Reconstruction of Sequences over Non-Identical Channels

Michal Horovitz
Dept. of Computer Science
Technion-Israel Institute of Technology
Haifa 32000, Israel
Email: *michalho@cs.technion.ac.il*

Eitan Yaakobi
Dept. of Computer Science
Technion-Israel Institute of Technology
Haifa 32000, Israel
Email: *yaakobi@cs.technion.ac.il*

***Abstract*—Motivated by the error behavior in DNA storage channels, in this work we extend the previously studied *sequence reconstruction problem* by Levenshtein. The reconstruction problem studies the model in which the information is read through multiple noisy channels, and the decoder, which receives all channel estimations, is required to decode the information. For the combinatorial setup, the assumption is that all the channels cause at most some t errors. However, since the channels do not necessarily have the same behavior, we generalize this model and assume that the channels are not identical and thus may cause a different maximum number of errors. For example, we assume that there are N channelsthat cause at most t_1 or t_2 errors, where $t_1 < t_2$, and the number of channels with at most t_1 errors is at least $\lceil pN \rceil$, for some fixed $0 < p < 1$. If the information codeword belongs to a code with minimum distance d, the problem is then to find the minimum number of channels that guarantees successful decoding in the worst case.**

I. Introduction

The *sequence reconstruction problem* was first proposed and studied by Levenshtein in [8], [9]. In this model, a codeword is transmitted over multiple channels and a decoder, which receives all channel outputs, decodes the transmitted word. The assumption is that all channels are the same and are uncorrelated, with the only exception that all channel outputs are different from each other. This model was originally motivated by chemical and biological processes where the information is replicated and can be read from different noisy sources. However, it was also shown to be relevant in storage technologies, where the stored information has multiple copies or a single copy is read by several different read heads. Specifically, the applicability of this model is most relevant to *DNA storage* [1], [2], [16], [17], [18]. Both for *in vitro* and *in vivo* storage systems, the information has a large number of copies stored in DNA strands and the goal is to read these strands and reconstruct the data, while every estimation of the data is erroneous.

The reconstruction model studied by Levenshtein and later by others was combinatorial. Suppose all words belong to some space V with distance function ρ. It is assumed that the information codeword x belongs to a code with minimum distance d and the number of errors in every channel is at most t. Then, the goal is to find the minimum number of channels that guarantees unique decoding in the worst case. Clearly, if $t < \lfloor (d-1)/2 \rfloor$, then a single channel is sufficient. Otherwise, it was shown that this number has to be greater than the largest intersection of two balls with radius t and minimum distance d between their centers, that is,

$$\max\{|B_t(x) \cap B_t(z)| : x, z \in V, \rho(x,z) \geq d\},$$

where $B_t(x) = \{y \in V : \rho(x,y) \leq t\}$. Later, this combinatorial problem was studied for several channels. In [8], Levenshtein studied the cases of substitution errors, the Johnson graphs, and several more general metric distances. More results for other general error graphs were given in [10], [11], and in [5], [6], [7], it was studied for permutations. The case of permutations with the Kendall's τ distance were investigated in [15] as well as the Grassmann graph case. Levenshtein's results for deletions and insertions in [9] were extended in [13] for insertions and in [3] for deletions. In [14], the connection between the reconstruction problem and associative memories was studied, and in [4] it was analyzed for the purpose of asymptotically improving the Gilbert-Varshamov bound.

Motivated by the error behavior in DNA storage, in this work, we generalize Levenshtein's model and assume a combinatorial model where the channels are not identical. When reading the data stored in DNA strands, it may happen that some estimations are more noisy than the others [16]. In the reconstruction model this is translated to channels that cause a different maximum number of errors. For example, it is known that for substitution errors, if the transmitted word belongs to a code with minimum Hamming distance 3 and there are at most 2 errors in every channel, then 7 channels are necessary and sufficient for successful decoding. However, if at most 2 channels cause two errors (and the rest 1 error), then we show that 5 channels are necessary and sufficient for successful decoding. In [12], a similar problem was studied for the setup in which every channel can cause a different number of insertions.

Formally, we define this model as follows. Let ℓ be the number of possible types of channels. For $T = (t_1, \ldots, t_\ell)$ and $P = (p_1, \ldots, p_\ell)$, where $t_1 < \cdots < t_\ell \in \mathbb{N}$ and $0 < p_1 < \cdots < p_{\ell-1} < p_\ell = 1$, we say that a system with N channels is a (T,P)-*channel system* if for $1 \leq i \leq \ell$, $\lceil p_i N \rceil$ of the channels cause at most t_i errors. For example, Levenshtein's model is a special case with $\ell = 1$ and $p_1 = 1$. Our goal in this work is to study the minimum number of channels N required for a (T,P)-channel system for successful decoding when the information is a codeword which belongs to a code with minimum distance d. Note that in this case there are two setups we can study, namely, given a (T,P)-channel system, the decoder may or may not know the type of each channel. Our main focus will be on substitution errors while other channels are left for future work.

The rest of the paper is organized as follows. In Section II, we formally define the models. In Section III we solve the reconstruction problem for the case $\ell = 2$, and we apply this solution for substitution errors in Section IV. Then, in Section V, we extend this analysis for arbitrary ℓ, and finally in Section VI, we consider special cases of $\ell = 2$. Due to the lack of space, some of the proofs in the paper are omitted or shortened.

