On the Biased Partial Word Collector Problem

Philippe Duchon[1] and Cyril Nicaud[2]

[1] Univ. Bordeaux, LaBRI, UMR 5800, F-33400 Talence, France
CNRS, LaBRI, UMR 5800, F-33400 Talence, France
[2] Université Paris-Est, LIGM (UMR 8049), CNRS, ENPC, ESIEE Paris, UPEM,
F-77454 Marne-la-Vallée, France

Abstract. In this article we consider the following question: N words of length L are generated using a biased memoryless source, i.e. each letter is taken independently according to some fixed distribution on the alphabet, and collected in a set (duplicates are removed); what are the frequencies of the letters in a typical element of this random set? We prove that the typical frequency distribution of such a word can be characterized by considering the parameter $\ell = L/\log N$. We exhibit two thresholds $\ell_0 < \ell_1$ that only depend on the source, such that if $\ell \leq \ell_0$, the distribution resembles the uniform distribution; if $\ell \geq \ell_1$ it resembles the distribution of the source; and for $\ell_0 \leq \ell \leq \ell_1$ we characterize the distribution as an interpolation of the two extremal distributions.

1 Introduction

The *coupon collector problem* is a classical topic in discrete probability; in its most basic form, the question is to determine how many independent draws from a uniform distribution on some fixed finite set E (say, of cardinality n) are needed, in expectation, to obtain each possible value at least once. The answer turns out to be exactly nH_n, where $H_n = 1 + 1/2 + \cdots + 1/n$ denotes the n-th harmonic number.

It is natural to consider non-uniform versions of the problem, where values have different probabilities. Typically, some structure is needed on the set of possible values to make the problem tractable. The *weighted word collector problem*, as studied in [1], corresponds to the case where E is a set of words of a fixed length L (possibly, all words over some finite alphabet A, *i.e.* A^L, but more generally for some language $\mathcal{L} \subseteq A^L$), each word has a probability proportional to its weight, and this weight is defined as a product of individual letter weights.

In this paper, we consider a process related to the weighted word collector process when $\mathcal{L} = A^L$. In this case, words are drawn according to a *memoryless source*: each letter a_i has a specific *a priori* probability $p(a_i)$, and words are composed of L independent letters drawn from this probability distribution. Instead of waiting for all possible words to appear, we consider a *partial word collector*: we repeatedly draw random words from the memoryless model, keeping track of the set of distinct words "collected".

In this setting, we try to answer the following somewhat informal question:

> When N independent draws have been made from the random word model, what does a "typical" member of the set of already-seen words look like?

More precisely, we study the likely *composition* of collected words; that is, the number of occurrences of each letter in a word drawn uniformly at random from the set of already collected words. At this point, it is only for convenience that we describe the process in terms of picking a random word from those already collected; our interest is in identifying what a typical collected word looks like.

This question is related to the *subword complexity* of a long random word in the memoryless model. The subword complexity function of a (finite or infinite) word w is the function that maps each integer L to the number of different factors of length L that appear in w; as shown in [4], the (random) subword complexity for factors of length L of a random word of length $N + L - 1$ in the memoryless model, is very close (in expectation, and in distribution) to the size of our partial word collection.

Our initial motivation for studying this problem follows the work of Rubinchik and Shur on the expected number of distinct palindromic factors in a uniform random word [6], which we extended to α-gapped patterns [3] (an α-gapped pattern is a factor uvu with $|uv| \leq \alpha|u|$ for given $\alpha \geq 1$). When trying to extend these results to words generated by a memoryless source, a problem very similar to the one investigated in this paper arises: typical palindromic factors have length in $\Theta(\log n)$, and a promising way to count them is to identify those who contribute the most, using (and extending) the methods presented in this paper.

Special cases of our question can be answered easily, at least informally. If all words have the same probability (*i.e.*, $p(a_i) = 1/k$ for each of the k letters), then drawing a uniform word from the set of collected words is equivalent to drawing a uniform word; by the law of large numbers, each letter is extremely likely (at least for large L) to have an observed frequency close to its *a priori* probability $1/k$.

When the letter probabilities are not uniform, the asymptotic regimes (with fixed L and variable N) are intuitively clear. For very small N, all N collected words are likely to be different, so the two-step sampling process is equivalent to drawing a single random word from the memoryless model; *i.e.* letter frequencies are likely to be close to the letter probabilities $p(a_i)$. At the other end of the spectrum, if N is large enough, with high probability all words have been collected, so that drawing a uniform word from the set of collected words is almost equivalent to drawing a uniform random word; letter frequencies should be close to the uniform $1/k$.

The interesting case lies in the intermediate regime. In the present paper, we exhibit an evolution for the likely composition of typical collected words; the precise statement is given by Theorem 1. The significant parameter is the ratio $\ell = L/\log(N)$. Informally, as N grows (as ℓ decreases), the composition goes continuously from the *a priori* composition to the *uniform* composition

along a predetermined curve: at all times, each letter a_i has typical frequency proportional to $p(a_i)^c$ for some constant c that depends on the ratio ℓ; as N grows from 1 to infinity (as ℓ decreases from infinity to 0), c decreases from 1 to 0. Our main theorem also gives explicit thresholds, one until which the typical letter frequencies fit the *a priori* frequencies, and one after which they fit the uniform frequencies. Unsurprisingly, the latter is significantly smaller than the time to a full collection: typical collected words "look like" uniform words long before all words have been collected.

2 Definitions and notations

If k is a positive integer, let $[k] = \{1, \dots, k\}$. For $\mathbf{x} = (x_1, \dots, x_k) \in \mathbb{R}^k$, let $\|\mathbf{x}\| = \sqrt{\sum_{i\in[k]} x_i^2}$ and if $d > 0$, let $\overline{B}(\mathbf{x}, d)$ be the closed ball of all vectors $\mathbf{y}$ such that $\|\mathbf{y} - \mathbf{x}\| \leq d$.

