The Entropy Rate of Some Pólya String Models

Ohad Elishco, *Member, IEEE*, Farzad Farnoud (Hassanzadeh), *Member, IEEE*, Moshe Schwartz, *Senior Member, IEEE*, and Jehoshua Bruck, *Fellow, IEEE*

***Abstract*— We study random string-duplication systems, which we call Pólya string models. These are motivated by a class of mutations that are common in most organisms and lead to an abundance of repeated sequences in their genomes. Unlike previous works that study the combinatorial capacity of string-duplication systems, or in a probabilistic setting, various string statistics, this work provides the exact entropy rate or bounds on it, for several probabilistic models. The entropy rate determines the compressibility of the resulting sequences, as well as quantifying the amount of sequence diversity that these mutations can create. In particular, we study the entropy rate of noisy string-duplication systems, including the tandem-duplication, end-duplication, and interspersed-duplication systems, where in all cases we study duplication of length 1 only. Interesting connections are drawn between some systems and the signature of random permutations, as well as to the beta distribution common in population genetics.**

***Index Terms*— DNA storage, string-duplication systems, entropy rate, Pólya string models.**

I. Introduction

SEVERAL mutation processes are known, which affect the genetic information stored in the DNA. Among these are transposon-driven repeats [22] and tandem repeats which are believed to be caused by slipped-strand mispairings [25]. In essence, these mutation processes take a substring of the DNA and insert a copy of it somewhere else (in the former case), or next to the original copy (in the latter). In human DNA, it is known that its majority consists of repeated sequences [22]. Moreover, certain repeats cause important phenomena such as chromosome fragility, expansion diseases, gene silencing [38], and rapid morphological variation [15].

Manuscript received August 20, 2018; revised May 15, 2019; accepted August 9, 2019. Date of publication August 22, 2019; date of current version November 20, 2019. This work was supported in part by the National Foundation under Grant CCF-1317694, Grant CCF-1755773, Grant CCF-1816409, Grant CCF-1816965, and Grant CCF-1717884 and in part by the United States–Israel Binational Science Foundation (BSF) under Grant 2017652. This article was presented in part at the 2016 IEEE International Symposium on Information Theory.

O. Elishco and M. Schwartz are with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer Sheva 8410501, Israel (e-mail: ohadeli@post.bgu.ac.il; schwartz@ee.bgu.ac.il).

F. Farnoud (Hassanzadeh) was with the Electrical Engineering Department, California Institute of Technology, Pasadena, CA 91125 USA. He is now with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903 USA (e-mail: farzad@virginia.edu).

J. Bruck is with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: bruck@paradise.caltech.edu).

Communicated by A. W. Eckford, Associate Editor for Communications.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2019.2936556

A formal mathematical model for studying these kinds of mutation processes is the notion of *string-duplication systems*. In such systems, a seed string (or strings) evolves over time by successive applications of mutating functions. For example, functions taking a substring of a string and copying it next to itself model mutation by tandem duplication. These string-duplication systems were studied in the context of formal languages (e.g., [24]) in an effort to place the resulting sets of mutated sequences within Chomsky's hierarchy of formal languages, as well as to derive closure properties.

In the context of coding theory, string-duplication systems were studied, motivated by applications to DNA storage in living organisms. In such a storage scheme, information is stored in the DNA of some organisms, and later read from them or their descendants [35]. This information, however, is corrupted by mutations. These include substitution errors, as well as insertions and deletions – all of which have already been extensively studied in the coding-theoretic community. However, another type of error is that of duplication, modeled mathematically by string-duplication systems.

Various aspects of string-duplication systems were studied, geared towards a comprehensive coding solution to duplication mutations. In [13], [16], the duplication mutation processes were treated as a source, and their exponential growth rate, i.e., their *combinatorial capacity*, studied. This provided insights into the structure of error balls in the string-duplication channel. Some error-correcting codes for tandem duplication were presented in [17], [23]. The confusability of strings under tandem duplication was studied in [6], the mutation distance was bounded in [1], and more recently, [40] developed reconstruction schemes for uniform tandem duplication.

A drawback of all the papers mentioned above is a combinatorial (adversarial) approach, whereas we suspect a scenario involving DNA storage in living organisms must be probabilistic. To address this gap, a probabilistic model was studied in [12]. This model is not concerned with which mutated strings are possible, but rather with which are *probable*. With appropriate distributions applied to the choice of the mutated point, the mutation length, and its final position, we obtain an induced distribution on resulting strings. However, [12] was not able to provide any exact entropy rate calculation nor bounds, and managed to study only peripheral properties of the resulting string distributions, namely the frequencies of symbols and substrings.

Thus, the goal of this paper is to find the exact entropy rate of probabilistic string-duplication systems, or bound it. We also generalize the process to include noisy duplication.

The entropy of biological sequences is of interest for several reasons. First, it provides a lower bound on the compressed size of sequence data, against which we can evaluate the performance compression methods. Second, entropy is a measure of diversity and complexity. The fact that biological diversity arises from sequence diversity motivates studying how sequence diversity is created through mutation. For example, we may explore how the diversity of the output of a string system is influenced by model parameters, such as the substitution rate. Third, entropy and other measures of sequence complexity have been used to determine the origin and/or the role of DNA sequences [11], [31], [39], for example to classify protein-coding and non-coding regions of a genome. While there have been many works studying entropy of biological sequences, including [11], [21], [26], [34], they have focused on *estimating* the entropy based on data rather than finding the entropy as a function of model parameters.

The main contributions of this paper are exact expressions for end-duplication systems and interspersed-duplication systems, for all noise parameters. Additionally, we find the exact entropy rate of noiseless tandem duplication and complement tandem duplication, and bound the entropy rate of the general noisy tandem-duplication system. Duplication is a major mechanism for the generation of genetic material [30], [37] and repeated sequences make up a significant part of many genomes, including that of humans [22]. Many of the repeated sequences, however, are not exact copies, pointing to the presence of other mutations, such as substitutions. Indeed, probabilistic models of tandem duplication studied in the literature also allow substitutions and other point mutations [5], [8], [14], [19], [20]. We thus study noisy duplication models, where copies can deviate from the original sequence. We note however that this is still an incomplete model and does not consider mutations such as insertion, deletion, and translocation, for the sake of simplicity. In all cases we study duplication of length 1 only. As we later see, even for very modest parameters this problem is challenging.

An important tool, widely used in the study of genetic drift in population genetics, is a *Pólya urn model*. It consists of an urn with balls of two different colors. In each step a ball is randomly (independently and uniformly) chosen and returned to the urn along with k new balls of the same color [9], [29], [32], [33]. There are many extensions to this model, where after each draw, a set of balls, whose number and composition depends on the color of the drawn ball, are put into the urn. However, in these models there is no structure on the balls in the urn and only the number of balls of each color matters. Thus, these models fail to apply to strings.

We therefore suggest extensions of the Pólya urn models to what we call *Pólya string models*, in which the balls form a string, which may be circular or linear, similar to bases of a DNA molecule. A step in this model typically involves choosing a random position (or equivalently a ball) in the string, where a modification to the string – the mutation – occurs. In this paper, we focus on models in which after the draw, a sequence of balls is inserted to the string whose composition and position depend on the local properties of the string around the chosen position.

The paper is organized as follows. In Section II we fix our notation and definitions that are used throughout the paper. In Section III we find the exact entropy rate of end duplication. In Section IV study tandem duplication. Section V presents the entropy rate of interspersed duplication. We conclude in Section VI by providing some insight and comparisons with the combinatorial capacity and Pólya urn models.

II. Preliminaries

Let $\Sigma \triangleq \{0, 1\}$ be the binary alphabet. Not all results may be readily generalized to any finite alphabet. Thus, for the sake of simplicity we focus on the binary case only, and leave the generalization to future work. The elements of Σ are referred to as letters (symbols). We use the notation common to formal languages to describe strings over Σ. The set of length-n strings (sequences) over Σ is denoted by Σ^n. We let Σ^* denote the set of all finite-length strings over Σ. The unique empty string is denoted by ε. The set of all finite-length non-empty strings is denoted by $\Sigma^+ \triangleq \Sigma^* \setminus \{\varepsilon\}$.

To help with readability, we shall use the first lowercase letters of the roman alphabet, e.g., $a, b, c, \ldots$, to denote single letters from the alphabet Σ. We shall use the last lowercase letters of the roman alphabet, e.g., $u, v, w, \ldots$, to denote strings from Σ^*.

Let $w \in \Sigma^*$ be a string. We use $|w|$ to denote the length of w, i.e., the number of letters it contains. Obviously, $|\varepsilon| = 0$. If $w' \in \Sigma^*$, the concatenation of w and w' is denoted ww'. For $i \in \mathbb{N}$, the ith letter of a string $w \in \Sigma^*$ (assuming $|w| \geqslant i$) will be denoted by w_i, i.e., $w = w_1\, w_2 \ldots w_{|w|}$ with $w_j \in \Sigma$ for all j.

The number of occurrences of a symbol $a \in \Sigma$ in the string w is denoted by $|w|_a$. If $w \neq \varepsilon$, then the frequency of $a \in \Sigma$ in w is defined by $\mathrm{fr}_a(w) \triangleq |w|_a / |w|$.

For a natural number $n \in \mathbb{N}$ we use $[n]$ to denote the set $[n] \triangleq \{1, 2, \ldots, n\}$. We also recall the definition of the binary entropy function, $H_2 : [0, 1] \to [0, 1]$ defined as

$$H_2(x) \triangleq -x \log_2(x) - (1 - x) \log_2(1 - x).$$

Example 1. *Let $w = 0011$ and $w' = 001$. We have that $w_4 = 1$, $ww' = 0011001$ with $|w|_0 = 2$ and $|ww'|_0 = 4$. Also, $\mathrm{fr}_0(w) = 1/2$ while $\mathrm{fr}_0(ww') = 4/7$.* □

The Pólya string model may be quite generally defined. Intuitively, the model takes a starting string, and in a sequence of steps, mutates it over time. A formal definition follows.

Definition 2. *A Pólya string model is defined by $S = (\Sigma, s, T)$, where Σ is a finite alphabet, $S(0) = s \in \Sigma^+$ is a* seed string, *and $T : \Sigma^* \to \Sigma^*$ is a non-deterministic* duplication rule. *The string model is the following discrete-time random process: For all $i \in \mathbb{N}$ set $S(i) = T(S(i - 1))$.*

Several rule choices parallel the combinatorial (deterministic) systems studied in [13], and are special cases of the general stochastic systems studied in [12]. In particular, we define the following three Pólya string models, which we study in the rest of the paper. All three models share the fact that the mutation rule chooses a random location in the string it is given, and duplicates the single symbol appearing in