II. Definitions and Problem Setup

For a positive integer h, we denote by $[h]$ the set $\{1, 2, \ldots, h\}$. Let V be a finite set with a distance function $\rho: V \times V \to \mathbb{N}$. For $x \in V$, the ball of radius t centered at x is the set $B_t(x) = \{y : \rho(x,y) \leq t\}$. A combinatorial channel C is called a *t-error channel*, if for any input $x \in V$ the output of C is in $B_t(x)$.

978-1-5090-4096-4/17/$31.00 ©2017 IEEE

A *channel system* is a system consisting of some N combinatorial channels $C_1, C_2, \ldots, C_N$. We say that a word $x \in V$ is *transmitted over the channel system* if x is transmitted over C_i for all $i \in [N]$, and y_i is the output of the ith channel. The receiver applies a decoding function $\mathcal{D}(y_1, \ldots, y_N)$ in order to reconstruct the transmitted word x, and exact reconstruction happens when $x = \mathcal{D}(y_1, \ldots, y_N)$. In this paper we only refer to the exact reconstruction problem and we assume that all channel outputs are different from each other.

Let $T = (t_1, \ldots, t_\ell)$ and $P = (p_1, \ldots, p_\ell)$ such that $t_1 < t_2 < \cdots < t_\ell \in \mathbb{N}$ and $0 < p_1 < p_2 < \cdots < p_{\ell-1} < p_\ell = 1$. A channel system with N combinatorial channels is called a (T, P)-*channel system* if for each $i \in [\ell]$, $\lceil p_i N \rceil$ of the channels are t_i-error channels. The *size* of a channel system is the number of channels comprised in it.

We consider two models, which depend upon whether the behavior of each specific channel is known or unknown to the decoder. In the first channel system, called the *sequenced-channel system*, the decoder *knows* the maximum number of errors in every channel. However, in the second channel system, called the *non-sequenced-channel system*, only the distribution of the errors in the channels is known to the decoder, but the number of errors in each individual channel is unknown. For example, the decoder may know that half of the channels are t_1-error channels, and the rest are t_2-error channels, but it does not know what the exact type of each channel is.

For $U \subseteq V$, we denote by $N^u(T, P, U)$ the minimum size of a (T, P)-non-sequenced-channel system such that every $x \in U$ has exact reconstruction. Similarly, $N^k(T, P, U)$ is defined for the sequenced-channel system. Note that $N^k(T, P, U) \leq N^u(T, P, U)$. In the rest of the paper, whenever we write g, we refer to $g \in \{k, u\}$.

The main problem we study in this paper is formulated as follows.

Problem 1. *Let V be a finite set with distance function $\rho : V \times V \to \mathbb{N}$, $T = (t_1, \ldots, t_\ell)$, and $P = (p_1, \ldots, p_\ell)$. For all $U \subseteq V$, find the values of $N^u(T, P, U)$ and $N^k(T, P, U)$.*

III. The case $\ell = 2$

In this section we study Problem 1 for two types of channels. This result generalizes the case studied by Levenshtein when all the channels are identical [8].

For $x, z \in V$ and $t_1 < t_2 \in \mathbb{N}$ we define

$$I(x, z, t_1, t_2) = B_{t_1}(x) \cap B_{t_2}(z),\ I(x, z, t_1) = B_{t_1}(x) \cap B_{t_1}(z),$$

and

$$N(x, z, t_1, t_2) = |I(x, z, t_1, t_2)|, N(x, z, t_1) = |I(x, z, t_1)|.$$

In the sequel, we assume that $x \neq z$. It is clear that

$$N^g(T, P, U) = \max\{N^g(T, P, \{x, z\}) : x, z \in U\}.$$

Hence, we focus on finding the value of $N^g(T, P, \{x, z\})$ for all $x, z \in U$. Recall that a $(T = (t_1, t_2), P = (p, 1))$-channel system of size N is a set of N combinatorial channels, where $\lceil pN \rceil$ of the channels are t_1-error channels and the others are t_2-error channels.

The following theorem solves Problem 1 for the sequenced model. We omit its proof since it is a simplified version of the non-sequenced case.

Theorem 1. *If $U = \{x, z\} \subseteq V$, $T = (t_1, t_2)$, and $P = (p, 1)$ then $N^k(T, P, U) = N + 1$, where*

$$N = \min\{\lfloor N(x, z, t_1)/p \rfloor, N(x, z, t_2)\}.$$

In the rest of this section, we present the solution for the non-sequenced model. We define

$$N'(x, z, t_1, p) = \min\{L : 2\lceil pL \rceil - L > N(x, z, t_1), L \geq 1\} - 1,$$

where $\min \emptyset = \infty$. This value will be used in calculating the value of $N^u(T, P, \{x, z\})$. The following proposition studies the value of $N'(x, z, t_1, p)$.

Proposition 2. *For $0 < p \leq 1/2$:*

$$N'(x, z, t_1, p) = \begin{cases} 0 & \text{if } N(x, z, t_1) = 0, \\ \infty & \text{otherwise.} \end{cases}$$

For $1/2 < p < 1$:

$$\left\lfloor \frac{N(x, z, t_1) - 2}{2p - 1} \right\rfloor \leq N'(x, z, t_1, p) \leq \left\lfloor \frac{N(x, z, t_1)}{2p - 1} \right\rfloor.$$

We note that if x is transmitted over a channel system with N channels, then at least $\lceil pN \rceil$ of the outputs are in $B_{t_1}(x)$, and all the N outputs in $B_{t_2}(x)$. Thus, to support exact reconstruction for x, we require that for every $z \in U$, there are no N outputs such that all the following three conditions hold simultaneously

(1) at least $\lceil pN \rceil$ of the outputs are in $B_{t_1}(x)$,
(2) at least $\lceil pN \rceil$ of the outputs are in $B_{t_1}(z)$,
(3) all the N outputs are in $B_{t_2}(x) \cap B_{t_2}(z)$.