Let $A = \{a_1, \dots, a_k\}$ be an alphabet with $k \geq 2$ letters, which we fix from now on. For any word $w \in A^*$, let $\mathbf{comp}(w) = (|w|_1, \dots, |w|_k)$ denote its *composition vector*, where $|w|_i$ is the number of occurrences of a_i in w. If w is not empty, let $\mathbf{freq}(w) = (\frac{|w|_1}{|w|}, \dots, \frac{|w|_k}{|w|})$ denote its *frequency vector*.

The (natural-based) entropy function on k positive variables is defined by

$$H_k(\mathbf{x}) = H_k(x_1, \dots, x_k) = -\sum_{i\in[k]} x_i \log x_i.$$

We will omit the index k in the sequel, as it is fixed in our settings, and write $H(\mathbf{x})$ instead of $H_k(\mathbf{x})$.

Throughout the article, we assume some probability vector $\mathbf{p} = (p_1, \dots, p_k)$ to be fixed, with $p_i \neq 0$ for every $i \in [k]$, and we consider statistics in the memoryless model where each letter a_i has probability p_i. We denote by $\mathbb{P}_L$ this probability measure on A^L. We also assume that $\mathbf{p}$ is not the uniform distribution, *i.e.* there exists $i \in [k]$ such that $p_i \neq \frac{1}{k}$. Let $p_{\min} = \min_{i\in[k]} p_i$ and $p_{\max} = \max_{i\in[k]} p_i$ be the minimal and maximal values of $\mathbf{p}$.

Remark that since the number of letters $k \geq 2$ and the probability distribution $\mathbf{p}$ are fixed throughout the article, the constants we use may implicitly depend on $\mathbf{p}$ and k.

In our statements and proofs below, we mainly work on frequency vectors of words of length L, which are very specific vectors of $\mathbb{R}^k$. Depending on our needs, we will see them as frequency vectors, probability vectors (going from discrete to continuous), or even just vectors. We therefore introduce the following notations:

– For given positive integer L, let $\mathcal{F}_L$ denote the set of frequency vectors of words of length L, defined by

$$\mathcal{F}_L = \left\{ (x_1, \dots, x_k) \in \mathbb{R}^k : \sum_{i\in[k]} x_i = 1 \text{ and } \forall i \in [k],\ x_i L \in \mathbb{Z}_{>0} \right\}.$$

– Let $\mathcal{P}$ denote the set of probability vectors, defined by

$$\mathcal{P} = \left\{ (x_1, \ldots, x_k) \in \mathbb{R}^k : \sum_{i \in [k]} x_i = 1 \text{ and } \forall i \in [k],\ 0 \leq x_i \leq 1 \right\}.$$

– We will also need a restriction of $\mathcal{P}$ to probability vectors that are not close to the border. Let $\tilde{\mathcal{P}}$ be the subset of $\mathcal{P}$ defined by

$$\tilde{\mathcal{P}} = \left\{ (x_1, \ldots, x_k) \in \mathbb{R}^k : \sum_{i \in [k]} x_i = 1 \text{ and } \forall i \in [k],\ \frac{p_{\min}}{2} \leq x_i \leq 1 \right\}.$$

3 Main result and proof sketch

In this section we define the problem that is studied in this paper, state our main result and provide a very informal proof sketch.

3.1 The biased partial coupon collector problem

For any two positive integers N and L, we are interested in the following two-steps process:

1. Generate, independently, N words of length L following the memoryless distribution of parameter $\mathbf{p}$ and collect them in a set $\mathcal{S}_{N,L}$, disregarding multiplicities; if a given word is generated several times, it contributes only once to $\mathcal{S}_{N,L}$ (consequently, the number of elements in $\mathcal{S}_{N,L}$ is itself random).
2. Draw uniformly at random an element of $\mathcal{S}_{N,L}$, which we call $U_{N,L}$.

More formally, let X_1, …, X_N be N i.i.d. random words, each of length L and chosen using the memoryless source of parameter $\mathbf{p}$. If $\boldsymbol{u} = (u_1, \ldots, u_N)$ is a tuple of elements, of A^L, let $\mathrm{Set}(\mathbf{u})$ be the set defined by

$$\mathrm{Set}(\mathbf{u}) = \left\{ u \in A^L :\ \exists i \in [N],\ u_i = u \right\}.$$

The random set $\mathcal{S}_{N,L}$ is defined by $\mathcal{S}_{N,L} = \mathrm{Set}(X_1, \ldots, X_n)$, and $U_{N,L}$ consists in choosing uniformly at random an element of $\mathcal{S}_{N,L}$ (which cannot be empty in our settings).

We mainly focus on the typical composition of letters within the result of our random process, *i.e.* we are interested in the random vector $\mathbf{freq}(U_{N,L})$.

Intuitively, our main theorem states that when (a) L is small compared to $\log N$, almost all words have been collected, hence our process is almost the same as selecting uniformly at random a word of A^L: the frequency vector resembles the uniform distribution. On the other hand, when (c) $\log N$ is small compared to L, only a few number of words have been collected, there are few duplicates, hence our process is almost the same as just generating one word with the source. The intermediate range (b) corresponds to an interpolation between the

two distributions. Also, note that, seeing the process as incrementally collecting words, regime (a) occurs well before a majority of possible words have been collected, and that regime (c) lasts well after duplicates have become numerous; the precise description of the regions for the various regimes is also a part of our results.

The statement is the following.

Theorem 1. *Consider a memoryless source for a fixed alphabet of size k, of probability vector $\mathbf{p}$ that is not the uniform distribution, and define*

$$\Phi(t) = \sum_{i\in[k]} p_i^t,\ \forall t \in \mathbb{R}.$$

Let $\ell_0 = \frac{-k}{\sum_{i\in[k]} \log p_i}$ and $\ell_1 = \frac{1}{H(\mathbf{p})}$. There exist a positive integer L_0 and a positive real λ such that for every integers $L \geq L_0$ and $N \geq 2$, if we set $\ell = \frac{L}{\log N}$ then the following results hold:
(a) If $\ell \leq \ell_0$, then

$$\mathbb{P}\left(\|\,\mathbf{freq}(U_{N,L}) - \overline{\mathbf{x}}\| \geq \frac{\log L}{\sqrt{L}}\right) \leq L^{-\lambda \log L},\ \textit{with}\ \overline{\mathbf{x}} = \left(\frac{1}{k},\ldots,\frac{1}{k}\right).$$

