Uncertainty of Reconstruction with List-Decoding from Uniform-Tandem-Duplication Noise

Yonatan Yehezkeally *Member, IEEE*, Moshe Schwartz *Senior Member, IEEE*

***Abstract*—We propose a list-decoding scheme for reconstruction codes in the context of uniform-tandem-duplication noise, which can be viewed as an application of the associative memory model to this setting. We find the uncertainty associated with $m > 2$ strings (where a previous paper considered $m = 2$) in asymptotic terms, where code-words are taken from an error-correcting code. Thus, we find the trade-off between the design minimum distance, the number of errors, the acceptable list size and the resulting uncertainty, which corresponds to the required number of distinct retrieved outputs for successful reconstruction. It is therefore seen that by accepting list-decoding one may decrease coding redundancy, or the required number of reads, or both.**

***Index Terms*—DNA storage, reconstruction, string-duplication systems, list decoding.**

I. Introduction

WITH recent improvements in DNA sequencing and synthesis technologies, and the advent of CRISPR/Cas gene editing technique [26], the case for DNA as a data-storage medium, specifically *in-vivo*, is now stronger than ever before. It offers a long-lasting and high-density alternative to current storage media, particularly for archival purposes [6]. Moreover, due to medical necessities, the technology required for data retrieval from DNA is highly unlikely to become obsolete, which as recent history shows, cannot be said of concurrent alternatives (e.g., the floppy disk, compact cassete, VHS tape, etc.).

In-vivo DNA storage has somewhat lower data density than in-vitro storage, but it provides a reliable and cost-effective propagation via replication, in addition to some protection to stored data (see [2], [12], [13] and references therein). It also has applications including watermarking genetically modified organisms [1], [10], [22] or research material [15], [30] and concealing sensitive information [7]. However, mutations introduce a diverse set of potential errors, including symbol- or burst-substitution/insertions/deletion, and duplication (including tandem- and interspersed-duplication).

This work was presented in part at ISIT'2020.

This work was supported in part by the Israel Science Foundation (ISF) under grant no. 270/18.

Moshe Schwartz is with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer Sheva 8410501, Israel (e-mail: schwartz@ee.bgu.ac.il).

Yonatan Yehezkeally was with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer Sheva 8410501, Israel. He is now with the Institute for Communications Engineering, Technical University of Munich, 80333 Munich, Germany (e-mail: yonatan.yehezkeally@tum.de).

The effects of duplication errors, specifically, were studied in a number of recent works including [11], [12], [16]–[20], [23], [25], [27]–[29] among others. These works provided some implicit and explicit constructions for uniform-tandem-duplication codes, as well as some bounds. In [33] the authors then argued that a classical error-correction coding approach is sub-optimal for the application, as it does not take advantage of the cost-effective data replication offered inherently by the medium of in-vivo DNA; instead, it was shown that re-framing the problem as a *reconstruction* scheme [21] reduces the redundancy required for any fixed number of duplication errors. In this setting, several (distinct) noisy channel outputs are assumed to be available to the decoder. Since its introduction, several applications of the reconstruction problem to storage technologies were found [3], [5], [31], [32]. Of these, [31] in particular extended the reconstruction model to *associative memory*, where one retrieves the set of all entries (or code-words) *associated* with every element of a given set. For a given size of entry set, the maximal number of entries being possibly associated with all of them was dubbed the *uncertainty* of the memory.

Study of this extended model for in-vivo DNA data storage is motivated by a list-decoding reconstruction scheme, whereby tolerance for decoding a list of possible inputs, given multiple channel outputs, enables coding with a lower minimum distance, thereby reducing the redundancy of the code. Alternatively, given the same code, it allows reducing the number of required outputs for reconstruction.

This paper focuses on uniform tandem-duplication noise; i.e., we assume throughout that the length of duplication window is fixed. In practical applications, a more complex model where that length is permitted to belong to some set, or perhaps is simply bounded, is more realistic; however, we focus on this model as a step towards that end. Our main goal is to analyze the uncertainty associated with codes which are subsets of a typical set of strings (consisting of most strings in Σ^n, a definition which is made precise in Lemma 4) as a function of the acceptable list size m and code minimum distance d. In our analysis, the number of tandem repeats t which channel outputs undergo is fixed.

The paper is organized as follows: In Section II we describe the main contribution of this paper, put it in context of related works, and discuss possible directions for future study. In Section III we present notations and definitions. Then, in Section IV, we find the uncertainty of the aforementioned typical set in asymptotic form, and develop an efficient decoding scheme. We then extend and repeat our analysis in Section V for error-correcting codes contained in that typical set.

II. Related works and main contribution

Associative memory was discussed in [31], where items are retrieved by association with other items; the human mind seems to operate in this fashion, one concept bringing up memories of other, related, concepts or events. The more items one considers together, the smaller the set of items associated with all of them. More precisely, one defines the uncertainty of an associative memory as the largest possible size of set $N(m)$ whose members are associated with all elements of an m-subset of the memory code-book.

This model is a generalization of the reconstruction problem posed by Levenshtein in [21], wherein a transmission model is assumed with the decoder receiving multiple channel outputs of the same input. N is then the largest size of intersection of balls of radius t about two distinct code-words, where at most t errors are assumed to have occurred in each transmission; if $N+1$ outputs are available to the decoder, the correct input can be deduced.

This can be viewed as a reduction of the associative memory model to the case of $m=2$, allowing a precise reconstruction of the unique $(m-1=1)$ input. When $m>2$, the decoder seeing $N(m)+1$ channel outputs can only unambiguously infer which list of $l<m$ code-words contains the correct input; thus, a list-decoding model is suggested.

In [33] the authors studied the reconstruction problem for uniform-tandem-duplication noise, which is applicable to in-vivo DNA data storage. An uncertainty which is sub-linear in the message length was assumed (as it represents the number of distinct reads required for decoding), and it was shown that the redundancy required for unique reconstruction was $(t-1)\log_q(n)+O(1)$ (compared to the $t\log_q(n)+O(1)$ redundancy required for unique decoding from a single output [16], [18]), where n is the message length, t the number of errors, and q the alphabet size.

In this paper, we apply the associative memory model from [31] (where binary vectors with the Hamming distance were considered) to the setting of uniform-tandem-duplication noise in finite strings, i.e., we consider list-decoding instead of a unique reconstruction. We shall restrict our attention to code-books contained in a typical subspace, asymptotically achieving the full space size.

Our goal is to find the trade-off between the code redundancy, the number of tandem-duplication errors, the uncertainty, and the decoded list size. We find the asymptotic behavior, as the message length n grows, of the uncertainty, or required number of reads (more precisely, that number minus one) N, where it is viewed as a function of the list size (plus one) m, the design minimum distance d, and the number of tandem-duplication errors t. Our main contribution (see Corollary 28) can informally be summarized in

$$\log_n N + \lceil \log_n(m) \rceil + d = t + \epsilon + o(1),$$

where $\epsilon \in \{0,1\}$ is a non-increasing function of m, which we find. Thus, such a trade-off is established.

This can be seen as an extension to the results in [33], where unique reconstruction $(m=2)$ was required, and it was seen that coding with minimum distance $d=t$ enables sub-linear uncertainty (i.e., $\log_n(N)=o(1)$).

In conclusion, we show that list-decoding is not only theoretically feasible, but may be efficiently performed. This is done using an isometric transform to integer vectors, and by utilizing combination generators; efficient list-decoding algorithms are developed, given a sufficient number of distinct channel outputs. If the code-book is restricted, then this task is reduced to that of decoding an error-correcting code.

In the future, we believe that a study of reconstruction schemes, with or without list-decoding, is of interest with other error models which affect in-vivo DNA data storage; related models to uniform tandem-duplication noise, which have recently been studied on their own and may now be easier to analyze in that setting, and therefore are a logical first step in this direction, may be bounded tandem-duplication (see, e.g., [11], [12], [17]) or combined uniform-tandem-duplication and substitution noise [27]–[29].

III. Preliminaries

Let Σ^* denote the set of finite strings over an alphabet Σ, which is assumed to be a finite unital ring of size q (e.g., $\mathbb{Z}_q$, or when q is a prime power, $\mathrm{GF}(q)$).

The length of a string $x\in\Sigma^*$ is denoted $|x|$, and the concatenation of $x,y\in\Sigma^*$ is denoted xy. A tandem-duplication (or tandem repeat) of fixed duplication-window length k (thus, *uniform* tandem-duplication noise) at index i is defined as follows, for $a\in\Sigma^*$ such that $a=xyz$, $x,y,z\in\Sigma^*$, $|x|=i$ and $|y|=k$:

$$\mathcal{T}_i(a) \triangleq xyyz.$$

Thus, uniform tandem-duplication noise with duplication-window length k acts only on strings of length $\geqslant k$, which we denote $\Sigma^{\geqslant k}$. In order to simplify our analysis, we assume throughout the paper that $k\geqslant 2$.

If $y\in\Sigma^{\geqslant k}$ can be derived from $x\in\Sigma^{\geqslant k}$ by a sequence of tandem repeats, i.e., if there exist $i_1,\ldots,i_t$ such that

$$y=\mathcal{T}_{i_t}(\cdots\mathcal{T}_{i_1}(x)),$$

then y is called a *t-descendant* (or simply *descendant*) of x (vice versa, x is an *ancestor* of y), and we denote $x \overset{t}{\Longrightarrow} y$. We say that x is a 0-descendant of itself. If $t=1$ we denote $x\Longrightarrow y$. Where the number of repeats is unknown or irrelevant, we may denote $x\overset{*}{\Longrightarrow} y$. We define the set of t-descendants of x as

$$D^t(x) \triangleq \left\{ y\in\Sigma^* : x \overset{t}{\Longrightarrow} y \right\},$$

and the *descendant cone* of x as

$$D^*(x) \triangleq \left\{ y\in\Sigma^* : x \overset{*}{\Longrightarrow} y \right\} = \bigcup_{t=0}^{\infty} D^t(x).$$

If there exists no $z\neq x$ such that $z\overset{*}{\Longrightarrow} x$, we say that x is *irreducible*. The set of irreducible strings of length n is denoted $\mathrm{Irr}(n)$. It can be shown (see, e.g., [12]) that for all $y\in\Sigma^{\geqslant k}$ there exists a unique irreducible x, called the *duplication root* of y and denoted $\mathrm{drt}(y)$, such that $y\in D^*(x)$. This induces a partition of $\Sigma^{\geqslant k}$ into descendant cones; i.e., it induces an equivalence relation, denoted herein $\sim_k$.