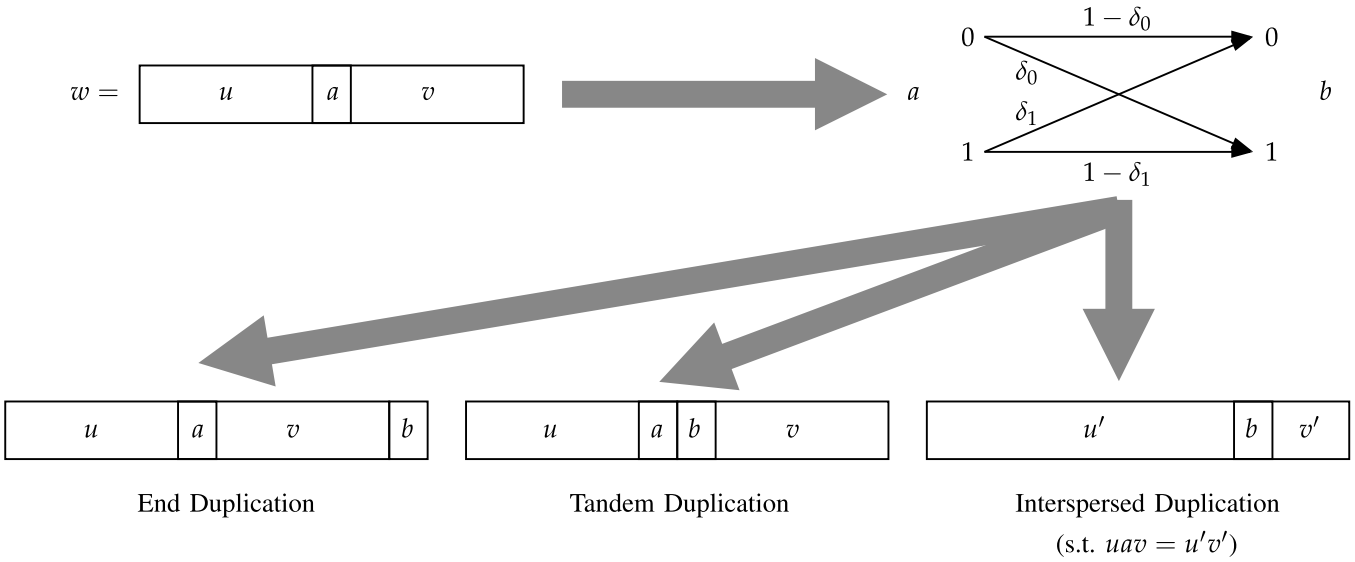

Fig. 1. A step in the three Pólya string models: A random position in the word w is chosen. The letter a in that position is fed to an asymmetric binary channel whose output is b. The letter b is either placed at the end (for end duplication), after the letter a (for tandem duplication), or in some random position (for interspersed duplication).

that location. The duplicate symbol however is noisy, namely, it may be seen as if having passed through a binary asymmetric channel. The rules differ in the location the new symbol is inserted. The three models are defined as follows:

a) ***End duplication****:* For any real numbers $\delta_0, \delta_1 \in [0, 1]$, the end-duplication system is defined as $S^{\text{end}}_{\delta_0,\delta_1} = (\Sigma, s, T^{\text{end}}_{\delta_0,\delta_1})$, where for all $w \in \Sigma^+$,

$$T^{\text{end}}_{\delta_0,\delta_1}(w) \triangleq uavb.$$

Here $u, v \in \Sigma^*$, $a, b \in \Sigma$, $uav = w$, the length $|ua|$ is chosen randomly independently and uniformly from $[|w|]$, and $\Pr(a = b|a = i) = 1 - \delta_i$. In essence, this non-deterministic rule chooses a uniformly random position in w, and duplicates the letter there at the end of the word. If the chosen bit is $a = 0$, the duplicated symbol is complemented with probability δ_0, and similarly, if $a = 1$ the duplicated bit is complemented with probability δ_1. The end-duplication model, while not corresponding to a common type of biological mutation, is a relatively simple point of departure that provides intuition into duplication systems.

b) ***Tandem duplication****:* Similarly, for any real numbers $\delta_0, \delta_1 \in [0, 1]$, the tandem-duplication system is defined as $S^{\text{tan}}_{\delta_0,\delta_1} = (\Sigma, s, T^{\text{tan}}_{\delta_0,\delta_1})$, where for all $w \in \Sigma^+$,

$$T^{\text{tan}}_{\delta_0,\delta_1}(w) \triangleq uabv.$$

Here $u, v \in \Sigma^*$, $a, b \in \Sigma$, $uav = w$, the length $|ua|$ is chosen randomly independently and uniformly from $[|w|]$, and $\Pr(a = b|a = i) = 1 - \delta_i$. This time, the $T^{\text{tan}}_{\delta_0,\delta_1}$ rule chooses a uniformly random position in w, and duplicates the letter there right after its original position. If the chosen bit is $a = 0$, the duplicated symbol is complemented with probability δ_0, and similarly, if $a = 1$ the duplicated bit is complemented with probability δ_1. The tandem-duplication system is motivated by tandem-duplication mutations, which are caused by polymerase slippage, also known as slipped-strand mispairing. During DNA replication, the polymerase responsible for constructing the new DNA strand may "slip," thereby creating an extra copy of a segment of the genome next to the original in the next generation [5], [25].

c) ***Interspersed duplication****:* Finally, for any real numbers $\delta_0, \delta_1 \in [0, 1]$, the interspersed-duplication system is defined as $S^{\text{int}}_{\delta_0,\delta_1} = (\Sigma, s, T^{\text{int}}_{\delta_0,\delta_1})$, where for all $w \in \Sigma^+$,

$$T^{\text{int}}_{\delta_0,\delta_1}(w) \triangleq u'bv'.$$

Here $u, v, u', v' \in \Sigma^*$, $a, b \in \Sigma$, $uav = w = u'v'$. The length $|ua|$ is chosen randomly independently and uniformly from $[|w|]$. Additionally, the length $|u'b|$ is also chosen randomly independently and uniformly from $[|w|+1]$. As for the inserted letter b, $\Pr(a = b|a = i) = 1 - \delta_i$. Intuitively, the $T^{\text{int}}_{\delta_0,\delta_1}$ rule chooses a uniformly random position in w, and duplicates the letter there to a uniformly chosen position. Like before, if the chosen bit is $a = 0$, the duplicated symbol is complemented with probability δ_0, and similarly, if $a = 1$ the duplicated bit is complemented with probability δ_1. This system is a simplified representation of biological interspersed duplications. These mutations result from mobile elements in the genome, called transposons or jumping genes, which can copy and paste themselves in different locations [22].

A step in each of the three Pólya string systems described above is depicted in Figure 1.

Given a Pólya string system S, the set of choices leading from $S(0)$ to $S(n)$ is denoted by $\mathcal{H}(n)$ and is referred to as the *history* of the sequence. More precisely, if $S(i)$ was obtained from $S(i-1)$ by taking the symbol in the jth position, passing it through the binary asymmetric channel to obtain the symbol b, and inserting it in the ℓth position, we encode this step using the tuple $\mathcal{H}_i \triangleq (j, b, \ell)$. The history is then the concatenation of steps, namely,

$$\mathcal{H}(n) \triangleq \mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_n.$$

The *entropy rate* of the process S is defined as

$$h(S) \triangleq \limsup_{n\to\infty} \frac{1}{n} H(S(n)),$$

where H is the entropy function,

$$H(S(n)) \triangleq -\sum_{w\in\Sigma^*} \Pr(S(n)=w)\log_2 \Pr(S(n)=w).$$

Since $H(S(n)|\mathcal{H}(n))=0$,

$$h(S)=\limsup_{n\to\infty}\frac{1}{n}I(S(n);\mathcal{H}(n)),$$

where I denotes mutual information. Thus $h(S)$ can be viewed as quantifying the amount of information that the sequence contains about its history. The problem of identifying the duplication history is of interest in the study of evolution and, in particular, has been investigated for tandem repeats [3], [10], [14], [36]. The mutual information $I(S(n);\mathcal{H}(n))$ can be used to find bounds on the error of identifying the duplication history through Fano's inequality [7].

Finally, we note that since S is a tuple $S=(\Sigma,s,T)$, its entropy rate, $h(S)$, depends not only on the channel parameters of T, but also on the choice of seed string s. This will become evident in the scenarios studied later in the paper.

III. End Duplication

We start our exploration of Pólya string systems with the end-duplication system. We distinguish between two cases that require different treatment. We first study the end-duplication system where the duplicated bit is unchanged (i.e., never complemented).

A. The Noiseless Channel: $\delta_0=\delta_1=0$

Theorem 3. *Let $\Sigma=\{0,1\}$, $s\in\Sigma^+$ be a seed string, and denote $t_0\triangleq|s|_0$, $t_1\triangleq|s|_1$. If $t_0,t_1\geqslant 1$, then the entropy rate of $S=S^{\text{end}}_{0,0}=(\Sigma,s,T^{\text{end}}_{0,0})$ is*

$$\begin{aligned}h(S^{\text{end}}_{0,0}) &= \int_0^1 \beta(p;t_0,t_1)H_2(p)\,\mathrm{d}p\\ &= \frac{\log_2 e}{t_0+t_1}\left((t_0+t_1)\mathsf{H}_{t_0+t_1}-t_0\mathsf{H}_{t_0}-t_1\mathsf{H}_{t_1}\right),\end{aligned}$$

where

$$\beta(p;t_0,t_1)\triangleq\frac{(t_0+t_1-1)!}{(t_0-1)!(t_1-1)!}p^{t_0-1}(1-p)^{t_1-1},$$

is the pdf for the $\mathrm{Beta}(t_0,t_1)$ *distribution, and where* H_m *denotes the mth harmonic number,*

$$\mathsf{H}_m\triangleq\sum_{i=1}^{m}\frac{1}{i}.$$

Proof: Fix any $w\in\Sigma^n$, and denote $k_0\triangleq|w|_0$, $k_1\triangleq|w|_1$, hence $k_0+k_1=n$. If we write $w=w_1\,w_2\ldots w_n$, with $w_i\in\Sigma$, we can now find that

$$\begin{aligned}&\Pr(S(n)=sw)\\ &=\prod_{i=1}^{n}\frac{|sw_1\ldots w_{i-1}|_{w_i}}{|sw_1\ldots w_{i-1}|}\\ &=\frac{t_0(t_0+1)\ldots(t_0+k_0-1)t_1(t_1+1)\ldots(t_1+k_1-1)}{(t_0+t_1)(t_0+t_1+1)\ldots(t_0+t_1+k_0+k_1-1)}\\ &=f(t_0,t_1,k_0,k_1)\\ &\triangleq\frac{(t_0+t_1-1)!(t_0+k_0-1)!(t_1+k_1-1)!}{(t_0-1)!(t_1-1)!(t_0+t_1+k_0+k_1-1)!}.\end{aligned}\tag{1}$$

We note that this probability does not depend on the order of bits in w. Thus, let us denote by A_{k_0} the event that $S(n)=sw$, and $|w|_0=k_0$. Obviously,

$$\Pr(A_{k_0})=\binom{n}{k_0}f(t_0,t_1,k_0,n-k_0).\tag{2}$$

We now have,

$$\begin{aligned}&h(S)\\ &=\limsup_{n\to\infty}\frac{1}{n}H(S(n))\\ &=-\limsup_{n\to\infty}\frac{1}{n}\sum_{w\in\Sigma^n}\Big(f(t_0,t_1,|w|_0,|w|_1)\\ &\qquad\qquad\cdot\log_2\ f(t_0,t_1,|w|_0,|w|_1)\Big)\\ &=-\limsup_{n\to\infty}\frac{1}{n}\sum_{k_0=0}^{n}\Pr(A_{k_0})\log_2\ f(t_0,t_1,k_0,n-k_0)\\ &=-\limsup_{n\to\infty}\frac{1}{n}\sum_{k_0=0}^{n}\Big(\Pr(A_{k_0})\\ &\qquad\qquad\cdot\log_2\frac{(t_0+k_0-1)!(t_1+n-k_0-1)!}{(t_0+t_1+n-1)!}\Big)\\ &\quad-\frac{1}{n}\log_2\frac{(t_0+t_1-1)!}{(t_0-1)!(t_1-1)!},\end{aligned}$$

and we note that the last term is $o(1)$. We also have,

$$\begin{aligned}&\frac{(t_0+k_0-1)!(t_1+n-k_0-1)!}{(t_0+t_1+n-1)!}\\ &\qquad=\frac{1}{t_0+t_1+n-1}\binom{t_0+t_1+n-2}{t_0+k_0-1}^{-1}.\end{aligned}$$

Thus,

$$h(S)=\limsup_{n\to\infty}\frac{1}{n}\sum_{k_0=0}^{n}\Pr(A_{k_0})\log_2\binom{t_0+t_1+n-2}{t_0+k_0-1}.$$

