Efficient Capacity-Achieving Codes for General Repeat Channels

Francisco Pernice
Stanford University
fpernice@stanford.edu

Ray Li
Stanford University
rayyli@stanford.edu

Mary Wootters
Stanford University
marykw@stanford.edu

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834386

Abstract—Given a probability distribution D over the non-negative integers, a D-repeat channel acts on an input symbol by repeating it a number of times distributed as D. For example, the binary deletion channel (D=Bernoulli) and the Poisson repeat channel (D=Poisson) are special cases. We say a D-repeat channel is square-integrable if D has finite first and second moments. In this paper, we construct explicit codes for all square-integrable D-repeat channels with rate arbitrarily close to the capacity, that are encodable and decodable in linear and quasi-linear time, respectively. We also consider possible extensions to the repeat channel model, and illustrate how our construction can be extended to an even broader class of channels capturing insertions, deletions, and substitutions.

Our work offers an alternative, simplified, and more general construction to the recent work of Rubinstein [3], who attains similar results to ours in the cases of the deletion channel and the Poisson repeat channel. It also slightly improves the runtime and decoding failure probability of the polar codes constructions of Tal et al. [1] and of Pfister and Tal [2] for the deletion channel and certain insertion/deletion/substitution channels. Our techniques follow closely the approaches of Guruswami and Li [4] and Con and Shpilka [5]; what sets apart our work is that to obtain our result, we show that a capacity-achieving code for the channels in question can be assumed to have an "approximate balance" in the frequency of zeros and ones of all sufficiently long substrings of all codewords. This allows us to attain near-capacity-achieving codes in a general setting. We consider this "approximate balance" result to be of independent interest, as it can be cast in much greater generality than just repeat channels.

A full version of this paper is available at https://arxiv.org/abs/2201.12746.

Index Terms—synchronization channels, efficient codes, explicit codes.

I. Introduction

Fixing a probability distribution D over the natural numbers $\mathbb{N}$, a D-repeat channel acts on an input bit by repeating it a number of times distributed like D. Special cases include the binary deletion channel, Poisson repeat channel, and the sticky channels (the latter two were introduced by Mitzenmacher et al [6], [7]). We say a D-repeat channel is square-integrable if D has finite first and second moments. In general, the output of a D-repeat channel has random length, and does not preserve synchronization; that is, one cannot see the index at the input of a given observed bit at the output. This introduces memory into the channel, making its analysis much more complicated than its memoryless counterparts. For example, in stark contrast with the memoryless case, even in the simplest case of the binary deletion channel (where D = Bernoulli(p)), the capacity is unknown, although several lower and upper bounds have been proved (see [8], [9] for two excellent surveys on synchronization channels).

More recently, progress has been made on constructing explicit and efficient codes whose rates approximate the state of the art lower bounds on capacity for certain simple special cases of repeat channels. Guruswami and Li [4] gave the first explicit and efficient codes for the deletion channel with $\Theta(1-d)$ rate, where d is the deletion probability, achieving a rate of $(1-d)/120$. This was later improved by Con and Shpilka [5] to $(1-d)/16$. Tal et al. [1] gave a construction using polar codes proved to achieve the capacity of the deletion channel by considering a sequence of hidden-markov input processes that approach the maximum mutual information. In [2], their construction was extended to a more general model of synchronization errors, which allows for simple insertions and bit flips. Very recently, Rubinstein [3] gave a black-box construction, which takes a general (inefficient and non-explicit) code for the deletion channel or the Poisson repeat channel of a given rate R and produces an efficient and explicit code of rate $R-\varepsilon$, for any $\varepsilon > 0$. In particular, this yields an efficient and explicit code achieving capacity on these channels. However, to our knowledge, no efficient and explicit code construction has been given of even non-trivial rate for general square-integrable repeat channels.

In this paper, we show that, by extending the techniques of [4], [5], we can obtain codes for any square-integrable repeat channel that are efficiently encodable and decodable and of rate within ε of the capacity, for any $\varepsilon > 0$. In the full version [10], we also illustrate how our construction can give explicit, efficient capacity achieving codes for an even broader class of channels capturing insertions, deletions, and substitutions.

As mentioned above, similar results appeared in the literature before, and our result differs in the following ways. First, the work [3] proves the same result for the deletion channel and the Poisson repeat channel. Our construction generalizes the result to general repeat channels, and we illustrate how it can be generalized further to channels capturing insertions, deletions, and substitutions (see the full version, [10]). We

FP is supported by CURIS 2021. RL is supported by NSF Grants DGE-1656518, CCF-1814629. MW is partially supported by NSF Grant CCF-1844628 and by a Sloan Research Fellowship. We thank Ido Tal for pointing out an error in our description of [1], [2] in an earlier version of this work, and anonymous reviewers for helpful comments.