A useful tool in studying uniform tandem-duplication noise is the *discrete derivative* ϕ defined for $x \in \Sigma^{\geqslant k}$:

$$\phi(x) \triangleq \hat{\phi}(x)\bar{\phi}(x),$$

where

$$\hat{\phi}(x) \triangleq x(1), x(2), \dots, x(k),$$
$$\bar{\phi}(x) \triangleq x(k+1) - x(1), \dots, x(|x|) - x(|x|-k).$$

Here, $x(i)$ denote the i^{th} letter of the string x; Note that $\hat{\phi}(x), \bar{\phi}(x)$, and consequently $\phi(x)$, are themselves strings in Σ^*. As seen, e.g., in [12], ϕ is injective, and if $\bar{\phi}(x) = uv$ for $u, v \in \Sigma^*$, $|u| = i$, then $\bar{\phi}(\mathcal{T}_i(x)) = u0^k v$. This was used in [33] to define the function $\psi_x \colon D^*(x) \to \mathbb{N}^{w+1}$ by

$$\psi_x(y) \triangleq (\lfloor u(1)/k \rfloor, \dots, \lfloor u(w+1)/k \rfloor),$$

if

$$\bar{\phi}(y) = 0^{u(1)} a_1 0^{u(2)} \dots a_w 0^{u(w+1)},$$

where $w = \mathrm{wt}(\bar{\phi}(x))$ and $a_1 \dots, a_w \in \Sigma \setminus \{0\}$. It was shown that ψ_x is a poset isomorphy, where $D^*(x)$ is ordered with $\overset{*}{\Longrightarrow}$ and $\mathbb{N}^{w+1}$ with the product order, which we denote by $\leqslant$. Further, when considering $\mathbb{N}^{w+1}$ as a poset with the product order, we shall use the notations $\vee, \wedge$ for the *supremum* and *infimum*, respectively; these are also the coordinatewise *maximum* and *minimum*, respectively.

A metric can be defined on $D^r(x)$ for each r (in particular, but not necessarily, when x is irreducible) in the following way:

Definition 1 *For any $r \in \mathbb{N}$, $x \in \Sigma^{\geqslant k}$, and $y_1, y_2 \in D^r(x)$, we define*

$$d(y_1, y_2) \triangleq \min\{t \in \mathbb{N} : D^t(y_1) \cap D^t(y_2) \neq \emptyset\}.$$

It is seen in [12] that this is well defined, in the sense that there does exist such t, for $y_1, y_2 \in D^r(x)$, such that $D^t(y_1) \cap D^t(y_2) \neq \emptyset$.

If we define on $\mathbb{N}^{w+1}$ the 1-norm

$$\|u\|_1 \triangleq \sum_{i=1}^{w+1} u(i),$$

and metric

$$d_1(u, v) \triangleq \tfrac{1}{2}\|u - v\|_1,$$

then ψ_x is also an isometry (see [33]) between $D^r(x)$, for each r, and its image in $\mathbb{N}^{w+1}$, which is the simplex

$$\Delta_r^w \triangleq \{u \in \mathbb{N}^{w+1} : \|u\|_1 = r\} = \psi_x(D^r(x)).$$

Here, $\psi_x(D^r(x))$ is the image of ψ_x; in more generality, for any code $C \subseteq D^r(x)$ we let $\psi_x(C) \triangleq \{\psi_x(y) : y \in C\}$.

To simplify analysis in the $\mathbb{N}^{w+1}$ domain, we make the following notation:

Definition 2 *For $w, r, s \in \mathbb{N}$ and $u \in \Delta_{r+s}^w$, denote the* lower-bounds set

$$A_r(u) \triangleq \{v \in \Delta_r^w : v \leqslant u\}.$$

The focus of this paper is to find the uncertainty, after t tandem repeats, as a function of the acceptable list size m. This is made precise by the following definition.

Definition 3 *Given $n, t \in \mathbb{N}$ and $x_1, \dots, x_m \in \Sigma^n$, we define*

$$S_t(x_1, \dots, x_m) \triangleq \bigcap_{i=1}^{m} D^t(x_i).$$

Then, the uncertainty *associated with a code $C \subseteq \Sigma^n$ is*

$$N_t(m, C) \triangleq \max_{\substack{x_1, \dots, x_m \in C \\ x_i \neq x_j}} |S_t(x_1, \dots, x_m)|.$$

Correspondingly, for $w, r \in \mathbb{N}$ and $u_1, \dots, u_m \in \Delta_r^w$ we define

$$\bar{S}_t(u_1, \dots, u_m) \triangleq \bigcap_{i=1}^{m} \{v \in \mathbb{N}^{w+1} : v \geqslant u_i,\ \|v - u_i\|_1 = t\};$$
$$\bar{N}_t(m, w, r) \triangleq \max_{u_1, \dots, u_m \in \Delta_r^w} |\bar{S}_t(u_1, \dots, u_m)|.$$

In the next section we describe a typical set of strings in Σ^n, then by ascertaining $\bar{N}_t(m, w, r)$ for that set we find an asymptotic expression (in the string length n) for the uncertainty associated with that set, as a function of m.

Finally, in our analysis we shall use the following asymptotic notation: for two sequences a_n, b_n we say that $a_n \sim b_n$ if $a_n = b_n(1 + o(1))$.

IV. Typical set

We observe that the sets introduced in the previous section have many parameters. A complete combinatorial analysis of those would be encumbered by extreme cases which occur in a vanishingly small fraction of the space; analysis of these cases is therefore not only more challenging, but also less enlightening. Since our main goal is an asymptotic analysis, we proceed by eliminating those rare pathological cases, and focus on the common typical ones. In particular, we would like to limit our attention to strings $x \in \Sigma^n$ for which the Hamming weight of $\bar{\phi}(x)$ and the 1-norm of $\psi_{\mathrm{drt}(x)}(x)$, as well as the difference between them, are asymptotically linearly proportional to the string length n. Those strings would form the code which we study. Thus, we start by presenting in the following lemma the code C for which it shall be our goal to find $N_t(m, C)$.

Lemma 4 *Define the family of codes*

$$\mathrm{Typ}^n \triangleq \left\{ x \in \Sigma^n : \begin{smallmatrix} \left|w(x) - \frac{q-1}{q}(n-k)\right| < n^{3/4} \\ \left|r(x) - \frac{q-1}{q(q^k-1)}(n-k)\right| < 2n^{3/4} \end{smallmatrix} \right\},$$

where $w(x) \triangleq \mathrm{wt}_H(\bar{\phi}(x))$ and $r(x) \triangleq \|\psi_{\mathrm{drt}(x)}(x)\|_1$. Then for sufficiently large n:

$$\frac{|\mathrm{Typ}^n|}{|\Sigma^n|} \geqslant 1 - 4e^{-\sqrt{n}/2} \underset{n \to \infty}{\longrightarrow} 1.$$

Proof: We note that if $x, y \in \Sigma^n$ differ only in a single coordinate, then $|w(x) - w(y)|, |r(x) - r(y)| \leqslant 2$. If the $x(i)$'s are thought of as independent and uniformly distributed random variables on Σ, then by McDiarmid's inequality [8] we have

$$\tfrac{1}{|\Sigma^n|}\left|\left\{x \in \Sigma^n : |w(x) - \mathbb{E}[w(x)]| \geqslant n^{3/4}\right\}\right| \leqslant 2e^{-\sqrt{n}/2},$$
$$\tfrac{1}{|\Sigma^n|}\left|\left\{x \in \Sigma^n : |r(x) - \mathbb{E}[r(x)]| \geqslant n^{3/4}\right\}\right| \leqslant 2e^{-\sqrt{n}/2}.$$

Further note that if $\mathbb{E}[r(x)] = \alpha(n-k) + o(n^{3/4})$ then for large enough n we also have

$$\tfrac{1}{|\Sigma^n|}\left|\left\{x \in \Sigma^n : |r(x) - \alpha(n-k)| \geqslant 2n^{3/4}\right\}\right| \leqslant 2e^{-\sqrt{n}/2},$$

and hence

$$\tfrac{1}{|\Sigma^n|}\left|\left\{x \in \Sigma^n : \substack{|w(x)-\mathbb{E}[w(x)]|<n^{3/4} \\ |r(x)-\alpha(n-k)|<2n^{3/4}}\right\}\right| \geqslant 1 - 4e^{-\sqrt{n}/2}.$$

Next, note that $u(i) \triangleq \big(\bar{\phi}(x)\big)(i)$ are also independent and uniformly distributed. Define the indicator functions $a(i) \triangleq \mathbb{1}_{\{u(i)\neq 0\}}$. Clearly

$$\mathbb{E}[w(x)] = \sum_{i=1}^{n-k} \mathbb{E}[a(i)] = \sum_{i=1}^{n-k} \Pr(u(i) \neq 0) = \tfrac{q-1}{q}(n-k).$$

See the Appendix for proof that $\mathbb{E}[r(x)] = \frac{q-1}{q(q^k-1)}(n-k) + O(1)$, which concludes the proof. ■

We remark that a similar concentration result (for $w(x)$ and $\mathrm{wt}_H(\psi_{\mathrm{drt}(x)}(x))$ instead of $r(x)$) was derived in [16, Lem. 3] using a different approach.

Before analyzing the uncertainty $N_t(m, \mathrm{Typ}^n)$, we note that the process of list-decoding given sufficiently many $(N_t(m, \mathrm{Typ}^n) + 1)$ distinct strings in Σ^{n+kt}, i.e., finding $x_1, \ldots, x_l \in \mathrm{Typ}^n$, $l < m$, such that these strings lie in $S_t(x_1, \ldots, x_l) \setminus \bigcup_{x \in \mathrm{Typ}^n \setminus \{x_1,\ldots,x_l\}} D^t(x)$, is straightforward:

Algorithm A Denote $N \triangleq N_t(m, \mathrm{Typ}^n)$ and assume as input distinct $y_1, \ldots, y_{N+1} \in \Sigma^{n+kt}$ such that there exists $x \in \mathrm{Typ}^n$ satisfying $y_1, \ldots, y_{N+1} \in D^t(x)$. (Note that, when such x exists, $\mathrm{drt}(x)$ may be determined by, e.g., $\mathrm{drt}(x) = \mathrm{drt}(y_1)$; hence, $\psi_{\mathrm{drt}(x)}$ in particular may be used at will.)

1) Apply $\psi_{\mathrm{drt}(y_1)}$ to map them to $v_1, \ldots, v_{N+1} \in \Delta^w_{r+t}$, where $w = \mathrm{wt}\big(\bar{\phi}(\mathrm{drt}(y_1))\big)$ and $r = \left\|\psi_{\mathrm{drt}(y_1)}(y_1)\right\|_1 - t$; note that prior computation of $\mathrm{drt}(y_1)$ is not required to perform this mapping, and that it may be found as a byproduct of finding any v_i.
2) Find $u \triangleq \bigwedge_{i=1}^{N+1} v_i \in \Delta^w_{r'}$ by calculating the minimum over each coordinate.
3) Calculate $A_r(u)$.
4) Return $\psi^{-1}_{\mathrm{drt}(y_1)}(A_r(u))$ as a list. □

We defer proving the validity of Algorithm A to the end of the section, since an asymptotic evaluation of its run-time complexity involves analysis of $N_t(m, \mathrm{Typ}^n)$, which we shall next tend to; before doing so, however, we shall present an example of the application of Algorithm A.