To obtain the desired form, we further simplify the expression for $h(S)$. First, by the well known approximation to the binomial coefficient (e.g., see [28]), we have

$$\binom{t_0+t_1+n-2}{t_0+k_0-1}=2^{n(H_2(k_0/n)+o(1))}.$$

Second, we note that for any fixed $\ell\in\mathbb{N}$,

$$\frac{(m+\ell)!}{m!}=m^{\ell}(1+O(1/m)).$$

Thus, we may rewrite (2) as

$$\begin{aligned}\Pr(A_{k_0}) &= \frac{(t_0+t_1-1)!}{(t_0-1)!(t_1-1)!}\cdot\frac{(k_0+t_0-1)!}{k_0!}\cdot\frac{(k_1+t_1-1)!}{k_1!}\\ &\quad\cdot\frac{n!}{(n+t_0+t_1-1)!}\\ &=\frac{(t_0+t_1-1)!}{(t_0-1)!(t_1-1)!}\cdot\frac{k_0^{t_0-1}k_1^{t_1-1}}{n^{t_0+t_1-1}}\\ &\quad\cdot(1+O(1/k_0+1/k_1+1/n))\\ &=\frac{1}{n}\cdot\beta\left(\frac{k_0}{n};t_0,t_1\right)\\ &\quad\cdot(1+O(1/k_0+1/(n-k_0)+1/n)).\end{aligned}$$

Putting this all together and using standard calculus techniques we obtain, for all $0 < \epsilon < \frac{1}{2}$,

$$\begin{aligned} &h(S) \\ &= \limsup_{n\to\infty} \frac{1}{n} \sum_{k_0=0}^{n} \Pr(A_{k_0}) \log_2 \binom{t_0+t_1+n-2}{t_0+k_0-1} \\ &= \limsup_{n\to\infty} \sum_{k_0=\epsilon n}^{(1-\epsilon)n} \frac{1}{n} \beta\left(\frac{k_0}{n}; t_0, t_1\right) H_2\left(\frac{k_0}{n}\right) (1+o(1)) + \phi(\epsilon) \\ &= \int_{\epsilon}^{1-\epsilon} \beta(p; t_0, t_1) H_2(p)\, \mathrm{d}p + \phi(\epsilon), \end{aligned}$$

where $0 \leqslant \phi(\epsilon) \leqslant 2\epsilon$. Taking the limit as $\epsilon \to 0^+$, we prove the first claim.

We continue to prove the second claim. Consider the following integral:

$$\int_0^1 p^{t_0+\epsilon}(1-p)^{t_1-1}\, \mathrm{d}p. \tag{3}$$

We use a Taylor series to obtain,

$$p^{\epsilon} = 2^{\epsilon \log_2 p} = \sum_{i=0}^{\infty} \frac{\epsilon^i (\ln 2)^i}{i!} (\log_2 p)^i.$$

Plugging this in (3) we get

$$\begin{aligned} &\int_0^1 p^{t_0+\epsilon}(1-p)^{t_1-1}\, \mathrm{d}p \\ &\quad = \sum_{i=0}^{\infty} \frac{\epsilon^i (\ln 2)^i}{i!} \int_0^1 p^{t_0}(1-p)^{t_1-1} (\log_2 p)^i\, \mathrm{d}p. \end{aligned} \tag{4}$$

We recall the definition of the gamma function (e.g., see [18, Ch. 11]),

$$\Gamma(x) \triangleq \int_0^{\infty} t^{x-1} e^{-t}\, \mathrm{d}t.$$

Additionally, $\Gamma(x+1) = x\Gamma(x)$, and in particular, for all $m \in \mathbb{N}$, $\Gamma(m+1) = m!$. We also recall the beta function,

$$B(x, y) \triangleq \int_0^1 t^{x-1}(1-t)^{y-1}\, \mathrm{d}t = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

for all $x, y \in \mathbb{R}$, $x, y > 0$. Thus, (3) becomes

$$\begin{aligned} &\int_0^1 p^{t_0+\epsilon}(1-p)^{t_1-1}\, \mathrm{d}p \\ &\quad = \frac{\Gamma(t_0+\epsilon+1)\Gamma(t_1)}{\Gamma(t_0+t_1+\epsilon+1)} \\ &\quad = \frac{(t_1-1)!}{(t_0+t_1+\epsilon)(t_0+t_1+\epsilon-1)\ldots(t_0+\epsilon+1)} \\ &\quad = \frac{(t_1-1)!}{(t_0+t_1)(1+\frac{\epsilon}{t_0+t_1})\ldots(t_0+1)(1+\frac{\epsilon}{t_0+1})} \\ &\quad = \frac{(t_1-1)!t_0!}{(t_0+t_1)!} \cdot \frac{1}{(1+\frac{\epsilon}{t_0+t_1})\ldots(1+\frac{\epsilon}{t_0+1})}. \end{aligned}$$

Using a Taylor series,

$$e^{\frac{\epsilon}{t_0+t_1}} = 1 + \frac{\epsilon}{t_0+t_1} + O(\epsilon^2).$$

Hence,

$$\begin{aligned} &\int_0^1 p^{t_0+\epsilon}(1-p)^{t_1-1}\, \mathrm{d}p \\ &\quad = \frac{(t_1-1)!t_0!}{(t_0+t_1)!} \cdot e^{-\left(\frac{1}{t_0+t_1}+\cdots+\frac{1}{t_0+1}\right)\epsilon} + O(\epsilon^2) \\ &\quad = \frac{(t_1-1)!t_0!}{(t_0+t_1)!} \cdot e^{-(\mathsf{H}_{t_0+t_1}-\mathsf{H}_{t_0})\epsilon} + O(\epsilon^2). \end{aligned}$$

Yet another Taylor series we get

$$e^{-(\mathsf{H}_{t_0+t_1}-\mathsf{H}_{t_0})\epsilon} = 1 - \left(\mathsf{H}_{t_0+t_1} - \mathsf{H}_{t_0}\right)\epsilon + O(\epsilon^2).$$

Plugging this back, we obtain

$$\begin{aligned} &\int_0^1 p^{t_0+\epsilon}(1-p)^{t_1-1}\, \mathrm{d}p \\ &\quad = \frac{(t_1-1)!t_0!}{(t_0+t_1)!}\left(1 - \left(\mathsf{H}_{t_0+t_1} - \mathsf{H}_{t_0}\right)\epsilon\right) + O(\epsilon^2). \end{aligned} \tag{5}$$

By equating the coefficient of ϵ^1 in (4) and (5) we get

$$\begin{aligned} &\frac{\ln 2}{1!} \int_0^1 p^{t_0}(1-p)^{t_1-1}(\log_2 p)\, \mathrm{d}p \\ &\quad = -\frac{(t_1-1)!t_0!}{(t_0+t_1)!}\left(\mathsf{H}_{t_0+t_1} - \mathsf{H}_{t_0}\right). \end{aligned}$$

We now repeat the same process, but instead of starting with (3), we take

$$\int_0^1 p^{t_0-1}(1-p)^{t_1+\epsilon}\, \mathrm{d}p,$$

and we get

$$\begin{aligned} &\frac{\ln 2}{1!} \int_0^1 p^{t_0-1}(1-p)^{t_1}(\log_2(1-p))\, \mathrm{d}p \\ &\quad = -\frac{(t_0-1)!t_1!}{(t_0+t_1)!}\left(\mathsf{H}_{t_0+t_1} - \mathsf{H}_{t_1}\right). \end{aligned}$$

Finally,

$$\begin{aligned} h(S) &= \int_0^1 \beta(p; t_0, t_1) H_2(p)\, \mathrm{d}p \\ &= \frac{(t_0+t_1-1)!}{(t_0-1)!(t_1-1)!}\left(\int_0^1 p^{t_0}(1-p)^{t_1-1}(\log_2 p)\, \mathrm{d}p \right. \\ &\qquad \left. + \int_0^1 p^{t_0-1}(1-p)^{t_1}(\log_2(1-p))\, \mathrm{d}p\right) \\ &= \frac{\log_2 e}{t_0+t_1}\left((t_0+t_1)\mathsf{H}_{t_0+t_1} - t_0\mathsf{H}_{t_0} - t_1\mathsf{H}_{t_1}\right), \end{aligned}$$

thus, proving the second claim as well. ∎

We comment that the case of either $t_0 = 0$ or $t_1 = 0$ in Theorem 3 is not interesting since then we have only strings of repeated symbols, and therefore, entropy rate of 0.

B. The Noisy Channel: $\delta_0 + \delta_1 > 0$

We move on to the case where the duplicated bit is passed through a noisy asymmetric binary channel. Calculating the entropy rate explicitly is not a simple task. This is due to the fact that in contrast to the previous case of $S^{\mathrm{end}}_{0,0}$, the probability of obtaining a specific sequence is not a function of the

frequency of symbols as in (1). This is demonstrated in the following example.

Example 4. *Consider* $S = S^{\text{end}}_{1,1}(\Sigma, s, T^{\text{end}}_{1,1})$, *with* $s = 01$. *Calculating the probability of the sequences* $S(3) = 01110$ *and* $S(3) = 01011$ *for we obtain*

$$\Pr(S^{\text{end}}_{1,1}(3) = 01110) = \frac{1}{2}\cdot\frac{1}{3}\cdot\frac{3}{4} \neq \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{2}{4} = \Pr(S^{\text{end}}_{1,1}(3) = 01011).$$

□

The following lemma will be instrumental in finding the entropy rate of $S^{\text{end}}_{\delta_0,\delta_1}$.

Lemma 5. *Let* $\Sigma = \{0,1\}$, $s \in \Sigma^+$ *be a seed string, and denote* $S = S^{\text{end}}_{\delta_0,\delta_1} = (\Sigma, s, T^{\text{end}}_{\delta_0,\delta_1})$. *If for any real* $\epsilon_1, \epsilon_2 > 0$, *there exists* $N \in \mathbb{N}$ *such that for all* $n \geqslant N$, $\Pr(|\text{fr}_0(S(n)) - \alpha| \leqslant \epsilon_1) \geqslant 1 - \epsilon_2$ *for some real* $\alpha \in [0,1]$, *then*

$$h(S^{\text{end}}_{\delta_0,\delta_1}) = H_2(\alpha(1-\delta_0) + (1-\alpha)\delta_1).$$

Proof: For our convenience, let $g : [0,1] \to [0,1]$ be defined as $g(x) \triangleq x(1-\delta_0) + (1-x)\delta_1$. Fix some real $\delta > 0$. Since $H_2(g(x))$ is continuous, by the Heine-Cantor Theorem $H_2(g(x))$ is uniformly continuous. Thus, there exists $\epsilon_1 > 0$ such that for all $x_1, x_2 \in [0,1]$, $|x_1 - x_2| \leqslant \epsilon_1$ implies

$$|H_2(g(x_1)) - H_2(g(x_2))| \leqslant \frac{1}{2}\delta.$$

We note that for $S = S^{\text{end}}_{\delta_0,\delta_1}$, and all $w \in \Sigma^{n+|s|}$, we have

$$\Pr(S(n+1) = w0 \mid S(n) = w) = g(\text{fr}_0(w)).$$

Additionally, by the theorem requirements we are assured we can find $N \in \mathbb{N}$ such that for all $n \geqslant N$ we have

$$\Pr(|\text{fr}_0(S(n)) - \alpha| \leqslant \epsilon_1) \geqslant 1 - \frac{1}{2}\delta. \tag{6}$$

For the rest of the proof, we consider the underlying sample space to be the space of all infinite sequences,

$$\Sigma^{\mathbb{N}} \triangleq \{a_1\ a_2\ a_3 \ldots \ : \ \forall i \in \mathbb{N}, a_i \in \Sigma\}.$$