Authorized licensed use limited to: Stanford University. Downloaded on February 16,2023 at 00:14:37 UTC from IEEE Xplore. Restrictions apply.

also believe our proof is simpler and contains techniques that may be of independent interest. Second, the works [1], [2] obtain similar results for the deletion channel [1] and insertion/deletion/substitution channels [2]. Compared to these works, we give slightly faster decoding algorithms and slightly smaller error probability: for any $0 < \nu' < \nu < 1/3$, [1], [2] give decoding error probability $e^{-\Omega(n^{\nu'})}$ in time $O(n^{1+3\nu})$, while we achieve decoding error probability $e^{-\Omega(n)}$ in time $O(n \operatorname{poly} \log n)$.

A. Organization

In Section II we review some background material needed for the proofs. In Section III we give the construction, and prove its correctness. We refer the reader to the full version of the paper for a more detailed comparison between our work and [3], who obtains similar results. In the full version of the paper, we also give an overview of how our results can be extended to a more general error model.

II. Preliminaries

A. Notation and Basic Concepts

In what follows, $\{0,1\}^n$ for $n \in \mathbb{N} \cup \{\infty\}$ denotes the set of bit strings of length n; we also let $\{0,1\}^* = \bigcup_{n \in \mathbb{N}} \{0,1\}^n$. For $x \in \{0,1\}^n$, we let x_j^k denote the substring of x starting at index j and ending at k, inclusive, and unless specified otherwise, we let $x_i := x_i^i$. For $n \in \mathbb{N}$, we let $[n] = \{1, 2, \ldots, n\}$; even if $n \in \mathbb{R}_+$, we let $[n] := [\lfloor n \rfloor]$. For two strings $x, y \in \{0,1\}^*$, we let xy denote their concatenation, and for $k \in \mathbb{N}$, $(x)^k$ denotes the k-wise concatenation of x with itself; we let $(x)^0$ be the empty string. For $x \in \{0,1\}^n$, $|x| = n$ denotes the length of x. We denote the capacity of an arbitrary channel Ch by $\operatorname{Cap}(Ch)$. All logs (hence entropies, etc.) in this paper are base 2. For a probability distribution D over $\mathbb{R}$, we let $\mu(D)$ denote the expectation; whenever D is understood we sometimes just write μ. Similarly we let $\sigma^2(D)$ denote the variance, and we sometimes just write σ^2. Throughout, "quasi-linear time" means $O(n \operatorname{poly}(\log n))$ time.

For completeness, we give a definition of a general binary communication channel, introducing further notation.

Definition II.1. For Ω a probability space, a binary communication channel is a map $Ch : \Omega \times \{0,1\}^* \to \{0,1\}^*$. For $x \in \{0,1\}^*$, we write $Ch(x)$ to denote the random variable $\omega \to Ch(\omega, x)$.

In this paper we deal specifically with square-integrable binary repeat channels, which we define next.

Definition II.2. For a probability distribution D over $\mathbb{N}$, let $\Omega = \mathbb{N}^\infty$ (the infinite product space), with a D^∞ measure (the infinite product measure). The binary D-repeat channel is defined as $RC_D(\omega, x) = (x_1)^{\omega_1} (x_2)^{\omega_2} \ldots (x_n)^{\omega_n}$ for $x \in \{0,1\}^n$. We say RC_D is square-integrable if $\mu(D) < \infty$ and $\sigma^2(D) < \infty$.

That is, each bit sent through the RC_D gets repeated $R \sim D$ times. We note that it is well-defined to speak of the index at the input that gave rise to a given bit at the output: the origin bit of the jth bit at the output is the $\min\{i \geq 1 : \sum_{k=1}^{i} \omega_k \geq j\}$'th bit at the input.

Finally we define the trimming repeat channels, which unlike the objects defined above are non-standard, but which are an important part of our construction. We note that "trimming versions" of synchronization channels appear in the works [1], [2] and play a role similar to the one in our construction.

Definition II.3. Let $TRIM$ be a (deterministic) channel which acts on $x \in \{0,1\}^n$ by deleting the longest possible substrings at the beginning and end of x consisting entirely of zeros. Specifically, $TRIM(x) = x_{\min\{i \in [n] : x_i = 1\}}^{\max\{i \in [n] : x_i = 1\}}$ (or the empty string if x is all zeros). Let RC_D be as in Definition II.2. We then define the trimming D-repeat channel by the composition $TRC_D := TRIM \circ RC_D$.