Example 5 *Let $q = 3, k = 2$, and take $n = 11, t = 3$. We shall read multiple distinct elements of $D^t(x)$ for some unknown $x \in \mathrm{Typ}^n$, and would like to decode a list of strings in Typ^n of which x is a member.*

The first read we make is

$$y_1 = 10101012122222222.$$

(This suffices to determine $\mathrm{drt}(y_1) = \mathrm{drt}(x) = 10122$.)

Further, Suppose that we are willing to accept a list of size at most 3, and therefore set $m = 4$. Observing that

$$\bar{\phi}(y_1) = 000002001000000,$$

and consequently

$$w = \mathrm{wt}_H(\bar{\phi}(y_1)) = 2; \quad r = \left\|\psi_{\mathrm{drt}(y_1)}(\bar{\phi}(y_1))\right\|_1 - t = 3,$$

it happens to be the case that 4 distinct reads will suffice for this purpose.

(The reader referring back to this example, after having read the analysis following it, will note that Lemma 13 and Example 15 establish that $\mu(w, r, 1) = 3$ and $\mu(w, r, 2) = 4$, respectively; consequently, Corollary 12 then implies $\sigma(4, w, r) = 2$, and Corollary 10 implies that $\bar{N}_t(m, w, r) = 3$.

It is also of interest to note that, indeed, $x \in \mathrm{Typ}^n = \mathrm{Typ}^{11}$.)

We therefore make 3 additional distinct reads of $D^t(x)$, obtaining:

$$y_2 = 10101010122222222,$$
$$y_3 = 10101012222222222,$$
$$y_4 = 10101012121222222.$$

We can now find

$$\bar{\phi}(y_1) = 000002001000000 = 0^{1+2k}20^k10^{3k},$$
$$\bar{\phi}(y_2) = 000000021000000 = 0^{1+3k}20^010^{3k},$$
$$\bar{\phi}(y_3) = 000002100000000 = 0^{1+2k}20^010^{4k},$$
$$\bar{\phi}(y_4) = 000002000010000 = 0^{1+2k}20^{2k}10^{2k},$$

which may be more succinctly represented by

$$v_1 = \psi_{\mathrm{drt}(y_1)}(y_1) = (2, 1, 3),$$
$$v_2 = \psi_{\mathrm{drt}(y_1)}(y_2) = (3, 0, 3),$$
$$v_3 = \psi_{\mathrm{drt}(y_1)}(y_3) = (2, 0, 4),$$
$$v_4 = \psi_{\mathrm{drt}(y_1)}(y_4) = (2, 2, 2).$$

This concludes Step 1. The coordinatewise minimum required in Step 2 is therefore

$$u = (2, 0, 2),$$

and for Step 3 we find $A_r(u) = A_3((2, 0, 2)) = \{u_1, u_2\}$, where

$$u_1 = (1, 0, 2),$$
$$u_2 = (2, 0, 1).$$

For Step 4, we therefore find $x_i = \psi^{-1}_{\mathrm{drt}(y_1)}(u_i)$, $i \in \{1, 2\}$, by

$$\bar{\phi}(x_1) = 000210000 = 0^{1+k}20^k10^{2k},$$
$$\bar{\phi}(x_2) = 000002100 = 0^{1+2k}20^k10^k,$$

and therefore

$$x_1 = 10101012222,$$
$$x_2 = 10101010122.$$

We note that the algorithm produced a list of size 2*, smaller than the design requirements; its guarantee of list size relies on maximal lower-bounds-set size, hence specific examples may well produce shorter lists.*

Next, for Typ^n we show that the uncertainty can be calculated by $\bar{N}_t$, which provides an expression we may more easily analyze.

Lemma 6 *For $C \subseteq \Sigma^n$, there exist $x \in \mathrm{Irr}$ and $u_1, \ldots, u_m \in \psi_x(C \cap D^*(x))$ such that*

$$N_t(m, C) = \left|\bar{S}_t(u_1, \ldots, u_m)\right|.$$

Proof: Take $x_1, \ldots, x_m \in C$ such that $|S_t(x_1, \ldots, x_m)| = N_t(m, C)$, and note that if there exist $x_i \not\sim_k x_j$, then $S_t(x_1, \ldots, x_m) = \emptyset$, in contradiction. Hence there exists $x = \mathrm{drt}(\{x_1, \ldots, x_m\})$. The claim now follows from the isometry ψ_x, i.e., $\psi_x(S_t(x_1, \ldots, x_m)) = \bar{S}_t(\psi_x(x_1), \ldots, \psi_x(x_m))$. ■

Corollary 7 *For $k \geqslant 2$ and sufficiently large n,*

$$N_t(m, \mathrm{Typ}^n) = \\ = \max\left\{\bar{N}_t(m, w, r) : \begin{smallmatrix} \left|w - \frac{q-1}{q}(n-k)\right| < n^{3/4} \\ \left|r - \frac{q-1}{q(q^k-1)}(n-k)\right| < 2n^{3/4} \end{smallmatrix}\right\}.$$

Proof: For convenience, we denote

$$M \triangleq \max\left\{\bar{N}_t(m, w, r) : \begin{smallmatrix} \left|w - \frac{q-1}{q}(n-k)\right| < n^{3/4} \\ \left|r - \frac{q-1}{q(q^k-1)}(n-k)\right| < 2n^{3/4} \end{smallmatrix}\right\}.$$

By Lemma 6 we have $x \in \mathrm{Irr}$ and $u_1, \ldots, u_m \in \psi_x(\mathrm{Typ}^n \cap D^*(x))$ such that $N_t(m, C) = \left|\bar{S}_t(u_1, \ldots, u_m)\right|$. If we take $y_1, \ldots, y_m \in \mathrm{Typ}^n \cap D^*(x)$ such that $\psi_x(y_i) = u_i$ for all $i \in \{1, \ldots, m\}$, then we have $w(y_1) = \ldots = w(y_m) = w(x)$. Furthermore, it follows from $|y_1| = \ldots = |y_m| = n$ that $r(y_1) = \ldots = r(y_m)$. Denote therefore $w \triangleq w(y_1)$ and $r \triangleq r(y_1)$, and since $y_1, \ldots, y_m \in \mathrm{Typ}^n$ we have

$$\left|w - \tfrac{q-1}{q}(n-k)\right| < n^{3/4}; \quad \left|r - \tfrac{q-1}{q(q^k-1)}(n-k)\right| < 2n^{3/4}.$$

Therefore, noting that $\bar{N}_t(m, w, r) \geqslant \left|\bar{S}_t(u_1, \ldots, u_m)\right| = N_t(m, C)$, we conclude that $N_t(m, C) \leqslant M$.

To show the the other direction, note that for every pair w, r satisfying

$$\left|w - \tfrac{q-1}{q}(n-k)\right| < n^{3/4}; \quad \left|r - \tfrac{q-1}{q(q^k-1)}(n-k)\right| < 2n^{3/4},$$

there exists $x \in \mathrm{Irr}(n-kr)$ (so that $D^r(x) \subseteq \mathrm{Typ}^n$) for which $w(x) = w$. This follows from counting the required number of zeros in $\bar{\phi}(x)$ for such x, which is $n - (1+r)k - w$; for large enough n this number is positive and no greater than $(k-1)(w+1)$. Hence, after arbitrarily choosing w non-zero elements of Σ, we may pad them with runs of $k-1$ zeros or less (obtaining a string in $\bar{\phi}(\mathrm{Irr})$) to achieve a total length of $n - (1+r)k$. The derived string is $\bar{\phi}(x)$ for the desired $x \in \mathrm{Irr}(n - rk)$, and by again arbitrarily choosing any k elements of Σ for the role of $\hat{\phi}(x)$, we may indeed find x as desired.

Now, taking w, r in the required ranges such that $M = \bar{N}_t(m, w, r)$, $u_1, \ldots, u_m \in \Delta_r^w$ such that $\left|\bar{S}_t(u_1, \ldots, u_m)\right| = \bar{N}_t(m, w, r) = M$, and $x \in \mathrm{Irr}(n - rk)$ as described, we may find $y_1, \ldots, y_m \in \mathrm{Typ}^n \cap D^*(x)$ by defining $y_i \triangleq \psi_x^{-1}(u_i)$ for all $i \in \{1, \ldots, m\}$, and note $N_t(m, C) \geqslant |S_t(y_1, \ldots, y_m)| = \left|\bar{S}_t(u_1, \ldots, u_m)\right| = M$. ■

Hence, the quantity one needs to assess is $\bar{N}_t(m, w, r)$. We do that next by exploiting the lattice structure of $\mathbb{N}^{w+1}$, and introducing the connection to supremum height and lower-bound-set size in that lattice.

Lemma 8 *Given $u_1, \ldots, u_m \in \Delta_r^w$, denote $u \triangleq \bigvee_{i=1}^m u_i$. Then,*

$$\left|\bar{S}_t(u_1, \ldots, u_m)\right| = \begin{cases} 0 & \|u\|_1 > r + t, \\ \binom{w+t+r-\|u\|_1}{w} & \textit{otherwise.} \end{cases}$$

Proof: The proposition follows from the lattice structure of $\mathbb{N}^{w+1}$, i.e.,

$$\bar{S}_t(u_1, \ldots, u_m) = \left\{v \in \mathbb{N}^{w+1} : v \geqslant u, \ \|v - u_1\|_1 = t\right\}.$$

When $\|u\|_1 - r = \|u - u_1\|_1 > t$, then, the set is empty. Otherwise, the size of the set corresponds to the number of ways to distribute $t - \|u - u_1\|_1 = t - (\|u\|_1 - r)$ balls into $w + 1$ bins. ■

Definition 9 *Denote for $m, w, r \in \mathbb{N}$ the* minimum supremum height

$$\sigma(m, w, r) \triangleq \min_{u_1, \ldots, u_m \in \Delta_r^w} \left\|\bigvee_{i=1}^m u_i\right\|_1 - r.$$

Conversely, for $w, r, s \in \mathbb{N}$ denote the maximal lower-bounds-set size

$$\mu(w, r, s) \triangleq \max\left\{|A_r(u)| : u \in \Delta_{r+s}^w\right\}.$$

(Recall that $A_r(u)$ is the lower-bounds set of u.)

Corollary 10 $\bar{N}_t(m, w, r) = \binom{w+t-\sigma(m,w,r)}{w}$.

Proof: The proposition follows from Lemma 8. ■

It is therefore seen that the main task is to find or estimate the minimum supremum height. We next show the duality between $\sigma(m, w, r)$ and $\mu(w, r, s)$, which we shall use to calculate the former.

