \ldots, n\}$ is denoted by $[n]$. For a word $\boldsymbol{x}$, its subword starting at the i-th index of length ℓ is denoted by $\boldsymbol{x}[i;\ell]$. The length of the word $\boldsymbol{x}$ is denoted by $|\boldsymbol{x}|$. The Hamming distance between two words $\boldsymbol{x}$ and $\boldsymbol{y}$ of the same length is denoted by $d(\boldsymbol{x},\boldsymbol{y})$ and the Hamming weight of $\boldsymbol{x}$ is $wt(\boldsymbol{x})$. Denote $\Sigma_2 = \{0,1\}$. The families of constraints which will be studied in this paper are formally defined in the next definition.

Definition 1.

1) *Let ℓ be an even positive integer and δ a nonnegative integer. A word $\boldsymbol{x}$ is said to satisfy the (ℓ,δ)-**locally-balanced constraint** (or is (ℓ,δ)-**locally balanced**) if for all $1 \leqslant i \leqslant |\boldsymbol{x}| - \ell + 1$, it holds that*
$$\ell/2 - \delta \leqslant wt(\boldsymbol{x}[i;\ell]) \leqslant \ell/2 + \delta.$$
2) *Let ℓ be an even positive integer and δ a nonnegative integer. A word $\boldsymbol{x}$ is said to satisfy the strong-(ℓ,δ)-**locally-balanced constraint** (or is strong-(ℓ,δ)-**locally balanced**) if for all even $\ell' \geqslant \ell$, the word $\boldsymbol{x}$ satisfies the (ℓ',δ)-locally-balanced constraint.*
3) *Let ℓ be a positive integer and δ an integer. A word $\boldsymbol{x}$ is said to satisfy the (ℓ,δ)-**locally-bounded constraint** (or is (ℓ,δ)-**locally bounded**) if for all $1 \leqslant i \leqslant |\boldsymbol{x}| - \ell + 1$, it holds that*
$$wt(\boldsymbol{x}[i;\ell]) \leqslant \ell/2 - \delta.$$

The set of all words (of any finite length) that are (ℓ,δ)-locally-balanced, strong-(ℓ,δ)-locally-balanced, (ℓ,δ)-locally-bounded is denoted by $\mathcal{S}^{bl}(\ell,\delta)$, $\mathcal{S}^{sbl}(\ell,\delta)$, $\mathcal{S}^{bd}(\ell,\delta)$, respectively. The capacity of the (ℓ,δ)-locally-balanced, strong-(ℓ,δ)-locally-balanced, (ℓ,δ)-locally-bounded constraint is defined to be
$$\mathbb{C}^{bl}(\ell,\delta) = \limsup_{n\to\infty} \frac{\log(|\mathcal{S}^{bl}(\ell,\delta) \cap \Sigma_2^n|)}{n},$$
$$\mathbb{C}^{sbl}(\ell,\delta) = \limsup_{n\to\infty} \frac{\log(|\mathcal{S}^{sbl}(\ell,\delta) \cap \Sigma_2^n|)}{n},$$
$$\mathbb{C}^{bd}(\ell,\delta) = \limsup_{n\to\infty} \frac{\log(|\mathcal{S}^{bd}(\ell,\delta) \cap \Sigma_2^n|)}{n},$$
respectively.

For $\delta = 0$, there are exactly $\binom{\ell}{\ell/2}$ words which satisfy the $(\ell,0)$-locally-balanced constraint for any length greater than $\ell-1$. Hence, for all ℓ, $\mathbb{C}^{bl}(\ell,0) = \mathbb{C}^{sbl}(\ell,0) = 0$. On the other hand, if $\delta \geqslant \ell/2$ then every word satisfies the (ℓ,δ)-locally-balanced constraint and therefore $\mathbb{C}^{bl}(\ell,\delta) = 1$. Similarly, for $\delta \leqslant -\ell/2$, $\mathbb{C}^{bd}(\ell,\delta) = 1$. For any fixed ℓ, the locally balanced constraint eliminates all subwords having either low or high weight. For example, for $\ell = 4$ and $\delta = 1$, the Hamming weight of every length-four subword has to be between one and three, that is, the forbidden subwords are 0000 and 1111. Similarly, for $\ell = 6$, the forbidden subwords are the length-6 words of Hamming weight $0, 1, 5$, and 6. Hence, for any fixed values of ℓ and δ this problem can be described as a constrained system that can be represented by a graph $G_{\ell,\delta}$. The graph in this case will be strongly connected, so the constrained system is irreducible. Therefore, by the Perron-Frobenius theorem, the capacity is given by $\log(\lambda)$, where λ is the largest eigenvalue of the adjacency matrix $A_{G_{\ell,\delta}}$ of the graph $G_{\ell,\delta}$ [11, Th. 3.4]. According to this analysis, we were able to calculate the capacity results for $4 \leqslant \ell \leqslant 14$ and $\delta = 1, 2$, which are listed in Table I. Similar calculations can be carried for the locally-bounded constraint.

