The weighted words collector

Jérémie Du Boisberranger, Danièle Gardy, Yann Ponty

THE WEIGHTED WORDS COLLECTOR

JÉRÉMIE DU BOISBERRANGER†, DANIÈLE GARDY†, AND YANN PONTY★

Abstract. Motivated by applications in bioinformatics, we consider the word collector problem, i.e. the expected number of calls to a random weighted generator of words of length n before the full collection is obtained. The originality of this instance of the non-uniform coupon collector lies in the, potentially large, multiplicity of the words/coupons of a given probability/composition. We obtain a general theorem that gives an asymptotic equivalent for the expected waiting time of a general version of the Coupon Collector. This theorem is especially well-suited for classes of coupons featuring high multiplicities. Its application to a given language essentially necessitates some knowledge on the number of words of a given composition/probability. We illustrate the application of our theorem, in a step-by-step fashion, on three exemplary languages, revealing asymptotic regimes in $\Theta(\mu(n) \cdot n)$ and $\Theta(\mu(n) \cdot \log n)$, where $\mu(n)$ is the sum of weights over words of length n.

1. Introduction

The choice of a suitable random model for the input instances of an algorithm is critical for its analysis. In an attempt to capture non-uniform distributions naturally arising in real-life data, Denise *et al* [6, 5] studied **weighted languages**, a natural generalization of context-free languages [11] where atomic weights are associated to each letter. The weight of a word is then simply the product of its letters' own weight. This naturally induces a probability distribution over the class of words of a given length n, where the probability of any given word is proportional to its weight. Aside from arguably being the simplest non-uniform generalization of combinatorial classes, such distributions naturally arise in statistical physics (Boltzmann partition function), with direct applications in algorithm design (Monte-Carlo Markov Chains) and bioinformatics [13]. Random generation algorithms were also proposed for these distributions [5], leading to an efficient multidimensional generalization of Boltzmann sampling [3].

In the field of Bioinformatics, RNA folding has been one of the leading problems of the past three decades. Given an RNA sequence of length n, composed of four types of nucleotides $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$, the goal is to predict the **secondary structure**, a non-crossing subset of experimentally-determined base-pairs (hydrogen bonds). This coarse-grain representation of the 3D conformation of RNA molecules has been extensively studied from a combinatorial perspective [18, 17]. A statistical sampling approach proposed by Ding and Lawrence [7] is one of the leading methods for tackling this problem. At the core of this method, one makes repeated calls to a random generation algorithm, generating secondary structures with probability proportional to their Boltzmann factor. Unfortunately, such a redundancy is arguably uninformative when the probability of each conformation can be exactly and efficiently estimated after each generation. One can thus interpret this redundancy as a degradation of the algorithm performance, and analyze the expected time-complexity of generating k *distinct* conformations. In the worst-case scenario, the targeted number k of secondary structures is the total number of secondary structures. Since energy-weighted secondary structures are in bijection with weighted *peakless*-Motzkin words, then the worst-case/average-case (resp. on k and n the length) complexity of the algorithm is exactly the waiting-time of completing the class of weighted Motzkin words of length n.

Generalizing on this question, the central problem addressed by this article is that of the **Weighted Words Collector**: Given a language $\mathcal{L}$ and a word length n, how many calls to a weighted generation algorithm must be made before all the words of length n are obtained? This problem is clearly a weighted instance of the ubiquitous **Coupon Collector problem** which, given a finite collection of m objects produced by a random source, studies the expected time $E[C_m]$ of the full collection, i.e. the expected number of generations before each object is generated at least once. This problem

naturally arises in a large variety of context, including the analysis of database [2] and network [12] probabilistic algorithms. In the specific context of weighted languages, the two main specificities are the non-uniform nature of the weighted distribution and the potentially large multiplicity of coupons.

