Quantized-Constraint Concatenation and the Covering Radius of Constrained Systems

Dor Elimelech
Electrical and Computer Engineering
Ben-Gurion University of the Negev,
Beer Sheva 8410501, Israel
doreli@post.bgu.ac.il

Tom Meyerovitch
Department of Mathematics
Ben-Gurion University of the Negev,
Beer Sheva 8410501, Israel
mtom@bgu.ac.il

Moshe Schwartz
Electrical and Computer Engineering
Ben-Gurion University of the Negev,
Beer Sheva 8410501, Israel
schwartz@ee.bgu.ac.il

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206533

***Abstract*—We introduce a novel framework for implementing error-correction in constrained systems. The main idea of our scheme, called Quantized-Constraint Concatenation (QCC), is to employ a process of embedding the codewords of an error-correcting code in a constrained system as a (noisy, irreversible) quantization process. This is in contrast to traditional methods, such as concatenation and reverse concatenation, where the encoding into the constrained system is reversible. The possible number of channel errors QCC is capable of correcting is linear in the block length n, improving upon the $O(\sqrt{n})$ possible with the state-of-the-art known schemes. For a given constrained system, the performance of QCC depends on a new fundamental parameter of the constrained system – its covering radius.**

Motivated by QCC, we study the covering radius of constrained systems in both combinatorial and probabilistic settings. We reveal an intriguing characterization of the covering radius of a constrained system using ergodic theory.

I. Introduction

Constrained codes are often employed in communication and storage systems in order to mitigate the occurrence of data-dependent errors. In many channels, some words are more prone to error than others, and therefore by avoiding them, the number of errors is reduced. Such codes are called constrained codes. While the use of constrained codes may significantly reduce the occurrence of data-dependent errors, in many realistic scenarios, the transmitted data may still be corrupted by data-independent errors.

A well-known strategy for handling the corruption of data is to combine error-correcting codes with constrained codes. This has been extensively studied during the past 40 years (see for example [2], [5], [8]–[10], [15]), and recently regained attention due to the increased interest in DNA storage systems. Over the last years, error-correcting constrained codes for DNA storage have been studied in numerous works [1], [3], [4], [12], [14], [18]–[21], with particular attention given to the GC-content constraint and the run-length (homopolymer) constraint.

Despite the considerable recent progress made in construction and analysis of error-correcting constrained codes for specific families of constraints, still, only a few general frameworks for implementing error correction in constrained systems are known (see [17, Ch. 8] for a survey). An important example for such a framework is the method of reverse concatenation, sometimes called modified concatenation (see [2], [8], [10], [15]), in which an error-correction encoding follows a constrained encoder. Recently, an improvement of the reverse-concatenation method, called segmented reverse concatenation, was suggested [9]. A principle limitation to these methods is in the error-correction capability. While the state-of-the-art method presented in [9] allows for a correction of $O(\sqrt{n})$ errors (where n is the block length), a general technique for correcting $\Theta(n)$ errors in constrained systems is unknown.

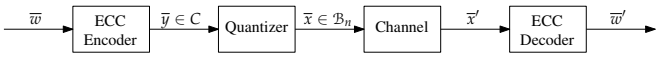

Fig. 1. A block diagram for quantized-constraint concatenation (QCC)

Motivated by this gap, we propose an alternative strategy, *quantized-constraint concatenation (QCC)*, for the implementation of error correction in constrained systems, which also works in the presence of $\Theta(n)$ errors. The basic idea behind our proposed method is simple: we suggest to consider the embedding process of information in the constrained media as a *quantization* process, rather than a coding process. In traditional methods (including concatenation and reverse concatenation) a constrained word represents the data to be transmitted and protected against errors. Thus, the constrained encoder is reversible, and it incurs a rate penalty, on top of the rate penalty for the error-correcting code. In QCC, we consider the constrained word as a corrupted version of the information, obtained by a quantization procedure. Thus, the constrained quantizer incurs no rate penalty. Instead, the parameters of the error-correcting code are designed to handle both errors caused by the channel and by the quantization process.

Let $\mathcal{B}_n \subseteq \Sigma^n$ be some set of constrained words of length n over some finite alphabet Σ. Assume furthermore that $r < n$ is an integer such that for any word $\overline{y} \in \Sigma^n$ there exists a corresponding word $\overline{x} \in \mathcal{B}_n$ with Hamming distance $d(\overline{x}, \overline{y}) \leqslant r$. Given an error-correcting code $C \subseteq \Sigma^n$ that can correct $t > r$ errors, we propose the following constrained error-correction procedure (see Figure 1):

- *Encoding:* Given an information word $\overline{w}$, use an error-correcting code encoder to map it to a codeword $\overline{y} \in C$.
- *Quantization:* Given $\overline{y} \in C$, find a constrained word $\overline{x} \in \mathcal{B}_n$ such that $d(\overline{y}, \overline{x}) \leqslant r$, and transmit $\overline{x}$.
- *Channel:* At the channel output, $\overline{x}' \in \Sigma^n$, a corrupted version of $\overline{x}$, is observed.