The following theorem establishes our result in calculating the value of $N^u(T, P, U)$.

Theorem 3. *If $U = \{x, z\} \subseteq V$, $T = (t_1, t_2)$, and $P = (p, 1)$ then $N^u(T, P, U) = N + 1$, where*

$$N = \min\{\ \lfloor N(x, z, t_1, t_2)/p \rfloor,\ N(x, z, t_2),\ \lfloor N(z, x, t_1, t_2)/p \rfloor,\ N'(x, z, t_1, p)\}.$$

Proof. If a (T, P)-channel system consists of $J = N + 1$[1] channels, then, by the definition of N, at least one of the following conditions holds:

(1) $J \geq \lfloor N(x, z, t_1, t_2)/p \rfloor + 1$, (3) $J \geq N(x, z, t_2) + 1$,
(2) $J \geq \lfloor N(z, x, t_1, t_2)/p \rfloor + 1$, (4) $2\lceil pJ \rceil - N(x, z, t_1) > J$.

The first condition implies

$$\lceil pJ \rceil \geq pJ \geq p \cdot (\lfloor N(x, z, t_1, t_2)/p \rfloor + 1) > p \cdot N(x, z, t_1, t_2)/p = N(x, z, t_1, t_2).$$

By the same computation for the second condition, we conclude that at least one of the following conditions holds:

(1) $\lceil pJ \rceil > N(x, z, t_1, t_2)$, (3) $J > N(x, z, t_2)$,
(2) $\lceil pJ \rceil > N(z, x, t_1, t_2)$, (4) $2\lceil pJ \rceil - N(x, z, t_1) > J$.

The above conditions are symmetric for x and z. Thus, without loss of generality, let x be the transmitted word. If Condition (1) or (3) holds, since $\lceil pJ \rceil$ of the outputs are in $B_{t_1}(x)$ and J outputs are in $B_{t_2}(x)$, then not all the outputs are in $B_{t_2}(z)$. If Condition (2) holds, there are no $\lceil pJ \rceil$ outputs in $B_{t_1}(z)$. Thus, if one of conditions (1), (2), or (3) holds, then z will not be decoded incorrectly. For Condition (4), assume that we have m outputs in $I(x, z, t_1)$, where $m \leq N(x, z, t_1)$. In order to decode z incorrectly we must have at least $\lceil pJ \rceil - m$ outputs in $I(z, x, t_1, t_2) \setminus I(x, z, t_1)$. Furthermore, since x was transmitted at least $\lceil pJ \rceil - m$ outputs are in $I(x, z, t_1, t_2) \setminus I(x, z, t_1)$. Thus, we must have that $2\lceil pJ \rceil - m \leq J$ in contradiction to Condition (4).

For the second direction we have to prove that N channels are not sufficient for exact reconstruction where $U = \{x, z\}$. The following four conditions hold simultaneously.

(1) $\lceil pN \rceil \leq N(x, z, t_1, t_2)$, (3) $N \leq N(x, z, t_2)$,
(2) $\lceil pN \rceil \leq N(z, x, t_1, t_2)$, (4) $2\lceil pN \rceil - N(x, z, t_1) \leq N$.

[1]Note that for $J > N + 1$, a (T, P)-non-sequenced-channel system of size J may not support exact reconstruction. That could happen only if $J \leq \min\{\lfloor N(x, z, t_1, t_2)/p \rfloor, \lfloor N(z, x, t_1, t_2)/p \rfloor, N(x, z, t_2)\}$.

The first condition is derived as follows. If $\lceil pN \rceil = pN$, then

$$\begin{aligned}\lceil pN \rceil = pN &\leq p \cdot \lfloor N(x,z,t_1,t_2)/p \rfloor \\ &\leq p \cdot N(x,z,t_1,t_2)/p = N(x,z,t_1,t_2).\end{aligned}$$

Otherwise, $\lceil pN \rceil = \lfloor pN \rfloor + 1$, and

$$\begin{aligned}\lceil pN \rceil = \lfloor pN \rfloor + 1 &< p \cdot \lfloor N(x,z,t_1,t_2)/p \rfloor + 1 \\ &\leq p \cdot (N(x,z,t_1,t_2)/p) + 1 = N(x,z,t_1,t_2) + 1.\end{aligned}$$

Thus, for both cases, $\lceil pN \rceil \leq N(x,z,t_1,t_2)$. The second condition is obtained by the same way.

For this part, we present a set of N outputs which can be the outcome when transmitting either x or z. Let $m = N(x,z,t_1)$. If $m < \lceil pN \rceil$ then m outputs are in $I(x,z,t_1)$, at least $\lceil pN \rceil - m$ outputs are in $I(x,z,t_1,t_2) \setminus I(x,z,t_1)$ (by Conditions (1) and (4)), at least $\lceil pN \rceil - m$ in $I(z,x,t_1,t_2) \setminus I(x,z,t_1)$ (by Conditions (2) and (4)), and all the outputs are in $I(x,z,t_2)$ (by Condition (3)). Otherwise, $m \geq \lceil pN \rceil$, and then $\lceil pN \rceil$ outputs are in $I(x,z,t_1)$ and all the rest are in $I(x,z,t_2)$ (by Condition (3)). Thus, at least $\lceil pN \rceil$ of the outputs are in $B_{t_1}(x)$, and all the N outputs are in $B_{t_2}(x)$, and the same holds for z. □

Note that the setup where all the channels are t-error channels is a special case of $N^u(T,P,U)$ and $N^k(T,P,U)$ for $T = (t,t_2)$ and $P = (1,1)$.