Lemma 11 *Take $w, r, s \in \mathbb{N}$. If $s \geqslant wr$ then*

$$\mu(w, r, s) = |\Delta_r^w| = \binom{r+w}{r} \quad \textit{and} \quad \sigma(|\Delta_r^w|, w, r) = wr.$$

For $s < wr$ we have

$$\sigma(\mu(w, r, s), w, r) = s.$$

Proof: The first part of the proposition is justified by $(r, r, \ldots, r) \in \Delta_{(w+1)r}^w$.

For the second, take $u \in \Delta_{r+s}^w$ satisfying $|A_r(u)| = \mu(w, r, s)$. Since $\bigvee A_r(u) \leqslant u$ we have (see Definition 9)

$$\sigma(\mu(w, r, s), w, r) \leqslant s.$$

However, if $\sigma(\mu(w,r,s),w,r) < s$, then we may find $v = \bigvee A_r(v)$ satisfying $|A_r(v)| \geqslant \mu(w,r,s)$ and $\|v\|_1 < r+s < (w+1)r$. Therefore, we know that $A_r(v) \neq \Delta_r^w$, hence there exist $v', v'' \in \Delta_r^w$, $v' \notin A_r(v)$ (thus $v' \nleqslant v$) and $v'' \in A_r(v)$, satisfying $d_1(v',v'') = 1$. It follows that $\|v \vee v'\|_1 = \|v\|_1 + 1 \leqslant r+s$, in contradiction to $|A_r(u)| = \mu(w,r,s)$. It follows that $\sigma(\mu(w,r,s),w,r) = s$. ■

Corollary 12 *If $\mu(w,r,s) < m \leqslant \mu(w,r,s+1)$ then*

$$\sigma(m,w,r) = s+1.$$

Proof: Firstly, since $m \mapsto \sigma(m,w,r)$ is non-decreasing by definition, then by Lemma 11

$$\begin{aligned} s = \sigma(\mu(w,r,s),w,r) \leqslant \sigma(m,w,r) \leqslant \\ \leqslant \sigma(\mu(w,r,s+1),w,r) = s+1. \end{aligned}$$

However, if $\sigma(m,w,r) = s$, by finding $u_1,\ldots,u_m \in \Delta_r^w$ with $\|\bigvee_{i=1}^m u_i\|_1 = r+s$ we deduce $\mu(w,r,s) \geqslant m$, in contradiction. ■

Since we now know that calculating $\mu(w,r,s)$ is sufficient for our purposes, we turn to that task; since our focus is Typ^n, we may do so for the relevant ranges of w,r, whenever that is simpler.

Lemma 13 *For $w,r,s \in \mathbb{N}$ there exists $u \in \Delta_{r+s}^w$ such that $|A_r(u)| = \mu(w,r,s)$ and for all $1 \leqslant i < j \leqslant w+1$ it holds that $|u(i) - u(j)| < 2$.*

Proof: Take $u \in \Delta_{r+s}^w$ satisfying $|A_r(u)| = \mu(w,r,s)$, and assume to the contrary that there exist i,j such that, w.l.o.g., $u(j) \geqslant u(i)+2$. Denote by u' the vector which agrees on u on all coordinates except $u'(j) = u(j)-1$ and $u'(i) = u(i)+1$.

Further, partition $A_r(u)$ and $A_r(u')$ by the projection on the subspace formed by all the coordinates except i and j. For any matching classes $C, C' \subseteq \Delta_r^w$ in the corresponding partitions, denote by $t(C) = t(C')$ the difference between r and the sum of all coordinates other than i,j; Note that $|C|$ is the number of ways to distribute $t(C)$ balls into two bins with capacities $u(i), u(j)$ (and correspondingly $u'(i), u'(j)$ for $|C'|$), hence

$$\begin{aligned} |C| &= \min\{t(C), u(i)\} - \max\{t(C) - u(j), 0\} + 1 \\ &\leqslant \min\{t(C), u(i)+1\} - \max\{t(C) - u(j) + 1, 0\} + 1 \\ &= \min\{t(C'), u'(i)\} - \max\{t(C) - u'(j), 0\} + 1 = |C'|, \end{aligned}$$

where the inequality is justified by cases for $t(C)$, and is strict only if $u(i) < t(C) < u(j)$. Thus, the proof is concluded. ■

Lemma 13 allows us to find $\mu(w,r,s)$ with relative ease; perhaps the most straightforward example of that is a precise calculation for the cases $s = 1,2$, which we present next; following the examples we conduct a more extensive evaluation, for $s > 2$ and the relevant ranges of w,r.

Example 14 *Any vector $u \in \Delta_{r+1}^w$ having $1 + \min\{w,r\}$ positive coordinates has precisely*

$$|A_r(u)| = 1 + \min\{w,r\},$$

since any lower bound in Δ_r^w is reached by subtracting 1 from a chosen positive coordinate. By Lemma 13 one such vector satisfies $\mu(w,r,1) = |A_r(u)|$, therefore

$$\mu(w,r,1) = 1 + \min\{w,r\}.$$

Example 15 *We define an injection*

$$\xi\colon \left\{v \in \mathbb{N}^{w+1} : v \leqslant u\right\} \to \mathbb{N}^{w+1}$$

by $\xi(v) \triangleq u - v$; then clearly, ξ is distance preserving, and in particular injective. Hence,

$$\mu(w,r,2) \leqslant |\Delta_2^w| = \binom{w+2}{2}.$$

This is achieved with equality when $r+2 \geqslant 2(w+1)$, since there exists $(2,2,\ldots,2) \leqslant v \in \Delta_{r+2}^w$, and it holds that $\xi(A_r(v)) = \Delta_2^2$). The inequality is strict, however, when $r < 2w$.

To examine the remaining cases, note first that increasing any coordinate of u above 2 has no effect on $|A_r(u)|$. Further, we again know by Lemma 13 that $\mu(w,r,2)$ is achieved when u has the greatest number of positive coordinates, and among such vectors, the greatest number greater than or equal to 2. Now, by counting the number of lower bounds for any such $u \in \Delta_{r+2}^w$ we see that

$$\mu(w,r,2) = \begin{cases} \binom{w+2}{2}, & r \geqslant 2w; \\ \binom{w+1}{2} + (r-w+1), & w-1 \leqslant r < 2w; \\ \binom{r+2}{2}, & r < w-1. \end{cases}$$

As can now be seen, a complete evaluation of $\mu(w,r,s)$ for $s > 2$ is possible using Lemma 13, but it involves application of the inclusion-exclusion principle and its results are not illuminating. We shall see instead that an asymptotic evaluation of $\mu(w,r,s)$ for typical ranges of w,r will suffice. To do so, we note the following proposition.

Lemma 16 *Fix t, and take w,r such that $r+t \leqslant w+1$. For all $s \leqslant t$ it holds that*

$$\mu(w,r,s) = \binom{r+s}{s}.$$

Proof: By Lemma 13 we know that $u \in \Delta_{r+s}^w$ achieving $|A_r(u)| = \mu(w,r,s)$ is such that $r+s$ of its coordinates equal 1, and the remaining $w+1-r-s$ equal 0. The proposition follows. ■

We can use Corollary 12 together with Lemma 16 to establish the main result of this section, in the following theorem. Before doing so, we note a consequence of, e.g., Lemma 16, namely that for any string $x \in \mathrm{Typ}^n$, and any $y \in D^t(x)$, it holds that

$$\left|\left\{x' \in \mathrm{Typ}^n : y \in D^t(x')\right\}\right| = O(n^t).$$

Hence, we have for $m_n = \omega(n^t)$ that $N_t(m_n, \mathrm{Typ}^n) = o(1)$; it is therefore only interesting to find an asymptotic expression for $N_t(m_n, \mathrm{Typ}^n)$ when $m_n = O(n^t)$.

Theorem 17 *Fix t and a sequence $m_n = O(n^t)$. Then*

$$N_t(m_n, \mathrm{Typ}^n) \sim \tfrac{1}{(e_t(m_n,n))!}\left(\tfrac{q-1}{q}n\right)^{e_t(m_n,n)},$$

where $e_t(m_n,n) = t - \lceil \log_n(m_n)\rceil - \delta(m_n,n)$ and $\delta(m,n) \in \{0,1\}$ is a non-decreasing function in m.

Proof: Let $s \triangleq \lceil \log_n(m_n)\rceil$.

Recall from Lemma 16 that for $w \geqslant r+t-1$

$$\mu(w,r,s-1) = \binom{r+s-1}{r} < \frac{(r+s-1)^{s-1}}{(s-1)!},$$

hence for r satisfying $\left|r - \frac{q-1}{q(q^k-1)}(n-k)\right| < 2n^{3/4}$ and sufficiently large n

$$\log_n \mu(w,r,s-1) < s-1.$$

On the other hand we have

$$\mu(w,r,s+1) = \binom{r+s+1}{r} > \frac{r^{s+1}}{(s+1)!},$$

and therefore, for such r,

$$\begin{aligned}\log_n \mu(w,r,s+1) &> \log_n\left(\frac{1+o(1)}{(s+1)!}\left(\frac{q-1}{q(q^k-1)}n\right)^{s+1}\right)\\ &= s+1+o(1).\end{aligned}$$

Since $s-1 < \log_n(m_n) \leqslant s$ it now follows from Corollary 12, for sufficiently large n (which does not depend on s, i.e., on m_n), and w, r satisfying

$$\begin{aligned}&\left|w - \tfrac{q-1}{q}(n-k)\right| < n^{3/4}\\ &\left|r - \tfrac{q-1}{q(q^k-1)}(n-k)\right| < 2n^{3/4},\end{aligned}$$

that

$$\sigma(m_n,w,r) = s + \delta(m_n,n,r),$$

where

$$\delta(m_n,n,r) = \begin{cases}1, & m_n > \mu(w,r,s) = \binom{r+\lceil \log_n(m_n)\rceil}{r};\\ 0, & \text{otherwise.}\end{cases}$$

Next, for such n, w, r we have

$$\begin{aligned}&\binom{w+t-\sigma(m_n,w,r)}{w}\\ &\quad= \tfrac{1+o(1)}{(t-(s+\delta(m_n,n,r)))!}\left(\tfrac{q-1}{q}n\right)^{t-(s+\delta(m_n,n,r))}.\end{aligned}$$

It therefore follows from Corollary 7 and Corollary 10 that

$$\begin{aligned}N_t(m_n,\mathrm{Typ}^n) &= \tfrac{1+o(1)}{(t-(s+\delta(m_n,n)))!}\left(\tfrac{q-1}{q}n\right)^{t-(s+\delta(m_n,n))}\\ &= \tfrac{1+o(1)}{e_t(m_n,n)!}\left(\tfrac{q-1}{q}n\right)^{e_t(m_n,n)},\end{aligned}$$

where $\delta(m_n,n) = 1$ if and only if $\delta(m_n,n,r) = 1$ for all r satisfying the above requirement, and $e_t(m_n,n)$ is as defined in the theorem's statement. ■

Finally, we conclude the section by referring back to Algorithm A, proving its validity, and analyzing its run-time complexity.