(b) If $\ell_0 \leq \ell \leq \ell_1$, then

$$\mathbb{P}\left(\|\,\mathbf{freq}(U_{N,L}) - \mathbf{x}_c\| \geq \frac{\log L}{\sqrt{L}}\right) \leq L^{-\lambda \log L},\ \textit{with}\ \mathbf{x}_c = \left(\frac{p_1^c}{\Phi(c)},\ldots,\frac{p_k^c}{\Phi(c)}\right),$$

where c is the unique solution in $[0,1]$ of the equation $\ell\Phi'(c) + \Phi(c) = 0$.
(c) If $\ell \geq \ell_1$, then

$$\mathbb{P}\left(\|\,\mathbf{freq}(U_{N,L}) - \mathbf{p}\| \geq \frac{\log L}{\sqrt{L}}\right) \leq L^{-\lambda \log L},\ \textit{with}\ \mathbf{p} = (p_1,\ldots,p_k).$$

Remark 1. Observe that if $\ell = \ell_0$, then $\mathbf{x}_c = \overline{\mathbf{x}}$ and that if $\ell = \ell_1$, then $\mathbf{x}_c = \mathbf{p}$. Observe also that if we allow $\mathbf{p} = \overline{\mathbf{x}}$, then $\ell_0 = \ell_1$ and everything collapses; in this case our random process, for any value of N, is just a complicated way to produce a uniform random word of length L.

Remark 2. A reviewer suggested a change of parameterization that looks promising, by setting $N = (k^L)^\alpha$, for $\alpha \in \mathbb{R}^+$. This way the parameter is α and not ℓ, and they are related by $\alpha = \frac{1}{\ell \log k}$. Thus, as N increases for fixed L, α also increases; notice that the case $\alpha = 1$ corresponds to the situation where one draws just enough words to possibly get each existing word exactly once (though this would happen with extremely small probability). In this parameterization, the first threshold α_0, corresponding to ℓ_1, is $\alpha_0 = H_k(\mathbf{p})$, where $H_k(\mathbf{x}) = \frac{1}{\log k} H(\mathbf{x})$ is the entropy in base k of $\mathbf{x}$. The second threshold α_1, corresponding to ℓ_0, is $\alpha_1 = -\frac{1}{k\sum_i \log_k p_i}$. The reviewer also observed that α_1 can be written $\alpha_1 = 1 + \frac{1}{\log k} D_{KL}(\mathbf{u}|\mathbf{p})$, using the Kullback-Leibler divergence [5,

p. 38], which is a classical notion in information theory defined for two positive vectors $\mathbf{x}$ and $\mathbf{y}$ of size k by

$$D_{KL}(\mathbf{x}|\mathbf{y}) = \sum_{i\in[k]} x_i \log \frac{x_i}{y_i}.$$

Here the 1 term in the expression for α_1 comes from the (base k) entropy of the uniform distribution vector $\mathbf{u}$. In this form, it is more readily apparent that $\alpha_0 < 1 < \alpha_1$ holds as soon as $\mathbf{p}$ is not the uniform vector, since the Kullback-Leibler divergence is positive.

3.2 Main steps of the proof

For this proof sketch, the reader must be aware that we do not mention some technical conditions that are necessary for some statements to be correct. Our aim here is only to present an informal guide to the technical sections where Theorem 1 is proved.

It is convenient to introduce the parameter ℓ defined by $L = \ell \log N$, as the main phase transitions appear when the number of words is exponential in their lengths[3]. The proof consists in identifying the frequency vectors corresponding to compositions of words that contribute the more to $\mathcal{S}_{N,L}$.

For a given $\mathbf{x} \in \mathcal{F}_L$, we first prove that there are roughly[4] $N^{\ell H(\mathbf{x})}$ words of length L whose frequency vector is $\mathbf{x}$. Let $\mathcal{W}_L(\mathbf{x})$ denote this set of words.

Observe that words that share the same frequency vector $\mathbf{x}$ have the same probability $p_L(\mathbf{x}) = N^{\ell \sum_{i\in[k]} x_i \log p_i}$ of being generated by the memoryless source. Hence, the probability that a word of $\mathcal{W}_L(\mathbf{x})$ appears in $\mathcal{S}_{N,L}$ is exactly $q_{N,L}(\mathbf{x}) := 1 - (1 - p_L(\mathbf{x}))^N$. There are two cases: If $1 + \ell \sum_{i\in[k]} x_i \log p_i < 0$ then $q_{N,L}(\mathbf{x}) \approx N^{1+\ell \sum_{i\in[k]} x_i \log p_i}$; otherwise $q_{N,L}(\mathbf{x}) \approx \Theta(1)$. For $w \in \mathcal{W}_L(\mathbf{x})$, this can be summarized as follows:

$$\mathbb{P}(w \in \mathcal{S}_{N,L}) = q_{N,L}(\mathbf{x}) \approx N^{\min\left(0, 1+\ell \sum_{i\in[k]} x_i \log p_i\right)}.$$

By linearity of expectation, we can thus estimate the expected contribution to $\mathcal{S}_{N,L}$ of $\mathcal{W}_L(\mathbf{x})$ for a given frequency vector $\mathbf{x}$:

$$\mathbb{E}\,|\mathcal{S}_{N,L} \cap \mathcal{W}_L(\mathbf{x})| \approx N^{\ell \min(H(\mathbf{x}), K_\ell(\mathbf{x}))}, \text{ with } K_\ell(\mathbf{x}) = H(\mathbf{x}) + \frac{1}{\ell} + \sum_{i\in[k]} x_i \log p_i.$$

To find the frequency vectors that contribute the most, we have to study the function $G_\ell(\mathbf{x}) = \min(H(\mathbf{x}), K_\ell(\mathbf{x}))$. This is the minimum of two strictly concave functions on $\mathcal{P}$, each of which has a maximum. There are two main situations for the location of the maximum of such a function, as depicted in Fig. 1 (for functions of a single variable).

[3] It is a more natural view of the process to consider L as fixed and N as varying; this, however, leads to the somewhat artificial parameterization $N = \exp(L/\ell)$.