- *Decoding:* Use the decoder for C on $\overline{x}'$ and obtain $\overline{w}'$.

If the channel does not introduce more than $t-r$ errors, i.e., $d(\overline{x},\overline{x}') \leqslant t-r$, then $d(\overline{y},\overline{x}') \leqslant t$. Since C can correct t errors, we have $\overline{w} = \overline{w}'$, namely, it is possible to correct $t-r$ channel errors. We are therefore interested in the minimal number r, such that any word in the space can be quantized to a word in $\mathcal{B}_n$ with at most r coordinates changed. In coding-theory terminology, this quantity, denoted by $R(\mathcal{B}_n)$, is called the *covering radius* of $\mathcal{B}_n$. Using this technique, it is now possible to correct $\Theta(n)$ errors: assume that we have a constrained system such that for all n we have $R(\mathcal{B}_n) \leqslant \rho \cdot n$, and $(C_n)_{n\in\mathbb{N}}$ is a sequence of codes capable of correcting $\delta \cdot n$ errors for some $\delta > \rho$. Using the scheme presented above, it is therefore possible to correct $(\delta-\rho)\cdot n$ channel errors, which is linear in n.

In order to further our understanding of the proposed scheme, it is crucial to study the covering radius of constrained systems, which is the goal of this paper. We outline the contributions we make. In Section III, we provide a combinatorial definition for the covering radius of a constrained system, and investigate some of its fundamental properties. We also observe an intriguing phenomenon: We present an example of a constrained system with positive capacity that has the same covering radius as the repetition code, which has zero capacity. Inspired by this phenomenon, in Section IV we take a probabilistic approach and define the *essential* covering radius. We show that this version disregards the extreme cases causing the unwanted phenomenon described above. We also use the framework of ergodic theory to give an alternative characterization of the essential covering radius. Due to space limitations, proofs are omitted, and may be found in [7].

II. Preliminaries

Throughout this paper we shall use lower-case letters, x, to denote scalars and symbols, overlined lower-case letters, $\overline{x}$, to denote finite-length words, and bold lower-case letters, $\mathbf{x}$, to denote bi-infinite sequences. We use upper-case letters, X, for constrained systems. For a bi-infinite sequence $\mathbf{x} = \ldots,\mathbf{x}_{-1},\mathbf{x}_0,\mathbf{x}_1,\ldots$ and $n \leqslant m$ we denote the subword $\mathbf{x}_n^m \triangleq \mathbf{x}_n,\ldots,\mathbf{x}_m$ (and similarly $\overline{x}_n^m$ for finite words). We use Σ to denote a finite alphabet, and $[n] \triangleq \{0,1,\ldots,n-1\}$.

The set of words of length n over Σ is denoted by Σ^n. If $\overline{u} \in \Sigma^n$, we shall index its letters by $[n]$, i.e., $\overline{u} = \overline{u}_0,\overline{u}_1,\ldots,\overline{u}_{n-1}$. For any $\overline{v},\overline{u} \in \Sigma^n$, we define the Hamming distance as $d(\overline{u},\overline{v}) \triangleq |\{i \in [n] \,|\, \overline{u}_i \neq \overline{v}_i\}|$. The ball of radius r (with respect to the Hamming distance) centered in $\overline{x}$ is denoted by $B_r(\overline{x})$. The covering radius of a code $C \subseteq \Sigma^n$ is the minimal integer r such that the union of balls of radius r, centered at the codewords of C, covers the whole space. That is,

$$R(C) \triangleq \min\left\{ r \in \mathbb{N}\cup\{0\} \;\middle|\; \bigcup_{\overline{c}\in C} B_r(\overline{c}) = \Sigma^n \right\}.$$

Elements in Σ^n whose distance to the closest codeword of C is $R(C)$, are called *deep holes* (e.g., see [6, Definition 2.1.3]).

We turn to discuss constrained systems. These are often studied in the framework of symbolic dynamics (see for example [13], [17]). In a typical setting we have a finite alphabet Σ, and the space of bi-infinite sequences of Σ, denoted $\Sigma^{\mathbb{Z}}$, is considered as a compact metrizable topological space, equipped with the product topology (where Σ has the discrete topology). The dynamics on the system $\Sigma^{\mathbb{Z}}$ are realized by the shift transformation, $T : \Sigma^{\mathbb{Z}} \to \Sigma^{\mathbb{Z}}$, $(T\mathbf{x})_n \triangleq \mathbf{x}_{n+1}$, which is a topological homeomorphism of the system.