TABLE I: Capacity Results for the Locally-Balanced Constraint

ℓ	4	6	8	10	12	14
$\delta = 1$	0.879	0.841	0.824	0.815	0.811	0.807
$\delta = 2$	1	0.975	0.958	0.947	0.939	0.933

While calculating the capacity of the locally-balanced and locally-bounded constraints can be accomplished for any fixed values of ℓ and δ, it will no longer be feasible when ℓ increases since the number of the states in the graph $G_{\ell,\delta}$ grows exponentially with ℓ. More than that, the strong-locally-balanced

[1]A strand of DNA is said to have approximately balanced GC-content if, in every substring of a prescribed length, approximately half the bases are either guanine or cytosine.

constraint cannot be solved this way since it has to satisfy the locally-balanced constraint for infinite values of ℓ. In the rest of the paper we will mostly be interested in studying the capacity of these constraints when the value of ℓ is large enough.

A simple construction of words satisfying both the locally-balanced and strong-locally-balanced constraints works as follows. Let us start with the code $\mathcal{B} = \{01, 10\}$ and let $\mathcal{C} = \mathcal{B}^* = \bigcup_{i=1}^{\infty} \mathcal{B}^i$. The asymptotic rate of the code $\mathcal{C}$ is 0.5 and all its codewords are $(\ell, 1)$-locally balanced for all $\ell \geqslant 4$. The next theorem summarizes several simple observations on the capacity values of these constraints, some of which have been explained above.

Theorem 2.

1) *For all* $\ell \geqslant 2$, $\mathbb{C}^{bl}(\ell, 0) = 0$.
2) *If* $\delta \geqslant \ell/2$ *then* $\mathbb{C}^{bl}(\ell, \delta) = 1$ *and* $\mathbb{C}^{bd}(\ell, -\delta) = 1$.
3) *For all* $\ell \geqslant 4$, $\mathbb{C}^{bl}(\ell, 1) \geqslant \mathbb{C}^{sbl}(\ell, 1) \geqslant 0.5$.
4) *If* $\delta_1 \leqslant \delta_2$ *then* $\mathbb{C}^{bl}(\ell, \delta_1) \leqslant \mathbb{C}^{bl}(\ell, \delta_2)$, $\mathbb{C}^{sbl}(\ell, \delta_1) \leqslant \mathbb{C}^{sbl}(\ell, \delta_2)$, *and* $\mathbb{C}^{bd}(\ell, \delta_2) \leqslant \mathbb{C}^{bd}(\ell, \delta_1)$.
5) *For all* ℓ, δ, *and* t, $\mathbb{C}^{bl}(t\ell, t\delta) \geqslant \mathbb{C}^{bl}(\ell, \delta)$, $\mathbb{C}^{sbl}(t\ell, t\delta) \geqslant \mathbb{C}^{sbl}(\ell, \delta)$, *and* $\mathbb{C}^{bd}(t\ell, t\delta) \geqslant \mathbb{C}^{bd}(\ell, \delta)$.

A special family of words which will play an important role in our construction and analysis is the set of words with bounded *running digital sum* (*RDS*). The RDS of a binary word $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is denoted by $\boldsymbol{s} = RDS(\boldsymbol{x}) = (s_0, s_1, \ldots, s_n)$, where $s_0 = 0$ and for all $1 \leqslant i \leqslant n$, s_i is defined by