B. Dobrushin's Theorem

For the square-integrable D-repeat channels, as well as a wide class of other synchronization channels, Dobrushin [11] showed that the capacity is given by a certain limit of the finite-length message maximum mutual information between input and output; this extended the fundamental result of Shannon [12] for memoryless channels. Here we state his theorem in our setting and notation. We refer the reader to the excellent survey of Cheraghchi and Ribeiro [8] for an illuminating discussion. Before the theorem we give a simple (non-general) definition of a stationary ergodic process, which will be important in our proof.

Definition II.4. A stochastic process $\{X_j\}_{j \geq 1}$ is stationary if for every $j, N \in \mathbb{N}$ we have $(X_1, \ldots, X_N) \stackrel{D}{=} (X_{j+1}, \ldots, X_{j+N})$, where $\stackrel{D}{=}$ denotes equality in distribution. Moreover, the process $\{X_j\}_{j \geq 1}$ is stationary ergodic if it is stationary and it satisfies Birkhoff's Pointwise Ergodic Theorem, i.e. for every $f \in L^1$ we almost surely have

$$Ef(X_1) = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} f(X_j).$$

Theorem II.5 (Dobrushin's Theorem [11]). Consider a channel Ch that acts independently on each input bit, and appends the corresponding outputs, i.e. we have $Ch(x) \stackrel{D}{=} (Ch^1(x_1))(Ch^2(x_2)) \ldots (Ch^n(x_n))$ for $x \in \{0,1\}^n$, where the Ch^i are independent copies of Ch. Suppose further that $E|Ch(b)| < \infty$ for $b \in \{0,1\}$, i.e. the channel output has finite expected length for each input bit. Then the capacity is given by

$$\operatorname{Cap}(Ch) = \lim_{n \to \infty} \frac{1}{n} \sup_{X^n} I(X^n; Y^n),$$

where the sup is taken over all random variables X^n supported on $\{0,1\}^n$ and $Y^n = Ch(X^n)$. Moreover, the capacity is achieved by a stationary ergodic input process.

We remark that Theorem II.5 applies to square-integrable repeat channels. When understood from context, we will drop the parameter n and just write X for a random variable supported on $\{0,1\}^n$, and let $Y = Ch(X)$. For channels

Authorized licensed use limited to: Stanford University. Downloaded on February 16,2023 at 00:14:37 UTC from IEEE Xplore. Restrictions apply.

for which Dobrushin's Theorem does not necessarily apply (like trimming D-repeat channels), we refer to the limit $\lim_{n\to\infty} \frac{1}{n} \sup_{X^n} I(X^n; Y^n)$ as the information rate of the channel. We emphasize that the fact that the capacity is attained by a stationary ergodic process in Theorem II.5 will be instrumental in our construction.

C. Worst-case insertion/deletion codes

We now state the result of Haeupler and Shahrasbi [13], which is an important component of our construction, as it was in [4] and [5].

Theorem II.6 ([13], [14]). For every $\varepsilon, \delta \in (0, 1)$ there exists a family of codes C_n of rate $1-\delta-\varepsilon$ over an alphabet Σ of size $O_\varepsilon(1)$ that can (deterministically) correct insertion/deletion (worst-case) errors resulting in an edit distance (the minimum number of insertions and/or deletions to convert the input into the ouput) at most δn. Moreover, the C_n have encoding and decoding algorithms that run in linear and quasi-linear time, respectively.

III. Main Result

Our main result is a proof of existence of efficient near-optimal codes for square-integrable repeat channels with rates approaching capacity. When restricted to the binary deletion channel or the Possion repeat channel, our construction streamlines the approach of [3]. Specifically, we prove the following:

Theorem III.1. Fix a square-integrable repeat channel RC_D. For every $\varepsilon > 0$, there exists a family of codes $\{C_n\}$ with rate R for the RC_D with $R \geq Cap(RC_D) - \varepsilon$ and linear and quasi-linear time encoding and decoding algorithms, respectively. Moreover, the decoder has probability of failure $e^{-\Omega(n)}$.

We organize the remaining of this section as follows: in Section III-A we give the construction, and in Section III-B we give the proof of correctness.