A subshift (or shift space) $X \subseteq \Sigma^{\mathbb{Z}}$ is a compact subspace, which is invariant under the shift transformation. For a subshift X, the language of X is the set of all finite words that appear as subwords of some element in X. That is

$$\mathcal{B}(X) \triangleq \left\{ \overline{x} = (x_0\ldots x_k) \;\middle|\; \begin{smallmatrix}\exists \mathbf{x}\in X, n\in\mathbb{Z} \text{ s.t.} \\ \mathbf{x}_n^{n+k} = \overline{x}, k\in\mathbb{N}\cup\{0\}\end{smallmatrix} \right\}.$$

The set of words of length n in the language is denoted by $\mathcal{B}_n(X) \triangleq \mathcal{B}(X) \cap \Sigma^n$. The topological entropy, also called capacity, of X is defined to be the following limit (which exists by Fekete's lemma)

$$h(X) \triangleq \lim_{n\to\infty} \frac{\log_{|\Sigma|} |\mathcal{B}_n(X)|}{n}.$$

In our setting, constrained systems are those shift spaces which can be realized by walks on some labeled graph.

Definition 1 *A shift space $X \subseteq \Sigma^{\mathbb{Z}}$ is called a constrained system (or a sofic shift) if there exists a finite directed graph $G = (V,E)$ and a labeling function $L : E \to \Sigma$ such that*

$$X = X_G \triangleq \left\{ (L(e_i))_{i\in\mathbb{Z}} \;\middle|\; \begin{smallmatrix}(e_i)_{i\in\mathbb{Z}} \text{ is a bi-infinite} \\ \text{directed path in } G\end{smallmatrix} \right\}.$$

A labeled graph $G = (V,E,L)$ is called irreducible if any two vertices are connected by a directed path. An irreducible graph is called primitive if the greatest common divisor of all cycle lengths is 1. It is well known (e.g., see [13, Theorem 4.5.8]) that an irreducible graph is primitive if and only if there exists $n \in \mathbb{N}$ such that for any two vertices $v,v' \in V$ there exists a directed path of length n from v to v'.

Definition 2 *A constrained system $X \subseteq \Sigma^{\mathbb{Z}}$ is called irreducible (respectively: primitive), if there exists an irreducible (respectively: primitive) labeled graph G such that $X = X_G$.*

A special family of constrained systems of particular interest is the family of systems defined by a finite set of local constraints. These are referred to as *systems of finite type*. A constrained system $X \subseteq \Sigma^{\mathbb{Z}}$ is said to be a system of finite type (SFT) if there exists some $m \in \mathbb{N}$ and a finite set of forbidden words $\mathcal{F} \subseteq \Sigma^m$ such that X is the set of all bi-infinite sequences not containing any forbidden pattern from $\mathcal{F}$. That is

$$X = X_{\mathcal{F}} \triangleq \left\{ \mathbf{x} \in \Sigma^{\mathbb{Z}} \;\middle|\; \forall n \in \mathbb{Z}, \mathbf{x}_n^{n+m-1} \notin \mathcal{F} \right\}.$$

III. The Covering Radius of a Constrained System

We begin with a definition of the covering radius of a set $B \subseteq \Sigma^n$ with respect to another set $A \subseteq \Sigma^n$.

Definition 3 *Let $A, C \subseteq \Sigma^n$, then the covering radius of C with respect to A is defined to be*

$$R(C,A) \triangleq \min\left\{ r \in \mathbb{N} \cup \{0\} \;\middle|\; A \subseteq \bigcup_{\overline{x}\in C} B_r(\overline{x}) \right\} = \max_{\overline{y}\in A} \min_{\overline{x}\in C} d(\overline{x},\overline{y}).$$

For constrained systems $X, Y \subseteq \Sigma^{\mathbb{Z}}$ we define the asymptotic covering radius of X with respect to Y to be the asymptotic normalized covering radius of n-tuples from X with respect to n-tuples from Y.

Definition 4 *Let $X, Y \subseteq \Sigma^{\mathbb{Z}}$ be shift spaces, then we define*

$$R(X,Y) \triangleq \liminf_{n\to\infty} \frac{R(\mathcal{B}_n(X), \mathcal{B}_n(Y))}{n}, \tag{1}$$

where we remind that $\mathcal{B}_n(X)$ and $\mathcal{B}_n(Y)$ are the subwords of length n from X and Y respectively.

In a typical coding-theoretic framework, the covering radius is considered as a property of a single code in the Hamming space of a finite length n. A constrained system on the other hand may be associated with a sequence of codes, which are the sets of constrained words of fixed lengths. The covering radius of the constrained system, as defined above, is in fact the asymptotic value of the (normalized) covering radii of this corresponding sequence of codes. An immediate question that comes up when considering our definition of the covering radius, is whether the limit from (1) exists. Under certain conditions, we can show it does:

Proposition 5 *Assume that $X, Y \subseteq \Sigma^{\mathbb{Z}}$ are constrained systems. If X or Y are primitive, then the* lim inf *in the definition of $R(X,Y)$ is actually a limit:*

$$R(X,Y) = \lim_{n\to\infty} \frac{R(\mathcal{B}_n(X), \mathcal{B}_n(Y))}{n}.$$

Remark 1 *Our proof of Proposition 5 gives in the case $Y = \Sigma^{\mathbb{Z}}$ that*

$$R(X,Y) = \lim_{n\to\infty} \frac{R(\mathcal{B}_n(X), \Sigma^n)}{n} = \sup_{n\in\mathbb{N}} \frac{R(\mathcal{B}_n(X), \Sigma^n)}{n}.$$