$$s_i = \sum_{j=1}^{i} (-1)^{1-x_j} = 2wt((x_1, \ldots, x_i)) - i.$$

For a word $\boldsymbol{x}$, let $RDS^{\max}(\boldsymbol{x}) = \max_{0 \leqslant i \leqslant n}\{s_i\}$ and similarly $RDS^{\min}(\boldsymbol{x}) = \min_{0 \leqslant i \leqslant n}\{s_i\}$. The *disbalance* of the word $\boldsymbol{x}$, denoted by $\mathsf{dis}(\boldsymbol{x})$, is defined by the value $\mathsf{dis}(\boldsymbol{x}) = RDS^{\max}(\boldsymbol{x}) - RDS^{\min}(\boldsymbol{x})$. For a given positive integer δ, a word $\boldsymbol{x}$ is said to satisfy the *δ-RDS constraint (or is called a δ-RDS word)*, if $\mathsf{dis}(\boldsymbol{x}) \leqslant \delta$. The set of all finite length words satisfying the δ-RDS constraint is denoted as $\mathcal{S}^{RDS}(\delta)$ and the capacity of this constraint is

$$\mathbb{C}^{RDS}(\delta) = \limsup_{n \to \infty} \frac{\log(|\mathcal{S}^{RDS}(\delta) \cap \Sigma_2^n|)}{n}.$$

The value of $\mathbb{C}^{RDS}(\delta)$ is related to the well-known problem of counting the number of Dyck paths of bounded height (for a complete survey of related problems, see Chapter 10 of [1]). It has been proved that for $\delta > 0$, $\mathbb{C}^{RDS}(\delta) = \log(2\cos\frac{\pi}{\delta+2})$. For example, $\mathbb{C}^{RDS}(2) = 0.5$ and $\mathbb{C}^{RDS}(3) = \log(\varphi)$, where $\varphi = (1+\sqrt{5})/2$ is the golden ratio and thus $\mathbb{C}^{RDS}(3) = \log((1+\sqrt{5})/2) \approx 0.694$. In the next section it will be studied how δ-RDS words can be used to generate locally-balanced and strong-locally-balanced words.

III. The Strong-Locally-Balanced Constraint

In this section, we completely solve the capacity of the strong-locally-balanced constraint for all values of ℓ and δ. More specifically, it is established that every $(2\delta+1)$-RDS word is (ℓ, δ)-locally balanced for all $\delta \geqslant 0$ and $\ell \geqslant 4$. This already provides a lower bound for the locally-balanced and the strong-locally-balanced constraints. However, it was surprising to observe that even though the opposite direction does not hold, we established that this lower bound is indeed asymptotically tight for the strong-locally-balanced constraint.

In the following theorem, it is proved that the capacity of the RDS constraint serves as a lower bound on the capacity of both the locally-balanced and the strong-locally-balanced constraints.

Theorem 3. *For all* $\delta > 0, \ell \geqslant 4$,

$$\mathcal{S}^{RDS}(2\delta+1) \subseteq \mathcal{S}^{sbl}(\ell, \delta) \subseteq \mathcal{S}^{bl}(\ell, \delta),$$

and thus

$$\mathbb{C}^{bl}(\ell, \delta) \geqslant \mathbb{C}^{sbl}(\ell, \delta) \geqslant \mathbb{C}^{RDS}(2\delta+1).$$

In particular, $\mathbb{C}^{bl}(\ell, 1) \geqslant \mathbb{C}^{sbl}(\ell, 1) \geqslant \mathbb{C}^{RDS}(3) \approx 0.694$.

Proof: By definition we have $\mathcal{S}^{sbl}(\ell, \delta) \subseteq \mathcal{S}^{bl}(\ell, \delta)$. Let $\boldsymbol{x} \in \mathcal{S}^{RDS}(2\delta+1)$ and assume in the contrary that it does not satisfy the (ℓ, δ)-locally-balanced constraint for some even integer $\ell \geqslant 4$. Then it has a subword $\boldsymbol{x}[i; \ell]$ of weight $wt(\boldsymbol{x}[i; \ell]) \leqslant \ell/2 - \delta - 1$ or $wt(\boldsymbol{x}[i; \ell]) \geqslant \ell/2 + \delta + 1$. Then in its corresponding RDS sequence, for

$$s_{i-1} = 2wt((x_1, \ldots, x_{i-1})) - (i-1)$$

and

$$s_{i-1+\ell} = 2wt((x_1, \ldots, x_{i-1+\ell})) - (i-1+\ell)$$

it holds that

$$\begin{aligned} \Delta &= s_{i-1+\ell} - s_{i-1} \\ &= 2wt((x_1, \ldots, x_{i-1+\ell})) - 2wt((x_1, \ldots, x_{i-1})) - \ell \\ &= 2wt(\boldsymbol{x}[i; \ell]) - \ell, \end{aligned}$$