[4] By "roughly" we mean up to some multiplicative power of L, with $L = \Theta(\log N)$ at our scale.

Fig. 1. The two possibilities for the location of the maximum value of g defined as the minimum of two concave functions g_1 and g_2 that both have a maximum. On the left, the situation where the maximum of g_1 (resp. g_2) is reached for some x_0 with $g_2(x_0) \geq g_1(x_0)$ (resp. $g_1(x_0) \geq g_2(x_0)$); in this case, the maximum of $g = \min(g_1, g_2)$ is $g(x_0) = g_1(x_0)$. On the right, the case were the maxima of both g_1 and g_2 do not satisfy the previous condition; the maximum of g is then located at the intersection of both curves. Note that when dealing with concave functions of several variables, this intersection is not reduced to a single point, but it still contains the maximum.

It is well known [5] that $H(\mathbf{x})$, the classical entropy function on $\mathcal{P}$, has its maximum for the uniform distribution $\overline{\mathbf{x}} = (\frac{1}{k}, \ldots, \frac{1}{k})$, with $H(\overline{\mathbf{x}}) = \log k$. Moreover, by Gibbs' inequality [5], $K_\ell(\mathbf{x})$ reaches its maximum on $\mathcal{P}$ at $\mathbf{p}$. Hence $\overline{\mathbf{x}}$ and $\mathbf{p}$ are two candidates for the first case of Fig. 1, corresponding to cases (a) and (c) of Theorem 1, respectively. Case (b) corresponds to the second case of Fig. 1, when $K_\ell(\overline{\mathbf{x}}) < H(\overline{\mathbf{x}})$ and $H(\mathbf{p}) < K_\ell(\mathbf{p})$: the maximum of $G_\ell(\mathbf{x})$ is reached for a value $\mathbf{x}$ such that $H(\mathbf{x}) = K_\ell(\mathbf{x})$, *i.e.* on the hyperplane of equation $\sum_{i\in[k]} x_i \log p_i = -\frac{1}{\ell}$. Standard techniques for multivariate differentiable functions yield that the maximum is located at $\mathbf{x}_c$ given in Theorem 1.

To turn these informal steps into a full proof of our main statement, we also need to prove that $U_{N,L}$ resemble the frequency vector that contributes the most in expectation, to $|\mathcal{S}_{N,L} \cap \mathcal{W}_L(\mathbf{x})|$, *i.e.* that the distribution of $\mathbf{freq}(U_{N,L})$ is concentrated around $\mathbf{x}_c$.

4 Proof of Theorem 1

4.1 Preliminary results

The following lemma establishes some simple bounds for the cardinality of $\mathcal{W}_L(\mathbf{x})$, justifying the rough estimate of $N^{\ell H(\mathbf{x})}$ discussed in Section 3.2.

Lemma 1. *There exists a positive real constant α such that for all $\mathbf{x} \in \mathcal{P}$,*

$$|\mathcal{W}_L(\mathbf{x})| \leq \alpha e^{L\,H(\mathbf{x})} = \alpha N^{\ell\,H(\mathbf{x})}. \tag{1}$$

There exists a positive real constant β such that for all $\mathbf{x} \in \tilde{\mathcal{P}}$, such that for all $\mathbf{y} \in \overline{B}(\mathbf{x}, \frac{k}{L}) \cap \mathcal{F}_L$ we have

$$|\mathcal{W}_L(\mathbf{y})| \geq \beta L^{-\frac{k-1}{2}} e^{L\,H(\mathbf{x})} = \beta L^{\frac{k-1}{2}} N^{\ell\,H(\mathbf{x})}. \tag{2}$$

Moreover $\overline{B}(\mathbf{x}, \frac{k}{L}) \cap \mathcal{F}_L$ is not empty.

Lemma 2 below will be used to prove that sufficiently many words of $\mathcal{S}_{N,L}$ have a (well chosen) frequency vector $\mathbf{y}$. It is proved using the study of negatively associated random variables by Dubhashi and Ranjan [2].

Lemma 2. *Let $\mathbf{y} \in \mathcal{F}_L$ such that $q_{N,L}(\mathbf{y}) \geq \gamma$, for some $\gamma > 0$. Then the following inequality holds:*

$$\mathbb{P}\left(|\mathcal{W}_L(\mathbf{y}) \cap \mathcal{S}_{N,L}| \leq \frac{\gamma}{2}|\mathcal{W}_L(\mathbf{y})|\right) \leq \exp\left(-\frac{\gamma^2|\mathcal{W}_L(\mathbf{y})|}{2}\right).$$

For $\mathbf{x} \in \mathcal{P}$, let $\mathcal{B}_L(\mathbf{x})$ be the set of probability vectors that are far from $\mathbf{x}$:

$$\mathcal{B}_L(\mathbf{x}) = \left\{\mathbf{y} \in \mathcal{P} :\ \|\mathbf{y} - \mathbf{x}\| \geq \frac{\log L}{\sqrt{L}}\right\}.$$

Proposition 1 is our main tool for proving Theorem 1. The technical conditions can be seen as follows. Condition (1) is used to obtain an upper bound on the number of elements of $\mathcal{S}_{N,L}$ whose frequency vectors are in $\mathcal{B}_L$, which holds with very high probability (using the Markov inequality from the bound on the expectation). Condition (2) ensures that with high probability we have a lot of elements of $\mathcal{S}_{N,L}$ whose frequency vectors are not in $\mathcal{B}_L$; the precise statement of the condition is chosen to fit the formula within the exponential in Lemma 2.

Proposition 1. *Let ℓ^- and ℓ^+ in $\mathbb{R} \cup \{-\infty, +\infty\}$ such that $\ell^- < \ell^+$. Let L_0 be a sufficiently large integer, and $\mathbf{x} \in \tilde{\mathcal{P}}$. Assume that there exist two positive constants λ_1 and λ_2, such that, for any $L \geq L_0$ and $N \geq 2$ for which $\ell = L/\log(N)$ satisfies $\ell^- \leq \ell \leq \ell^+$; then for any $\mathbf{y} \in \overline{B}(\mathbf{x}, k/L)$ the following two conditions hold:*

(1) $\mathbb{E}|\mathcal{W}_L(\mathcal{B}_L(\mathbf{x})) \cap \mathcal{S}_{N,L}| \leq L^{-\lambda_1 \log L} N^{\ell H(\mathbf{x})} q_{N,L}(\mathbf{y})$;
(2) $N^{\ell H(\mathbf{x})} q_{N,L}(\mathbf{y})^2 \geq L^{\lambda_2 \log L}$.