In the uniform distribution, either probabilistic or combinatorial arguments can be used to establish that $E[C_m] = m \cdot \mathcal{H}(m) \in \Theta(m \log m)$, where $\mathcal{H}(m) = \sum_{i \geq 1} 1/i$ is the m-th harmonic number. For general distributions,where the i-th object is generated with probability p_i, Flajolet, Gardy and Thimonier [9] gave a general expression for the waiting time of the full collection:

$$E[C_m] = \int_0^\infty \left(1 - \prod_{i=1}^{m} \left(1 - e^{-p_i t}\right)\right) dt. \tag{1.1}$$

However, specializing this formula for a given probability distribution seldom leads to spectacular simplifications, and the derivation of asymptotic estimates for parameterized families of objects usually remains challenging. To overcome this limitation, many efforts have focused on providing closed-form approximations [2], asymptotic equivalents [4, 14] and algorithms for computing the waiting time over non-uniform distributions of diverse degrees of generality. Unfortunately, as will be discussed in Section 3, these results either fail to apply to classes of coupons of high multiplicity, lead to bounds on the asymptotic behavior that are not tight, or require extensive *a priori* knowledge on the distribution, motivating further studies in the context of languages.

After a brief introduction, this extended abstract states, in Section 2, a general theorem for weighted families of coupons. More precisely, Theorem 2.1 relates the asymptotic behavior of a general Weighted Coupon Collector Problem to the multiplicity and weight of the i-th class of coupons. Section 3 compares the scope of the theorem with previous works addressing a similar problem. Section 4 develops a methodology to ease the verification of the conditions of Theorem 2.1 in the case of context-free languages, and applies it on illustrative examples. Finally, we conclude in Section 5, summarizing the contribution and describing future developments.

2. A general theorem for coupons of large multiplicities

2.1. **Definitions and notations.** Given a sequence $\{w_i\}_{i=1}^m$ of positive numbers, or **weights**, associated with a collection of m objects, one defines a **weighted probability distribution** $\{p_i\}_{i=1}^m$ over the collection as:

$$p_i = \frac{w_i}{\mu(m)}, \forall\ i \leq m \qquad \text{where} \qquad \mu(m) = \sum_{i=1}^{m} w_i.$$

In this work, we are interested in distributions with high multiplicity, in the sense that many coupons may share the same weight/probability. Let us then introduce $\mathbf{W}_m = \{W_{m,i}\}_i$ the increasing finite sequence that contains all distinct weights in $\mathbf{w}$. Finally, for each $i \in [1, |\mathbf{W}_m|]$, let us denote by $M_{m,i}$ the number of coupons of weight $W_{m,i}$. Let us finally observe that:

$$m = \sum_{i=1}^{|\mathbf{W}_m|} M_{m,i} \qquad \text{and} \qquad \mu(m) = \sum_{i=1}^{|\mathbf{W}_m|} M_{m,i} W_{m,i}.$$

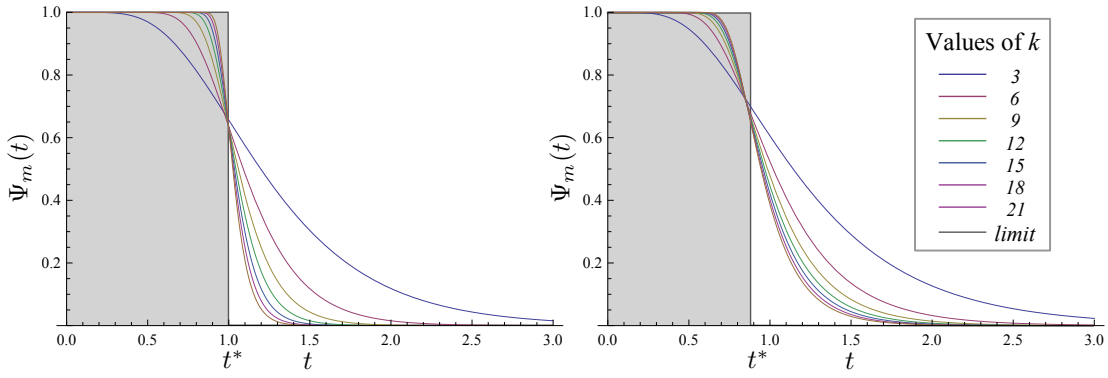
2.2. **Main result.** Our main result gives an asymptotical expression for the expected time to get a full collection of m objects when $m \to \infty$. Since the accessible weights $\mathbf{W}_m$ and their multiplicity $\mathbf{M}_m$ can drastically vary for different values of m, let us start by restricting their asymptotic behavior:

H1. *Separability.* There exist functions $f_1, ..., f_p$ ($F := f_1$), $g_1, ..., g_p$ ($G := g_1$), h and H, for some $p > 0$, such that

$$M_{m,i} \underset{m \to \infty}{\sim} \frac{e^{\sum_{j=1}^{p} f_j(i) g_j(m)}}{h(i)}$$

and for all $m \geq 1$ and $i \leq |\mathbf{W}_m|$,

$$M_{m,i} \leq \frac{e^{\sum_{j=1}^{p} f_j(i) g_j(m)}}{H(i)},$$



Uniform distribution Weighted distribution

FIGURE 1. Plots of the $\Psi_m(t)$ functions as appearing for a uniform (Left) and weighted (Right, $\pi(a)/\pi(b) = 2/3$) distribution over the $(a+b)^*$ language. We consider $m = 2^k$ coupons/words, for several values of $k \in \{3, 6, 9, 12, 15, 18, 21\}$. The convergence of $\Psi_m(t)$ towards a step function when $m \to \infty$ featuring a transition at t^* ($t^* = 1$ in the uniform distribution and $t^* = 8/9$ in the weighted one) is crucial to our approach.

where F is positive, non zero everywhere, and for all $j \leq p$, $g_j(m) = o(G(m))$, when $m \to \infty$, and $G(m) \to +\infty$. Moreover, $\sum_{i \in [1, |\mathbf{W}_m|]} \frac{1}{H(i)}$ converges.

H2. *Regularity.* There exist functions $c > 0$ and $\omega > 0$ such that $\forall k > 0$, $\exists m_k > 0$ such that $\forall m \geq m_k$, $\forall i \leq k$,

$$W_{m,i} = c(i) \cdot \omega(m)$$

and $W_{m,i} \geq c(i) \cdot \omega(m)$ for all $m, i \geq 1$.

H3. *Sufficient growth.* If $|\mathbf{W}_m| \underset{m \to \infty}{\to} \infty$ then one has

$$\lim_{i \to \infty} \frac{c(i)}{f_j(i)} = +\infty, \forall j \leq p.$$

Theorem 2.1. *Assume that, for all $m > 0$, the weights $\boldsymbol{W}_m$ and multiplicities $\boldsymbol{M}_m$ of the coupon collection satisfy conditions **H1**, **H2**, and **H3**. Then, as $m \to \infty$, one has:*

$$E[C_m] \sim t^*(F, c) \frac{\mu(m)}{\omega(m)} G(m) \tag{2.1}$$

where:

- *$\mu(m)$ is the total weight of all coupons*
- *F and G (cf **H1**) define the leading term of the growth of $M_{m,i}$ as $m \to \infty$*
- *$\omega(m)$ is, up to a constant factor, the smallest weight within the collection (cf **H2**)*
- *$t^*(F, c)$ is the largest value of t such that there exists $x \in \mathbb{N}$ such that $F(x) - t \cdot c(x) > 0$.*

Sketch of proof.

Page-length restrictions lead us to omit the full proof of this theorem in this extended abstract. However let us sketch the proof strategy.

Applying the substitution $u \frac{\omega(m)}{\mu_m \sum_{j=1}^{p} g_j(m)} \to t$ on Equation 1.1 gives

$$E[C_m] = \frac{\mu_m}{\omega(m)} \sum_{j=1}^{p} g_j(m) \int_0^\infty \Psi_m(t)\, dt \quad \text{where} \quad \Psi_m(t) := \left[1 - \prod_{i=1}^{|\mathbf{W}_m|} \left(1 - e^{-t \frac{W_{m,i}}{\omega(m)} \sum_{j=1}^{p} g_j(m)} \right)^{M_{m,i}} \right].$$

Focusing on this expression, one shows that the integral of $\Psi_m(t)$ converges towards some constant. Indeed, numerical analysis, as illustrated by Figure 2.2, suggests that Ψ_m converges toward a step function when $m \to \infty$. This can be rigorously shown under the conditions **H1**, **H2** and **H3**, and the integral from 0 to $t^*(f_1, c)$ converges to $t^*(f_1, c)$, while the remaining integral converges to 0. □