Theorem 18 *Algorithm A operates in $O(n^t) = \mathrm{poly}(N)$ steps, and produces $x_1,\ldots,x_l \in \mathrm{Typ}^n$, $l < m$, such that*

$$y_1,\ldots,y_{N+1} \in S_t(x_1,\ldots,x_l) \setminus \bigcup_{\substack{x\in\mathrm{Typ}^n\\ x\notin\{x_1,\ldots,x_l\}}} D^t(x).$$

Proof: First, note that the existence of an ancestor for all $y_1,\ldots,y_{N+1}$ implies that $y_i \in D^*(\mathrm{drt}(y_1))$ for all i. Moreover, note that finding any v_i may be done in $O(n)$ steps (by calculating $\bar{\phi}(y_i)$ and recording lengths of runs of zeros in the process). Any one of these can also produce $\mathrm{drt}(y_1)$. Hence Step 1 concludes in $O(Nn)$ steps.

Step 2 can also be performed in $O(Nw) = O(Nn)$ steps.

Now, note that since an ancestor of all y_i's exists in Σ^n, $r' \geqslant r$. It is hence possible to compute $A_r(u)$. This may be achieved by finding all ways of distributing $r' - r < t$ balls into $w+1$ bins with capacities $u(j)$, e.g., by utilizing combination generators for all $\binom{w+r'-r}{w}$ combinations, then discarding combination which violate the bin-capacity restriction. Combination generating algorithms exist which generate all combinations in $O\left(\binom{w+r'-r}{w}\right) = O(n^{t-1})$ steps (e.g., see [24]), and pruning illegal combinations can be done in $O(w)$ steps each. Step 3 can therefore be performed in $O(n^t)$ steps.

Finally, the pre-image $\psi^{-1}_{\mathrm{drt}(y_1)}(A_r(u))$ is a set of ancestors of $y_1,\ldots,y_{N+1}$, which is a subset Typ^n, and no other element of Typ^n is an ancestor of $y_1,\ldots,y_{N+1}$. We also know that $|A_r(u)| < m$, otherwise a contradiction is reached to the definition of N. Computing $\psi^{-1}_{\mathrm{drt}(y_1)}(A_r(u))$ given $\mathrm{drt}(y_1)$ requires $O(|A_r(u)|w) \leqslant O(mn)$ steps. ■

By an examination of the proof, we note that Algorithm A may be applied to list-decode elements of any $C \subseteq \Sigma^n$, and not necessarily Typ^n; the proof remains unchanged, except that one cannot deduce from the existence of a single $x \in C$ such that $y_1,\ldots,y_{N+1} \in D^t(x)$, that $\psi^{-1}_{\mathrm{drt}(y_1)}(A_r(u)) \subseteq C$. Instead, in the general case, that verification (and discardment of invalid outputs) must be performed as an additional step. This may be unnecessary in some cases (e.g., if $C = \Sigma^n$).

V. Uncertainty with underlying ECC

In the previous section, a reconstruction problem with a list-decoding algorithm was considered, when the underlying message space was unconstrained (more precisely, constrained only to a typical set). However, one is naturally interested in a more general setting, in which the message space may be a code with a given minimum distance. Thus, in this section, we shall consider the uncertainty associated with codes $C \subseteq \mathrm{Typ}^n$ such that for all distinct $c, c' \in C$, $d(c,c') \geqslant d$, for some $d > 0$. We start with a definition of a typical set with a minimum distance.

Definition 19 *Given $m,n,t,d \in \mathbb{N}$, the* uncertainty associated with the minimum distance d in the typical sense *is defined as*

$$N_t^{\mathrm{Typ}}(m,n,d) \triangleq \max_{\substack{x_1,\ldots,x_m\in\mathrm{Typ}^n\\ d(x_i,x_j)\geqslant d}} |S_t(x_1,\ldots,x_m)|.$$

Correspondingly, for $w, r \in \mathbb{N}$,

$$\bar{N}_t(m,w,r,d) \triangleq \max_{\substack{u_1,\dots,u_m \in \Delta_r^w \\ d_1(u_i,u_j)\geqslant d}} \left|\bar{S}_t(u_1,\dots,u_m)\right|,$$

$$\mu(w,r,s,d) \triangleq \max_{u \in \Delta_{r+s}^w} \max\left\{|C| : {}^{C \subseteq A_r(u)}_{\forall v \neq v' \in C: d_1(v,v') \geqslant d}\right\}$$

$$\sigma(m,w,r,d) \triangleq \min_{\substack{u_1,\dots,u_m \in \Delta_r^w \\ d_1(u_i,u_j)\geqslant d}} \left\| \bigvee\nolimits_{i=1}^m u_i \right\|_1 - r.$$

It should be noted that if $d > t$ then $N^{\mathrm{Typ}}(2,n,d) = 0$, meaning that unique decoding from a single noisy output is possible. It was seen in [33] that $d = t$ suffices for unique reconstruction ($m = 2$) with sub-linear uncertainty (in fact, $N = 1$, which corresponds to receiving two distinct noisy outputs, suffices). We shall incidentally see that again while considering $d \leqslant t$.

As in the previous section, we defer study of the uncertainty $N_t^{\mathrm{Typ}}(m,n,d)$, and begin by presenting a list-decoding scheme given sufficiently many ($N_t^{\mathrm{Typ}}(m,n,d)+1$) distinct strings in

$$D^t(C) \triangleq \bigcup_{c \in C} D^t(c),$$

for some given code $C \subseteq \mathrm{Typ}^n$ with minimum distance d. We shall assume that a decoding scheme for recovering from at most $d-1$ errors is known for C, which we denote by $\mathcal{D}\colon \Sigma^{n+k(d-1)} \to C$.

Algorithm B Fix n, m and $d \leqslant t$; take $C \subseteq \mathrm{Typ}^n$ with minimum $d(\cdot,\cdot)$ distance d (see Definition 1), and assume a decoding scheme for recovering up to $d-1$ tandem-duplication errors is provided. Denote $N \triangleq N_t^{\mathrm{Typ}}(m,n,d)$ and assume as input distinct $y_1,\dots,y_{N+1} \in \Sigma^{n+kt}$ such that there exists $x \in C$ satisfying $y_1,\dots,y_{N+1} \in D^t(x)$.

1) Apply Algorithm A to obtain $z_1,\dots,z_l \in \Sigma^{n+k(d-1)}$ such that

$$y_1,\dots,y_{N+1} \in S_{t-d+1}(z_1,\dots,z_l) \setminus \bigcup_{\substack{z \in \Sigma^{n+k(d-1)} \\ z \notin \{z_1,\dots,z_l\}}} D^{t-d+1}(z).$$

2) Decode each z_i with the provided algorithm to produce $x_i \triangleq \mathcal{D}(z_i) \in C$; if $z_i \notin D^{d-1}(x_i)$, discard x_i.
3) Return every x_i that was not discarded in the last step, as a list. $\square$

Proof of correctness and analysis of run-time complexity will be presented in Theorem 29, at the end of the section. At this point we will instead present an example of the algorithm's application.

Example 20 *We continue the discussion of Example 5. In particular, we let $q = 3, k = 2$, but this time take $n = 9, t = 4$. As before, we shall read multiple distinct elements of $D^t(x)$, where x is now an unknown element of a code $C \subseteq \mathrm{Typ}^n$ correcting a single tandem-duplication (i.e., with minimum $d(\cdot,\cdot)$ distance $d = 2$). We would like to decode a list of strings in C, of which x is a member. We also make the arbitrary decision to require a list of size at most 2, setting $m = 3$ (which may be justified by our desire to do at least as well as we did in the previous example, since the added redundancy of an error-correcting code can be expected to offset the additional duplication error we allowed for; the more cynical reader will note that it is also easier to analyze).*

We make the same first read as in Example 5:

$$y_1 = 10101012122222222.$$

(Recall, $\mathrm{drt}(y_1) = \mathrm{drt}(x) = 10122$.)

We still have $w = \mathrm{wt}_H(\bar{\phi}(y_1)) = 2$, but in this case $r = \left\|\psi_{\mathrm{drt}(y_1)}\left(\bar{\phi}(y_1)\right)\right\|_1 - t = 2$. Now, it suffices to obtain 2 distinct reads.

(Again, a reader looking back at this example will note that we may use Lem. 23 to determine–after a short exhaustive search–that $\mu(w,r,3,d) = 2$ and $\mu(w,r,4,d) = 3$. This implies, by Lem. 22, that $\sigma(m,w,r,d) = 4$, and Corollary 22 determines that $\bar{N}_t(m,w,r,d) = 1$.)

We therefore make one other distinct read; we'll use y_2 from Example 5,

$$y_2 = 10101010122222222,$$

and we've already seen that

$$v_1 = \psi_{\mathrm{drt}(y_1)}(y_1) = (2,1,3),$$
$$v_2 = \psi_{\mathrm{drt}(y_1)}(y_2) = (3,0,3).$$

To conclude the application of Algorithm A in Step 1, we find $u = v_1 \wedge v_2 = (2,0,3)$, and $A_{r+d-1}(u) = A_3(u) = \{u_1,u_2,u_3\}$, where

$$u_1 = (1,0,2),$$
$$u_2 = (2,0,1),$$
$$u_3 = (0,0,3).$$

Denoting $z_i = \psi^{-1}_{\mathrm{drt}(y_1)}(u_i)$, $i \in \{1,2,3\}$, we might now find

$$\phi(z_1) = 10000210000,$$
$$\phi(z_2) = 10000002100,$$
$$\phi(z_3) = 10021000000,$$

and consequently

$$z_1 = 10101222222,$$
$$z_2 = 10101012222,$$
$$z_3 = 10122222222.$$

However, Step 2 calls for an application of the decoding function $\mathcal{D}$, which is more easily conceptualized in $\mathbb{N}^3$. Recall, $\psi_{\mathrm{drt}(y_1)}(C \cap D^*(\mathrm{drt}(y_1)) \subseteq \Delta_r^w = \Delta_2^2$, and has minimum d_1 distance 2 (correcting a single tandem duplication). It may be the following (optimal) such code:

$$\{(2,0,0),(0,2,0),(0,0,2)\},$$

in which case the decoder outputs

$$u_1' \triangleq \mathcal{D}(u_1) = (0,0,2),$$
$$u_2' \triangleq \mathcal{D}(u_1) = (2,0,0),$$
$$u_3' \triangleq \mathcal{D}(u_1) = (0,0,2) = u_1'.$$