A distribution μ on $\Sigma^{\mathbb{N}}$ is induced by evolving from the seed s according to S. Thus, $S(n)$ is a random variable taking values from $\Sigma^{|s|+n}$, whose distribution is the marginal of μ on the first $|s|+n$ coordinates (sometimes called the $(|s|+n)$-length cylinder). Namely, the event $S(n) = w$ is the set

$$\left\{v \in \Sigma^{\mathbb{N}} \ : \ v_i = w_i \text{ for all } i \in [n + |s|]\right\}.$$

Similarly, we define S_i to be the projection of μ on the $(|s|+i)$th coordinate, i.e., the event $S_i = a$ is the set

$$\left\{v \in \Sigma^{\mathbb{N}} \ : \ v_{i+|s|} = a\right\}.$$

Let us define the event,

$$F \triangleq \left\{v \in \Sigma^{\mathbb{N}} \ : \ \forall n \geqslant N, |\text{fr}_0(v_1 \ldots v_n) - \alpha| \leqslant \epsilon_1\right\},$$

and denote by F^c its complement. We obtain that,

$$\begin{aligned}
h(S) &= \limsup_{n\to\infty} \frac{1}{n} H(S(n)) \\
&\leqslant \limsup_{n\to\infty} \frac{1}{n}\left(H(S(n) \mid F) + \frac{1}{2}\delta H(S(n) \mid F^c)\right) \\
&\leqslant \limsup_{n\to\infty} \frac{1}{n} H(S(n) \mid F) + \frac{1}{2}\delta \\
&\overset{(a)}{=} \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H(S_i \mid S(i-1), F) + \frac{1}{2}\delta \\
&\overset{(b)}{\leqslant} \limsup_{n\to\infty} \frac{1}{n}\Bigg(\sum_{i=1}^{N} H(S_i) \\
&\qquad + \sum_{i=N+1}^{n} H(S_i \mid S(i-1), F)\Bigg) + \frac{1}{2}\delta \\
&\overset{(c)}{\leqslant} \limsup_{n\to\infty} \frac{1}{n}\left(N + (n-N)\left(H_2(g(\alpha)) + \frac{\delta}{2}\right)\right) + \frac{1}{2}\delta \\
&= H_2(g(\alpha)) + \delta
\end{aligned}$$

where (a) follows from the chain rule for entropy, (b) follows since conditioning reduces entropy, and (c) follows since

$$H(S_i \mid S(i-1), F) = H_2(g(\text{fr}_0(S(i-1))))$$

and from (6).

Using similar reasoning,

$$\begin{aligned}
h(S) &= \limsup_{n\to\infty} \frac{1}{n} H(S(n)) \\
&\geqslant \limsup_{n\to\infty} \frac{1}{n}\left(1 - \frac{1}{2}\delta\right) H(S(n) \mid F) \\
&= \limsup_{n\to\infty} \frac{1}{n}\left(1 - \frac{1}{2}\delta\right)\sum_{i=1}^{n} H(S_i \mid S(i-1), F) \\
&\geqslant \limsup_{n\to\infty} \frac{1}{n}\left(1 - \frac{1}{2}\delta\right)\sum_{i=N+1}^{n} H(S_i \mid S(i-1), F) \\
&\geqslant \limsup_{n\to\infty} \frac{1}{n}\left(1 - \frac{1}{2}\delta\right)(n-N)(H_2(g(\alpha)) - \frac{\delta}{2}) \\
&= (1 - \frac{1}{2}\delta)(H_2(g(\alpha)) - \frac{\delta}{2}) \\
&\geqslant H_2(g(\alpha)) - \delta,
\end{aligned}$$

where the last inequality follows from the fact that $\frac{1}{2}\delta(H(\alpha) - \frac{\delta}{2}) \leqslant \frac{1}{2}\delta$.

We now have

$$H_2(g(\alpha)) - \delta \leqslant h(S) \leqslant H_2(g(\alpha)) + \delta.$$

Taking the limit as $\delta \to 0^+$ gives the claimed result. ■

The next step in finding the entropy rate of $S^{\text{end}}_{\delta_0,\delta_1}$ is to find the (almost sure) limit of the frequency of symbols. We make use of the following definition.

Definition 6. *Let* $(x_n)_{n\in\mathbb{N}}$ *be a sequence of real numbers, evolving according to the equation* $x_{n+1} = x_n + a \cdot f(x_n)$ *for some function* $f : \mathbb{R} \to \mathbb{R}$ *and a constant* $a \in \mathbb{R}$, $a > 0$. *We say that* x' *is an* equilibrium point *of the recursion* $x_{n+1} = x_n + a \cdot f(x_n)$ *if* $f(x') = 0$.

We prove the next lemma using stochastic approximation (for a comprehensive study see [4]).

Lemma 7. *Let $\Sigma = \{0, 1\}$, $s \in \Sigma^+$ be a seed string, and denote $S = S^{\text{end}}_{\delta_0,\delta_1} = (\Sigma, s, T^{\text{end}}_{\delta_0,\delta_1})$, where $\delta_0 + \delta_1 > 0$. Then*

$$\lim_{n\to\infty} \text{fr}_0\left(S(n)\right) = \frac{\delta_1}{\delta_0 + \delta_1}$$

almost surely.

Proof: Let $t_0 \triangleq |s|_0$ and $t_1 \triangleq |s|_1$. We further define

$$x_n \triangleq |S(n)|_0\,, \qquad z_n \triangleq \text{fr}_0(S(n)) = \frac{x_n + t_0}{n + t_0 + t_1}.$$

Let $g : [0, 1] \to [0, 1]$ be defined as

$$g(x) \triangleq x(1 - \delta_0) + (1 - x)\delta_1.$$

Note that for any $w \in \Sigma^{n+|s|}$,

$$\Pr\left(S(n + 1) = w0 \mid S(n) = w\right) = g(z_n),$$

and that $z_0 = \frac{t_0}{t_0+t_1}$. We write

$$x_{n+1} = x_n + \xi_{n+1}$$

where $\xi_{n+1} = 1$ if the $(n + 1)$st appended symbol (due to mutation) is a 0, and $\xi_{n+1} = 0$ otherwise. A simple calculation yields

$$\begin{aligned} z_{n+1} &= z_n + \frac{1}{n + 1 + t_0 + t_1}(\xi_{n+1} - z_n) \\ &= z_n + \frac{(g(z_n) - z_n) + (\xi_{n+1} - g(z_n))}{n + 1 + t_0 + t_1}. \end{aligned}$$

The main goal is to find the limit points of the sequence z_n.

Let $M_n \triangleq \xi_n - g(z_{n-1})$, and note that M_n is a martingale difference sequence. Indeed, if $\mathcal{F}_n$ is the σ-algebra generated by $\sigma(z_m, M_m,\ m \leqslant n)$ then

$$\begin{aligned} E\left[M_{n+1} \mid \mathcal{F}_n\right] &= E\left[\xi_{n+1} \mid \mathcal{F}_n\right] - g(z_n) \\ &= g(z_n) - g(z_n) \\ &= 0. \end{aligned}$$

Hence, the limiting differential equation z_n is expected to track is given by

$$\dot{z}_t = g(z_t) - z_t. \tag{7}$$

In order for the differential equation to have a unique solution for any z_0, we need to show that $g(z) - z$ is Lipschitz [4, Ch. 11, Theorem 5]. Indeed,

$$|(g(z) - z) - (g(y) - y)| = |(\delta_0 + \delta_1)(z - y)|\,,$$

which means that $g(z) - z$ is $(\delta_0 + \delta_1)$-Lipschitz. Solving the differential equation we obtain the solution

$$z_t = \frac{\delta_1}{\delta_0 + \delta_1} + \left(\frac{t_0}{t_0 + t_1} - \frac{\delta_1}{\delta_0 + \delta_1}\right) e^{-t(\delta_0+\delta_1)}.$$

From the solution of the differential equation, it is clear that the set $[0, 1]$ is an invariant set (any trajectory starting at $[0, 1]$ and evolves according to z_t will remain in the set). Also, we see that the point $z^* \triangleq \frac{\delta_1}{\delta_0+\delta_1}$ is an equilibrium point and since $g(z)$ is contraction (i.e., $|g(z_1) - g(z_2)| \leqslant |z_1 - z_2|$) it has only one equilibrium point (this is due to the Banach fixed-point theorem [2]). Hence, using [4, Corollary 4][1], z_n converges almost surely to z^*. ∎

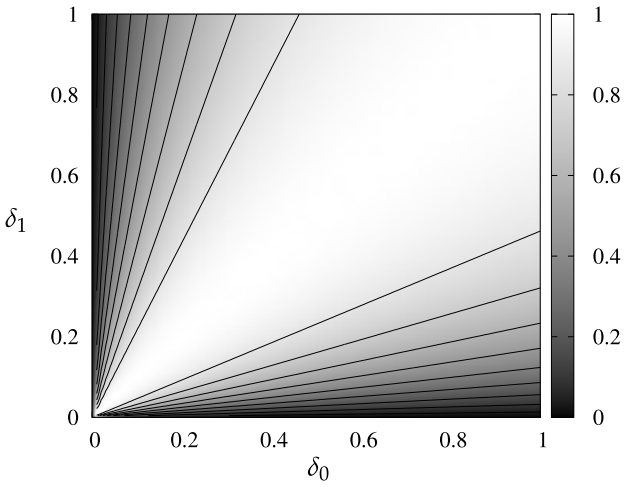

Fig. 2. A contour plot of $h(S^{\text{end}}_{\delta_0,\delta_1})$.

We remark that for $\delta_0 = \delta_1 = 0$, we obtain in (7) that $\dot{z}_t = 0$, which means that there is no singular attraction point (there is no stable equilibrium point). Hence, in order to use the same method, we need to evaluate the probability of every possible limiting point. This, as we know from the formula for $h(S^{\text{end}}_{0,0})$ from Theorem 3, is a function of the seed string, and is related to the beta distribution.

We can now state the entropy rate for $S^{\text{end}}_{\delta_0,\delta_1}$ with $\delta_0+\delta_1 > 0$.

Theorem 8. *Let $\Sigma = \{0, 1\}$, $s \in \Sigma^+$ be a seed string, and denote $S = S^{\text{end}}_{\delta_0,\delta_1} = (\Sigma, s, T^{\text{end}}_{\delta_0,\delta_1})$, where $\delta_0 + \delta_1 > 0$. Then*

$$h\left(S^{\text{end}}_{\delta_0,\delta_1}\right) = H_2\left(\frac{\delta_1}{\delta_0 + \delta_1}\right) = H_2\left(\frac{\delta_0}{\delta_0 + \delta_1}\right).$$

Proof: By Lemma 7 we obtain the limiting frequencies of $S(n)$. Then, by using Lemma 5 we obtain the desired result. ∎

Figure 2 shows a contour plot of the entropy rate of $S^{\text{end}}_{\delta_0,\delta_1}$.

IV. Tandem Duplication

We turn our attention in this section to tandem-duplication Pólya string models. We again consider several cases separately, depending on the parameters of the binary asymmetric channel, δ_0 and δ_1. We find the exact entropy rate of $S^{\text{tan}}_{0,0}$, and relate the entropy rate of $S^{\text{tan}}_{1,1}$ to a combinatorial property of permutations. Finally, we upper bound the entropy rate of the general $S^{\text{tan}}_{\delta_0,\delta_1}$.

A. The Noiseless Channel: $\delta_0 = \delta_1 = 0$

The entropy rate of the noiseless case is simple.

Theorem 9. *Let $\Sigma = \{0, 1\}$, $s \in \Sigma^+$ be a seed string, and denote $S = S^{\text{tan}}_{0,0} = (\Sigma, s, T^{\text{tan}}_{0,0})$. Then*

$$h(S^{\text{tan}}_{0,0}) = 0.$$

[1] Note that [4, Corollary 4] uses the notion of internally chain transitive. In our case, since z^* is a unique equilibrium point we obtain that the singleton $\{z^*\}$ is the internally chain transitive set in $[0, 1]$.