Example 1 *Consider the binary alphabet, $[2] \triangleq \{0,1\}$. Let $X_{0,k} \subseteq [2]^{\mathbb{Z}}$ be the $(0,k)$-RLL system, which comprises of all the binary sequences that do not contain $k+1$ consecutive zeros. That is, $X_{0,k} = X_{\mathcal{F}}$, where $\mathcal{F} = \{\overline{0}_{k+1} = (0,\ldots,0)\} \subseteq [2]^{k+1}$. We claim that*

$$R(X_{0,k}, [2]^{\mathbb{Z}}) = \frac{1}{k+1}.$$

Indeed, by Remark 1

$$R(X_{0,k}, [2]^{\mathbb{Z}}) \geqslant \frac{R(\mathcal{B}_{k+1}(X_{0,k}), [2]^{k+1})}{k+1} = \frac{R([2]^{k+1} \setminus \{\overline{0}_{k+1}\}, [2]^{k+1})}{k+1} = \frac{1}{k+1}.$$

We now show that the obtained lower bound is tight. Let $\overline{y} \in [2]^n$ be any binary word. Consider $\overline{x}$ given by

$$\overline{x}_i \triangleq \begin{cases} \overline{y}_i & i \bmod (k+1) \neq 0, \\ 1 & i \bmod (k+1) = 0. \end{cases}$$

Clearly $\overline{x}$ does not contain any subword of $k+1$ consecutive zeros and therefore $\overline{x} \in \mathcal{B}_n(X_{0,k})$. Since $d(\overline{x},\overline{y}) \leqslant \lceil \frac{n}{k+1} \rceil$ we conclude that $\frac{1}{n} R(\mathcal{B}_n(X_{0,k}), [2]^n) \leqslant \frac{1}{n} \lceil \frac{n}{k+1} \rceil$, and by taking the limit,

$$R(X_{0,k}, [2]^{\mathbb{Z}}) = \lim_{n\to\infty} \frac{R(\mathcal{B}_n(X_{0,k}), [2]^n)}{n} \leqslant \frac{1}{k+1}.$$

Using a ball-covering argument, we lower-bound the covering radius in terms of the capacities of the systems. We recall that $H_q : [0,1] \to [0,1]$ denotes the q-ary entropy function defined by

$$H_q(x) \triangleq x \log_q(q-1) - x \log_q(x) - (1-x)\log_q(1-x),$$

and for continuity, $H_q(0) \triangleq 0$ as well as $H_q(1) \triangleq \log_q(q-1)$. We also use $H_q^{-1} : [0,1] \to [0, 1-\frac{1}{q}]$ to denote its inverse.

Proposition 6 *Let $X, Y \subseteq \Sigma^{\mathbb{Z}}$ be constrained systems with capacities $h(X) \leqslant h(Y)$, and let us denote $|\Sigma| = q$. Then*

$$R(X,Y) \geqslant H_q^{-1}(h(Y) - h(X)).$$

Example 2 *Fix $\Sigma = [q] \triangleq \{0,\ldots,q-1\}$ and consider the repetition shift $X_{\text{rep}} \triangleq \{(\ldots,a,a,a,\ldots) \,|\, a \in [q]\}$. Clearly, X_{rep} is the SFT defined by the forbidden patterns $\mathcal{F} = \{ab \,|\, a,b \in [q], a \neq b\} \subseteq [q]^2$. Since $h(X_{\text{rep}}) = 0$, by Proposition 6*

$$R(X_{\text{rep}}, [q]^{\mathbb{Z}}) \geqslant H_q^{-1}(1-0) = 1 - \frac{1}{q}.$$

On the other hand, for any $n \in \mathbb{N}$ and for any $\overline{y} \in [q]^n$, it is clear that there exists at least one symbol $a \in [q]$ which appears in at least $\lceil \frac{n}{q} \rceil$ coordinates of $\overline{y}$, and in particular $d(\overline{y}, (a,\ldots,a)) \leqslant \lfloor \frac{q-1}{q} n \rfloor$. This proves that

$$R(\mathcal{B}_n(X_{\text{rep}}), [q]^n) \leqslant \left\lfloor \frac{q-1}{q} n \right\rfloor.$$

Taking the limit and combining with the lower bound we obtain

$$R(X_{\text{rep}}, [q]^{\mathbb{Z}}) = 1 - \frac{1}{q}.$$

At this point we have reached a curious situation. For the sake of illustrating it, fix the binary alphabet $\Sigma = [2]$. If we consider $X_{0,1}$, the $(0,1)$-RLL system from Example 1, then its capacity is known to be $h(X_{0,1}) = \log_2((1+\sqrt{5})/2) \approx 0.694$, and we have shown that its covering radius (with respect to $[2]^{\mathbb{Z}}$) is $R(X_{0,1}, [2]^{\mathbb{Z}}) = \frac{1}{2}$. However, in Example 2 we have seen that the binary repetition shift, X_{rep}, has the same covering radius $R(X_{\text{rep}}, [2]^{\mathbb{Z}}) = \frac{1}{2}$, but zero capacity, $h(X_{\text{rep}}) = 0$. From a coding perspective, even though $\mathcal{B}_n(X_{0,1})$ has exponentially more words than $\mathcal{B}_n(X_{\text{rep}})$, the worst-case covering

scenario, namely, a deep hole, is asymptotically within the same distance from the constrained code.