Having $x_i' = \psi^{-1}_{\mathrm{drt}(y_1)}(u_i')$, $i \in \{1,2\}$, *we find*

$$\phi(x_1') = 100210000,$$
$$\phi(x_2') = 100000021,$$

and therefore

$$x_1' = 101222222,$$
$$x_2' = 101010122.$$

Clearly, $z_i \in D^{d-1}(x_i) = D^1(x_i)$, $i \in \{1,2\}$, *hence Algorithm B is concluded.*

We might remark, however, that it is also possible that the underlying error-correcting code being used satisfies, e.g.,

$$\psi_{\mathrm{drt}(y_1)}(C \cap D^*(\mathrm{drt}(y_1)) = \{(2,0,0),(0,1,1)\},$$

in which case $u_1, u_3 \notin D^1(C \cap D^*(\mathrm{drt}(y_1))$*; hence, we cannot know* $u_1'' = \mathcal{D}(u_1)$ *and* $u_3'' = \mathcal{D}(u_3)$*. It is possible that either equals* $u_2'' = \mathcal{D}(u_2) = (2,0,0)$*, but it is just as feasible (depending on the chosen implementation of* $\mathcal{D}$*) that either might be* $(0,1,1)$*.*

Nevertheless, we may verify (concluding Step 2) that $z_2 \in D^1(x_2'')$ *(where we denote* $x_i'' = \psi^{-1}_{\mathrm{drt}(y_1)}(u_i'')$, $i \in \{1,2,3\}$*), but* $z_1, z_3 \notin D^1((0,1,1))$*, hence after discarding invalid outputs we output only* x_2''*.*

Next, we show that $N_t^{\mathrm{Typ}}(m,n,d)$ may be analyzed in terms of $\bar{N}_t(m,w,r,d)$.

Corollary 21 *For all sufficiently large* n*,*

$$N_t^{\mathrm{Typ}}(m,n,d) = \\ = \max\left\{ \bar{N}_t(m,w,r,d) : \substack{\left|w - \frac{q-1}{q}(n-k)\right| < n^{3/4} \\ \left|r - \frac{q-1}{q(q^k-1)}(n-k)\right| < 2n^{3/4}} \right\}.$$

Proof: Similarly to the proof of Lemma 6, a choice of $x_1, \dots, x_m \in \mathrm{Typ}^n$ satisfying $d(x_i, x_j) \geqslant d$ and $|S_t(x_1, \dots, x_m)| = N_t^{\mathrm{Typ}}(m,n,d)$ must also satisfy $x_i \sim_k x_j$ (otherwise $S_t(x_1, \dots, x_m) = \emptyset$), hence we may find $x \triangleq \mathrm{drt}(x_1) = \ldots = \mathrm{drt}(x_m)$. In addition

$$\left|\bar{S}_t(\psi_x(x_1), \dots, \psi_x(x_m))\right| = |S_t(x_1, \dots, x_m)|$$

and $d_1(\psi_x(x_i), \psi_x(x_j)) \geqslant d$. The other direction follows as in the proof of Corollary 7. ∎

We shall continue using an analogous approach to that of the previous section, in finding $\bar{N}_t(m,w,r,d)$ in order to estimate $N_t^{\mathrm{Typ}}(m,n,d)$.

Corollary 22 $\bar{N}_t(m,w,r,d) = \binom{w+t-\sigma(m,w,r,d)}{w}$.

Proof: This proposition follows from Lemma 8 in similar fashion to Corollary 10. ∎

Lemma 23 *Take some* $m,w,r,s,d \in \mathbb{N}$*. If*

$$\mu(w,r,s,d) < m \leqslant \mu(w,r,s+1,d)$$

then

$$\sigma(m,w,r,d) = s+1.$$

Proof: The proof follows the same arguments as in the proofs of Lemma 11 and Corollary 12. ∎

Lemma 24 *For* $0 < w,r,s,d \in \mathbb{N}$ *there exist* $u \in \Delta^w_{r+s}$*, and* $C \subseteq A_r(u)$ *with minimum* d_1 *distance* d*, satisfying* $|C| = \mu(w,r,s,d)$*, such that for no pair* $1 \leqslant i,j \leqslant w+1$*,* $i \neq j$*, it holds that* $u(i) \geqslant 2$ *and* $u(j) = 0$*.*

Proof: Take $u \in \Delta^w_{r+s}$ and $C \subseteq A_r(u)$ satisfying $|C| = \mu(w,r,s,d)$, and assume to the contrary that there exist such i,j; denote by u' the vector which agrees with u on all coordinates except $u'(j) = 1$ and $u'(i) = u(i) - 1$. The proposition is justified by finding any isometric injection $\rho\colon A_r(u) \to A_r(u')$.

Indeed, define $\rho(v) \triangleq v$ if $v(i) < u(i)$, otherwise

$$(\rho(v))(l) \triangleq \begin{cases} u(i) - 1, & l = i; \\ 1, & l = j; \\ v(l), & \text{otherwise.} \end{cases}$$

Then ρ is well defined. Moreover, take any $v_1, v_2 \in A_r(u)$. If $v_1(i), v_2(i) < u(i)$ then clearly $d_1(\rho(v_1), \rho(v_2)) = d_1(v_1, v_2)$. The same trivially holds when $v_1(i) = v_2(i) = u(i)$. If, w.l.o.g. $v_1(i) < v_2(i) = u(i)$, then

$$\begin{aligned} |(\rho(v_1))(i) - (\rho(v_2))(i)| &= |v_1(i) - (\rho(v_2))(i)| \\ &= |v_1(i) - v_2(i)| - 1 \end{aligned}$$

but

$$\begin{aligned} |(\rho(v_1))(j) - (\rho(v_2))(j)| &= |v_1(j) - (\rho(v_2))(j)| = |0 - 1| \\ &= 1 = |v_1(j) - v_2(j)| + 1, \end{aligned}$$

hence, once again, $d_1(\rho(v_1), \rho(v_2)) = d_1(v_1, v_2)$. ∎

As in Section IV, Lemma 24 allows us to find $\mu(w_n, r_n, s)$ for typical ranges of w_n, r_n, using binary constant-weight codes. This is given precise meaning in the following definition and lemma.

Definition 25 *Denote by* $\mathrm{GF}(2)$ *the field of size 2, and by* d_H *the Hamming metric. Denote by* $A(\nu, 2\delta, \omega)$ *the size of the largest length* ν *binary code with minimum Hamming distance* 2δ *and constant Hamming weight* ω*.*

Lemma 26 *Fix* t*, and take* w,r *such that* $r+t \leqslant w+1$*. For all* $s \leqslant t$ *it holds that*

$$\mu(w,r,s,d) = A(r+s, 2d, s).$$

Proof: By Lemma 24 we know that there exist $u \in \Delta^w_{r+s}$ and $C \subseteq A_r(u)$ satisfying

- $|C| = \mu(w,r,s,d)$.
- For all $v_1, v_2 \in C$, $v_1 \neq v_2$, it holds that $d_1(v_1, v_2) \geqslant d$.
- u has $r+s$ of its coordinates equal 1, and the remaining $w+1-r-s$ equal 0.

Define $\rho\colon A_r(u) \to \mathrm{GF}(2)^{r+s}$ by restricting $u - v$ to the support of u (and identifying $\mathrm{GF}(2)$ with $\{0,1\} \subseteq \mathbb{N}$). Then ρ is a bijection onto constant-Hamming-weight s elements of $\mathrm{GF}(2)^{r+s}$. Further, for all $v_1, v_2 \in A_r(u)$ it holds that

$$d_H(\rho(v_1), \rho(v_2)) = 2d_1(v_1, v_2).$$

Hence, there's a size-preserving one-to-one correspondence between codes $C' \subseteq A_r(u)$ with minimum d_1 distance d, and codes in $\mathrm{GF}(2)^{r+s}$ with minimum Hamming distance $2d$ and constant Hamming weight s. The proposition follows. ■

We can now summarize our observations in the following theorem.

Theorem 27 *Fix $d \leqslant t$ and a sequence $m_n = O(n^{t-d+1})$. Then*

$$N_t^{\mathrm{Typ}}(m_n, n, d) \sim \tfrac{1}{(e_t(m_n,n,d))!}\left(\tfrac{q-1}{q}n\right)^{e_t(m_n,n,d)},$$

where $e_t(m_n, n, d) = t - \lceil \log_n(m_n) \rceil - d + \epsilon(m_n, n, d)$ and $\epsilon(m, n, d) \in \{0, 1\}$ is a non-increasing function of m.

Proof: The proof follows the same lines as that of Theorem 17. Let $s \triangleq \lceil \log_n(m_n) \rceil + d - 1$.

Recall from the first Johnson bound [14, Th. 2] that

$$\begin{aligned} A(r+s-1, 2d, s-1) &\leqslant \binom{r+s-1}{s-d} \Big/ \binom{s-1}{s-d} \\ &< \frac{(d-1)!}{(s-1)!}(r+s-1)^{s-d}, \end{aligned}$$

hence for r satisfying $\left| r - \tfrac{q-1}{q(q^k-1)}(n-k) \right| < 2n^{3/4}$ and sufficiently large n

$$\log_n A(r+s-1, 2d, s-1) < s - d.$$

On the other hand, by [9, Th. 6] we have

$$A(r+s+1, 2d, s+1) \geqslant \frac{1}{p^{d-1}}\binom{r+s+1}{s+1}$$

for any prime power p, $p > r+s$. By the prime number theorem (a weaker version, or even Bertrand's postulate, suffices. See, e.g., [4]) there exists in fact such prime number p satisfying $r + s < p \leqslant n$ for sufficiently large n and r satisfying $\left| r - \tfrac{q-1}{q(q^k-1)}(n-k) \right| < 2n^{3/4}$, hence in particular

$$\begin{aligned} A(r+s+1, 2d, s+1) &\geqslant \frac{1}{n^{d-1}}\binom{r+s+1}{s+1} \\ &> \frac{r^{s+1}}{n^{d-1}(s+1)!}, \end{aligned}$$

and therefore

$$\begin{aligned} &\log_n A(r+s+1, 2d, s+1) \\ &\quad > \log_n\left(\frac{1+o(1)}{n^{d-1}(s+1)!}\left(\frac{q-1}{q(q^k-1)}n\right)^{s+1}\right) \\ &\quad = s - d + 2 + o(1). \end{aligned}$$

Since $s - d < \log_n(m_n) \leqslant s - d + 1$ it now follows from Lemma 23 and Lemma 26, for sufficiently large n (which does not depend on s, i.e., on m_n), and w, r satisfying

$$\begin{aligned} &\left| w - \tfrac{q-1}{q}(n-k) \right| < n^{3/4} \\ &\left| r - \tfrac{q-1}{q(q^k-1)}(n-k) \right| < 2n^{3/4}, \end{aligned}$$

that

$$\sigma(m_n, w, r, d) = s + \delta(m_n, n, r, d),$$

where

$$\delta(m_n, n, r, d) = \begin{cases} 1, & m_n > A(r+s, 2d, s); \\ 0, & \text{otherwise.} \end{cases}$$

(Note that that s is a function of m_n, n.)