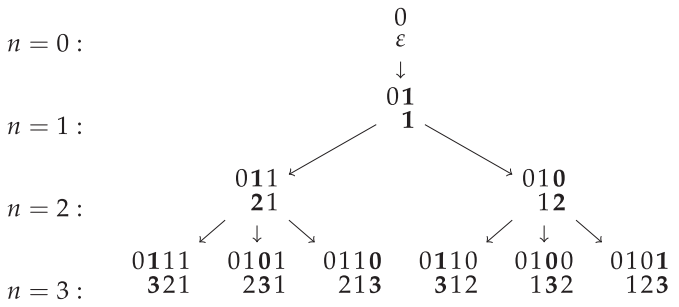

Fig. 3. The tree of sequences that can be obtained starting from $s = 0$ using the $T^{\tan}_{1,1}$ rule for $n \leqslant 3$. The first line in each node is the sequence and the second line is its history permutation.

Proof: A crude counting argument suffices for the proof. Consider the initial string $S(0)$, and denote the number of runs in it by r. Obviously any tandem-duplication operation extends existing runs and never creates new runs. Thus, obtaining $S(n)$ may be viewed as an action of throwing n balls into r bins. The total number of resulting strings (regardless of probability) is given exactly by $\binom{n+r-1}{r-1} \leqslant (n+r-1)^{r-1}$. Maximum entropy will be attained by a uniform distribution over those strings, and even in that case we get

$$h(S^{\tan}) \leqslant \limsup_{n\to\infty} \frac{1}{n} \log_2 (n+r-1)^{r-1} = 0.$$

A lower bound of 0 is trivial since we have at least one string for each length $n \geqslant |s|$. ∎

B. The Complementing Channel: $\delta_0 = \delta_1 = 1$

Next, we consider $S^{\tan}_{1,1}$, where the duplicated bit is always complemented. For simplicity, in what follows we assume that the seed string is $S(0) = s = 0$. We note then that $S(1) = 01$ always. As an example, a possible history leading to $S(3) = 0110$ is

$$0 \to 0\mathbf{1} \to 01\mathbf{0} \to 0\mathbf{1}10, \tag{8}$$

where in each step the new symbol is in bold.

The history of $S(n)$ can be encoded as a permutation of length n, called its *history permutation*, as follows: Replace each 0 or 1 with the number of the turn in which they were added to the sequence. For example, the history given in (8) corresponds to the history permutation 312:

$$\begin{aligned} 0 &\to 0\mathbf{1} \to 01\mathbf{0} \to 0\mathbf{1}10, \\ \varepsilon &\to \ \mathbf{1} \to \ 1\mathbf{2} \to \ \mathbf{3}12. \end{aligned}$$

Note that since 0 is always in the starting position, we drop it to obtain a permutation of $[n]$. It is clear that this provides us with a bijection between permutations of $[n]$ and a history resulting in a sequence $S(n) = 01w$, $w \in \{0,1\}^{n-1}$. This bijection will be useful in what follows.

The tree in Fig. 3 illustrates the history permutations and the sequences arising from them for $n \leqslant 3$. Since all histories are equally likely, all leaves at the same level in the tree are equally likely. Note however that not all sequences are equally likely as multiple histories may lead to the same sequence. For example, from Fig. 3, it is clear that $\Pr(S(3) = 0101) = 2 \cdot \Pr(S(3) = 0100)$.

The following definitions will be useful. For $n \in \mathbb{N}$ let $\mathbb{S}_n$ denote the symmetric group of permutations over $[n]$. Recall that the ith letter of $S(n)$, for $i \in [n+1]$, is denoted by $S_i(n)$. Furthermore, if $w \in \Sigma^*$, and $1 \leqslant i \leqslant j \leqslant |w|$, then we denote $w_i^j \triangleq w_i w_{i+1} \ldots w_j$. For $S(n)$, this notation becomes $S_i^j(n)$.

For a permutation $\pi \in \mathbb{S}_n$, define its signature $\mathrm{sig}(\pi) = u \in \{0,1\}^{n-1}$ such that

$$u_i \triangleq \begin{cases} 0, & \text{if } \pi_i > \pi_{i+1}, \\ 1, & \text{if } \pi_i < \pi_{i+1}, \end{cases}$$

for $i \in [n-1]$, i.e., ascents are marked by 1 and descents by 0. We also define, for each $u \in \{0,1\}^{n-1}$,

$$\Psi_u \triangleq \{\pi \in \mathbb{S}_n \ : \ \mathrm{sig}(\pi) = u\}.$$

The following lemma is useful in computing the entropy rate of the system.

Lemma 10. *Let $\Sigma = \{0,1\}$, and denote $S = S^{\tan}_{1,1} = (\Sigma, s = 0, T^{\tan}_{1,1})$. Then for all $u \in \Sigma^{n-1}$,*

$$\Pr(S(n) = 01u) = \frac{|\Psi_u|}{n!}.$$

Proof: Let the set of history permutations in S that lead to $01w$ be denoted by Π_{01w}. For technical reasons, we will need to consider also $S' = (\Sigma, s = 1, T^{\tan}_{1,1})$ (which differs from S by starting with the seed string 1 instead of 0). Obviously S and S' are isomorphic, by simply complementing all bits. Similarly, we denote the set of history permutations in S' that lead to $10w$ by Π_{10w}.

To prove the claim, it suffices to show that for all $w \in \Sigma^*$,

$$|\Pi_{01w}| = |\Pi_{10w}| = |\Psi_w|. \tag{9}$$

We show this by proving that the sizes of all sets satisfy the same recursion with the same initial values. The initial conditions for all recursions are

$$|\Pi_{01\varepsilon}| = |\Pi_{10\varepsilon}| = |\Psi_\varepsilon| = 1,$$

where ε is the empty string.

We start by providing two recursions for $|\Psi_w|$. For $v \in \Sigma^n$, let

$$\begin{aligned} T_v \triangleq \{i \in [n+1] \ : \ &(v_{i-1} = 1 \text{ or } i = 1) \\ &\text{and } (v_i = 0 \text{ or } i = n+1)\}, \\ U_v \triangleq \{i \in [n+1] \ : \ &(v_{i-1} = 0 \text{ or } i = 1) \\ &\text{and } (v_i = 1 \text{ or } i = n+1)\}, \end{aligned}$$

be the set of positions where 1 to 0 and 0 to 1 transitions occur (except at the boundaries). For example for $v = 0011010$, we have $T_v = \{1,5,7\}$ and $U_v = \{3,6,8\}$.

For $u \in \Sigma^n$, we can construct a permutation of $[n+1]$ with the signature u recursively by first determining the position of $n+1$. The set of valid positions for $n+1$ is precisely the set T_u. Suppose we place $n+1$ in position $i \in T_u$. We now need to construct two permutations with signatures u_1^{i-2} and u_{i+1}^n, each with a subset of $[n]$. We can choose the set of

elements for each of these two permutations in $\binom{n}{i-1}$ ways. Hence,

$$|\Psi_u| = \sum_{i \in T_u} \binom{n}{i-1} \left|\Psi_{u_1^{i-2}}\right| \left|\Psi_{u_{i+1}^n}\right|.$$

Similarly, by deciding where to place 1 (instead of $n+1$), we can show that

$$|\Psi_u| = \sum_{i \in U_u} \binom{n}{i-1} \left|\Psi_{u_1^{i-2}}\right| \left|\Psi_{u_{i+1}^n}\right|.$$

We now return to Π_{01u} and Π_{10u}. Note that (9) holds trivially if u is the empty string. Suppose (9) holds for all $u \in \Sigma^{n-1}$. Fix $u \in \Sigma^n$ and consider the sequence $01u$ as the result of the Pólya string model. In the permutations in Π_{01u}, the set of valid positions for 1 is precisely the set of positions in T_u. To see this note that in a permutation describing the history of $01u$, the element 1 can only correspond to the last element in a run of 1s in the string $01u$. Specifically, the element 1 can be placed in position 1 iff u starts with a 0 (since the bold 1 in $0\mathbf{1}u$ is the last 1 in a run); 1 can be placed in position $2 \leqslant i \leqslant n$ iff $u_{i-1}u_i = 10$; and finally, 1 can be placed in position $n+1$ iff $u_n = 1$ (again, the last 1 in a run of 1s).

Hence, we can construct these permutations recursively by first determining the position of 1 in them, and

$$\begin{aligned}|\Pi_{01u}| &= \sum_{i \in T_u} \binom{n}{i-1} \left|\Pi_{01u_1^{i-2}}\right| \left|\Pi_{10u_{i+1}^n}\right| \\ &= \sum_{i \in T_u} \binom{n}{i-1} \left|\Psi_{u_1^{i-2}}\right| \left|\Psi_{u_{i+1}^n}\right|.\end{aligned}$$

Similarly, for Π_{10u}, $u \in \Sigma^n$, the possible positions for 1 are precisely those in U_u as now 1 in the history permutation should correspond to the last 0 in a run of 0s in the string $10u$. So 1 can be placed in position 1 iff u starts with a 1; it can be placed in position $2 \leqslant i \leqslant n$ iff $u_{i-1}u_i = 01$; and finally it can be placed in position $n+1$ if $s_u = 0$. We thus have

$$\begin{aligned}|\Pi_{10u}| &= \sum_{i \in T_u} \binom{n}{i-1} \left|\Pi_{10u_1^{i-2}}\right| \left|\Pi_{01u_{i+1}^n}\right| \\ &= \sum_{i \in T_u} \binom{n}{i-1} \left|\Psi_{u_1^{i-2}}\right| \left|\Psi_{u_{i+1}^n}\right|.\end{aligned}$$

This completes the proof of (9) for all $u \in \Sigma^*$. ∎

As Lemma 10 shows, in order to find $h(S_{1,1}^{\tan})$ with seed string $s = 0$, we need to find the asymptotics of the probability that a uniformly chosen permutation from $\mathbb{S}_n$ has a given signature, as $n \to \infty$. We do not yet know how to attain this goal, and instead, use simplified versions of it to obtain bounds on the aforementioned entropy rate.

Theorem 11. *Let $\Sigma = \{0,1\}$, and denote $S = S_{1,1}^{\tan} = (\Sigma, s = 0, T_{1,1}^{\tan})$. Then,*

$$\frac{5\log_2 e - 2}{6} \leqslant h(S_{1,1}^{\tan}) \leqslant H_2\left(\frac{1}{3}\right).$$

Proof: Define the process $\overline{S}$ as follows. Suppose we uniformly and independently choose random reals in $[0,1]$ denoted by $X_1, X_2, \ldots$. We note that for any $i \neq j$, $\Pr[X_i = X_j] = 0$, and so with probability 1 the sequence $X_1, \ldots, X_n$ induces a uniformly chosen permutation from $\mathbb{S}_n$. Let

$$\overline{S}_i = \begin{cases} 1, & \text{if } X_i < X_{i+1} \\ 0, & \text{if } X_i > X_{i+1} \end{cases} \tag{10}$$

for $i \in \mathbb{N}$. Thus, $\overline{S}_1 \ldots \overline{S}_{n-1}$ form the signature of a uniformly chosen permutation from $\mathbb{S}_n$. It follows from Lemma 10 that for any n and $u \in \Sigma^{n-1}$, we have

$$\Pr(S(n) = 01u) = \Pr(\overline{S}_1^{n-1} = u).$$

Note that the strings in S evolve by changing at a random position, but $\overline{S}$ can be viewed as evolving by changing at the end, and thus is easier to analyze.