A. Construction

We prove in Lemma III.2 that the information rates of the RC_D and TRC_D are the same. In Proposition III.4, we further show that we can assume the existance of a general (non-explicit and inefficient) code C_{in} for the TRC_D such that each sufficiently long substring of each codeword in C_{in} is approximately balanced in zeros and ones (see Proposition III.4), with rate $R \geq Cap(RC_D) - \varepsilon$, for any $\varepsilon > 0$. This will be the inner code in our construction, which we assume has (not necessarily efficient) encoding and decoding algorithms Enc_{in} and Dec_{in}, respectively. Then, as in the work of Con and Shpilka [5], for a codeword length m to be fixed later, we take 2^m as the desired alphabet size for the [13], [14] code (i.e. $|\Sigma| = 2^m$ in Theorem II.6), making sure to take m large enough for the code of [13], [14] to be effective. Our encoding procedure $Enc : \{0,1\}^{km} \to \{0,1\}^n$ for some $x \in \{0,1\}^{km}$ works as follows:

1) We split x into $x_1, \ldots, x_k$, with $|x_j| = m$, and we view each x_j as a member of Σ, hence $x \in \Sigma^k$. We then use the encoder of Theorem II.6 (call it Enc_{out}) to encode x. This yields $\mathbf{z} = Enc_{out}(x) \in \Sigma^{k/(1-\delta-\varepsilon)}$.
2) We again split $\mathbf{z}$ into $\mathbf{z}_1, \ldots, \mathbf{z}_{k'}$ where $\mathbf{z}_j \in \Sigma$, $k' = k/(1-\delta-\varepsilon)$, and view each $\mathbf{z}_j$ as an element in $\{0,1\}^m$. We then encode each $\mathbf{z}_j$ with our inner code to produce $\mathbf{b}_j = Enc_{in}(\mathbf{z}_j) \in \{0,1\}^{m/(R-\varepsilon)}$, where, by taking m large enough, we have made the rate of the inner code $R - \varepsilon$. We note that since $m = O(1)$, this can be done in time $O(1)$.
3) Finally we concatenate the $\mathbf{b}_j$ and put buffers of all zeros in between. Specifically, our final encoding of x is

$$Enc(x) = \mathbf{b}_1 0^b \mathbf{b}_2 0^b \ldots 0^b \mathbf{b}_{k'},$$

where $b = b(m) = \eta m$ is a constant independent of $n = k' \cdot (\frac{m}{R-\varepsilon} + b) = km/(Cap(RC_D) - \psi(\varepsilon, \delta, \eta, k, m))$ with $\psi \to 0$ as $\varepsilon, \delta, \eta \to 0$ and $k, m \to \infty$, so by taking $\varepsilon, \delta, \eta$ small enough and m large enough, we can make the rate of the code get arbitrarily close to $Cap(RC_D)$.

We note that since Enc_{out} runs in linear time, so does our encoding Enc. For the decoding Dec of a received string $y \in \{0,1\}^*$, we reverse the steps above:

1) We identify the buffers of zeros by interpreting any maximal contiguous block of $\geq \frac{\mu}{2}\eta m$ zeros as a buffer. We remove the buffers, producing the received inner strings $y_1, \ldots, y_\ell$.
2) We decode each y_j with our inner code to produce $\mathbf{z}_j = Dec_{in}(y_j) \in \{0,1\}^m$ for $j \leq \ell$.
3) We interpret each $\mathbf{z}_j$ as a symbol in Σ, and we decode the concatenation $\mathbf{z} = \mathbf{z}_1 \ldots \mathbf{z}_\ell \in \Sigma^\ell$ with the outer code, to produce our final decoding of y:

$$Dec(y) = Dec_{out}(\mathbf{z}) \in \{0,1\}^{km}.$$

We note that the identification of the buffers runs in linear time and Dec_{out} runs in quasi-linear time, hence our overall decoding Dec runs in quasi-linear time as well.

B. Proof of Correctness

We organize the proof of Theorem III.1 as follows. First we prove that the information rates of the repeat channels are unchanged if we trim off the zeros at the ends of the output. Second, we argue that we can assume there exist capacity-achieving codes with a sufficiently balanced distribution of zeros and ones in all its codewords. Finally, we put these results together into our proof of correctness of the construction given in Section III-A. We begin with the first required result.

Lemma III.2. Let RC_D be a square-integrable repeat channel. Then the information rate of TRC_D is $Cap(RC_D)$.

Proof. Let X be supported on $\{0,1\}^n$ and $Y = RC_D(X)$. Let $L = \min\{i \in [n] : Y_i = 1\}, R = \max\{i \in [n] : Y_i = 1\}$ be as in Definition II.3 for the random string Y : they are the (random) indices that mark the all-zero substrings that would be trimmed if Y were passed through $TRIM$. Now let L, R be the indices of the bits in X that, when X is passed through the RC_D, end up at indices L, R in Y.

Authorized licensed use limited to: Stanford University. Downloaded on February 16,2023 at 00:14:37 UTC from IEEE Xplore. Restrictions apply.