Next, for such n, w, r we have

$$\begin{aligned} &\binom{w + t - \sigma(m_n, w, r, d)}{w} \\ &\quad = \tfrac{1+o(1)}{(t-(s+\delta(m_n,n,r,d)))!}\left(\tfrac{q-1}{q}n\right)^{t-(s+\delta(m_n,n,r,d))}. \end{aligned}$$

It therefore follows from Corollary 21 and Corollary 22 that

$$\begin{aligned} N_t^{\mathrm{Typ}}(m_n, n, d) &= \tfrac{1+o(1)}{(t-(s+\delta(m_n,n,d)))!}\left(\tfrac{q-1}{q}n\right)^{t-(s+\delta(m_n,n,d))} \\ &= \tfrac{1+o(1)}{e_t(m_n,n,d)!}\left(\tfrac{q-1}{q}n\right)^{e_t(m_n,n,d)}, \end{aligned}$$

where $\delta(m_n, n, d) = 1$ if and only if $\delta(m_n, n, r, d) = 1$ for all r satisfying the above requirement, $\epsilon(m_n, n, d) \triangleq 1 - \delta(m_n, n, d)$, and $e_t(m_n, n, d)$ is as defined in the theorem's statement. ■

It is again remarked here that in the case that coding is performed with $d = t$, we observe that unique reconstruction ($m = 2$) is possible with just two reads ($N = 1$); To see that, note that $\delta(2, r, d) = 0$ for all $r \geqslant d$, hence for sufficiently large n we have $\epsilon(2, d) = 1$ and therefore $e_t(2, n, d) = 0$. This result, as mentioned above, was already observed in [33].

The trade-off established in Theorem 27 between the code minimum distance d (equivalently, its redundancy, since as seen in [16], [18] and mentioned above, a code with minimum distance d has optimal redundancy $(d-1)\log_q(n) + O(1)$), the number of tandem-duplication errors t, the decoded list size m_n, and the resulting uncertainty $N_t^{\mathrm{Typ}}(m_n, n, d)$, is perhaps better visualized in the following corollary.

Corollary 28 *Fix $d \leqslant t$ and a sequence $m_n = O(n^{t-d+1})$. Then*

$$\begin{aligned} &\log_n N_t^{\mathrm{Typ}}(m_n, n, d) + \lceil \log_n(m_n) \rceil + d = \\ &\qquad = t + \epsilon(m_n, d) + o(1), \end{aligned}$$

where $\epsilon(m, d) \in \{0, 1\}$ is a non-increasing function of m.

Finally, we conclude the section by proving correctness for Algorithm B, and analyzing its run-time complexity.

Theorem 29 *Algorithm B produces $x_1, \ldots, x_{l'} \in C$, $l' < m$, such that*

$$y_1, \ldots, y_{N+1} \in S_t(x_1, \ldots, x_{l'}) \setminus \bigcup_{\substack{x \in \mathrm{Typ}^n \\ x \notin \{x_1, \ldots, x_l\}}} D^t(x).$$

Further, it operates in $O(n^t + n^{t-d+1}\mathcal{C})$ steps, where $\mathcal{C}$ is the run-time complexity of $\mathcal{D}$.

Proof: There is one assumption to Algorithm A and Theorem 18 which may now not be satisfied, that indeed there exists $z \in \Sigma^{n+k(d-1)}$ such that $y_1, \ldots, y_{N+1} \in D^t(z)$. If there does not, then Step 1 might fail because Algorithm A finds $\hat{z} \triangleq \bigwedge_{i=1}^{N} y_i$ with $|\hat{z}| = n + ks$ and $s < (d-1)$. If that is the case, however, such $\hat{z}$ may still be passed on to the next step, since we may still decode it to a unique $x \in C$ for which $z \in D^s(x)$ (since C has minimum distance d, there cannot exist two distinct ancestors of z in C), which justifies the claim. Otherwise, Theorem 18 proves that the first step produces what is claimed, and we may assume w.l.o.g. that $s \geqslant d-1$.

This assumption now implies that for each $x \in C$ such that $y_1, \dots, y_{N+1} \in D^t(x)$ there exists $z \in D^{d-1}(x)$ such that $y_1, \dots, y_{N+1} \in D^t(z)$, hence $z \in \{z_1, \dots, z_l\}$; this is because one may arbitrarily choose such $x \leqslant z \leqslant \hat{z}$. On the other hand, each $z \in \Sigma^{n+k(d-1)}$ can be decoded to at most a single $x \in C$ for which $z \in D^{d-1}(x)$ (again, due to the code's minimum distance), and that x satisfies $y_1, \dots, y_{N+1} \in D^t(x)$. We remark that it is possible that the first step produces $z_i \notin D^{d-1}(C)$, hence $x_i = \mathcal{D}(z_i)$ may be erroneous (as the decoder receives invalid input); however, as $y_1, \dots, y_{N+1} \in D^{t-d+1}(z_i)$, such results can indeed be discarded by testing if $z_i \in D^{d-1}(x_i)$.

Note that if distinct $x_1, \dots, x_m \in C$ are produced by Step 2, we have $|S_t(x_1, \dots, x_m)| \geqslant |\{y_1, \dots, y_{N+1}\}| = N+1$ and therefore a contradiction. Hence, $l' < m$.

Finally, we know that Step 1 operates in $O(n^t) = \mathrm{poly}(N)$ steps. Since testing whether $z_i \in D^{d-1}(x_i)$ may be done in $O(n)$ steps, Step 2 clearly operates in $O(l(\mathcal{C}+n))$ steps. Hence, it now suffices to show that $l = O(n^{t-d+1})$ to conclude the proof.

To that end, note that the number of t-ancestors of $y \in \Sigma^{n+kt}$ is bound from above by $\mu(w,r,t)$, where $w = \mathrm{wt}_H(\bar{\phi}(y)) \leqslant n-k$ and $r = \left\|\psi_{\mathrm{drt}(y)}(y)\right\|_1 - t$. As in Example 15, using ξ we note that

$$\begin{aligned}\mu(w,r,t) &\leqslant |\Delta_t^w| = \binom{w+t}{w} < \frac{1}{t!}(w+t)^t \\ &\leqslant \frac{1}{t!}(n-k+t)^t < (n+t)^t.\end{aligned}$$

Hence $N_t((n+t)^t, \mathrm{Typ}^n) = 0$; this in particular implies that for $\hat{m} \triangleq (n+t+(k-1)(d-1))^{t-d+1}$ we have

$$N_{t-d+1}\Big(\hat{m}, \mathrm{Typ}^{n+k(d-1)}\Big) = 0 \leqslant N_t^{\mathrm{Typ}}(m,n,d).$$

Note, then, that $l < \hat{m}$. This result can be considerably improved by noting that for all m' satisfying

$$N_{t-d+1}(m', \mathrm{Typ}^{n+k(d-1)}) \leqslant N_t^{\mathrm{Typ}}(m,n,d)$$

it holds that $l < m'$, but for our purposes $\hat{m}$ does suffice. ■

Appendix
Conclusion of proof of Lemma 4

As in the proof of Lemma 4, we define $u(i) \triangleq \big(\bar{\phi}(x)\big)(i)$. Further define for all $1 \leqslant i \leqslant n-k$ and $1 \leqslant j < n-k-i+1$ the indicator $I_i(j)$ of the event of a run of precisely j zeros starting in u at index i. Then

$$\begin{aligned}\mathbb{E}[r(x)] &= \sum_{i=1}^{n-k}\sum_{j=1}^{n-k-i+1} \left\lfloor \tfrac{j}{k} \right\rfloor \Pr(I_i(j)=1) \\
&= \left\lfloor \tfrac{n-k}{k} \right\rfloor \Pr(I_1(n-k)=1) + \sum_{j=1}^{n-k-1} \left\lfloor \tfrac{j}{k} \right\rfloor \Pr(I_1(j)=1) \\
&\quad + \sum_{i=2}^{n-k} \left\lfloor \frac{n-k-i+1}{k} \right\rfloor \Pr(I_i(n-k-i+1)=1) \\
&\quad + \sum_{i=2}^{n-k-1}\sum_{j=1}^{n-k-i} \left\lfloor \tfrac{j}{k} \right\rfloor \Pr(I_i(j)=1) \\
&= \left\lfloor \frac{n-k}{k} \right\rfloor \frac{1}{q^{n-k}} + \sum_{j=1}^{n-k-1} \left\lfloor \frac{j}{k} \right\rfloor \frac{q-1}{q^{j+1}} \\
&\quad + \sum_{i=2}^{n-k} \left\lfloor \frac{n-k-i+1}{k} \right\rfloor \frac{q-1}{q^{n-k-i+2}} \\
&\quad + \sum_{i=2}^{n-k-1}\sum_{j=1}^{n-k-i} \left\lfloor \frac{j}{k} \right\rfloor \frac{(q-1)^2}{q^{j+2}} \\
&= \frac{\lfloor n/k \rfloor - 1}{q^{n-k}} + 2\frac{q-1}{q}\sum_{j=1}^{n-k-1} \frac{\lfloor j/k \rfloor}{q^j} \\
&\quad + \frac{(q-1)^2}{q^2}\sum_{i=2}^{n-k-1}\sum_{j=1}^{n-k-i} \frac{\lfloor j/k \rfloor}{q^j}\end{aligned}$$

We note that

$$\begin{aligned}\sum_{j=1}^{p} \frac{\lfloor j/k \rfloor}{q^j} &= \sum_{j=k}^{p} \frac{\lfloor j/k \rfloor}{q^j} \\
&= \sum_{j=k\lfloor p/k \rfloor}^{p} \frac{\lfloor p/k \rfloor}{q^j} + \sum_{i=1}^{\lfloor p/k \rfloor - 1}\sum_{j=0}^{k-1} \frac{i}{q^{ik+j}} \\
&= \frac{q}{q-1}\Bigg[\lfloor p/k \rfloor \left(\frac{1}{q^{k\lfloor p/k \rfloor}} - \frac{1}{q^{p+1}}\right) \\
&\quad + \left(1 - \frac{1}{q^k}\right)\sum_{i=1}^{\lfloor p/k \rfloor - 1} \frac{i}{q^{ik}}\Bigg] \\
&= \frac{q}{q-1}\Bigg[\lfloor p/k \rfloor \left(\frac{1}{q^{k\lfloor p/k \rfloor}} - \frac{1}{q^{p+1}}\right) \\
&\quad + \frac{1}{q^k-1}\left(1 - \frac{1}{q^{k(\lfloor p/k \rfloor - 1)}}\right) \\
&\quad - \frac{\lfloor p/k \rfloor - 1}{q^{k\lfloor p/k \rfloor}}\Bigg] \\
&= \frac{q}{q-1}\Bigg[\frac{1}{q^k-1}\left(1 - \frac{1}{q^{k(\lfloor p/k \rfloor - 1)}}\right) \\
&\quad + \frac{1}{q^{k\lfloor p/k \rfloor}} - \frac{\lfloor p/k \rfloor}{q^{p+1}}\Bigg]\end{aligned}$$