which leads to either $\Delta \leqslant -2\delta - 2$ or $\Delta \geqslant 2\delta + 2$. However, since $\boldsymbol{x}$ is a $(2\delta+1)$-RDS word it holds that $|s_{i-1+\ell} - s_{i-1}| \leqslant 2\delta + 1$, which results with a contradiction. Therefore, $\mathcal{S}^{RDS}(2\delta+1) \subseteq \mathcal{S}^{sbl}(\ell, \delta)$ and the theorem follows. ∎

Theorem 3 assures that every $(2\delta+1)$-RDS word is also a strong-(ℓ, δ)-locally-balanced word. Hence, every code for the RDS constraint can also be used in the construction of codes for the strong-locally-balanced and locally-balanced constraints. The next theorem shows that, for the strong-(ℓ, δ)-locally-balanced constraint, the set of $(2\delta+1)$-RDS words is indeed asymptotically optimal.

Theorem 4. *For all* $\delta > 0, \ell \geqslant 4$, $\mathbb{C}^{sbl}(\ell, \delta) = \mathbb{C}^{RDS}(2\delta+1)$. *In particular,* $\mathbb{C}^{sbl}(\ell, 1) = \mathbb{C}^{RDS}(3) \approx 0.694$.

Proof: According to the previous theorem, $\mathbb{C}^{sbl}(\ell, \delta) \geqslant \mathbb{C}^{RDS}(2\delta+1)$. Thus we only need to show that $\mathbb{C}^{sbl}(\ell, \delta) \leqslant \mathbb{C}^{RDS}(2\delta+1)$. For convenience we only prove for $\delta = 1$. The proof for a general δ is essentially the same.

1) When $\ell = 4$, for any $\boldsymbol{x} \in \mathcal{S}^{sbl}(4, 1)$, in its corresponding $\boldsymbol{s} = RDS(\boldsymbol{x})$ there are never two symbols s_i and s_j of difference 4 and hence $\boldsymbol{x}$ satisfies the 3-RDS constraint. That is, $\mathcal{S}^{sbl}(4, 1) \subseteq \mathcal{S}^{RDS}(3)$ and $\mathbb{C}^{sbl}(4, 1) = \mathbb{C}^{RDS}(3)$.

2) When $\ell = 6$, compared to the previous case, now we may have consecutive 0000 or 1111 inside a word $\boldsymbol{x} \in \mathcal{S}^{sbl}(6,1)$. The key is that the number of occurrences of 0000 and 1111 is at most once each. Suppose otherwise, say a word $\boldsymbol{x} \in \mathcal{S}^{sbl}(6,1)$ is of the form

$$(\cdots, 0000, x_{t+1}, \cdots, x_{t+k}, 0000, x_{t+k+5}, \cdots).$$

Note that since every subword of length 6 is locally balanced then we must have $x_{t+1} = x_{t+2} = x_{t+k-1} = x_{t+k} = 1$. If $k \leqslant 5$, then the only possible choices of $(x_{t+1}, \cdots, x_{t+k})$ are $(11), (111), (1111), (11011), (11111)$, where each choice will lead to a subword violating the strong-$(6,1)$-locally-balanced constraint; if $k \geqslant 6$ is even, then the weight of the subword $(x_{t+1}, \cdots, x_{t+k})$ is at most $\frac{k}{2}+1$ and thus the subword $(0000, x_{t+1}, \cdots, x_{t+k}, 0000)$ has weight at most $\frac{k}{2}+1 < \frac{k+8}{2}-1$, thus violating the locally-balanced constraint of length $k+8$; if $k \geqslant 6$ is odd, then the weight of the subword $(x_{t+1}, \cdots, x_{t+k})$ is at most $\frac{k-1}{2}+2$ and thus the subword $(0000, x_{t+1}, \cdots, x_{t+k}, 0000, x_{t+k+5})$ has weight at most $\frac{k-1}{2}+2+1 < \frac{k+9}{2}-1$, thus violating the locally-balanced constraint of length $k+9$. To sum up, there is at most one subword 0000 inside a word $\boldsymbol{x} \in \mathcal{S}^{sbl}(6,1)$. Similarly there is at most one subword 1111.