As a final comment for this section, we would like to comment on the relation of $R(X,\Sigma^{\mathbb{Z}})$ to the QCC framework. Since we are interested in asymptotics, assume that the sequence of error-correcting codes in the QCC scheme is $(C_n)_{n\in\mathbb{N}}$, where C_n is of length n. The expression $R(X,\Sigma^{\mathbb{Z}}) = \lim_{n\to\infty}\frac{1}{n}R(\mathcal{B}_n(X),\Sigma^n)$ is an upper bound on the worse-case quantization error rate using a sequence of codes $(C_n)_{n\in\mathbb{N}}$, which is actually $\lim_{n\to\infty}\frac{1}{n}R(\mathcal{B}_n(X),C_n)$. The bound $R(X,\Sigma^{\mathbb{Z}})$ is pessimistic twice: once for allowing deep holes to determine the covering radius, and twice, for assuming they reside in C_n. Since $\lim_{n\to\infty}\frac{1}{n}R(\mathcal{B}_n(X),C_n)$ may be hard to compute and depends on the sequence of error-correcting codes, we may use $R(X,\Sigma^{\mathbb{Z}})$ as an upper bound on the worse-case quantization error, which is independent of the sequence of codes.

IV. The Essential Covering Radius

The covering radius that was studied in the previous section may be perhaps too pessimistic in the sense that it is determined by the worst-case quantization distance. In this section we study a different definition of the covering radius, which we call the essential covering radius. Given $\varepsilon > 0$, the ϵ-covering radius of a constraint system is, loosely speaking, the smallest r such that $(1-\varepsilon)$-fraction the words in the space can be quantized to the constraint system. In what follows, we further generalize this to a probabilistic definition.

We begin by stating some basic definitions and well known results from ergodic theory. For any finite alphabet Σ, we consider $\Sigma^{\mathbb{Z}}$ as a measurable space, together with the Borel Σ-algebra induced by the product topology on $\Sigma^{\mathbb{Z}}$. Similarly, any subshift $Y \subseteq \Sigma^{\mathbb{Z}}$ is considered as a measurable space.

Definition 7 (Invariant and ergodic measures) *Let $Y \subseteq \Sigma^{\mathbb{Z}}$ be a subshift. A probability measure μ on Y is called* shift invariant *if $\mu(T^{-1}B) = \mu(B)$ for any measurable set B. A shift-invariant measure μ is further said to be* ergodic *if $T^{-1}B = B$ implies $\mu(B) = 0$ or $\mu(Y \setminus B) = 0$. The set of shift-invariant probability measures on Y is denoted by $M(Y)$, and the set of ergodic measures in $M(Y)$ is denoted by $M_{\mathcal{E}}(Y)$.*

For a measure $\mu \in M(Y)$ we denote by μ_n the marginal measure of μ on the first n coordinates, which is a probability measure on Σ^n. To avoid cumbersome notation, throughout this work we shall use $\mathbb{P}_\mu[A]$ in order to denote the measure $\mu(A)$, and $\mathbf{Y}$ for a random bi-infinite sequence on Y. Throughout this article we use bold upper-case letters for bi-infinite sequences of random variables, not to be confused with non-bold capital letters used to denote constrained systems.

We are now ready to define the essential covering radius.

Definition 8 *For any real $\varepsilon > 0$, two sets $A, C \subseteq \Sigma^n$, and η, a probability measure on A, we define $R_\varepsilon(C,A,\eta)$ to be*

$$\min\left\{ r \in \mathbb{N}\cup\{0\} \;\middle|\; \eta\left(A \cap \left(\bigcup_{\overline{x}\in C} B_r(\overline{x})\right)\right) \geqslant 1-\varepsilon \right\}.$$

We remark that when η is the uniform measure on A, $R_\varepsilon(C,A,\eta)$ is the ε-covering radius of A, namely the smallest r such that at least $(1-\epsilon)$-fraction of the words in C are at distance at most r from A, as desired.