Now

$$\frac{\lfloor n/k \rfloor - 1}{q^{n-k}} + 2\frac{q-1}{q}\sum_{j=1}^{n-k-1} \frac{\lfloor j/k \rfloor}{q^j} = O(1).$$

Hence, it suffices to find

$$\frac{(q-1)^2}{q^2}\sum_{i=2}^{n-k-1}\sum_{j=1}^{n-k-i}\frac{\lfloor j/k\rfloor}{q^j}=\frac{(q-1)^2}{q^2}\sum_{p=1}^{n-k-2}\sum_{j=1}^{p}\frac{\lfloor j/k\rfloor}{q^j}$$
$$=\frac{q-1}{q(q^k-1)}\sum_{p=1}^{n-k-2}\left(1-\frac{1}{q^{k(\lfloor p/k\rfloor-1)}}\right)$$
$$-\frac{q-1}{q}\sum_{p=1}^{n-k-2}\left[\frac{\lfloor p/k\rfloor}{q^{p+1}}-\frac{1}{q^{k\lfloor p/k\rfloor}}\right].$$

Again, note that $\sum_{p=1}^{n-k-2}\frac{\lfloor p/k\rfloor}{q^{p+1}}=O(1)$; in addition, we note that $\sum_{p=1}^{n-k-2}\frac{1}{q^{k\lfloor p/k\rfloor}}=O(1)$.

We therefore find $\mathbb{E}[r(x)]=\frac{q-1}{q(q^k-1)}(n-k)+O(1)$.

Acknowledgments

The authors gratefully acknowledge the two anonymous reviewers and associate editor, whose careful reading and suggestions helped shape this paper. We also offer special thanks to Prof. Jehoshua Bruck for his illuminating insight, and the conversation that originally turned our attention to the problem explored in this paper.

References

[1] M. Arita and Y. Ohashi, "Secret signatures inside genomic DNA," *Biotechnology Progress*, vol. 20, no. 5, pp. 1605–1607, 2004.

[2] F. Balado, "Capacity of DNA data embedding under substitution mutations," *IEEE Trans. on Inform. Theory*, vol. 59, no. 2, pp. 928–941, Feb. 2013.

[3] Y. Cassuto and M. Blaum, "Codes for symbol-pair read channels," *IEEE Trans. on Inform. Theory*, vol. 57, no. 12, pp. 8011–8020, Dec. 2011.

[4] P. L. Chebyshev, "Mémoire sur les nombres premiers," *J. Math. Pures Appl.*, vol. 17, pp. 366–390, 1852.

[5] Y. M. Chee, H. M. Kiah, A. Vardy, V. K. Vu, and E. Yaakobi, "Coding for racetrack memories," *IEEE Trans. on Inform. Theory*, vol. 64, no. 11, pp. 7094–7112, Nov. 2018.

[6] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.

[7] C. T. Clelland, V. Risca, and C. Bancroft, "Hiding messages in DNA microdots," *Nature*, vol. 399, no. 6736, pp. 533–534, 1999.

[8] J. L. Doob, "Regularity properties of certain families of chance variables," *Transactions of the American Mathematical Society*, vol. 47, no. 3, pp. 455–486, 1940.

[9] R. L. Graham and N. J. A. Sloane, "Lower bounds for constant weight codes," *IEEE Trans. on Inform. Theory*, vol. 26, no. 1, pp. 37–43, Jan. 1980.

[10] D. Heider and A. Barnekow, "DNA-based watermarks using the DNA-Crypt algorithm," *BMC Bioinformatics*, vol. 8, no. 1, pp. 176–185, May 2007.

[11] S. Jain, F. Farnoud, and J. Bruck, "Capacity and expressiveness of genomic tandem duplication," *IEEE Trans. on Inform. Theory*, vol. 63, no. 10, pp. 6129–6138, Oct. 2017.

[12] S. Jain, F. Farnoud, M. Schwartz, and J. Bruck, "Duplication-correcting codes for data storage in the DNA of living organisms," *IEEE Trans. on Inform. Theory*, vol. 63, no. 8, pp. 4996–5010, Aug. 2017.

[13] ——, "Noise and uncertainty in string-duplication systems," in *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT'2017), Aachen, Germany*, Jun. 2017, pp. 3120–3124.

[14] S. M. Johnson, "A new upper bound for error-correcting codes," *IRE Trans. on Inform. Theory*, vol. 8, no. 3, pp. 203–207, Apr. 1962.

[15] D. C. Jupiter, T. A. Ficht, J. Samuel, Q.-M. Qin, and P. de Figueiredo, "DNA watermarking of infectious agents: Progress and prospects," *PLoS pathogens*, vol. 6, no. 6, p. e1000950, 2010.

[16] M. Kovačević and V. Y. F. Tan, "Asymptotically optimal codes correcting fixed-length duplication errors in DNA storage systems," *IEEE Communications Letters*, vol. 22, no. 11, pp. 2194–2197, Nov. 2018.

[17] M. Kovačević, "Codes correcting all patterns of tandem-duplication errors of maximum length 3," *arXiv preprint arXiv:1911.06561*, 2019.

[18] A. Lenz, N. Jünger, and A. Wachter-Zeh, "Bounds and constructions for multi-symbol duplication error correcting codes," in *Proceedings of the 16th International Workshop on Algebraic and Combinatorial Coding Theory (ACCT), Svetlogorsk, Russia*, Svetlogorsk, Russia, Sep. 2018, pp. 129–133.

[19] A. Lenz, A. Wachter-Zeh, and E. Yaakobi, "Bounds on codes correcting tandem and palindromic duplications," in *Proceedings of the 2017 International Workshop on Coding and Cryptography (WCC'2017), Saint-Petersburg, Russia*, Sep. 2017.

[20] ——, "Duplication-correcting codes," *Designs, Codes and Cryptography*, vol. 87, no. 2, pp. 277–298, Mar. 2019.

[21] V. I. Levenshtein, "Efficient reconstruction of sequences," *IEEE Trans. on Inform. Theory*, vol. 47, no. 1, pp. 2–22, Jan. 2001.

[22] M. Liss, D. Daubert, K. Brunner, K. Kliche, U. Hammes, A. Leiherer, and R. Wagner, "Embedding permanent watermarks in synthetic genes," *PLoS ONE*, vol. 7, no. 8, p. e42465, 2012.

[23] H. Mahdavifar and A. Vardy, "Asymptotically optimal sticky-insertion-correcting codes with efficient encoding and decoding," in *Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT'2017), Aachen, Germany*, Jun. 2017, pp. 2683–2687.

[24] F. Ruskey, J. Sawada, and A. Williams, "De bruijn sequences for fixed-weight binary strings," *SIAM J. Discrete Math.*, vol. 26, no. 2, pp. 605–617, 2012.

[25] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, "Exact reconstruction from insertions in synchronization codes," *IEEE Trans. on Inform. Theory*, vol. 63, no. 4, pp. 2428–2445, Apr. 2017.

[26] S. L. Shipman, J. Nivala, J. D. Macklis, and G. M. Church, "CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria," *Nature*, vol. 547, p. 345, Jul. 2017.

[27] Y. Tang and F. Farnoud, "Error-correcting codes for noisy duplication channels," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep. 2019, pp. 140–146.

[28] ——, "Error-correcting codes for short tandem duplication and substitution errors," in *Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT'2020), Los Angeles, CA, USA*, Jun. 2020, pp. 734–739.

[29] Y. Tang, Y. Yehezkeally, M. Schwartz, and F. Farnoud, "Single-error detection and correction for duplication and substitution channels," *IEEE Trans. on Inform. Theory*, vol. 66, no. 11, pp. 6908–6919, Nov. 2020.

[30] P. C. Wong, K. kwok Wong, and H. Foote, "Organic data memory using the DNA approach," *Communications of the ACM*, vol. 46, no. 1, pp. 95–98, Jan. 2003.

[31] E. Yaakobi and J. Bruck, "On the uncertainty of information retrieval in associative memories," *IEEE Trans. on Inform. Theory*, vol. 65, no. 4, pp. 2155–2165, Apr. 2019.

[32] E. Yaakobi, J. Bruck, and P. H. Siegel, "Constructions and decoding of cyclic codes over b-symbol read channels," *IEEE Trans. on Inform. Theory*, vol. 62, no. 4, pp. 1541–1551, Apr. 2016.

[33] Y. Yehezkeally and M. Schwartz, "Reconstruction codes for DNA sequences with uniform Tandem-Duplication errors," *IEEE Trans. on Inform. Theory*, vol. 66, no. 5, pp. 2658–2668, May 2020.

Yonatan Yehezkeally (S'12–M'20) is a postdoctoral researcher in the Associate Professorship of Coding and Cryptography (Prof. Wachter-Zeh), Institute for Communications Engineering, TU Munich Department of Electrical and Computer Engineering. His research interests include coding for DNA storage, combinatorial structures, algebraic coding, and finite group theory.

Yonatan received the B.Sc. (*cum laude*) degree in Mathematics, and the M.Sc. (*summa cum laude*) and Ph.D. degrees in Electrical and Computer Engineering, in 2013, 2017 and 2020 respectively, all from Ben-Gurion University of the Negev, Beer-Sheva, Israel.

Moshe Schwartz (M'03–SM'10) is a professor at the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Israel. His research interests include algebraic coding, combinatorial structures, and digital sequences.

Prof. Schwartz received the B.A. (*summa cum laude*), M.Sc., and Ph.D. degrees from the Technion – Israel Institute of Technology, Haifa, Israel, in 1997, 1998, and 2004 respectively, all from the Computer Science Department. He was a Fulbright post-doctoral researcher in the Department of Electrical and Computer Engineering, University of California San Diego, and a post-doctoral researcher in the Department of Electrical Engineering, California Institute of Technology. While on sabbatical 2012–2014, he was a visiting scientist at the Massachusetts Institute of Technology (MIT).

Prof. Schwartz received the 2009 IEEE Communications Society Best Paper Award in Signal Processing and Coding for Data Storage, and the 2020 NVMW Persistent Impact Prize. He has also been serving as an Associate Editor for Coding Techniques for the IEEE Transactions on Information Theory since 2014, and an Editorial Board Member for the Journal of Combinatorial Theory Series A since 2021.