Claim III.3. Let $Y = RC_D(X)$ and $Y' = TRC_D(X)$. Then

$$|I(X;Y) - I(X;Y|\mathcal{L},\mathcal{R})| = o(n) \quad (1)$$

and

$$\lim_{n\to\infty} \frac{1}{n} \sup_X I(X;Y|\mathcal{L},\mathcal{R}) = \lim_{n\to\infty} \frac{1}{n} \sup_X I(X;Y') \quad (2)$$

By Dobrushin's Theorem, Claim III.3 proves the lemma. For the proof of Claim III.3, we will assume that that there exists deterministic $B > 0$ such that if $R \in D$, then $R \le Bn$ with probability 1. This assumption is without loss; if D has unbounded support, we consider the "truncation of D at Bn," denoted D_n: for $R \in D$, $R' \in D_n$, we define

$$P(R' = k) := \begin{cases} P(R = k)/(\sum_{\ell \le Bn} P(R = \ell)) & \text{if } k \le Bn \\ 0 & \text{otherwise.} \end{cases}$$

Note that for each $x \in \{0,1\}^n$, we have $TRC_{D_n}(x) \overset{D}{=} TRC_D(x)$ and $RC_{D_n}(x) \overset{D}{=} RC_D(x)$ outside the set $\bigcup_{j=1}^n \{R_i > Bn\}$, where $R_i \in D$ is the number of repetitions of x_i when passed through TRC_D or RC_D. By Chebyshev's inequality and a union bound, this set has probability $O(n^{-1})$. It's easy to show that this implies that the information rates of TRC_{D_n} and TRC_D are the same, as are the information rates of RC_{D_n} and RC_D.[1] We refer the reader to the appendix of the full version [10] for a complete proof of this simple fact.

Now assuming D is bounded by $Bn > 0$ as above, for (1), we have

$$I(X;Y|\mathcal{L},\mathcal{R}) = H(X|\mathcal{L},\mathcal{R}) - H(X|Y,\mathcal{L},\mathcal{R}).$$

We also have $H(X|\mathcal{L},\mathcal{R}) \le H(X)$, and by the chain rule,

$$\begin{aligned} H(X|\mathcal{L},\mathcal{R}) &= H(X,\mathcal{L},\mathcal{R}) - H(\mathcal{L},\mathcal{R}) \\ &\ge H(X) - H(\mathcal{L},\mathcal{R}), \end{aligned}$$

so $H(X|\mathcal{L},\mathcal{R}) - H(X) \le 0$ and $H(X|\mathcal{L},\mathcal{R}) - H(X) \ge -H(\mathcal{L},\mathcal{R})$, hence $|H(X) - H(X|\mathcal{L},\mathcal{R})| \le H(\mathcal{L},\mathcal{R})$ and by an identical derivation also $|H(X|Y) - H(X|Y,\mathcal{L},\mathcal{R})| \le H(\mathcal{L},\mathcal{R})$. Hence by the triangle inequality $|I(X;Y) - I(X;Y|\mathcal{L},\mathcal{R})| \le 2H(\mathcal{L},\mathcal{R}) \le 4\log n = o(n)$ since $(\mathcal{L},\mathcal{R})$ is supported in $[n]^2$, proving (1). For (2), we have

$$\begin{aligned} I(X;Y|\mathcal{L},\mathcal{R}) &= I(X_1^{\mathcal{L}}, X_{\mathcal{L}+1}^{\mathcal{R}-1}, X_{\mathcal{R}}^n; Y_1^L, Y_{L+1}^{R-1}, Y_R^n) \\ &= I(X_1^{\mathcal{L}}, X_{\mathcal{L}+1}^{\mathcal{R}-1}, X_{\mathcal{R}}^n; Y_{L+1}^{R-1}) + I(X_1^{\mathcal{L}}, X_{\mathcal{L}+1}^{\mathcal{R}-1}, X_{\mathcal{R}}^n; Y_1^L, Y_R^n) \\ &\le I(X; Y_{L+1}^{R-1}|\mathcal{L},\mathcal{R}) + H(Y_1^L, Y_R^n) \\ &\le I(X; Y_{L+1}^{R-1}|\mathcal{L},\mathcal{R}) + H(L,R,|Y|) \\ &= I(X; Y_{L+1}^{R-1}|\mathcal{L},\mathcal{R}) + o(n), \end{aligned}$$

where the penultimate inequality is because by definition, Y_1^L and Y_R^n are strings of all zeros, so they are uniquely specified if the length of Y and the indices L and R are given, and the last equality is because $(L,R,|Y|)$ is supported on $[Bn]^3$. Now by the same argument as in (1), we again obtain $|I(X;Y_{L+1}^{R-1}|\mathcal{L},\mathcal{R}) - I(X;Y_{L+1}^{R-1})| = o(n)$, and since $Y_{L+1}^{R-1} = TRIM(Y)$, we get $|I(X;Y|\mathcal{L},\mathcal{R}) - I(X;Y')| = o(n)$, where $Y' = TRC_D(X)$. This then gives

$$\lim_{n\to\infty} \frac{1}{n} \sup_X I(X;Y|\mathcal{L},\mathcal{R}) = \lim_{n\to\infty} \frac{1}{n} \sup_X I(X;Y'),$$

where $Y = RC_D(X)$ and $Y' = TRC_D(X)$, proving (2). This proves Claim III.3 and hence the lemma. □