Then, there exists $\lambda > 0$ such that $\mathbb{P}\left(\|U_{N,L} - \mathbf{x}\| \geq \frac{\log L}{\sqrt{L}}\right) \leq L^{-\lambda \log L}$ holds for any (L, N) satisfying the same conditions.

Lemma 3 will be used to prove that Condition (1) of Proposition 1 holds in certain cases. Its proof heavily relies on concavity in order to extend a local bound globally.

Lemma 3. *Let f be a concave continuous function on a convex domain $\mathcal{C} \subseteq \mathbb{R}^k$ that has a maximum on $\mathcal{C}$ at $\mathbf{y}$. Assume furthermore that there exist two positive real constants ρ and η such that for every $\mathbf{x} \in \overline{B}(\mathbf{y}, \rho) \cap \mathcal{C}$ the following inequality holds:*

$$f(\mathbf{x}) \leq f(\mathbf{y}) - \eta\|\mathbf{x} - \mathbf{y}\|^2.$$

Then for every positive real $r \leq \rho$, for every $\mathbf{x} \in \mathcal{C}$ such that $\|\mathbf{x} - \mathbf{y}\| > r$ we have $f(\mathbf{x}) \leq f(\mathbf{y}) - \eta\, r^2$.

Finally, we will use the following bounds for $p_L(\mathbf{y})$ and $q_{N,L}(\mathbf{y})$ in terms of $p_L(\mathbf{x})$, which are obtained using basic computations.

Lemma 4. *Let $\mathbf{x} \in \tilde{\mathcal{P}}$ and let $\mathbf{y} \in \overline{B}(\mathbf{x}, k/L) \cap \mathcal{P}$. There exist two positive constants κ_1 and κ_2 such that $\kappa_1 p_L(\mathbf{x}) \leq p_L(\mathbf{y}) \leq 2\kappa_2 p_L(\mathbf{x})$ and*

$$1 - \exp(-\kappa_1 N\, p_L(\mathbf{x})) \leq q_{N,L}(\mathbf{y}) \leq 1 - \exp(-\kappa_2 N\, p_L(\mathbf{x})).$$

4.2 Proof for range (a): $\ell \leq \ell_0$

In this section, we are in the case where $\ell \leq \ell_0 = \frac{-k}{\sum_{i\in[k]} \log p_i}$. Our goal is to apply Proposition 1 for $\mathbf{x} = \overline{\mathbf{x}} = (\frac{1}{k}, \ldots, \frac{1}{k})$, with $\ell^- = -\infty$ and $\ell^+ = \ell_0$. Observe that $N^{\ell H(\overline{\mathbf{x}})} = k^L$ and that, by Lemma 4, for any $\mathbf{y} \in \overline{B}(\mathbf{x}, k/L)$ we have

$$q_{N,L}(\mathbf{y}) \geq 1 - \exp(-\kappa_1 N p_L(\mathbf{x})) = 1 - \exp\left(-\kappa_1 N^{1+\frac{\ell}{k}\sum_{i\in[k]} \log p_i}\right).$$

As $\ell \leq \ell_0$, we have $1 + \frac{\ell}{k}\sum_{i\in[k]} \log p_i \geq 0$ and thus $q_{N,L}(\mathbf{y}) \geq 1 - e^{-\kappa_1}$.

To verify Condition (1) of Proposition 1, we rely on the following result on the entropy function H, which can be obtained by standard techniques of analysis in several variables:

Lemma 5. *The exists a neighborhood $\mathcal{V}_{\overline{\mathbf{x}}}$ of $\overline{\mathbf{x}} = (\frac{1}{k}, \ldots, \frac{1}{k})$ such that, for every $\mathbf{x} \in \mathcal{P} \cap \mathcal{V}_{\overline{\mathbf{x}}}$ the following inequalities hold:*

$$\log k - k\|\mathbf{x} - \overline{\mathbf{x}}\|^2 \leq H(\mathbf{x}) \leq \log k - \frac{k\|\mathbf{x} - \overline{\mathbf{x}}\|^2}{3}. \tag{3}$$

When L is sufficiently large, $\overline{B}(\overline{\mathbf{x}}, \log L/\sqrt{L}) \subseteq \mathcal{V}_{\overline{\mathbf{x}}}$ of Lemma 5. So we can apply Lemma 3 with $\mathbf{y} = \overline{\mathbf{x}}$ and $r = \log L/\sqrt{L}$, to obtain that for every $\mathbf{x} \in \mathcal{B}_L(\overline{\mathbf{x}})$ we have

$$H(\mathbf{x}) \leq H(\overline{\mathbf{x}}) - \frac{\eta \log^2 L}{L} = \log k - \frac{k \log^2 L}{3L}.$$

By Lemma 1, we can bound the number of words in $\mathcal{W}_L(\mathbf{x})$ for $\mathbf{x} \in \mathcal{B}_L(\overline{\mathbf{x}})$:

$$|\mathcal{W}_L(\mathbf{x})| \leq \alpha N^{\ell\, H(\mathbf{x})} \leq \alpha N^{\ell \log k - \ell \frac{k \log^2 L}{3L}} = \alpha k^L\, L^{-k \log L/3}.$$

Observe that $|\mathcal{B}_L \cap \mathcal{F}_L| \leq |\mathcal{F}_L| \leq L^k$, since there are at most L^k compositions of letters for words of A^L. Therefore,

$$\mathbb{E}|\mathcal{W}_L(\mathcal{B}_L(\overline{\mathbf{x}})) \cap \mathcal{S}_{N,L}| \leq |\mathcal{W}_L(\mathcal{B}_L(\overline{\mathbf{x}})) \cap \mathcal{F}_L| \leq \alpha L^k\, k^L\, L^{-k \log L/3}.$$

This proves that Condition (1) holds in our case, since for any $\lambda_1 < \frac{k}{3}$, the right term is at most $q_{N,L}(\mathbf{y}) k^L L^{-\lambda_1 \log L}$, as we saw that $q_{N,L}(\mathbf{y}) \geq 1 - e^{-\kappa_1}$ at the beginning of the section.