The following corollary is deduced immediately by Proposition 2 and Theorem 3.

Corollary 4. *$N^u(T,P,\{x,z\}) = N+1$ where N is defined as follows. For $0 < p \leq 1/2$:*

$$N = \begin{cases} 0 & \text{if } N(x,z,t_1) = 0 \\ \min\{\lfloor N(x,z,t_1,t_2)/p \rfloor, & \text{otherwise.} \\ \quad \lfloor N(z,x,t_1,t_2)/p \rfloor, & \\ \quad N(x,z,t_2)\}. & \end{cases}$$

For $1/2 < p < 1$:

$$\begin{aligned}N = \min\{&\lfloor N(x,z,t_1,t_2)/p \rfloor, \quad N(x,z,t_2), \\ &\lfloor N(z,x,t_1,t_2)/p \rfloor, \quad N'(x,z,t_1,p)\}.\end{aligned}$$

In the following section we show how to apply the result from Corollary 4 to explicitly solve Problem 1 with $\ell = 2$ for substitution errors over the binary alphabet.

IV. Substitution errors

Let $V = \{0,1\}^n$ be the set of all length n words over the binary alphabet. The Hamming distance function $\rho : V \times V \to \mathbb{N}$ is defined by $\rho(x,z) = |\{i : x_i \neq z_i\}|$.

Note, that for all $x,z \in V$, $N(x,z,t_1,t_2)$ and $N(x,z,t)$ depend only on $d = \rho(x,z)$. Thus, for $x,z \in V$ such that $d = \rho(x,z)$, we denote by $N(d,t_1,t_2)$ and $N(d,t)$ the values $N(x,z,t_1,t_2)$ and $N(x,z,t)$, respectively. Let $N^g(T,P,d)$ be defined as the maximum value of $N^g(T,P,U)$, for all U such that $d(U) \geq d$, where $d(U) = \min\{\rho(x,z) : x,z \in U\}$. As before we get that

$$N^g(T,P,d) = \max\{N^g(T,P,\{x,z\}) : x,z \in V, \rho(x,z) \geq d\}.$$

The next theorem proves that for all $d \geq 1$, $N^g(T,P,d) \geq N^g(T,P,d+1)$. This desirable property, known as the *monotonicity by intersection* [8], holds also in our case.

Theorem 5. *For fixed p, $0 < p < 1$, $d \geq 1$, and $t_1 < t_2$, $N^g(T,P,d) \geq N^g(T,P,d+1)$, where $T = (t_1,t_2)$ and $P = (p,1)$.*

According to Theorem 5, in order to calculate the value of $N^g(T,P,d)$ it is enough to find the value of $N^g(T,P,\{x,z\})$, where $\rho(x,z) = d$. Therefore, according to Theorem 1 and Theorem 3, for $T = (t_1,t_2)$ and $P = (p,1)$, we conclude that

$$\begin{aligned}N^k(T,P,d) &= \min\{\lfloor N(d,t_1)/p \rfloor, N(d,t_2)\}, \text{ and} \\ N^u(T,P,d) &= \min\{\lfloor N(d,t_1,t_2)/p \rfloor, N(d,t_2), N'(d,t_1,p)\},\end{aligned}$$

where $N'(d,t_1,p) = \min\{L : 2\lceil pL \rceil - L > N(d,t_1), L \geq 1\} - 1$.

In the sequel, we find explicitly the value of $N^u(T,P,d)$. We focus on the non-sequenced model, since the sequenced one can be easily derived from Theorem 1 and Levenshtein's results in [8].

The following lemma was shown in [8].

Lemma 6. *For $t,d \geq 1$,*

$$N(d,t) = \sum_{i=0}^{\lfloor t - \frac{d}{2} \rfloor} \binom{n-d}{i} \cdot \sum_{k=d-t+i}^{t-i} \binom{d}{k},$$

where $\binom{a}{b} = 0$ if $a < b$ or $b < 0$.

Note that $t - \lceil \frac{d}{2} \rceil = \lfloor t - \frac{d}{2} \rfloor$. By similar combinatorial computation, we can compute the value of $N(d,t_1,t_2)$.

Lemma 7. *For $t_1 \leq t_2$:*

$$N(d,t_1,t_2) = \sum_{i=0}^{\lfloor \frac{t_1+t_2-d}{2} \rfloor} \binom{n-d}{i} \cdot \sum_{k=d-t_2+i}^{t_1-i} \binom{d}{k},$$

The following two lemmas compare between the three components which determine the value of $N^u(T,P,d)$, for $d \geq 1$, $t_1 < t_2 \in \mathbb{N}$, and fixed $0 < p < 1$. Lemma 8 compares between $\lfloor N(d,t_1,t_2)/p \rfloor$ and $N(d,t_2)$.