Remark 2.2. *Theorem 2.1 applies in the special case of the uniform distribution. Considering the weighted distribution* $\{w_i\}_{i=1}^m = \{1\}_{i=1}^m$*, one has* $p_i = 1/m$ *and* $\mu_m = m$*. The set of weights is then restricted to* $\boldsymbol{W}_m = (1)$ *and therefore* $[1, |\boldsymbol{W}_m|] = \{1\}$ *is finite for any value of* m*.*
One has $M_{m,1} = m = e^{\log m}$*, so taking* $F(1) := f_1(1) = 1$*,* $G(m) := g_1(m) = \log m$*,* $h(1) = 1$ *and* $H(1) = 1$ *satisfies* ***H1****, since* $\sum_{1 \le i \le |\boldsymbol{W}_m|} 1/H(i)$ *reduces to* $1/H(1)$ *and obviously converges. Moreover, one has* $W_{m,1} = 1$*, therefore* ***H2*** *is satisfied with* $c(1) = 1$ *and* $\omega(m) = 1$ *(no assumption on the behavior of* ω *is required). Since* $[1, |\boldsymbol{W}_m|]$ *is finite,* ***H3*** *does not need to be verified. One can easily see that* $t^*(f_1, c) = t^*(1, 1) = 1$*, and then*

$$E[C_m] \sim m \log m$$

where one recognizes the well-known asymptotics of the uniform coupon collector.

3. Comparison with existing results

Let us compare the scope of our result with previous work on the subject; we recall that m is the number of coupons/words, growing exponentially with n the length of words. Given the rich literature dealing with variations on the Coupon Collector problem (e.g. waiting time of first occurrence of a k-duplicated collection [1]), we will restrict our comparison to three results that are representative of the main approaches used to tackle the problem.

Berenbrink and Sauerwald [2]*:* $\mathcal{O}(\log\log(m))$ *and* $\mathcal{O}(\log\log\log(m))$ *approximations for general distributions.* Building on previous results [15], Berenbrink and Sauerwald consider the two expressions

$$\mathcal{U}_2 := \sum_{i=1}^{m} \frac{1}{ip_i} \qquad \text{and} \qquad \mathcal{U}_4 := \sum_{i=1}^{\log\log m} \frac{1}{i} \frac{1}{e^{j_i - 1} p_1} \mathcal{H}_{g_{j_i}}$$

where g_i is the number of coupons c such that $e^{i-1} < p_c/p_1 < e^i$, and $\{j_i\}_{i=1}^{\log\log m}$ is a sequence of indices such that $\{\frac{1}{e^{j_i} p_1} \mathcal{H}_{g_{j_i}}\}_{i=1}^{\log\log m}$ is decreasing. They show that $\mathcal{U}_2$ and $\mathcal{U}_4$ approximate $E[C_m]$ within $\mathcal{O}(\log\log(m))$ and $\mathcal{O}(\log\log\log(m))$ ratios respectively. More precisely, they show that

$$\frac{\mathcal{U}_2}{3e \log\log m} \le E[C_m] \le 2 \cdot \mathcal{U}_2 \qquad \text{and} \qquad \frac{\mathcal{U}_4}{\log\log\log m} \le E[C_m] \le 35 \cdot \mathcal{U}_4.$$

Furthermore, $\mathcal{U}_2$ can be computed in polynomial-time (on n), since there exists at most $n^{|\Sigma|}$ compositions/weights. However, the exponential growth of m on n limits the final precision of the approximation ratio to $\mathcal{O}(\log n)$. Finally, an efficient evaluation of $\mathcal{U}_4$ would yield a $\mathcal{O}(\log\log\log m)$ approximation in time $\Theta(\log n)$. Unfortunately, figuring out a suitable sequence $\{j_i\}_{i=1}^{\log\log m}$ remains challenging, and seems to require comparable knowledge over the multiplicity of coupons as for the application of Theorem 2.1.