Condition (2) trivially holds as we have, for any $\lambda_2 > 0$:

$$N^{\ell H(\mathbf{x})} q_{N,L}(\mathbf{y})^2 \geq (1 - e^{-\kappa_1})^2 k^L \geq L^{\lambda_2 \log L}.$$

We can therefore apply Proposition 1, concluding the proof for range (a).

4.3 Proof for range (b): $\ell_0 \leq \ell \leq \ell_1$

In this section, we are in the case where $\frac{-k}{\sum_{i\in[k]} \log p_i} \leq \ell \leq \frac{1}{H(\mathbf{p})}$. Our goal is still to apply Proposition 1, but we first need to find the vector $\mathbf{x}$ that concentrates the frequency vectors of the output of our process.

Recall that $K_\ell(\mathbf{x}) = H(\mathbf{x}) + \frac{1}{\ell} + \sum_{i\in[k]} x_i \log p_i$. We have the following bound on the expected number of words of given frequency vectors that appear in $\mathcal{S}_{N,L}$. It is obtained by linearity of the expectation, and by obtaining bounds on $q_{N,L}(\mathbf{x})$ as in Lemma 4.

Lemma 6. *Let $\mathbf{x} \in \mathcal{P}$. The expected cardinality of $\mathcal{S}_{N,L} \cap \mathcal{W}_L(\mathbf{x})$ satisfies*

$$\mathbb{E}\,|\mathcal{S}_{N,L} \cap \mathcal{W}_L(\mathbf{x})| \leq 2\alpha N^{\ell \min(H(\mathbf{x}), K_\ell(\mathbf{x}))},$$

with the same α as in Lemma 1.

Recall that $G_\ell(\mathbf{x}) = \min(H(\mathbf{x}), K_\ell(\mathbf{x}))$, so we can rewrite the bound in Lemma 6 into $2\alpha N^{G_\ell(\mathbf{x})}$. We now study $G_\ell(\mathbf{x})$ for the range corresponding to case (b), *i.e.* $\ell_0 \leq \ell \leq \ell_1$. Recall also that Φ is the mapping defined by $\Phi(t) = \sum_{i\in[k]} p_i^t$.

Lemma 7. *For any ℓ such that $\ell_0 \leq \ell \leq \ell_1$, as a function on $\mathcal{P}$, the function $G_\ell(\mathbf{x})$ admits a unique maximum at $\mathbf{x}_c = \frac{1}{\Phi(c)}(p_1^c, \ldots, p_k^c)$, where $c \in [0,1]$ is the unique solution of $\ell\Phi'(c) + \Phi(c) = 0$. Moreover, $\mathbf{x}_c$ is in the hyperplane defined by the equation $H(\mathbf{x}) = K_\ell(\mathbf{x})$.*

Now that the maximum $\mathbf{x}_c$ is located, we want to provide an upper bound the expect cardinality of $\mathcal{B}_L(\mathbf{x}_c) \cap \mathcal{S}_{N,L}$, to fulfill Condition (1) of Proposition 1. For this, we want to use Lemma 3, and therefore need an upper bound of $G_\ell(\mathbf{x})$ around its maximum $G_\ell(\mathbf{x}_c)$. This is the purpose of Lemma 8 below, whose proof relies on classical analysis of functions of several variables.

Lemma 8. *Let c be the unique solution of $\ell\Phi'(c) + \Phi(c) = 0$. There exists a real constant $\rho > 0$ such that for every $\mathbf{x} \in \overline{B}(\mathbf{x}_c, \rho) \cap \mathcal{P}$, $G_\ell(\mathbf{x}) \leq G_\ell(\mathbf{x}_c) - \|\mathbf{x} - \mathbf{x}_c\|^2$.*

We will also need the following technical lemma, to prove that the cardinality of $\mathcal{W}_L(\mathbf{x}_c)$, which is roughly $N^{\ell H(\mathbf{x}_c)}$, is at least some power of N. Its proof consists in studying $\ell \mapsto \ell H(\mathbf{x}_c)$, where c is viewed as a function of ℓ given by the implicit solution of $\ell\Phi'(c) + \Phi(c) = 0$.

Lemma 9. *For ℓ within range (b), let c be the solution of $\ell\Phi'(c) + \Phi(c) = 0$, the quantity $\ell H(\mathbf{x}_c)$ satisfies the following inequalities:*

$$0 < d \leq \ell H(\mathbf{x}_c) \leq 1, \text{ with } d = \frac{k \log k}{-\sum_{i\in[k]} \log p_i}.$$

Since $H(\mathbf{x}_c) = K_\ell(\mathbf{x}_c)$, we have $N^{\ell H(\mathbf{x}_c)} = N^{\ell G_\ell(\mathbf{x}_c)}$. Moreover, by Lemma 4, for any $\mathbf{y} \in \overline{B}(\mathbf{x}_c, k/L)$ we have

$$q_{N,L}(\mathbf{y}) \geq 1 - \exp(-\kappa_1 N p_L(\mathbf{x}_c)) = 1 - \exp\left(-\kappa_1 N^{1+\ell \sum_{i\in[k]} \frac{p_i^c}{\Phi(c)} \log p_i}\right).$$

But $\ell \sum_{i\in[k]} \frac{p_i^c}{\Phi(c)} \log p_i = \ell \frac{\Phi'(c)}{\Phi(c)} = -1$. Therefore, $q_{N,L}(\mathbf{y}) \geq 1 - e^{-\kappa_1}$.