Definition 9 *Let $X, Y \subseteq \Sigma^{\mathbb{Z}}$ be constrained systems, and $\mu \in M_{\mathcal{E}}(Y)$ be an ergodic measure. We define the* ε-covering radius *of X with respect to (Y,μ) by*

$$R_\varepsilon(X,Y,\mu) \triangleq \liminf_{n\to\infty} \frac{R(\mathcal{B}_n(X),\mathcal{B}_n(Y),\mu_n)}{n},$$

and the essential covering radius *by*

$$R_0(X,Y,\mu) \triangleq \lim_{\varepsilon\to 0} R_\varepsilon(X,Y,\mu).$$

We comment that the limit in the previous definition exists due to the monotonicity of $R_\varepsilon(X,Y,\mu)$ in ε. We also observe that, trivially, the essential covering radius is upper bounded by the (worst-case) covering radius, and we have

$$R_\varepsilon(X,Y,\mu) \leqslant R_0(X,Y,\mu) \leqslant R(X,Y).$$

We now revisit the examples from the previous section and consider their essential covering radius.

Proposition 10 *Consider the q-ary repetition system $X_{\text{rep}} \subseteq [q]^{\mathbb{Z}}$ from Example 2, and assume $Y = [q]^{\mathbb{Z}}$ is equipped with the uniform Bernoulli i.i.d measure, denoted by μ^{u}. Then the essential covering radius is equal to the covering radius, i.e.,*

$$R_0(X_{\text{rep}},[q]^{\mathbb{Z}},\mu^{\mathsf{u}}) = R(X_{\text{rep}},[q]^{\mathbb{Z}}) = 1-\frac{1}{q}.$$

As we have seen, the repetition system, whose capacity is zero, has the same covering radius and essential covering radius. The $(0,k)$-RLL system has positive capacity. While its covering radius is $\frac{1}{k+1}$, the following theorem asserts that its essential covering radius decays exponentially fast with k, in stark contrast to the repetition system.

Theorem 11 *Let $X_{0,k} \subseteq [2]^{\mathbb{Z}}$ be the $(0,k)$-RLL system from Example 1, and let $Y = [2]^{\mathbb{Z}}$ be equipped with the uniform Bernoulli i.i.d measure μ^{u}. Then*

$$R_0(X_{0,k},[2]^{\mathbb{Z}},\mu^{\mathsf{u}}) = \frac{1}{2(2^{k+1}-1)}.$$

It is desirable to have alternative expressions for the essential covering radius, which could assist in calculating or estimating its value. Inspired by tools used in the proof of Theorem 11, we give an ergodic-theoretic characterization.

Definition 12 *Let $X, Y \subseteq \Sigma^{\mathbb{Z}}$ be shift spaces, we consider $X \times Y$ as shift space, with the left shift acting as $T(\mathbf{x},\mathbf{y}) =$*

$(T\mathbf{x}, T\mathbf{y})$. *For an ergodic measure $\mu \in M_{\mathcal{E}}(Y)$, an extension of μ over $X \times Y$ is a shift-invariant measure ν on the product space $X \times Y$ whose Y-marginal is μ. Namely, ν satisfies that for any measurable $A \subseteq Y$, $\nu(X \times A) = \mu(A)$. An extension on $X \times Y$ is said to be ergodic if it is an ergodic measure with respect to the shift transformation on the product space. We let $M(X,Y,\mu)$ denote the set of all extensions of μ, and $M_{\mathcal{E}}(X,Y,\mu)$ denote the set of all ergodic extensions in $M(X,Y,\mu)$.*

We are now ready to state the main result of the section that gives an equivalent formulation of the essential covering radius via a minimization problem over invariant extensions.

Theorem 13 *Let $X, Y \subseteq \Sigma^{\mathbb{Z}}$ be constrained systems, and let $\mu \in M_{\mathcal{E}}(Y)$ be an ergodic measure. Then*

$$\begin{aligned} R_0(X,Y,\mu) &= \inf\{\mathbb{P}_\nu[\mathbf{X}_0 \neq \mathbf{Y}_0] \,|\, \nu \in M_{\mathcal{E}}(X,Y,\mu)\} \\ &= \inf\{\mathbb{P}_\nu[\mathbf{X}_0 \neq \mathbf{Y}_0] \,|\, \nu \in M(X,Y,\mu)\}. \end{aligned}$$

In the following example, we explicitly describe a sequence of extensions in $M_{\mathcal{E}}(X,Y,\mu)$ which approximates the essential covering radius of the $(0,k)$-RLL system from Example 1 with respect to the full-shift.

Example 3 *Let $X_{0,k} \subseteq [2]^{\mathbb{Z}}$ denote the $(0,k)$-RLL shift as in Example 1. Let $\overline{y} \in [2]^n$ be a finite binary word. We define $c(\overline{y})$ to be the length of longest zero suffix of $\overline{y}$, i.e.,*

$$c(\overline{y}) \triangleq \max\left\{ i \,\middle|\, \overline{y} = \overline{y}_0^{n-i-1}\overline{0}_i \right\}.$$

We fix $N \in \mathbb{N}$ and consider the map $f^{(N)} : [2]^{\mathbb{Z}} \to X_{0,k}$

$$f^{(N)}(\mathbf{y})_m = \begin{cases} 1 & c\left(\mathbf{y}_{m-(N(k+1)-1)}^{m-1}\right) \equiv k \pmod{k+1}, \\ \mathbf{y}_m & \text{otherwise.} \end{cases}$$