Thus the number of strong-$(6,1)$-locally-balanced words can be upper bounded as follows: 1) we first select the positions of the unique 0000 and 1111, for which the number of choices is at most n^2; 2) then the word is of the form $(\boldsymbol{x}_1, 0000, \boldsymbol{x}_2, 1111, \boldsymbol{x}_3)$, where $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ are strong-$(4,1)$-locally-balanced words of lengths n_1, n_2, n_3 (since initially they are strong-$(6,1)$-locally-balanced and furthermore there is no subword 0000 or 1111 inside them). If any n_i if finite, then the number of choices for $\boldsymbol{x}_i$ is only a constant. If n_i goes to infinity (no matter how slow compared to the growth of n), then asymptotically the number of choices for $\boldsymbol{x}_i$ is $2^{0.694 n_i}$. To sum up, the number of strong-$(6,1)$-locally-balanced words is upper bounded by $C \cdot n^2 \cdot 2^{0.694n}$ where C is a constant, and thus $\mathbb{C}^{sbl}(6,1) = \mathbb{C}^{RDS}(3)$.

3) Due to lack of space we only briefly describe the idea of the case with arbitrary ℓ, which is essentially similar as the previous case. Inside a word $\boldsymbol{x} \in \mathcal{S}^{sbl}(\ell,1)$ we may have subwords of length $\ell' \in \{4, 6, \ldots, \ell-2\}$ with weight less than $\frac{\ell'}{2}-1$ and we call these subwords as *negative subwords*. We may also have some subwords of length $\ell' \in \{4, 6, \ldots, \ell-2\}$ with weight more than $\frac{\ell'}{2}+1$ and we call these subwords as *positive subwords*. The key is again the limitations on these negative and positive subwords, in the sense that two negative subwords (or two positive subwords) can only have less than ℓ coordinates between them. Thus again most part of the word should behave as strong-$(4,1)$-locally-balanced words of some certain length and the asymptotic result follows.

IV. The Locally-Balanced Constraint

This section presents an upper bound on the capacity of the locally-balanced constraint that holds for all values of ℓ. Let us consider the set of all length-(2ℓ) $(\ell,1)$-locally-balanced words, which will be denoted by $A_{2\ell}$. That is,

$$A_{2\ell} = \mathcal{S}^{bl}(\ell,1) \cap \Sigma_2^{2\ell}.$$

First, the following lemma is shown.

Lemma 5. *For all $\ell \geqslant 4$,*

$$|A_{2\ell}| \leqslant 3 \cdot \left(2^{\ell \log(2+\sqrt{2})} + 2^{\frac{3\ell-1}{2}}\right) \approx 2^{1.77\ell}.$$

Proof: We will consider every word $\boldsymbol{a}$ in $A_{2\ell}$ both as a length-2ℓ word as well as a $2 \times \ell$ array of the form $\boldsymbol{a} = (\boldsymbol{a}_1, \boldsymbol{a}_2)$, where $\boldsymbol{a}_1, \boldsymbol{a}_2$ denotes the first, second row of the array, respectively. Clearly, it holds that $\ell/2 - 1 \leqslant wt(\boldsymbol{a}_1), wt(\boldsymbol{a}_2) \leqslant \ell/2 + 1$. Furthermore, for $1 \leqslant i \leqslant \ell+1$, we have that

$$\frac{\ell}{2} - 1 \leqslant wt(\boldsymbol{a}[i;\ell]) = wt(\boldsymbol{a}_1) + \sum_{j=1}^{i-1}(a_{2,j} - a_{1,j}) \leqslant \frac{\ell}{2} + 1,$$

where the subtraction $a_{2,j} - a_{1,j}$ is over the integers. Let $\boldsymbol{d}$ be the difference word between the words $\boldsymbol{a}_2$ and $\boldsymbol{a}_1$, over the integers, that is, $\boldsymbol{d} = \boldsymbol{a}_2 - \boldsymbol{a}_1 \in \{-1, 0, 1\}^\ell$. If we extend the definition of the RDS constraint also for non-binary words, we get that the word $\boldsymbol{d}$ satisfies the 2-RDS constraint. Furthermore, if $d_i \in \{-1, 1\}$ then the value of $a_{1,i}, a_{2,i}$ is uniquely determined, while for $d_i = 0$, these two bits have two options. Assume that the number of 2-RDS words with $0 \leqslant i \leqslant \ell$ zeros is n_i, then we conclude that the size of the set $A_{2\ell}$ is at most

$$\sum_{i=0}^{\ell} 2^i \cdot n_i.$$

Hence, we are only left with finding the value of n_i.