We can now prove that Condition (1) holds. By Lemma 8 and Lemma 3, when L is sufficiently large, for all $\mathbf{x} \in \mathcal{B}_L(\mathbf{x}_c)$ we have $G_\ell(\mathbf{x}) \leq G_\ell(\mathbf{x}_c) - \log^2 L/L$. Thus, by Lemma 6 we have

$$\mathbb{E}|\mathcal{W}_L(\mathbf{x}) \cap \mathcal{S}_{N,L}| \leq 2\alpha N^{\ell G_\ell(\mathbf{x})} \leq 2\alpha N^{\ell G_\ell(\mathbf{x}_c) - \ell \log^2 L/L} = 2\alpha N^{\ell H(\mathbf{x}_c)} L^{-\log L}.$$

Since $|\mathcal{B}_L(\mathbf{x}_c) \cap \mathcal{S}_{N,L}| \leq |\mathcal{F}_L| \leq L^k$, by linearity of the expectation we have

$$\mathbb{E}|\mathcal{W}_L(\mathcal{B}_L(\mathbf{x}_c)) \cap \mathcal{S}_{N,L}| \leq 2\alpha L^k N^{\ell H(\mathbf{x}_c)} L^{-\log L}.$$

As $q_{N,L}(\mathbf{y}) \geq 1 - e^{-\kappa_1}$, Condition (1) holds since

$$\mathbb{E}|\mathcal{W}_L(\mathcal{B}_L(\mathbf{x}_c)) \cap \mathcal{S}_{N,L}| \leq N^{\ell H(\mathbf{x}_c)} q_{N,L}(\mathbf{y}) L^{-\frac{1}{2}\log L}.$$

Condition (2) also holds: by Lemma 9, we have

$$N^{\ell H(\mathbf{x}_c)} q_{N,L}(\mathbf{y})^2 \geq N^d (1 - e^{-\kappa_1})^2 \leq L^{-\log L},$$

since the condition $\ell \leq \ell_1$ implies that N grows exponentially in L. Therefore, we can apply Proposition 1, concluding the proof for range (b).

4.4 Proof for range (c): $\ell \geq \ell_1$

For this range we cannot only rely on Proposition 1, as when ℓ is very large, N is small towards $\log L$, or even towards L, and the conditions do not hold anymore. We therefore split the proof into several subranges for ℓ.

To provide an upper bound for the expected cardinality of $|\mathcal{B}_L(\mathbf{p}) \cap \mathcal{S}_{N,L}|$, we use the classical large deviation results for the multinomial distribution [7, p. 462] to obtain the following lemma:

Lemma 10. *The following inequality holds:* $\mathbb{P}_L\left(\mathcal{B}_L(\mathbf{p})\right) \leq 2^k\, L^{-\frac{1}{2}\log L}$.

As a consequence, since we generate N random words with the source,

$$\mathbb{E}|\mathcal{B}_L(\mathbf{p}) \cap \mathcal{S}_{N,L}| \leq 2^k N\, L^{-\frac{1}{2}\log L}.$$

Observe also that $p_L(\mathbf{p}) = N^{-\ell H(\mathbf{p})}$, thus by Lemma 4, for $\mathbf{y} \in \mathcal{B}_L(\mathbf{p}, k/L)$,

$$q_{N,L}(\mathbf{y}) \geq 1 - \exp(-\kappa_1 N p_L(\mathbf{p})) = 1 - \exp\left(-\kappa_1 N^{1-\ell H(\mathbf{p})}\right).$$

As $e^{-t} \leq 1 - \frac{t}{2}$ for $t \in [0,1]$, we have $q_{N,L}(\mathbf{y}) \geq \frac{\kappa_1}{2} N^{1-\ell H(\mathbf{p})}$. Thus we have $N^{\ell H(\mathbf{p})} q_{N,L}(\mathbf{y}) \geq \frac{\kappa_1}{2} N$ and Condition (1) of Proposition 1 holds for any $\lambda_1 \in (0, \frac{1}{2})$. Unfortunately, Condition (2) does not always hold, and we have to change the proof when ℓ is too large.

▶ **case $\frac{1}{H(\mathbf{p})} \leq \ell \leq \frac{3}{2H(\mathbf{p})}$:** in this case, Condition (2) holds. Indeed,

$$N^{\ell H(\mathbf{p})} q_{N,L}(\mathbf{y})^2 \geq \frac{\kappa_1}{2} N^{2-\ell H(\mathbf{p})} \geq \frac{\kappa_1}{2}\sqrt{N}.$$

But the condition $\ell \leq \frac{3}{2H(\mathbf{p})}$ implies that N grows exponentially in N. Hence, Proposition 1 applies and the result holds for this subrange.

▶ **case $\frac{3}{2H(\mathbf{p})} \leq \ell \leq \ell_2$:** where $\ell_2 = \frac{2}{-\log \lambda}$, for some $\lambda \in (p_{\max}, 1)$. To complete the proof, we have to establish that $|\mathcal{S}_{N,L}|$ is large with very high probability. By Lemma 1, there exists $\mathbf{y} \in \overline{B}(\mathbf{p}, \frac{1}{L})$ such that $|\mathcal{W}_L(\mathbf{y})| \geq \beta L^{(1-k)/2} N^{\ell H(\mathbf{p})}$. By Lemma 4, $p_L(\mathbf{y}) \geq \kappa_1 N^{\ell H(\mathbf{p})}$ and therefore $\mathbb{P}_L(\mathcal{W}_L(\mathbf{y})) \geq \kappa L^{(1-k)/2}$, with $\kappa = \kappa_1 \beta$.