Lemma 8. *For any fixed p and n sufficiently large the following holds. If d is odd, $p \leq 1/2$, and $t_2 = t_1 + 1$, then*

$$N(d,t_2) < \lfloor N(d,t_1,t_2)/p \rfloor.$$

Otherwise,

$$N(d,t_2) \geq \lfloor N(d,t_1,t_2)/p \rfloor.$$

Proof. Note that $N(d,t_2) = \Theta(n^{\lfloor \frac{2t_2-d}{2} \rfloor})$ and $N(d,t_1,t_2) = \Theta(n^{\lfloor \frac{t_1+t_2-d}{2} \rfloor})$. Thus, we compare between the powers $\lfloor \frac{2t_2-d}{2} \rfloor$ and $\lfloor \frac{t_1+t_2-d}{2} \rfloor$. If $t_2 = t_1 + 1$ and d is odd then $\lfloor \frac{2t_2-d}{2} \rfloor = \lfloor \frac{t_1+t_2-d}{2} \rfloor$. In all other cases, $\lfloor \frac{2t_2-d}{2} \rfloor > \lfloor \frac{t_1+t_2-d}{2} \rfloor$, and hence $N(d,t_2) > \lfloor N(d,t_1,t_2)/p \rfloor$.

For the case of $t_2 = t_1 + 1$ and odd d, we compare the coefficients of the dominant powers. Denote $d = 2m+1$.

$$\begin{aligned}N(d,t_2) &= \left(\binom{d}{m} + \binom{d}{m+1}\right) \cdot \binom{n-d}{t_1-m} \\ &\quad + \sum_{k=m-1}^{m+2} \binom{d}{k} \cdot \binom{n-d}{t_1-m-1} + \Theta(n^{t_1-m-2}), \\ N(d,t_1,t_2) &= \binom{d}{m} \cdot \binom{n-d}{t_1-m} \\ &\quad + \sum_{k=m-1}^{m+1} \binom{d}{k} \cdot \binom{n-d}{t_1-m-1} + \Theta(n^{t_1-m-2}).\end{aligned}$$

Thus, the coefficient of the dominant powers in $N(d,t_2)$ is twice the coefficient of the corresponding term in $N(d,t_1,t_2)$. But, $N(d,t_1,t_2)$ is multiplied by $1/p$. Thus, for $p > 1/2$ we have

$$\lfloor N(d,t_1,t_2)/p \rfloor \leq N(d,t_2),$$

and for $p < 1/2$,

$$\lfloor N(d,t_1,t_2)/p \rfloor > N(d,t_2).$$

For $p = 1/2$, we compare the coefficient of the second dominant powers in these two terms and get that $\sum_{k=m-1}^{m+2} \binom{d}{k} < 2 \cdot \sum_{k=m-1}^{m+1} \binom{d}{k}$. Thus, we conclude that for this case $\lfloor N(d,t_1,t_2)/p \rfloor > N(d,t_2)$. □

The following lemma compares between the values of $N'(d,t_1,p)$ and $\min\{\lfloor N(d,t_1,t_2)/p \rfloor, N(d,t_2)\}$. Recall that according to Proposition 2, for $0 < p \leq 1/2$, $N'(d,t_1,p) \in \{0,\infty\}$, and by Lemma 8 if $1/2 < p < 1$ then $N(d,t_1,t_2)/p \leq N(d,t_2)$. Thus, in Lemma 9 we compare only between $\lfloor N(d,t_1,t_2)/p \rfloor$ and $\left\lfloor \frac{N(d,t_1)}{2p-1} \right\rfloor$ for $1/2 < p < 1$.

Lemma 9. *For any fixed p and n sufficiently large the following holds. If d is even, $t_2 = t_1 + 1$, and ($1/2 < p \leq 2/3$ or ($2/3 < p < 3/4$ and $d < \frac{2-2p}{3p-2}$)), then*

$$\left\lfloor \frac{N(d,t_1)}{2p-1} \right\rfloor > \lfloor N(d,t_1,t_2)/p \rfloor .$$

Otherwise,

$$\left\lfloor \frac{N(d,t_1)}{2p-1} \right\rfloor \leq \lfloor N(d,t_1,t_2)/p \rfloor .$$

Proof. Note that $N(d,t_1) = \Theta(n^{\lfloor \frac{2t_1-d}{2} \rfloor})$ and $N(d,t_1,t_2) = \Theta(n^{\lfloor \frac{t_1+t_2-d}{2} \rfloor})$. Thus, we compare the powers $\lfloor \frac{2t_1-d}{2} \rfloor$ and $\lfloor \frac{t_1+t_2-d}{2} \rfloor$. If $t_2 = t_1+1$ and d is even then $\lfloor \frac{2t_1-d}{2} \rfloor = \lfloor \frac{t_1+t_2-d}{2} \rfloor$. In all other cases, $\lfloor \frac{2t_1-d}{2} \rfloor < \lfloor \frac{t_1+t_2-d}{2} \rfloor$, and hence, $\left\lfloor \frac{N(d,t_1)}{2p-1} \right\rfloor < \lfloor N(d,t_1,t_2)/p \rfloor$.