Assume that the word $\boldsymbol{d}$ has i zeros and let $I \subseteq [\ell]$ be the set of the zero positions in $\boldsymbol{d}$. Let $\boldsymbol{d}'$ be the projection of $\boldsymbol{d}$ on the set $I^C = [\ell] \setminus I$, so $\boldsymbol{d}' = \boldsymbol{d}_{I^C}$. Now, we can consider $\boldsymbol{d}'$ as a length-$(\ell - i)$ binary word that satisfies the 2-RDS constraint. The number of such words is at most $3 \cdot 2^{\lceil \frac{\ell-i}{2} \rceil}$, i.e., $n_i \leqslant 3\binom{\ell}{i} \cdot 2^{\lceil \frac{\ell-i}{2} \rceil}$. Next, since

$$2^i \cdot n_i \leqslant 2^i \cdot 3\binom{\ell}{i} \cdot 2^{\lceil \frac{\ell-i}{2} \rceil} = 3\binom{\ell}{i} \cdot 2^{\frac{\ell}{2} + \lceil \frac{i}{2} \rceil},$$

we get that

$$\begin{aligned}
|A_{2\ell}| &\leqslant 3\sum_{i=0}^{\ell} \binom{\ell}{i} \cdot 2^{\frac{\ell}{2} + \lceil \frac{i}{2} \rceil} = 3 \cdot 2^{\frac{\ell}{2}} \sum_{i=0}^{\ell} \binom{\ell}{i} \cdot 2^{\lceil \frac{i}{2} \rceil} \\
&= 3 \cdot 2^{\frac{\ell}{2}} \left((1+\sqrt{2})^\ell + 2^{\ell - 0.5}\right) \\
&= 3 \cdot \left(2^{\frac{\ell}{2} + \ell \log(1+\sqrt{2})} + 2^{\frac{3\ell-1}{2}}\right) \\
&= 3 \cdot \left(2^{\ell \log(2+\sqrt{2})} + 2^{\frac{3\ell-1}{2}}\right).
\end{aligned}$$

■

According to Lemma 5, the following corollary derives an upper bound on the capacity of the locally-balanced constraint.

■

Corollary 6. *For all $\ell \geqslant 4$ it holds thst*

$$\mathbb{C}^{bl}(\ell,1) \leqslant \frac{\log(2+\sqrt{2})}{2} + \frac{\log\left(3+\frac{(2\sqrt{2}-2)^{\ell}}{\sqrt{2}}\right)}{2\ell} \approx 0.885.$$

This last result can also be extended for other values of δ, as proved in the next theorem.

Theorem 7. *For any fixed $\delta \geqslant 1$ it holds that*

$$\begin{aligned}\limsup_{\ell\to\infty} \mathbb{C}^{bl}(\ell,\delta) &\leqslant \frac{\log(2+2^{\mathbb{C}^{RDS}(2\delta)})}{2} \\ &= \frac{\log(2+2^{\log(2\cos\frac{\pi}{2\delta+2})})}{2} < 1.\end{aligned}$$

Proof: Following the proof of Lemma 5, we first define $A_{2\ell,\delta} = \mathcal{S}^{bl}(\ell,\delta) \cap \Sigma_2^{2\ell}$ and repeat the same steps in the proof. Now we conclude that the word $\boldsymbol{d}$ satisfies the 2δ-RDS constraint and the size of the set $A_{2\ell,\delta}$ is at most

$$C\sum_{i=0}^{\ell}\binom{\ell}{i}2^i \cdot (2^{\mathbb{C}^{RDS}(2\delta)})^{\ell-i} = C(2+2^{\mathbb{C}^{RDS}(2\delta)})^{\ell},$$

for some constant C. Finally, it is concluded that

$$\begin{aligned}\mathbb{C}^{bl}(\ell,\delta) &\leqslant \frac{\log(|A_{2\ell,\delta}|)}{2\ell} \leqslant \frac{\log(C(2+2^{\mathbb{C}^{RDS}(2\delta)})^{\ell})}{2\ell} \\ &= \frac{\log(2+2^{\mathbb{C}^{RDS}(2\delta)})}{2} + \frac{\log C}{2\ell},\end{aligned}$$

which verifies the statement in the theorem. ∎

It is concluded from Theorem 7 that for any fixed δ the capacity of $\mathbb{C}^{bl}(\ell,\delta)$ cannot approach 1 even when ℓ is large enough. On the hand, in case δ is not fixed and can grow with ℓ then this capacity value indeed approaches 1. This property is proved in the next theorem.