Boneh and Papanicolaou [4]*: Asymptotic estimates for truncated sequences of weighted coupons.* The authors derive general results for the asymptotic of the coupon collector problem under fairly general distributions of coupons. They consider a fixed sequence of strictly positive weights $\boldsymbol{\alpha} = \{a_k\}_{k=1}^{\infty}$, and study the truncation, $\boldsymbol{\alpha}_m$, of $\boldsymbol{\alpha}$ to its first m terms. They obtain two distinct theorems, depending on whether or not there exists $\xi \in]0, 1]$ such that $S := \sum_{k=1}^{\infty} \xi^{a_k} < \infty$. Under the hypotheses of our main Theorem 2.1, there exists a weight of unbounded multiplicity as m goes to the infinity, and S therefore diverges for any value of ξ. Consequently, only one of their results may apply, based on the assumption of a decreasing sequence $\boldsymbol{\alpha}$.

Many weighted distributions that satisfy hypothesis **H1** to **H3** cannot be defined by truncating a fixed decreasing sequence. For instance, suppose that for all m, the weighted distribution is

$\{\frac{2k-1}{k}\}_{k=1}^{m} \cup \{2\}$, each weight appearing m times. The fixed sequence such that the weighted distribution is obtained by truncating it to $\{1, 2, 3/2, 3/2, 5/3, 3/2, 5/3, 7/4, ...\}$, which is not decreasing.

Conversely, distributions with low multiplicity, satisfying their conditions are not covered by our Theorem 2.1. Therefore, both their result and ours are complementary, with few overlapping use-cases.

Neal [14]*: The limiting distribution.* Neal studied the distribution of the waiting time. Although the results described in this article can in principle be used to assess the expectation of the waiting time, verifying the prerequisites of its main theorem turns out to be considerably more involved than those of Theorem 2.1. In particular, one has to figure out suitable sequences $\{b_n\}$ and $\{k_n\}$, respectively related to the expectation and variance of the distribution, from which the limiting distribution follows. This result is therefore mostly suitable to prove a conjectured distribution from a limited list of its moments. Conversely, knowledge of the expectation, as obtained from our contribution, can help figuring out suitable sequences to apply their results.

4. Applications to languages: the word collector

4.1. **Weighted Languages.** Let $\mathcal{L}$ be a language defined on an alphabet Σ, and let $\mathcal{L}_n$ be its restriction to words of size n. A positive weight π_t is assigned to each letter t of Σ. One extends these weights multiplicatively on any word $\omega \in \mathcal{L}$ such that the weight of a word ω is

$$\pi(\omega) = \prod_{t \in \omega} \pi_t.$$

This defines a natural probability distribution on $\mathcal{L}_n$, given by

$$\mathbb{P}[\omega] = \frac{\pi(\omega)}{\sum\limits_{\omega' \in \mathcal{L}_n}}.$$

With these definitions, $\mathcal{L}_n$ is an example of a coupon collection where each coupon is a word of $\mathcal{L}_n$. The number m of coupons is the number of words of $\mathcal{L}_n$. As m is now function of n, all the characteristics of the weight distribution, such as $\mathbf{W}_m$, will be indexed by n instead of m.

The definition of the distribution of weights implies that all words having the same composition (i.e each letter occurs the same amount of times no matter of their order) have the same weight. It follows that this distribution may present high multiplicity. In order to apply the result of the previous section, we have to verify conditions **H1**, **H2** and **H3** in the context of π, which involves building $\mathbf{W}_n$, the vector of all distinct weights, or at least gathering sufficient information on it.

4.2. **Verifying preconditions H1, H3 and H3 in the context of weighted languages.** Let us outline a systematic method to verify the preconditions **H1**, **H2** and **H3** for a language $\mathcal{L}$ defined over an alphabet $\Sigma = (a_1, ..., a_k)$. The idea is, firstly, to classify the words of the language according to their weight and find the number of words having a given weight (Step 1). Then one has to find an ordering of the different weights (Step 2). If the order cannot be found explicitly, one has to find a sufficient approximation of it (Step 3). Once steps one through three are done, the verification of the hypotheses follows.

Step 1: Characterize the set of distinct weights.

The weight of a word is directly related with its composition (or sub-composition).

Definition 4.1 (Compositions and sub-compositions)**.** *The composition of a word is the vector of occurrences of each letter within the word. More precisely, if a word ω has x_1 occurrence of the letter a_1, ..., x_k times the letter a_k, irrespectively of their order, then its composition is $(x_1, ..., x_k)$.*
Suppose that $1 = \pi_{a_1} = ... = \pi_{a_l}$ for some l, then the sub-composition of a word of composition $(x_1, ..., x_k)$ is the vector $(x_{l+1}, ..., x_k)$, in a $(k-l)$-dimensional space, sometimes denoted $\boldsymbol{x}$.