For the case of $t_2 = t_1+1$ and even d, we compare the coefficients of the dominant powers.

$$N(d,t_1) = \tbinom{d}{d/2} \cdot \tbinom{n-d}{t_1-d/2} + \Theta(n^{t_1-d/2-1}),$$
$$N(d,t_1,t_2) = \left(\tbinom{d}{d/2-1} + \tbinom{d}{d/2} \right) \cdot \tbinom{n-d}{t_1-d/2} + \Theta(n^{t_1-d/2-1}).$$

Thus, the coefficient of the dominant term in $\left\lfloor \frac{N(d,t_1)}{2p-1} \right\rfloor$ is

$$\frac{1}{2p-1}\binom{d}{d/2},$$

while the corresponding coefficient in $\lfloor N(d,t_1,t_2)/p \rfloor$ is

$$\frac{1}{p}\left(\binom{d}{d/2} + \binom{d}{d/2-1} \right) = \frac{2d+2}{(d+2)p}\binom{d}{d/2}.$$

The inequality

$$\frac{2d+2}{(d+2)p} < \frac{1}{2p-1}$$

holds if and only if

$$(p \leq 2/3) \text{ or } (2/3 < p < 3/4 \text{ and } d < \frac{2-2p}{3p-2}).$$

Therefore, we conclude that

$$\left\lfloor \frac{N(d,t_1,t_2)}{p} \right\rfloor < \left\lfloor \frac{N(d,t_1)}{2p-1} \right\rfloor$$

if and only if d is even, $t_2 = t_1+1$, and $((1/2 < p \leq 2/3) \text{ or } (2/3 < p < 3/4 \text{ and } d < \frac{2-2p}{3p-2}))$. □

According to Corollary 4, Lemma 8, and Lemma 9, we can now summarize the results for the binary substitutions case.

Corollary 10. *For any fixed p and n sufficiently large the following holds.*

- *For $0 < p \leq 1/2$:*

$$N^u(T,P,d) = \begin{cases} 1 & \text{if } d > 2t_1, \\ \Theta(n^{\lfloor \frac{t_1+t_2-d}{2} \rfloor}) & \text{otherwise.} \end{cases}$$

- *For $1/2 < p < 1$:*

$$N^u(T,P,d) = \Theta(n^{\lfloor \frac{2t_1-d}{2} \rfloor}).$$

More specifically,

- *For $0 < p \leq 1/2$:*

$$N^u(T,P,d) = \begin{cases} 1 & \text{if } d > 2t_1, \\ N(d,t_2)+1 & \text{otherwise, if } d \text{ is odd} \\ & \text{and } t_2 = t_1+1, \\ \lfloor N(d,t_1,t_2)/p \rfloor + 1 & \text{otherwise.} \end{cases}$$

- *For $1/2 < p < 1$:*

$$N^u(T,P,d) = \begin{cases} \left\lfloor \frac{N(d,t_1,t_2)}{p} \right\rfloor & \text{if } d \text{ is even, } t_2 = t_1+1, \\ & \text{and } \left(\left(\frac{1}{2} < p \leq \frac{2}{3}\right) \vee \right. \\ & \left. \left(\frac{2}{3} < p < \frac{3}{4} \wedge d < \frac{2-2p}{3p-2}\right)\right), \\ N'(d,t_1,p) & \text{otherwise.} \end{cases}$$

We note that we can generalize Lemma 8 for non fixed values of p, i.e, for p which is a function of n, however this part is omitted due to lack of space.

Lastly, we discuss some special cases of this model. Let $L_1 = N(d,t_2)+1$ be the solution for the case where all the channels are identical, and $L_2 = N^u(T=(t_1,t_2), P=(p,1), d)$ be the solution for our general problem.

- For fixed p, $0 < p \leq 1/2$, $d = 1$, and $T = (2,4)$, $L_2 = \Theta(n^2)$, while $L_1 = \Theta(n^3)$,
- For fixed p, $0 < p \leq 1/2$, $d = 1$, and $T = (2,8)$, $L_2 = \Theta(n^4)$, while $L_1 = \Theta(n^7)$,
- For fixed p, $1/2 < p \leq 2/3$, $d = 2$, and $T = (4,5)$, $L_2 = \Theta(n^3)$, while $L_1 = \Theta(n^4)$.

V. Problem 1 - The General Case

In this section, we extend the solution from Section III. We provide a combinatorial translation for the general case of Problem 1, where $T = (t_1, \ldots, t_\ell)$ and $P = (p_1, \ldots, p_{\ell-1}, p_\ell)$, $t_1 < t_2 < \ldots < t_\ell \in \mathbb{N}$, and $0 < p_1 < p_2 < \ldots < p_{\ell-1} < p_\ell = 1$. A (T,P)-channel system of size N consists of N channels, where for each $i \in [\ell]$, $\lceil p_i N \rceil$ channels are t_i-error channels.

Theorem 11 and Theorem 12 generalize Theorem 1 and Theorem 3 for arbitrary ℓ, respectively.

Theorem 11. *For $x, z \in V$, $N^k(T,P,\{x,z\}) = N+1$, where*

$$N = \min\{\lfloor N(x,z,t_i)/p_i \rfloor : i \in [\ell]\}.$$

Now, we consider the non-sequenced case. Recall that if x is transmitted over a (T,P)-channel system of size N, then for all $i \in [\ell-1]$ at least $\lceil p_i N \rceil$ of the outputs are in $B_{t_i}(x)$, and all the N outputs are in $B_{t_\ell}(x)$. Then, x does not have exact reconstruction if there exists a different word z, where for all $i \in [\ell-1]$ at least $\lceil p_i N \rceil$ of the outputs are in $B_{t_i}(z)$, and all the N outputs are in $B_{t_\ell}(z)$.