Next, we show that we may assume an approximately balanced distribution of zeros and ones in all sufficiently long substrings of all codewords in an information-rate-achieving code. The following lemma, though simple, constitutes the substantial improvement in our argument as compared to those of [5] or [4]. We remark that this result is much more general than just the setting of repeat channels, and in particular applies to all channels to which Dobrushin's Theorem II.5 applies; for simplicity we state the lemma in the context relevant to our proof.

Proposition III.4. Fix a square-integrable D-repeat channel $Ch = RC_D$, or the trimming version $Ch = TRC_D$, with information rate I. For every $\zeta, \varepsilon \in (0,1)$ there exists $\gamma \in (0, \frac{1}{2})$ and a family of codes $C_n \subseteq \{0,1\}^n$ for Ch with rate $R \ge I - \varepsilon$ such that for every $c \in C_n$ and $i \in [n - \zeta n]$, we have $\gamma\zeta n \le w(x_i^{i+\zeta n}) \le (1-\gamma)\zeta n$, where $w : \{0,1\}^* \to \mathbb{N}$ is the Hamming weight (number of ones).

Proof. The result follows from the fact that in Dobrushin's Theorem, we may assume that the process which achieves the information rate is stationary ergodic (see Theorem II.5). Even if we deal with the trimming version of such a channel, by Lemma III.2, the same statement holds.[2] Now let $\{X_j\}_{j\ge 1}$, with $X_j \in \{0,1\}$, be a stationary ergodic process such that

$$I = \lim_{n\to\infty} \frac{1}{n} I(X_1^n; Y^n),$$

where $Y^n = Ch(X_1^n)$ and where X_1^n denotes the first n symbols of the process $\{X_j\}_{j\ge 1}$. Let $P := P(X_1 = 1)$, and note that by stationarity we have $P \in (0,1)$ or else $\{X_j\}$ is a trivial process, hence does not achieve the information rate. By Birkhoff's pointwise ergodic theorem, almost surely

$$P = \lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{X^t} 1\{X_j = 1\} = \lim_{t\to\infty} \frac{1}{t} w(X_1^t),$$

so in particular setting $t = \zeta n$, for any $\delta > 0$, with probability $p_n \to 1$, we have $(P-\delta)\zeta n \le w(X_1^{\zeta n}) \le (P+\delta)\zeta n$. Picking δ, γ small enough, we can ensure that $\gamma\zeta n \le w(X_1^{\zeta n}) \le (1-\gamma)\zeta n$ with probability p_n. Now to extend to the substrings,

[1] More precisely, we consider the of channel Ch which acts on $x \in \{0,1\}^n$ as $Ch(x) = RC_{D_n}(x)$, (or $Ch'x = TRC_{D_n}(x)$), where D_n is the described truncation of D. By the "information rate" of e.g. RC_{D_n} we mean the information rate of Ch.

[2] In fact, the statement of Lemma III.2 is that the information rates coincide; but by looking at the proof it is clear that we prove the stronger statement that each fixed process $\{X_j\}_{j\ge 1}$ satisfies $\lim_{n\to\infty} \frac{1}{n} I(X;Y) = \lim_{n\to\infty} \frac{1}{n} I(X;Y')$ for $Y = RC_D(X)$ and $Y' = TRC_D(X)$. Hence the information rate of the TRC_D is again attained by the stationary ergodic processes.

Authorized licensed use limited to: Stanford University. Downloaded on February 16,2023 at 00:14:37 UTC from IEEE Xplore. Restrictions apply.