Clearly, $\mathrm{Im}(f) \subseteq X_{0,k}$ since no run of $k+1$ zeroes may appear in $f^{(N)}(\mathbf{y})$. We note that the map $(f^{(N)}, \mathrm{Id}) : [2]^{\mathbb{Z}} \to X_{0,k} \times [2]^{\mathbb{Z}}$ is a sliding-block-code function (i.e., a function such that the value in each coordinate is determined by a finite block of adjacent coordinates), and therefore it is measurable and commutes with the shift transformation. Let μ^{u} be the uniform measure over $[2]^{\mathbb{Z}}$, and let ν_N be its push-forward measure on $X_{0,k} \times [2]^{\mathbb{Z}}$ using $f^{(N)}$. Clearly ν_N is an invariant measure, which is also ergodic (as a factor of an ergodic measure). Therefore, $\nu_N \in M_{\mathcal{E}}(X_{0,k}, [2]^{\mathbb{Z}}, \mu^{\mathsf{u}})$. We note that

$$\begin{aligned} &\mathbb{P}_{\nu_N}[\mathbf{X}_0 \neq \mathbf{Y}_0] \\ &= \mathbb{P}_{\mu^{\mathsf{u}}}\left[c\left(\mathbf{Y}_{-(N(k+1)-1)}^{-1}\right) \equiv k \pmod{k+1} \text{ and } \mathbf{Y}_0 = 0\right] \\ &= \sum_{i=0}^{N-1} \mathbb{P}_{\mu^{\mathsf{u}}}\left[c\left(\mathbf{Y}_{-(N(k+1)-1)}^{-1}\right) = i(k+1)+k \text{ and } \mathbf{Y}_0 = 0\right] \\ &= \mathbb{P}_{\mu^{\mathsf{u}}}\left[\mathbf{Y}_{-(N(k+1)-1)}^{0} = \overline{0}\right] + \sum_{i=1}^{N-1} \mathbb{P}_{\mu^{\mathsf{u}}}\left[\mathbf{Y}_{-i(k+1)}^{0} = 1\overline{0}\right] \\ &= \frac{1}{2^{N(k+1)}} + \frac{1}{2}\sum_{i=1}^{N-1} \frac{1}{2^{i(k+1)}}. \end{aligned}$$

Taking $N \to \infty$ we obtain

$$\lim_{N\to\infty} \mathbb{P}_{\nu_N}[\mathbf{X}_0 \neq \mathbf{Y}_0] = \frac{1}{2(2^{k+1}-1)} = R_0(X_{0,k}, [2]^{\mathbb{Z}}, \mu^{\mathsf{u}}).$$

To conclude this section we briefly discuss the essential covering radius in the context of the QCC scheme. Loosely speaking, asymptotically, all but a vanishing fraction of Σ^n may be quantized to $\mathcal{B}_n(X)$ by changing an $R_0(X, \Sigma^{\mathbb{Z}}, \mu^{\mathsf{u}})$-fraction of the positions. This fraction may be significantly lower than the worst-case fraction $R(X, \Sigma^{\mathbb{Z}})$. In a finite-length setting, at least a $(1-\varepsilon)$-fraction of Σ^n may be quantized to $\mathcal{B}_n(X)$ by changing at most $r_\varepsilon = R_\varepsilon(\mathcal{B}_n(X), \Sigma^n, \mu_n^{\mathsf{u}})$ positions. A small obstacle we need to overcome is the fact that in the QCC scheme we do not quantize any word from Σ^n, but rather only codewords of the error-correcting code C. The ε-fraction of words from Σ^n that are a long distance from $\mathcal{B}_n(X)$ may disproportionately reside in C. However, if we further assume that C is a linear error-correcting code, by a simple averaging argument there exists at least one coset of the code, C', such that the fraction of codewords whose distance to the language of X is at most r_ε. This means that there exists $C'' \subseteq C'$ with $|C''| \geqslant (1-\epsilon)|C'|$ such that $R(\mathcal{B}_n(X), C'') \leqslant r_\epsilon$.

V. Conclusion

While the covering radius of constrained systems is of independent intellectual merit, let us put our results in the context of the QCC scheme. Consider $X_{0,k}$, the $(0,k)$-RLL system described in Example 1. Using the coding scheme presented in [9, Theorem 1], it is possible to correct up to $O(\sqrt{n})$ errors. However, using QCC with the combinatorial covering radius (which in that case is $\frac{1}{k+1}$), since there exist error-correcting codes with non-vanishing rate capable of correcting up to $(\frac{1}{4} - \delta)n$ errors (for every $\delta > 0$), we obtain codes with non-vanishing rate capable of correcting up to $(\frac{1}{4} - \frac{1}{k+1} - \delta)n$ channel errors. On the other hand, we may use the essential covering radius of $X_{0,k}$ to bound the *essential* quantization noise. In that case, since