Theorem 12. *For $x, z \in V$, $N^u(T,P,\{x,z\}) = N+1$, where*

$$\begin{aligned} N = \min \quad & \{\lfloor N(x,z,t_i,t_\ell)/p_i \rfloor : i \in [\ell-1]\} \\ & \cup \{\lfloor N(z,x,t_i,t_\ell/p_i) \rfloor : i \in [\ell-1]\} \\ & \cup \{N(x,z,t_\ell)\} \\ & \cup \{N'(x,z,t_i,t_j,p_i,p_j) : i,j \in [\ell-1]\}, \end{aligned}$$

$$N'(x,z,t_i,t_j,p_i,p_j) = \min\{L : \lceil p_i L \rceil + \lceil p_j L \rceil - L > N(x,z,t_i,t_j), L \geq 1\} - 1,$$

and $\min \emptyset = \infty$.

VI. Special systems for $T = (t_1, t_2)$

In this section we study special cases of two types of channels. For $T = (t_1,t_2)$, $t_1 < t_2$, and a constant integer a, a channel system with N combinatorial channels is called a (T,i,a)-*channel system*, $i \in \{1,2\}$, if a of the channels are t_i-error channels, while the others are t_{3-i}-error channels. If the size of a system is smaller than a, then all the channels are t_i-error.

In this model, we consider both cases, sequenced and non-sequenced. For $U \subseteq V$, we denote by $N^u(T,i,a,U)$,$N^k(T,i,a,U)$ the minimum size of a (T,i,a)-non-sequence,(T,i,a)-sequenced -channel system such that each $x \in U$ has exact reconstruction, respectively.

In this section we solve the following problem for $i \in \{1,2\}$.

Problem 2. *Let V be a finite set with some distance function $\rho : V \times V \to \mathbb{N}$, for all $U \subseteq V$, find the values of $N^u(T,i,a,U)$ and $N^k(T,i,a,U)$.*

As before, we focus on sets of the form $U = \{x,z\}$ since $N^g(T,i,a,U) = \max\{N^g(T,i,a,\{x,z\}) : x,z \in U\}$.

The solution for this problem is presented in the next three theorems. The first theorem solves the problem for constant number of t_1-error channels. In this case, the minimum number of channels which are required for exact reconstruction does not depend on knowing the behavior of each channel.

The last two theorems present solutions for constant number of t_2-error channels; Theorem 15 for non-sequence system, and Theorem 14 for the sequence one.

Theorem 13. *For $U = \{x, z\} \subseteq V$ and $T = (t_1, t_2)$, $N^k(T, 1, a, U) = N^u(T, 1, a, U) = N + 1$, where*

$$N = \begin{cases} N(x, z, t_2) & \text{if } N(x, z, t_1) \geq a, \\ N(x, z, t_1) & \text{otherwise.} \end{cases}$$

Note, that in almost all the cases

$$N^k(T, 1, a, U) = N^u(T, 1, a, U) = N(x, z, t_2) + 1.$$

Proof. If $N(x, z, t_1) < a$, then a $(T, 2, a)$-channel system of size at most $N(x, z, t_1) + 1$ contains only t_1-channels. Thus, according to Levenshtein [8], $N^k(T, 1, a, U) = N^u(T, 1, a, U) = N(x, z, t_1) + 1$. If $N(x, z, t_1) \geq a$, then it is clear that $N(x, z, t_2) + 1$ channels are sufficient.

For the second direction, without loss of generality, let us assume that x is transmitted over the system. If a outputs are in $I(x, z, t_1)$ and all the $N(x, z, t_2)$ in $I(x, z, t_2)$, then z may be decoded incorrectly. □

In the second case $i = 2$ and a is the number of channels with maximum t_2 errors. First, we state the solution for the case where the type of the channels is known.

Theorem 14. *For $U = \{x, z\} \subseteq V$ and $T = (t_1, t_2)$, $N^k(T, 2, a, U) = N + 1$, where*

$$N = \min\{N(x, z, t_1) + a, N(x, z, t_2)\}.$$

Note, that in almost all the cases

$$N^k(T, 2, a, U) = N(x, z, t_1) + a + 1.$$

Lastly, we solve Problem 2 for the non-sequenced model.

Theorem 15. *For $U = \{x, z\} \subseteq V$ and $T = (t_1, t_2)$, $N^u(T, 2, a, U) = N + 1$, where*

$$N = \min\{\ N(x, z, t_1, t_2) + a,\ \ N(x, z, t_2),\ \ N(z, x, t_1, t_2) + a,\ \ N(x, z, t_1) + 2a\}.$$

Note, that in almost all the cases

$$N^u(T, 2, a, U) = N(x, z, t_1) + 2a + 1.$$

Proof. The proof is similar to the one of Theorem 3. If a $(T, 2, a)$-channel system consists of $J = N + 1$ channels, then, by the definition of N, at least one of the following conditions exists:

(1) $J - a > N(x, z, t_1, t_2)$, (3) $J > N(x, z, t_2)$,
(2) $J - a > N(z, x, t_1, t_2)$, (4) $2(J - a) - N(x, z, t_1) > J$.