We now consider the process from the start, when the N words of length L are repeatedly generated by the source. Let $Y_{N,L}$ be the random variable that counts the number of words of $\mathcal{W}_L(\mathbf{y})$ generated during this process. The random variable $Y_{N,L}$ is distributed as a binomial distribution of coefficients N and $\mathbb{P}_L(\mathcal{W}_L(\mathbf{y}))$: $\mathbb{E}[Y_{N,L}] = N\mathbb{P}_L(\mathcal{W}_L(\mathbf{y}))$ and there is concentration around the mean, by Chernoff-Hoeffding inequality:

$$\mathbb{P}\left(Y_{N,L} \leq \frac{\kappa}{2} L^{(1-k)/2} N\right) \leq \exp\left(\frac{-\kappa^2}{2} L^{1-k} N\right) \leq L^{-\log L}, \tag{4}$$

for L sufficiently large, since N is exponential in L when $\ell \leq \ell_2$. In the process of repeatedly generating N words, a word u is called a $\mathbf{y}$-*duplicate* if $\mathbf{freq}(u) = \mathbf{y}$ and u has already been generated. Let $D_{N,L}$ be the random variable that counts the number of $\mathbf{y}$-duplicates. We have $|\mathcal{S}_{N,L} \cap \mathcal{W}_L(\mathbf{y})| = Y_{N,L} - D_{N,L}$. If we only look at what happens inside $\mathcal{W}_L(\mathbf{y})$, we are considering the process of choosing $Y_{N,L}$ times an element of $\mathcal{W}_L(\mathbf{y})$ uniformly at random, as they all have the same probability of being generated. We will use the following classical lemma.

Lemma 11. *Let E be a set of cardinality $n \geq 1$. Let $D_E(n)$ denote the number of duplicates obtained when generating m times an element of E uniformly at random. Then*

$$\mathbb{E}[D_E(n)] \leq \frac{m^2}{2n}.$$

Since $Y_{N,L} \leq N$, the expected number of duplicates in $\mathcal{S}_{N,L}$ satisfies

$$\mathbb{E}[D_{N,L}] \leq \frac{N^2}{2|\mathcal{W}_L(\mathbf{y})|} \leq \frac{L^{(k-1)/2}}{2\beta} N^{2-\ell H(\mathbf{p})} \leq \frac{L^{(k-1)/2}}{2\beta}\sqrt{N}.$$

By Markov inequality, as N is exponential in L, we get that for any positive ν

$$\mathbb{P}\left(D_{N,L} \geq L^{2\nu \log L}\sqrt{N}\right) \leq \mathbb{P}\left(D_{N,L} \geq \frac{L^{(k-1)/2}}{2\beta} L^{\nu \log L}\sqrt{N}\right) \leq L^{-\nu \log L}. \tag{5}$$

Equation 4 and Equation 5 ensure by the union bound that

$$\mathbb{P}\left(Y_{N,L} - D_{N,L} \leq \frac{\kappa}{2} L^{(1-k)/2} N - L^{2\nu \log L}\sqrt{N}\right) \leq L^{-\log L} + L^{-\nu \log L}.$$

Recall that $Y_{N,L} - D_{N,L} = |\mathcal{W}(\mathbf{y}) \cap \mathcal{S}_{N,L}|$. As N is exponential in L, for any positive constant μ we have

$$\mathbb{P}\left(|\mathcal{W}(\mathbf{y}) \cap \mathcal{S}_{N,L}| \leq L^{-\mu \log L} N\right) \leq 2L^{-\nu \log L}.$$

Such a precise estimation of the cardinality of $\mathcal{W}(\mathbf{y}) \cap \mathcal{S}_{N,L}$ can be used as a substitute for Condition (2) of Proposition 1, concluding the proof for this subrange.

▶ **case $\ell \geq \ell_2$:** In this case, we simply rely on this classical result deduced from the Birthday Paradox problem, to establish that with very high probability, there are no duplicates in $\mathcal{S}_{N,L}$.

Lemma 12. *If $\ell \geq \ell_2$, then the probability that $|\mathcal{S}_{N,L}| < N$ is exponentially small in L.*

As a consequence, this proves our result for $\ell \geq \ell_2$: if the N elements generated by the source are pairwise distinct, then $U_{N,L}$ is distributed as a random element of the source. By Lemma 12, this happen with probability at least $1 - L^{-\log L}$.

5 Conclusions

In this article, we exhibited two thresholds for the typical frequency vector of a word in the partial word collector problem, for a memoryless source of parameter $\mathbf{p}$. We were able to establish that with high probability, it resembles either the uniform distribution, or a word generated by the source, or some vector in between, which we fully characterize.

We want to make the observation that, though two words with the same frequency vector have the same probability to be the result $U_{N,L}$ of our process, $U_{N,L}$ is not distributed as the output of a memoryless source: the computations for 3 words of length 2 on $\{a, b\}$ show that the probability that $U_{N,L}$ ends by a is not the same if we condition by starting by a or by b.

A natural continuation of this work, is to study the case where the N words are not independent anymore, but are the factors of length L of a random word of length $N + L - 1$. As shown for instance in [4], the correlation between the factors is small, so we expect a similar result, even if it is still ongoing work.

Another possible extension would be to study a similar word collection process when words are not produced by a memoryless source; say, using a Markov source instead. One would expect at least equivalents of regimes (a) and (c) in our theorem, though regime (c) could have asymptotic frequencies that are not uniform (as the support of the distribution may not be A^L); and any intermediate regime(s) could be much more difficult to determine.

Acknowledgments: the authors are grateful to Arnaud Carayol for his precious help when preparing this article, and an anonymous reviewer for suggesting the promising alternative α-parametrization of the problem.

References

1. J. Du Boisberranger, D. Gardy, and Y. Ponty. The weighted words collector. In *AOFA - 23rd International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms - 2012*, pages 243–264. DMTCS, 2012.
2. D. Dubhashi and D. Ranjan. Balls and Bins: A Study in Negative Dependence. *Random Struct. Algorithms*, 13(2):99–124, Sept. 1998.
3. P. Duchon, C. Nicaud, and C. Pivoteau. Gapped pattern statistics. In J. Kärkkäinen, J. Radoszewski, and W. Rytter, editors, *28th Annual Symposium on Combinatorial Pattern Matching, CPM 2017, July 4-6, 2017, Warsaw, Poland*, volume 78 of *LIPIcs*, pages 21:1–21:12. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017.
4. I. Gheorghiciuc and M. D. Ward. On Correlation Polynomials and Subword Complexity. *Discrete Mathematics & Theoretical Computer Science*, DMTCS Proceedings vol. AH, 2007 Conference on Analysis of Algorithms (AofA 07), Jan. 2007.
5. D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.
6. M. Rubinchik and A. M. Shur. The number of distinct subpalindromes in random words. *Fundam. Inform.*, 145(3):371–384, 2016.
7. A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996.