We now have,

$$\begin{aligned}h(S) &= \limsup_{n\to\infty} \frac{1}{n} H(S(n)) = \limsup_{n\to\infty} \frac{1}{n} H\left(\overline{S}_1^{n-1}\right) \\ &= \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n-1} H\left(\overline{S}_i | \overline{S}_1^{i-1}\right)\end{aligned} \tag{11}$$

Before proceeding with the proof, we show a simpler lower bound than the one given in the theorem. For $i \in \mathbb{N}$, since $\overline{S}_1^{i-1} \to X_i \to S_i$, i.e., they form a Markov chain, we have $H(\overline{S}_i | \overline{S}_1^{i-1}) \geqslant H(\overline{S}_i | X_i)$. Furthermore, $\Pr(\overline{S}_i = 0 | X_i = x) = x$. Thus from (11) we find

$$h(S) \geqslant H\left(\overline{S}_i | X_i\right) = \int_0^1 H_2(x)\,\mathrm{d}x = \frac{\log e}{2} \geqslant 0.7213.$$

With the same approach we can prove the stronger lower bound in the theorem. Note that $\overline{S}_1^{i-2} \to X_{i-1} \to \bar{S}_{i-1}^{i}$. So

$$\begin{aligned}H(\overline{S}_i | \overline{S}_1^{i-1}) &\geqslant H(\overline{S}_i | \overline{S}_{i-1}, X_{i-1}) \\ &= \int_0^1 x h_0(x)\,\mathrm{d}x + \int_0^1 (1-x)\, h_1(x)\,\mathrm{d}x,\end{aligned}$$

where

$$\begin{aligned}h_0(x) &= H\left(\overline{S}_i | \overline{S}_{i-1} = 0, X_{i-1} = x\right), \\ h_1(x) &= H\left(\overline{S}_i | \overline{S}_{i-1} = 1, X_{i-1} = x\right).\end{aligned}$$

We have

$$\begin{aligned}h_0(x) &= H_2\left(\frac{1}{x}\int_0^x y\,\mathrm{d}y\right) = H_2\left(\frac{x}{2}\right), \\ h_1(x) &= H_2\left(\frac{1}{1-x}\int_x^1 (1-y)\,\mathrm{d}y\right) = H_2\left(\frac{1-x}{2}\right).\end{aligned}$$

Hence,

$$\begin{aligned}H(\overline{S}_i | \overline{S}_1^{i-1}) &= \int_0^1 x H_2\left(\frac{x}{2}\right)\mathrm{d}x + \\ &\int_0^1 (1-x)\, H_2\left(\frac{1-x}{2}\right)\mathrm{d}x = \frac{5\log e - 2}{6} \geqslant 0.8689.\end{aligned}$$

Now we turn to proving the upper bound. Note that

$$\begin{aligned} h(S) &= \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n-1} H\left(\overline{S}_i|\overline{S}_1^{i-1}\right) \\ &\leqslant \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n-1} H\left(\overline{S}_i|\overline{S}_{i-1}\right) \\ &= H(\overline{S}_2|\overline{S}_1) \\ &= \frac{1}{2}\left(H(\overline{S}_2|\overline{S}_1=0)+H(\overline{S}_2|\overline{S}_1=1)\right) \\ &= \frac{1}{2}\cdot 2\cdot H_2\left(\frac{1}{3}\right) \leqslant 0.9183, \end{aligned}$$

since by integrating over the values of X_1^3, we find

$$\Pr\left(\overline{S}_2=0|\overline{S}_1=0\right) = \frac{\int_0^1 \mathrm{d}x_1 \int_0^{x_1}\mathrm{d}x_2\int_0^{x_2}\mathrm{d}x_3}{\int_0^1 \mathrm{d}x_1\int_0^{x_1}\mathrm{d}x_2} = \frac{1/6}{1/2} = \frac{1}{3}$$

as well as $\Pr\left(\overline{S}_2=1|\overline{S}_1=1\right)=\frac{1}{3}$. ∎

Both methods used in the proof of the preceding theorem can be extended to obtain better bounds, at the cost of more tedious proofs. For example, for the upper bound we can have

$$h(S_{1,1}^{\tan}) \leqslant H(\overline{S}_4|\overline{S}_2,\overline{S}_3)$$

Let $P_{ijk}=\Pr(\overline{S}_2=i,\overline{S}_3=j,\overline{S}_4=k)$. By integration, we find

$$(P_{000},P_{001},\ldots,P_{111}) = \frac{1}{24}(1,3,5,3,3,5,3,1).$$

Hence

$$H(\overline{S}_4|\overline{S}_2=0,\overline{S}_3=0) = H(\overline{S}_4|\overline{S}_2=1,\overline{S}_3=1) = H_2\left(\frac{2}{8}\right),$$
$$H(\overline{S}_4|\overline{S}_2=0,\overline{S}_3=1) = H(\overline{S}_4|\overline{S}_2=1,\overline{S}_3=0) = H_2\left(\frac{3}{8}\right).$$

So

$$h(S_{1,1}^{\tan}) \leqslant 2\cdot\frac{1}{6}H_2\left(\frac{2}{8}\right)+2\cdot\frac{1}{3}H_2\left(\frac{3}{8}\right) \leqslant 0.9067.$$

C. The Noisy Channel: $\delta_0+\delta_1>0$

Lastly, we address the general noisy case of $S_{\delta_0,\delta_1}^{\tan}$, with $\delta_0+\delta_1>0$. The methods used for finding the entropy rate of $S_{\delta_0,\delta_1}^{\mathrm{end}}$ need to be extended: instead of studying the frequencies of letters, we shall study the frequencies of pairs of adjacent letters. To that end, we need to extend some definitions.

Let $w\in\Sigma^n$, $n\in\mathbb{N}$, and let $u\in\Sigma^k$, $k\in\mathbb{N}$, where $k\leqslant n$. The number of occurrences of u in w as a substring is denoted by $|w|_u$, formally defined as

$$|w|_u \triangleq \left|\left\{i\in[n]\ :\ w_i^{i+n-1}=u\right\}\right|,$$

where indices are taken cyclically, i.e., w_n is followed by w_1. We also extend the definition of frequency,

$$\mathrm{fr}_u(w) \triangleq \frac{|w|_u}{|w|}.$$

Lemma 12. *Let $\Sigma=\{0,1\}$, $s\in\Sigma^+$ a seed string, and denote $S=S_{\delta_0,\delta_1}^{\tan}=(\Sigma,s,T_{\delta_0,\delta_1}^{\tan})$, where $\delta_0+\delta_1>0$. Then*

$$\lim_{n\to\infty}\begin{pmatrix}\mathrm{fr}_{00}(S(n))\\ \mathrm{fr}_{01}(S(n))\\ \mathrm{fr}_{10}(S(n))\\ \mathrm{fr}_{11}(S(n))\end{pmatrix} = \frac{1}{(1+\delta_0+\delta_1)(\delta_0+\delta_1)}\begin{pmatrix}(1-\delta_0+\delta_1)\delta_1\\ 2\delta_0\delta_1\\ 2\delta_0\delta_1\\ (1-\delta_1+\delta_0)\delta_0\end{pmatrix},$$

almost surely.

Proof: To avoid cumbersome notation, let us denote

$$x_n^u \triangleq |S(n)|_u, \qquad z_n \triangleq \begin{pmatrix}\mathrm{fr}_{00}(S(n))\\ \mathrm{fr}_{01}(S(n))\\ \mathrm{fr}_{10}(S(n))\\ \mathrm{fr}_{11}(S(n))\end{pmatrix}.$$

Let $\mathcal{F}_n$ be the filtration generated by z_n.

We first find the expected change in the multiplicities x_{n+1}^u for $u\in\{00,01,10,11\}$. To do so, we need to find the number of new occurrences of 00 and the number of occurrences that are eliminated by a mutation. First, we consider $u=00$. A new occurrence of u appears if 0 is duplicated or if the 1 in an occurrence of 10 is complement-duplicated (i.e., resulting in 100). An occurrence of 00 is eliminated if its first 0 is complement-duplicated. Thus

$$\begin{aligned} E[x_{n+1}^{00}-x_n^{00}|\mathcal{F}_n] &= z_n^0(1-\delta_0)+z_n^{10}\delta_1-z_n^{00}\delta_0 \\ &= z_n^{00}(1-2\delta_0)+z_n^{01}(1-\delta_0)+z_n^{10}\delta_1. \end{aligned}$$

Similarly, we have

$$\begin{aligned} E\left[x_{n+1}^{01}-x_n^{01}|\mathcal{F}_n\right] &= z_n^{00}\delta_0+z_n^{11}\delta_1, \\ E\left[x_{n+1}^{10}-x_n^{10}|\mathcal{F}_n\right] &= z_n^{00}\delta_0+z_n^{11}\delta_1, \\ E\left[x_{n+1}^{11}-x_n^{11}|\mathcal{F}_n\right] &= z_n^{01}\delta_0+z_n^{10}(1-\delta_1)+z_n^{11}(1-2\delta_1). \end{aligned}$$

By stacking these equations, we find A' such that

$$E\left[x_{n+1}-x_n|\mathcal{F}_n\right] = A'z_n.$$

By letting $A\triangleq A'-I$, we find

$$A=\begin{pmatrix}-2\delta_0 & 1-\delta_0 & \delta_1 & 0\\ \delta_0 & -1 & 0 & \delta_1\\ \delta_0 & 0 & -1 & \delta_1\\ 0 & \delta_0 & 1-\delta_1 & -2\delta_1\end{pmatrix}.$$

Using stochastic approximation, We can relate the behavior of z_n to the ODE $\dot{z}_t=Az_t$ (see [4]). In particular, z_n converges almost surely to the null space of A. From this, the theorem follows. ∎

The entropy rate of a source of strings whose limiting substring frequencies are known, was studied in [27], and an upper bound provided. We use this result to upper bound the entropy rate.

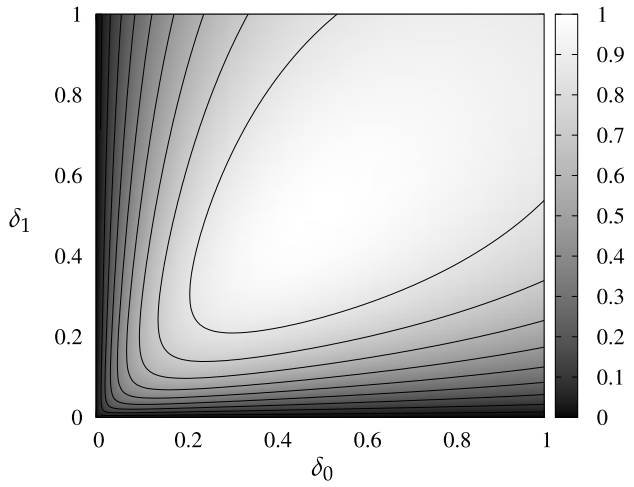

Fig. 4. A contour plot of the upper bound on $h(S^{\tan}_{\delta_0,\delta_1})$ of Theorem 13.

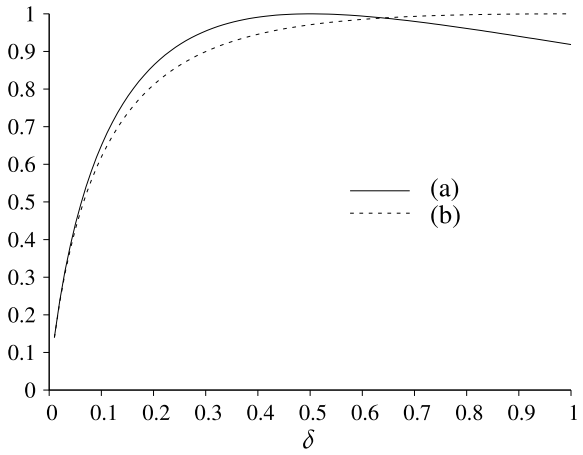

Fig. 5. (a) An upper bound on $h(S^{\tan}_{\delta,\delta})$, and (b) an upper bound on $h(S^{\text{tsb}}_{\delta})$.