Theorem 8. *If $\lim_{\ell\to\infty}\delta(\ell) = \infty$ then*

$$\limsup_{\ell\to\infty} \mathbb{C}^{bl}(\ell,\delta(\ell)) = 1.$$

Proof: For simplicity of the proof, we assume that ℓ is a multiple of $2\delta(\ell)$. In this case, we let

$$B_{\ell} = \{\boldsymbol{x} \in \Sigma_2^{2\delta(\ell)} \mid wt(\boldsymbol{x}) = \delta(\ell)\}$$

and $\mathcal{C}_{\ell} = B_{\ell}^*$. The Hamming weight w of every length-ℓ word in $\mathcal{C}_{\ell}$ satisfies

$$\left(\frac{\ell}{2\delta(\ell)} - 1\right)\delta(\ell) \leqslant w \leqslant \left(\frac{\ell}{2\delta(\ell)} - 1\right)\delta(\ell) + 2\delta(\ell),$$

and therefore $w \in [\ell/2 - \delta(\ell), \ell/2 + \delta(\ell)]$, that is, the word $\boldsymbol{x}$ is $(\ell, \delta(\delta))$-locally balanced. Lastly, the asymptotic rate of the code $\mathcal{C}_{\ell}$ is

$$\frac{\log\binom{2\delta(\ell)}{\delta(\ell)}}{2\delta(\ell)},$$

which approaches 1 for any $\delta(\ell)$ such that $\lim_{\ell\to\infty}\delta(\ell) = \infty$. ∎

V. The Locally-Bounded Constraint

In this section we study the locally-bounded constraint. Our main result is proved in the following theorem. Note that a similar technique has been applied in [16] when studying this constraint by using subwords of fixed weight. Despite these similarities, this construction is presented here for the completeness of the capacity results of this constraint.

Theorem 9. *For all $\delta < \sqrt{\ell}/2$ (negative or nonnegative) it holds that*

$$\lim_{\ell\to\infty} \mathbb{C}^{bd}(\ell,\delta) = 1.$$

Proof: For simplicity of the proof, let us assume that $\sqrt{\ell}$ is an integer. For all ℓ and n, which is a multiple of $\sqrt{\ell}$, we construct the following code

$$C_n = \left\{\boldsymbol{c} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{\frac{n}{\sqrt{\ell}}}) \mid wt(\boldsymbol{c}_i) \leqslant \frac{\sqrt{\ell}}{2} - 1, 1 \leqslant i \leqslant \frac{n}{\sqrt{\ell}}\right\}.$$

We will show that every codeword $\boldsymbol{c} \in C_n$ is (ℓ,δ)-locally bounded. For all $i \in [n-\ell+1]$, the subword $\boldsymbol{c}[i;\ell]$ contains at least $\sqrt{\ell}-1$ complete blocks from $\boldsymbol{c}$. Thus, it holds that

$$wt(\boldsymbol{c}[i;\ell]) \leqslant (\sqrt{\ell}-1)\left(\frac{\sqrt{\ell}}{2} - 1\right) + \sqrt{\ell} = \frac{\ell}{2} - \frac{\sqrt{\ell}}{2} + 1 \leqslant \frac{\ell}{2} - \delta.$$

On the other hand, for ℓ large enough, the cardinality of the code C_n is at least

$$(2^{\sqrt{\ell}-2})^{\frac{n}{\sqrt{\ell}}},$$

and thus it holds that

$$\lim_{n\to\infty}\frac{\log(|C_n|)}{n} \geqslant \lim_{n\to\infty}\frac{\log((2^{\sqrt{\ell}-2})^{\frac{n}{\sqrt{\ell}}})}{n} = 1 - \frac{2}{\sqrt{\ell}}.$$

Hence, we conclude that $\lim_{\ell\to\infty}\mathbb{C}^{bd}(\ell,\delta) = 1$. ∎