we first look at disjoint consecutive blocks: by stationarity we have $X_{i\zeta n+1}^{(i+1)\zeta n} \stackrel{D}{=} X_1^{\zeta n}$ for all $i \in [1/\zeta]$, so by a union bound over a constant $1/\zeta$ number of substrings, with probability $\to 1$ we have $\gamma\zeta n \le w(x_{i\zeta n+1}^{(i+1)\zeta n}) \le (1-\gamma)\zeta n$ simultaneously for all $i \in [1/\zeta]$. But we note that each substring $x_i^{i+3\zeta n}$ fully contains at least one block substring of the form $x_{j\zeta n+1}^{(j+1)\zeta n}$; hence $\frac{\gamma}{3}3\zeta n \le w(x_i^{i+3\zeta n}) \le (1-\frac{\gamma}{3})3\zeta n$ for all $i \in [n-\zeta n]$ simultaneously with probability $\to p_n$. Then resetting $\zeta' = 3\zeta$ and $\gamma' = \gamma/3$ yields the property of the lemma with probability $\to p_n$. Finally we note that we may extract a family of codes $\mathcal{C}_n$ of rate $R \ge \lim_{n\to\infty} \frac{1}{n} I(X_1^n; Y^n) - \varepsilon$ from $\{X_j\}$ via sampling, as in the standard proof of Shannon's theorem, and as extended by Dobrushin [11] (see also [15], Theorem 7.7.1). Since with high probability this process satisfies the required property, we may discard any codewords from $\mathcal{C}_n$ that don't satisfy it to obtain our desired family of codes C_n of the same rate. This concludes the proof. □

Proof of Theorem III.1. It remains to show that the decoding algorithm Dec described in Section III-A succeeds with high probability, for properly chosen (independent of n) inner code blocklength m. There are four potential sources of error in the decoding; the first three pertain to identifying the buffers of zeros, and the fourth to the inner code failures.

1) For a given buffer 0^b at the sender, less than $\frac{\mu}{2}b = \frac{\mu}{2}\eta m$ zeros survive, so the buffer is not identified.
2) All ones in a given inner codeword are deleted, so two adjacent buffers are incorrectly merged during decoding.
3) A substring of a received inner word longer than $\frac{\mu}{2}\eta m$ arrives with all zeros, so that a spurious buffer appears.
4) For a given correctly identified received inner word, the inner code decoding fails.

We note that error (1) results in the merging of two inner codewords in the decoding process. Since this merged codeword is not the output of the TRC_D with an inner codeword as input, we have no guarantee of a small probability of decoding error of the inner code. We consider the worst-case scenario: assume the inner decoding always fails in this string. At the outer code level, this then results in the deletion of two letters, and the insertion of another in the same location, i.e. an edit distance of 3. For error (2), we clearly have a deletion at the outer code level, i.e. an edit distance of 1. For error (3), we again cannot assume the inner code will succeed in decoding these two halves of a received codeword, and hence we assume the worst case scenario: one deletion and two insertions, i.e. edit distance 3. Finally for error (4) we clearly have a substitution at the outer code level, (which is equivalent to a deletion followed by an insertion), i.e. edit distance 2.

Now suppose that each of these errors occurs at most $k\delta/9$ times. Then the total edit distance is at most $k\delta/9 \cdot (3+1+3+2) = k\delta$. Hence to conclude the proof we must show that each error occurs more than $k\delta/9$ times with vanishing probability, for properly chosen m. This then implies that our outer code has to correct from an edit distance more than $k\delta$ with vanishing probability, i.e. the outer code succeeds with probability approaching 1 as $k \to \infty$ (hence $n \to \infty$).

Error (1) occurs with probability $P(|RC_D(0^{m\eta})| < \frac{\mu}{2}\eta m) = O(m^{-1})$ by Chebyshev's inequality. Hence for any η, taking m a large enough constant we can make this probability less than $\delta/10$. Since this error can happen independently for each of the $k-1$ buffers, the number of buffers that suffer from error (1) is given by a $Binomial(k-1, p)$ random variable, where $p \le \delta/10$. Again by a standard concentration bound, the probability that there are more than $k\delta/9$ errors vanishes as $k \to \infty$, as desired.

By Proposition III.4, each inner codeword has at least γm ones, for some $\gamma > 0$ independent of m. Hence error (2) occurs with probability $d^{\gamma m}$. As before, we take m large enough such that $d^{\gamma m} < \delta/10$, and then as $k \to \infty$, the probability of having more than $k\delta/9$ errors vanishes.

We now consider error (3). Consider the event that we receive a string s of all zeros with $|s| \ge \frac{\mu}{2}\eta m$ as part of the output of the channel for a codeword $x \in C_{in}$ as input. This implies one of two things: (a) that some substring $\tilde{s}$ of length $> \frac{1}{4}\eta m$ of the input had all its one bits deleted and gave rise to s, or (b) that some substring $\tilde{s}$ of length $\le \frac{1}{4}\eta m$ at the input gave rise to any string of length $\ge \frac{1}{2}\mu\eta m$ at the output. We analyze each case separately. For (a), by Proposition III.4, choosing $\zeta = \frac{1}{2}\mu\eta$, we must have $w(\tilde{s}) \ge \gamma\zeta m$. But then the probability that such a substring s, say at the beginning of the received word, exists in the first place is less than $d^{\gamma\zeta m}$, and by a union bound the probability that any such substring exists is less than $O(1) \cdot d^{\gamma\zeta m}$ (since the received word has length $\le m$, and hence we can discretize it into $O(1)$ substrings of size $\ge \frac{\mu}{2}\eta m$) which can be made less than $\delta/20$ for m chosen large enough. For (b), note that a substring of length $\le \frac{1}{4}\eta m$ at the input giving length $\ge \frac{1}{2}\mu\eta m$ at the output implies that there's a substring of length exactly $\frac{1}{4}\eta m$ giving an output of length $\ge \frac{1}{2}\mu\eta m$ (since a bigger input can only give a bigger output). But if $Z = X_1 + \cdots + X_t$, for $t = \frac{1}{4}\eta m$ and $X_j \in D$, the probability of this happening is