Theorem 13. *Let $\Sigma = \{0, 1\}$, $s \in \Sigma^+$ a seed string, and denote $S = S^{\tan}_{\delta_0,\delta_1} = (\Sigma, s, T^{\tan}_{\delta_0,\delta_1})$, where $\delta_0 + \delta_1 > 0$. Then*

$$h(S^{\tan}_{\delta_0,\delta_1}) \leqslant \frac{\delta_1}{\delta_0+\delta_1} H_2\left(\frac{1-\delta_0+\delta_1}{1+\delta_0+\delta_1}\right) + \frac{\delta_0}{\delta_0+\delta_1} H_2\left(\frac{1-\delta_1+\delta_0}{1+\delta_0+\delta_1}\right).$$

Proof: Let $z_\infty \triangleq (z^{00}_\infty, z^{01}_\infty, z^{10}_\infty, z^{11}_\infty)^T$ be the limit given by Lemma 12. From [27], the entropy rate is upper bounded above by

$$h(S) \leqslant -\sum_{u_1u_2} z^{u_1u_2}_\infty \log \frac{z^{u_1u_2}_\infty}{z^{u_1u_2}_\infty + z^{u_1\bar{u}_2}_\infty},$$

where $u_1, u_2 \in \{0, 1\}$ and $\bar{u}_i = 1 - u_i$. From this, by substituting the expression for z_∞ given in Lemma 12, the claim follows. ∎

The upper bound on the entropy rate of $S^{\tan}_{\delta_0,\delta_1}$ is shown in a contour plot in Figure 4.

We briefly discuss two extreme cases. For $\delta_0 = \delta_1 = \frac{1}{2}$, the upper bound states that $h(S^{\tan}_{1/2,1/2}) \leqslant 1$, which holds trivially. Indeed, it is not difficult to see that in fact $h(S^{\tan}_{1/2,1/2}) = 1$ since random bits are inserted at random positions in the sequence.

For $\delta_0 = \delta_1 = 1$, this upper bound equals $H_2(1/3) = 0.9183$, which is the same as the upper bound given by Theorem 11. The lower bound given by that theorem is 0.8689, which indicates that for this case, the gap between the upper bound and the true value is small.

We also discuss a similar string-duplication system that has already been studied in [14], [27]. In general, such comparisons can be useful to decide between proposed mutation models for a given sequence, especially biological sequences. In that system, instead of tandem duplications that are probabilistically noisy, independent tandem duplications and substitutions are allowed. We compare the behavior of that system with $S^{\tan}_{\delta,\delta}$ for some $\delta \in [0, 1]$. Specifically, we compare the bound of Theorem 13 for $\delta = \delta_0 = \delta_1$,

$$h(S^{\tan}_{\delta,\delta}) \leqslant H_2\left(\frac{1}{1+2\delta}\right),$$

with an upper bound for the system in which tandem duplications and substitutions occur with probabilities $1-\delta$ and δ, respectively, at a random position in the sequence. We refer to this system as S^{tsb}_{δ}. The definition of the entropy rate for S^{tsb}_{δ} is slightly different, to accommodate the fact that the length of the sequence does not necessarily grow in each step. It is shown in [27] that the entropy rate of this system is bounded from above by

$$h(S^{\text{tsb}}_{\delta}) \leqslant H_2\left(\frac{2\delta}{1+3\delta}\right).$$

The bounds are compared in Fig. 5. The bounds suggest that the systems behave differently when δ is away from 0. In particular, $h(S^{\text{tsb}}_{\delta}) \leqslant 0.9709$ and $h(S^{\tan}_{\delta,\delta}) = 1$ for $p = 1/2$. For this value of p, in S^{tsb}_{δ} half of the mutations are duplications, which make substrings 00 and 11 more likely than what is expected in a random sequence, leading to an entropy rate of less than 1.

V. Interspersed Duplication

Finally, we consider the case of interspersed duplication. While seemingly a more elaborate duplication rule (probabilistic both when choosing the bit to duplicate, as well as the insertion position), we now show that it has the same entropy rate as end duplication.

Theorem 14. *Let $\Sigma = \{0, 1\}$, $s \in \Sigma^+$ be a seed string, and denote*

$$S^{\text{end}} = S^{\text{end}}_{\delta_0,\delta_1} = (\Sigma, s, T^{\text{end}}_{\delta_0,\delta_1}),$$
$$S^{\text{int}} = S^{\text{int}}_{\delta_0,\delta_1} = (\Sigma, s, T^{\text{int}}_{\delta_0,\delta_1}).$$

Then

$$h(S^{\text{int}}_{\delta_0,\delta_1}) = h(S^{\text{end}}_{\delta_0,\delta_1}).$$

Proof: We first require some general arguments, in preparation for the proof for the entropy rate. Consider an

interspersed-duplication process, starting with the seed s, and running for n mutation steps. Denote the bit generated in the ith mutation step, $1 \leqslant i \leqslant n$, by $b_{|s|+i}$. We also use $b_1\, b_2 \cdots b_{|s|} = s$ to denote the bits of the seed string.

The bits, however, do not appear in the order of generation (as they do in end duplication), since they are inserted in random places. Thus, if after n mutations we reach a string $w \in \Sigma^{|s|+n}$, then

$$w = b(\pi) \triangleq b_{\pi(1)} b_{\pi(2)} \cdots b_{\pi(|s|+n)},$$

for some permutation $\pi \in \mathbb{S}_{|s|+n}$ that satisfies

$$\pi^{-1}(1) < \pi^{-1}(2) < \cdots < \pi^{-1}(|s|), \tag{12}$$

since the order of the bits of the seed string is maintained. For example, we may have

$$\begin{aligned} s = S(0) &= b_1 b_2 b_3 b_4 &&= 0011,\\ S(1) &= b_1 b_2 b_3 \boldsymbol{b_5} b_4 &&= 001\mathbf{0}1,\\ S(2) &= b_1 \boldsymbol{b_6} b_2 b_3 b_5 b_4 &&= 0\mathbf{1}0101, \end{aligned}$$

and $\pi = [1, 6, 2, 3, 5, 4]$.

Let us denote the set of permutations satisfying (12) by P_n, and hence, $|P_n| = (n + |s|)!/\,|s|!$. Since the insertion position at each mutation step is chosen independently and uniformly, the probability of each permutation is exactly,

$$\frac{1}{|s|+1} \cdot \frac{1}{|s|+2} \cdot \cdots \cdot \frac{1}{|s|+n} = \frac{|s|!}{(n+|s|)!},$$

i.e., the overall permutation is chosen uniformly from P_n.

Let us now denote,

$$\begin{aligned} t_0 &\triangleq |s|_0\ , & t_1 &\triangleq |s| - t_0,\\ k_0 &\triangleq \left|b_{|s|+1} \cdots b_{|s|+n}\right|_0\ , & k_1 &\triangleq n - k_0, \end{aligned}$$

namely, t_0 and t_1 denote the number of zeros and ones (respectively) in the seed string, and k_0 and k_1 denote the number of zeros and ones (respectively) in the bits generated due to mutations.

We say $\pi_1, \pi_2 \in P_n$ are equivalent, denoted $\pi_1 \sim \pi_2$, if $b(\pi_1) = b(\pi_2)$. This is clearly an equivalence relation. For $\pi \in P_n$, let E_π denote the equivalence class of π. Computing $|E_\pi|$ is hard, but it suffices for us to bound it by

$$k_0! k_1! \leqslant |E_\pi| \leqslant (t_0 + k_0)!(t_1 + k_1)!.$$

For the lower bound, we permute only the newly generated zeros between themselves, and similarly the ones, while keeping the bits of the seed in their place. For the upper bound, we permute all zeroes between themselves, and similarly the ones, thus, perhaps reaching some permutations that are not in P_n.

Lastly, denote by A_{k_0} the event that that among the n bits generated due to mutations, exactly k_0 are zeros, and the rest, $k_1 = n - k_0$ are ones. Also, let $B^{\text{int}}(n, k_0)$ denote the set of strings $w \in \Sigma^{|s|+n}$, $|w|_0 = k_0 + t_0$, that may be obtained from s using n interspersed-duplication mutations. It then follows that if $w \in B^{\text{int}}(n)$, then

$$\begin{aligned} &\frac{(t_0+t_1)!k_0!k_1!}{(t_0+t_1+k_0+k_1)!} \Pr(A_{k_0}) \leqslant \Pr(S^{\text{int}}(n) = w)\\ &\quad\leqslant \frac{(t_0+t_1)!(t_0+k_0)!(t_1+k_1)!}{(t_0+t_1+k_0+k_1)!} \Pr(A_{k_0}). \end{aligned}$$

This means that

$$\Pr(S^{\text{int}}(n) = w) = \Pr(A_{k_0}) \cdot 2^{-n(H_2(k_0/n)+o(1))},$$

as well as

$$\left|B^{\text{int}}(n, k_0)\right| = 2^{n(H_2(k_0/n)+o(1))}.$$

We are now ready to prove our claims. First, we look at the noiseless case, $\delta_0 = \delta_1 = 0$. The probability, $\Pr(A_{k_0})$, has already been given in (2). We therefore get,

$$\begin{aligned} h(S^{\text{int}}) &= \limsup_{n\to\infty} \frac{1}{n} H(S^{\text{int}}(n))\\ &= \limsup_{n\to\infty} \frac{1}{n} \sum_{k_0=0}^{n} \Pr(A_{k_0}) \log_2 \left|B^{\text{int}}(n, k_0)\right|\\ &= \int_0^1 \beta(p; t_0, t_1) H_2(p)\, \mathrm{d}p\\ &= h(S^{\text{end}}), \end{aligned}$$

exactly as in the proof of Theorem 3.

The second (and last) case is $\delta_0 + \delta_1 > 0$. Denote $\alpha \triangleq \delta_1/(\delta_0 + \delta_1)$. By Lemma 7, for any $\epsilon_1, \epsilon_2 > 0$, there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$, $\Pr(\left|\mathrm{fr}_0(S^{\text{int}}(n)) - \alpha\right| \leqslant \epsilon_1) \geqslant 1 - \epsilon_2$. Then,

$$\begin{aligned} &h(S^{\text{int}})\\ &\quad= \limsup_{n\to\infty} \frac{1}{n} H(S^{\text{int}}(n))\\ &\quad= \limsup_{n\to\infty} \frac{1}{n} \Bigg(\sum_{\left|\frac{k_0+t_0}{n+|s|} - \alpha\right| \leqslant \epsilon_1} \Pr(A_{k_0}) \log_2 \left|B^{\text{int}}(n, k_0)\right|\\ &\qquad + \sum_{\left|\frac{k_0+t_0}{n+|s|} - \alpha\right| > \epsilon_1} \Pr(A_{k_0}) \log_2 \left|B^{\text{int}}(n, k_0)\right| \Bigg)\\ &\quad\leqslant \max_{x \in [\alpha-\epsilon_1, \alpha+\epsilon_1]} H_2(x) + \epsilon_2. \end{aligned} \tag{13}$$

On the other hand,

$$\begin{aligned} &h(S^{\text{int}})\\ &\quad= \limsup_{n\to\infty} \frac{1}{n} H(S^{\text{int}}(n))\\ &\quad= \limsup_{n\to\infty} \frac{1}{n} \Bigg(\sum_{\left|\frac{k_0+t_0}{n+|s|} - \alpha\right| \leqslant \epsilon_1} \Pr(A_{k_0}) \log_2 \left|B^{\text{int}}(n, k_0)\right|\\ &\qquad + \sum_{\left|\frac{k_0+t_0}{n+|s|} - \alpha\right| > \epsilon_1} \Pr(A_{k_0}) \log_2 \left|B^{\text{int}}(n, k_0)\right| \Bigg)\\ &\quad\geqslant (1 - \epsilon_2) \min_{x \in [\alpha-\epsilon_1, \alpha+\epsilon_1]} H_2(x). \end{aligned} \tag{14}$$

Taking the limit of (13) and (14) as $\epsilon_1, \epsilon_2 \to 0^+$, we obtain

$$h(S^{\text{int}}) = H_2(\alpha) = h(S^{\text{end}}),$$

as claimed. ∎

VI. Conclusion

In this paper we defined and studied three Pólya string models. We determined the exact entropy rate of end duplication, $S^{\text{end}}_{\delta_0,\delta_1}$, and interspersed duplication, $S^{\text{int}}_{\delta_0,\delta_1}$, both for any noise parameters δ_0 and δ_1. We also found the exact entropy rate of noiseless tandem duplication, $S^{\text{tan}}_{0,0}$, as well as we connected the entropy rate of complement tandem duplication, $S^{\text{tan}}_{1,1}$, with the signatures of random permutations. Finally, we upper bounded the entropy rate of general noisy tandem duplication, $S^{\text{tan}}_{\delta_0,\delta_1}$.