VI. Conclusion

The study of constraints with local limitations on the Hamming weight of subwords has been initiated in this paper. We showed how to fully solve the strong-locally-balanced constraint. For the locally-balanced constraint we showed upper and lower bounds on the capacity which hold for all values of ℓ. Lastly, it was shown that the capacity of the locally-bounded constraint approaches 1 when ℓ increases and $\delta < \sqrt{\ell}/2$. While the results in the paper already established many of the capacity values of these constraints, there are still several interesting problems which are left open. In particular, an interesting observation from the capacity values listed in Table I, indicates that for fixed δ the capacity $\mathbb{C}^{bl}(\ell,\delta)$ decreases when ℓ increases. Under this assumption we can deduce that the capacity sequence $\mathbb{C}_1(\ell,\delta)$ converges to some value, which will be denoted by $\mathbb{C}^{bl}(\delta)$, that is,

$$\mathbb{C}^{bl}(\delta) = \lim_{\ell\to\infty}\mathbb{C}^{bl}(\ell,\delta).$$

Proving that indeed $\mathbb{C}^{bl}(\ell,\delta)$ decreases with ℓ and studying this capacity limit are among the problems we aim to continue solving as part of our future research.

REFERENCES

[1] M. Bóna, "Handbook of enumerative combinatorics," vol. 87, CRC Press, 2015.

[2] E. Chargaff, "Chemical specificity of nucleic acids and mechanism of their enzymatic degradation," *Experientia*, vol. 6, no. 6, pp. 201–209, Jun. 1950.

[3] E. Chargaff, "Structure and function of nucleic acids as cell constituents," *Federation Proceedings*, vol. 10, no. 3, pp. 654–659, Sep. 1951.

[4] F. Crick and J.D. Watson, "Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid," *Nature*, vol. 171, pp. 737–738, 1953.

[5] Y. Erlich and D. Zielinski, "DNA fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950–954, 2017.

[6] A.M. Fouladgar, O. Simeone, and E. Erkip, "Constrained codes for joint energy and information transfer," *IEEE Transactions on Communications*, vol. 62, no. 6, pp. 2121–2131, Jun. 2014.

[7] R.N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W.J. Stark, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angewandte Chemie International Edition*, vol. 54, no. 8, pp. 2552–2555, 2015.

[8] K.A.S. Immink, *Codes for Mass Data Storage*, Eindhoven: Shannon Foundation Publishers, 2004.

[9] S. Jain, N. Raviv, and J. Bruck, "Attaining the 2nd Chargaff rule by tandem duplications," *Proc. International Symposium on Information Theory*, pp. 2241–2245, Vail, Colorado, Jun. 2018.

[10] D.E. Knuth, "Efficient balanced codes," *IEEE Transactions on Information Theory*, vol. 32, no. 1, pp. 51–53, Jan. 1986.

[11] B.H. Marcus, R.M. Roth, and P.H. Siegel, *An introduction to coding for constrained systems*, Lecture notes, 2001.

[12] D. Mitchell and R. Bridge, "A test of Chargaff's secong rule," *Biochemical Biophys. Res. Commun.*, vol. 340, no. 1, pp. 90–94, 2006.

[13] Potomac Institute for Policy Studies, *The Future of DNA Data Storage*, Arlington, VA, retrievable at *https://potomacinstitute.org/publications/reports*, Sep. 2018.

[14] R. Rudner, J.D. Karkas, and E. Chargaff, "Separation of B. subtilis DNA into complementary strands: Direct analysis," *Proc. National Acad. Sci.*, vol. 60, pp. 921–922, 1968.

[15] A. Tandon, M. Motani, and L. R. Varshney, "On code design for simultaneous energy and information transfer," *Information Theory and Applications Workshop (ITA)*, pp. 1-6, San Diego, CA, Feb. 2014.

[16] A. Tandon, M. Motani, and L. R. Varshney, "Subblock-constrained codes for real-time simultaneous energy and information transfer," *IEEE Transactions on Information Theory*, vol. 62, no. 7, pp. 4212–4225, Apr. 2016.

[17] H.T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Scientific Reports*, vol. 7, no. 1, 2017.

[18] H.T. Yazdi, H.M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao, and O. Milenkovic, "DNA-based storage: Trends and methods," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 1, no. 3, pp. 230–248, 2015.