The above conditions are symmetric for x and z. Thus, without loss of generality, let x be the transmitted word. If Condition (1) or (3) holds, since $J - a$ of the outputs are in $B_{t_1}(x)$ and J outputs in $B_{t_2}(x)$, then not all the outputs are in $B_{t_2}(z)$. If Condition (2) holds, there are no $J - a$ outputs in $B_{t_1}(z)$. Thus, if one of the conditions (1), (2), or (3) holds, then z will not be decoded incorrectly. Regarding Condition (4), assume that we have m outputs in $I(x, z, t_1)$, where $m \leq N(x, z, t_1)$. In order to decode z incorrectly we must have at least $J - a - m$ outputs in $I(z, x, t_1, t_2) \setminus I(x, z, t_1)$. Furthermore, since x was transmitted at least $J - a - m$ outputs are in $I(x, z, t_1, t_2) \setminus I(x, z, t_1)$. Thus, we must have that $2(J - a) - m \leq J$ in contradiction to Condition (4).

For the second direction we have to prove that N channels are not sufficient to exact reconstruction where $U = \{x, z\}$. The following four conditions hold simultaneously.

(1) $N - a \leq N(x, z, t_1, t_2)$, (3) $N \leq N(x, z, t_2)$,
(2) $N - a \leq N(z, x, t_1, t_2)$, (4) $2(N - a) - N(x, z, t_1) \leq N$.

For this part, we present a set of N outputs which can be the output of both x and z. Let $m = N(x, z, t_1)$. If $m < N - a$ then, m outputs are in $I(x, z, t_1)$, at least $N - a - m$ in $I(x, z, t_1, t_2) \setminus I(x, z, t_1)$ (by Conditions (1) and (4)), at least $N - a - m$ in $I(z, x, t_1, t_2) \setminus I(x, z, t_1)$ (by Conditions (2) and (4)), and all the others in $I(x, z, t_2)$ (by Condition (3)). Otherwise, $m \geq N - a$, and then at least $N - a$ outputs are in $I(x, z, t_1)$ and a in $I(x, z, t_2)$ (by Condition (3)). Thus, at least $N - a$ of the outputs are in $B_{t_1}(x)$, and all the N outputs in $B_{t_2}(x)$, and the same holds for z. □

According to the previous theorem, one can verify that for the Hamming case with $a = 2$, $t_1 = 1, t_2 = 2$, and $\rho(x, z) = 3$, we get that $N^u(T, 2, a, U) = 5$, while if all channels cause at most 2 errors, then the number of channels for exact reconstruction is 7 [8].

Note that Theorem 15 can be also derived by a slight modification in Theorem 3. We denote $m = N(x, z, t_1)$ and we define here

$$N'(x, z, t_1, p) = \min\{L : 2\lceil pL \rceil - L > m, \lceil pL \rceil > m, L \geq 1\} - 1,$$

instead of the previous definition, where

$$N'(x, z, t_1, p) = \min\{L : 2\lceil pL \rceil - L > m, L \geq 1\} - 1.$$

This change has no affect on Theorem 3, since for fixed p, $0 < p < 1$, $2\lceil pL \rceil - L \leq \lceil pL \rceil$. Then, by substituting $\lceil pL \rceil = L - a$ in Theorem 3 we can conclude Theorem 15.

References

[1] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A DNA-based archival storage system," *ASPLOS*, pp. 637–649, Atlanta, GA, Apr. 2016.

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

[3] R. Gabrys and E. Yaakobi, "Sequence reconstruction over the deletion channel," *Proc. Int. Symp. Inform. Theory*, pp. 1596–1600, Barcelona, Spain, Jul. 2016.

[4] T. Jiang and A. Vardy, "Asymptotic improvement of the Gilbert-Varshamov bound on the size of binary codes," *IEEE Trans. on Inform. Theory*, vol. 50, no. 8, pp. 1655–1664, Aug. 2004.

[5] E. Konstantinova, "Reconstruction of permutations distorted by single reversal errors," *Discrete Applied Math.*, vol. 155, pp. 2426–2434, 2007.

[6] E. Konstantinova, V. Levenshtein, and J. Siemons, "Reconstruction of permutations distorted by single transposition errors," *http://arxiv.org/abs/math/0702191v1*, Feb. 2007.

[7] E. Konstantinova, "On reconstruction of signed permutations distorted by reversal errors," *Discrete Mathematics*, vol. 308, pp. 974–984, 2008.

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

[9] V. I. Levenshtein, "Efficient reconstruction of sequences from their subsequences or supersequences" *Journal of Combin. Theory, Ser. A*, vol. 93, no. 2, pp. 310–332, 2001.

[10] V.I. Levenshtein, E. Konstantinova, E. Konstantinov, and S. Molodtsov, "Reconstruction of a graph from 2-vicinities of its vertices," *Discrete Applied Math.*, vol. 156, pp. 1399–1406, 2008.

[11] V.I. Levenshtein and J. Siemons, "Error graphs and the reconstruction of elements in groups," *Journal of Combin. Theory, Ser. A*, vol. 116, pp. 795–815, 2009.

[12] K. Mazooji, Personal Communication, 2017.

[13] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, "Three novel combinatorial theorems for the insertion/deletion channel," *Proc. Int. Symp. Inform. Theory*, Hong Kong, 2015.

[14] E. Yaakobi and J. Bruck, "On the Uncertainty of Information Retrieval in Associative Memories," *Proc. Int. Symp. Inform. Theory*, pp. 106–110, Jul. 2012.

[15] E. Yaakobi, M. Schwartz, M. Langberg, J. Bruck, "Sequence reconstruction for Grassmann graphs and permutations," *Proc. Int. Symp. Inform. Theory*, pp. 874–878, Jul. 2013.

[16] S. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Cold Spring Harbor Labs Journals*, 2016.

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

[18] S. M. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Nature Scientific Reports*, vol. 5, no. 14138, Aug. 2015.