$$P(Z \ge \tfrac{1}{2}\mu\eta m) \le P\left[|Z - EZ| \ge \tfrac{1}{4}\mu\eta m\right] \le \frac{t\sigma^2}{(\frac{1}{4}\mu\eta m)^2} = \frac{\sigma^2}{\frac{1}{4}\mu\eta m} = O(m^{-1})$$

by Chebyshev's inequality. Again by a union bound over $O(1)$ possible initial substrings s, making m large enough we can make this $\le \delta/20$. Hence, the probability of error (3) is $\le \delta/20 + \delta/20 = \delta/10$, and by concentration of measure, more than $\delta/9$ errors occur with vanishing probability.

Error (4) occurs with vanishing probability as $m \to \infty$ by soundness of the inner code for the TRC_D. For m large enough this probability is less than $\delta/10$, and hence as above when $k \to \infty$ we get $k\delta/9$ errors with vanishing probability.

Finally, the error probability is $e^{-\Omega(n)}$ because, as was mentioned, the frequency of each error type (1-4) is a $Binomial(t, p)$ random variable with $t = k-1$ or $t = k$ and $p \le \delta/10$. Hence by a standard Chernoff bound, and union bounding over errors (1-4), we obtain the desired $e^{-\Omega(n)}$ probability of edit distance greater than $k\delta/9$, i.e. a $e^{-\Omega(n)}$ probability of failure. This concludes the proof. □

Authorized licensed use limited to: Stanford University. Downloaded on February 16,2023 at 00:14:37 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] I. Tal, H. Pfister, A. Fazeli, and A. Vardy, "Polar codes for the deletion channel: Weak and strong polarization," 07 2019, pp. 1362–1366.

[2] H. Pfister and I. Tal, "Polar codes for channels with insertions, deletions, and substitutions," 02 2021.

[3] I. Rubinstein, "Explicit and efficient construction of (nearly) optimal rate codes for binary deletion channel and the poisson repeat channel," 2021.

[4] V. Guruswami and R. Li, "Polynomial time decodable codes for the binary deletion channel," IEEE Transactions on Information Theory, 2019.

[5] R. Con and A. Shpilka, "Explicit and efficient constructions of coding schemes for the binary deletion channel," 2020 IEEE International Symposium on Information Theory (ISIT), pp. 84–89, 2020.

[6] M. Mitzenmacher and E. Drinea, "A simple lower bound for the capacity of the deletion channel," IEEE Transactions on Information Theory, 2006.

[7] M. Mitzenmacher, "Capacity bounds for sticky channels," IEEE Transactions on Information Theory, 2008.

[8] M. Cheraghchi and J. L. Ribeiro, "An overview of capacity results for synchronization channels," IEEE Transactions on Information Theory, 2019.

[9] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," Probability Surveys, pp. 1 – 33, 2009.

[10] F. Pernice, R. Li, and M. Wootters, "Efficient near-optimal codes for general repeat channels," 2022, available at arXiv:2201.12745.

[11] R. L. Dobrushin, "Shannon's theorems for channels with synchronization errors," Problemy Peredachi Informatsii, 1967.

[12] C. E. Shannon, "A mathematical theory of communication," The Bell System Technical Journal, 1948.

[13] B. Haeupler and A. Shahrasbi, "Synchronization strings: Codes for insertions and deletions approaching the singleton bound," ser. STOC 2017, 2017, p. 33–46.

[14] B. Haeupler, A. Rubinstein, and A. Shahrasbi, "Near-linear time insertion-deletion codes and (1 + ε)-approximating edit distance via indexing," in Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing. Association for Computing Machinery, 2019.

[15] T. M. Cover and J. A. Thomas, Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). USA: Wiley-Interscience, 2006.

Authorized licensed use limited to: Stanford University. Downloaded on February 16,2023 at 00:14:37 UTC from IEEE Xplore. Restrictions apply.