We make several interesting observations. First, had we used a Pólya urn model instead of a string model, then no difference would have been observed between tandem and end duplication. Indeed, the distribution of 0's and 1's in both cases is the same. However, when considering the structure of a string, the difference between the two comes to light.

Many other differences are apparent between the combinatorial capacity (found in [13]) and the entropy rate studied here, and we point a few:

- While the combinatorial capacity of (noiseless) end duplication is known to be 1, in the probabilistic model, the entropy rate varies depending on the starting string.
- Similarly, for the complement tandem-duplication model, it is easy to show that the combinatorial capacity is 1, while the entropy rate is bounded away from both 0 and 1.
- The entropy rate of $S^{\text{end}}_{\delta_0,\delta_1}$ is equal to that of $S^{\text{int}}_{\delta_0,\delta_1}$, which is not generally the case when using the combinatorial capacity.

Many open questions remain. Obvious ones include the determination of $h(S^{\text{tan}}_{\delta_0,\delta_1})$ for all values of δ_0 and δ_1. We also note that the systems studied in the current paper are limited to duplications of length 1, while genomic duplication mutations are observed for a large range of duplication lengths. Thus, extending the results to longer duplication lengths is an important open task. Other noise models are also of interest. For example, one might be interested in models in which mutation steps either duplicate or substitute a letter (e.g., see [14], [27]). Finally, more elaborate distributions may be studied, including context-sensitive duplication rules.

Acknowledgments

The authors would like to thank the associate editor and the anonymous reviewers, whose comments improved the presentation of the paper.

References

[1] N. Alon, J. Bruck, F. F. Hassanzadeh, and S. Jain, "Duplication distance to the root for binary sequences," *IEEE Trans. Inf. Theory*, vol. 63, no. 12, pp. 7793–7803, Dec. 2017.

[2] S. Banach, "Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales," *Fund. Math.*, vol. 3, no. 1, pp. 133–181, 1922.

[3] G. Benson and L. Dong, "Reconstructing the duplication history of a tandem repeat," in *Proc. ISMB*, 1999, pp. 44–53.

[4] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[5] P. P. Calabrese, R. T. Durrett, and C. F. Aquadro, "Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models," *Genetics*, vol. 159, no. 2, pp. 839–852, Oct. 2001.

[6] Y. M. Chee, J. Chrisnata, H. M. Kiah, and T. T. Nguyen, "Deciding the confusability of words under tandem repeats," 2017, *arXiv:1707.03956*. [Online]. Available: https://arxiv.org/abs/1707.03956

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[8] R. Durrett and S. Kruglyak, "A new stochastic model of microsatellite evolution," *J. Appl. Probab.*, vol. 36, no. 3, pp. 621–631, Sep. 1999.

[9] F. Eggenberger and G. Pólya, "Über die statistik verketteter vorgänge," *ZAMM-J. Appl. Math. Mech./Zeitschrift Angew. Math. Mech.*, vol. 3, no. 4, pp. 279–289, 1923.

[10] O. Elemento and O. Gascuel, "An efficient and accurate distance based algorithm to reconstruct tandem duplication trees," *Bioinformatics*, vol. 18, no. 2, pp. S92–S99, Oct. 2002.

[11] M. Farach, M. Noordewier, S. Savari, L. Shepp, A. Wyner, and J. Ziv, "On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence," in *Proc. 6th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*. Philadelphia, PA, USA: SIAM, 1995, pp. 48–57.

[12] F. Farnoud, M. Schwartz, and J. Bruck, "A stochastic model for genomic interspersed duplication," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 1731–1735.

[13] F. F. Hassanzadeh, M. Schwartz, and J. Bruck, "The capacity of string-duplication systems," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 811–824, Feb. 2016.

[14] F. Farnoud, M. Schwartz, and J. Bruck, "Estimation of duplication history under a stochastic model for tandem repeats," *BMC Bioinform.*, vol. 20, no. 1, 2019, Art. no. 64.

[15] J. W. Fondon, III, and H. R. Garner, "Molecular origins of rapid and continuous morphological evolution," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 52, pp. 18058–18063, 2004.

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

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

[18] A. Jeffrey and H. Dai, *Handbook of Mathematical Formulas and Integrals*, 4th ed. Amsterdam, The Netherlands: Elsevier, 2008.

[19] S. Kruglyak, R. T. Durrett, M. D. Schug, and C. F. Aquadro, "Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations," *Proc. Nat. Acad. Sci. USA*, vol. 95, no. 18, pp. 10774–10778, 1998.

[20] Y. Lai and F. Sun, "The relationship between microsatellite slippage mutation rate and the number of repeat units," *Mol. Biol. Evol.*, vol. 20, no. 12, pp. 2123–2131, 2003.

[21] J. K. Lanctot, M. Li, and E.-H. Yang, "Estimating DNA sequence entropy," in *Proc. 11th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*. Philadelphia, PA, USA: SIAM, 2000, pp. 409–418.

[22] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[23] A. Lenz, N. Jünger, and A. Wachter-Zeh, "Bounds and constructions for multi-symbol duplication error correcting codes," 2018, *arXiv:1807.02874*. [Online]. Available: https://arxiv.org/abs/1807.02874

[24] P. Leupold, V. Mitrana, and J. M. Sempere, "Formal languages arising from gene repeated duplication," in *Aspects of Molecular Computing*. Berlin, Germany: Springer, 2004, pp. 297–308.

[25] G. Levinson and G. A. Gutman, "Slipped-strand mispairing: A major mechanism for DNA sequence evolution," *Mol. Biol. Evol.*, vol. 4, no. 3, pp. 203–221, 1987.

[26] P. Liò, A. Politi, M. Buiatti, and S. Ruffo, "High statistics block entropy measures of DNA sequences," *J. Theor. Biol.*, vol. 180, no. 2, pp. 151–160, May 1996.

[27] H. Lou, M. Schwartz, and F. F. Hassanzadeh, "Evolution of N-gram frequencies under duplication and substitution mutations," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, Jun. 2018, pp. 2246–2250.

[28] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. Amsterdam, The Netherlands: North Holland, 1978.

[29] H. Mahmoud, *Pólya Urn Models*, 1st ed. London, U.K.: Chapman & Hall, 2008.

[30] S. Ohno, *Evolution by Gene Duplication*. Heidelberg, Germany: Springer-Verlag, 1970.

[31] Y. L. Orlov and V. N. Potapov, "Complexity: An Internet resource for analysis of DNA sequence complexity," *Nucleic Acids Res.*, vol. 32, pp. W628–W633, Jul. 2004.

[32] G. Pólya, "Sur quelques points de la théorie des probabilités," *Ann. Institut Henri Poincaré*, vol. 1, no. 2, pp. 117–161, 1930.

[33] G. Pólya and F. Eggenberger, "Sur l'interprétation de certaines courbes de fréquences," *Comptes Rendus De L'Académie Des Sci.*, vol. 187, pp. 870–872, Jul./Dec. 1928.
[34] A. O. Schmitt and H. Herzel, "Estimating the entropy of DNA sequences," *J. Theor. Biol.*, vol. 188, no. 3, pp. 369–377, 1997.
[35] 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, pp. 345–349, Jul. 2017.
[36] M. Tang, M. Waterman, and S. Yooseph, "Zinc finger gene clusters and tandem gene duplication," *J. Comput. Biol.*, vol. 9, no. 2, pp. 429–446, 2002.
[37] D. Tautz and T. Domazet-Lošo, "The evolutionary origin of orphan genes," *Nature Rev. Genet.*, vol. 12, no. 10, pp. 692–702, Oct. 2011.
[38] K. Usdin, "The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases," *Genome Res.*, vol. 18, no. 7, pp. 1011–1019, 2008.
[39] H. Wan, L. Li, S. Federhen, and J. C. Wootton, "Discovering simple regions in biological sequences associated with scoring schemes," *J. Comput. Biol., J. Comput. Mol. Cell Biol.*, vol. 10, no. 2, pp. 171–185, 2003.
[40] Y. Yehezkeally and M. Schwartz, "Reconstruction codes for DNA sequences with uniform tandem-duplication errors," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, Jun. 2018, pp. 2535–2539.

Ohad Elishco (S'12–M'18) is a post-doctoral researcher at the department of electrical engineering, University of Maryland at College Park. He received his B.Sc. in mathematics and his B.Sc. in electrical engineering in 2012 from Ben-Gurion University of the Negev, Israel; the M.Sc. and Ph.D degrees in electrical engineering from Ben-Gurion University if the Negev, Israel, in 2013, 2017 respectively. He was a post-doctoral researcher in the department of electrical engineering, Massachusetts institute of technology. His research interests are coding and dynamical systems.

Farzad Farnoud (Hassanzadeh) (M'13) is an Assistant Professor in the Department of Electrical and Computer Engineering and the Department of Computer Science at the University of Virginia. Previously, he was a postdoctoral scholar at the California Institute of Technology.

He received his MS degree in Electrical and Computer Engineering from the University of Toronto in 2008. From the University of Illinois at Urbana-Champaign, he received his MS degree in mathematics and his Ph.D. in Electrical and Computer Engineering in 2012 and 2013, respectively. His research interests include the information-theoretic and probabilistic analysis of genomic evolutionary processes; rank aggregation and gene prioritization; and coding for flash memory and DNA storage. He is the recipient of the 2013 Robert T. Chien Memorial Award from the University of Illinois for demonstrating excellence in research in electrical engineering and the recipient of the 2014 IEEE Data Storage Best Student Paper Award.

Moshe Schwartz (M'03–SM'10) is an associate 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.

Jehoshua Bruck (S'86–M'89–SM'93–F'01) is the Gordon and Betty Moore Professor of computation and neural systems and electrical engineering at the California Institute of Technology (Caltech). His current research interests include information theory and systems and the theory of computation in nature.

Dr. Bruck received the B.Sc. and M.Sc. degrees in electrical engineering from the Technion-Israel Institute of Technology, in 1982 and 1985, respectively, and the Ph.D. degree in electrical engineering from Stanford University, in 1989. His industrial and entrepreneurial experiences include working with IBM Research where he participated in the design and implementation of the first IBM parallel computer; cofounding and serving as Chairman of Rainfinity (acquired in 2005 by EMC), a spin-off company from Caltech that created the first virtualization solution for Network Attached Storage; as well as cofounding and serving as Chairman of XtremIO (acquired in 2012 by EMC), a start-up company that created the first scalable all-flash enterprise storage system.

Dr. Bruck is a recipient of the Feynman Prize for Excellence in Teaching, the Sloan Research Fellowship, the National Science Foundation Young Investigator Award, the IBM Outstanding Innovation Award and the IBM Outstanding Technical Achievement Award.