$$R_{\frac{1}{2}}(X_{0,k}, [2]^{\mathbb{Z}}, \mu^{\mathsf{u}}) \leqslant R_0(X_{0,k}, [2]^{\mathbb{Z}}, \mu^{\mathsf{u}}) = \frac{1}{2(2^{k+1}-1)},$$

using the QCC, it is possible to find error-correcting codes such that by removing at most half of the codewords (which asymptotically does not effect the rate), it is possible to improve our error correction capability to $(\frac{1}{4} - \frac{1}{2(2^{k+1}-1)} - \delta)n$ channel errors. Previous lower bounds on the possible rates for error-correcting constrained codes have been established in previous works [11], [16], via somewhat non-constructive methods. A certain advantage of our scheme is its simplicity and constructive nature.

Finally, we mention two major goals for subsequent work: The first goal is to find efficient methods to compute or estimate the essential covering radius. The second goal is to find efficient quantization algorithms that will enable the use of the QCC scheme.

REFERENCES

[1] K. G. Benerjee and A. Banerjee, "On homopolymers and secondary structures avoiding, reversible, reversible-complement and gc-balanced dna codes," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 204–209.

[2] W. Bliss, "Circuitry for performing error correction calculations on baseband encoded data to eliminate error propagation," *IBM Tech. Discl. Bul.*, vol. 23, pp. 4633–4634, 1981.

[3] K. Cai, Y. M. Chee, R. Gabrys, H. M. Kiah, and T. T. Nguyen, "Correcting a single indel/edit for dna-based data storage: Linear-time encoders and order-optimality," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3438–3451, 2021.

[4] K. Cai, H. M. Kiah, M. Motani, and T. T. Nguyen, "Coding for segmented edits with local weight constraints," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 1694–1699.

[5] Y. M. Chee, H. M. Kiah, and H. Wei, "Efficient and explicit balanced primer codes," *IEEE Transactions on Information Theory*, vol. 66, no. 9, pp. 5344–5357, 2020.

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

[7] D. Elimelech, T. Meyerovitch, and M. Schwartz, "Quantized-constraint concatenation and the covering radius of constrained systems," *arXiv preprint arXiv:2302.02363*, 2023.

[8] J. L. Fan and A. R. Calderbank, "A modified concatenated coding scheme, with applications to magnetic data storage," *IEEE Transactions on Information Theory*, vol. 44, no. 4, pp. 1565–1574, 1998.

[9] R. Gabrys, P. H. Siegel, and E. Yaakobi, "Segmented reverse concatenation: a new approach to constrained ecc," in *2020 International Symposium on Information Theory and Its Applications (ISITA)*. IEEE, 2020, pp. 254–258.

[10] K. A. S. Immink, "A practical method for approaching the channel capacity of constrained channels," *IEEE Transactions on Information Theory*, vol. 43, no. 5, pp. 1389–1399, 1997.

[11] V. D. Kolesnik and V. Y. Krachkovsky, "Generating functions and lower bounds on rates for limited error-correcting codes," *IEEE transactions on information theory*, vol. 37, no. 3, pp. 778–788, 1991.

[12] X. Li, M. Chen, and H. Wu, "Multiple errors correction for position-limited dna sequences with gc balance and no homopolymer for dna-based data storage," *Briefings in Bioinformatics*, 2022.

[13] D. Lind and B. Marcus, *An introduction to symbolic dynamics and coding*. Cambridge university press, 2021.

[14] X. Lu and S. Kim, "Design of nonbinary error correction codes with a maximum run-length constraint to correct a single insertion or deletion error for dna storage," *IEEE Access*, vol. 9, pp. 135 354–135 363, 2021.

[15] M. Mansuripur, "Enumerative modulation coding with arbitrary constraints and postmodulation error correction coding for data storage systems," in *Optical Data Storage'91*, vol. 1499. SPIE, 1991, pp. 72–86.

[16] B. H. Marcus and R. M. Roth, "Improved gilbert-varshamov bound for constrained systems," *IEEE transactions on information theory*, vol. 38, no. 4, pp. 1213–1221, 1992.

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

[18] T. T. Nguyen, K. Cai, K. A. S. Immink, and H. M. Kiah, "Constrained coding with error control for dna-based data storage," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 694–699.

[19] ——, "Capacity-approaching constrained codes with error correction for dna-based data storage," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 5602–5613, 2021.

[20] W. H. Press, J. A. Hawkins, S. K. Jones Jr, J. M. Schaub, and I. J. Finkelstein, "Hedges error-correcting code for dna storage corrects indels and allows sequence constraints," *Proceedings of the National Academy of Sciences*, vol. 117, no. 31, pp. 18 489–18 496, 2020.

[21] J. H. Weber, J. A. De Groot, and C. J. Van Leeuwen, "On single-error-detecting codes for dna-based data storage," *IEEE Communications Letters*, vol. 25, no. 1, pp. 41–44, 2020.