Let us denote by $M(\mathbf{x})$ the number of words of $\mathcal{L}_n$ having a given sub-composition $\mathbf{x}$. By definition, any words having the same sub-composition share the same weight. The reverse is not true in general, and words having different sub-compositions can have the same weight.

Notation 4.2. *$\Gamma_n \subset \mathbb{N}^{k-l}$ is the set of all distinct sub-compositions appearing in $\mathcal{L}_n$.*

Step 2 : Find a suitable ordering of weights.

Defining an ordering function over $\mathcal{L}_n$ will greatly help us characterize $\mathbf{W}_n$.

Definition 4.3 (Ordering function)**.** *Let ϕ_n be the application that assigns, to each sub-composition of Γ_n, the position of its weight in $\boldsymbol{W}_n$. One has*

$$\phi_n : \left\{ \begin{array}{ccc} \Gamma_n & \to & |\boldsymbol{W}_n| \\ \boldsymbol{x} & \mapsto & i, \text{ if } \pi(\boldsymbol{x}) = W_{n,i}. \end{array} \right. \tag{4.1}$$

In general, this function is not bijective, therefore let us define the generalized inversed ordering function $\tilde{\boldsymbol{\phi}}_n$ as follows :

$$\tilde{\boldsymbol{\phi}}_n : \left\{ \begin{array}{ccc} |\boldsymbol{W}_n| & \to & \Gamma_n \\ i & \mapsto & \boldsymbol{x}, \text{ if } W_{n,i} = \pi(\boldsymbol{x}) \text{ and } |\boldsymbol{x}| = \min(|(\boldsymbol{x}')|, W_{n,i} = \pi(\boldsymbol{x}')), \end{array} \right. \tag{4.2}$$

where $|\boldsymbol{x}| = x_{l+1} + ... + x_k$ if $\boldsymbol{x}$ is the sub-composition $(x_{l+1}, ..., x_k)$.

With these definitions, $W_{n,i}$ and $M_{n,i}$ can be written in terms of ϕ_n and $\tilde{\boldsymbol{\phi}}_n$ as

$$W_{n,i} = \pi(\tilde{\boldsymbol{\phi}}_n(i)) \qquad \text{and} \qquad M_{n,i} = \sum_{\mathbf{x} \in \mathcal{L}_n, \phi_n(\mathbf{x}) = i} \; \sum_{x_1 + ... + x_l = n - |\mathbf{x}|} M(\mathbf{x}). \tag{4.3}$$

Sub-compositions are vectors in a $(k-l)$-dimensional space. It is easily checked that the weight of any sub-composition, found underneath the $(k-l-1)$-plane $H(\mathbf{x})$ of equation $\sum_{j=l+1}^{k} x_j \log \pi_{a_i} = 0$, is smaller than $\pi(\mathbf{x})$, and that any sub-composition above has larger weight.

Definition 4.4. *Let $\Lambda_n(\boldsymbol{x}) \subset \Gamma_n$ be the set of sub-compositions below $H(\boldsymbol{x})$ (all the sub-compositions that belong to $H(\boldsymbol{x})$ have the same weight), and $S_n(\boldsymbol{x})$ be the number of sub-compositions that belong to $H(\boldsymbol{x})$.*

Then one has the following expression for ϕ_n :

$$\phi_n(\mathbf{x}) = \sum_{\mathbf{x'} \in \Lambda_n(\mathbf{x})} \frac{1}{S_n(\mathbf{x}')}. \tag{4.4}$$

Indeed, ϕ_n counts the number of sub-compositions, with distinct weights, under $H(\mathbf{x})$. If each weight matches a unique sub-composition, then $S_n(\mathbf{x}) = 1$ for all $\mathbf{x}$, and $\phi_n(\mathbf{x}) = |\Lambda_n(\mathbf{x})|$.

Step 3 : Approximate the ordering functions ϕ_n and $\tilde{\boldsymbol{\phi}}_n$.

Condition **H3** directly follows from steps 1 and 2. However, preconditions **H1** and **H2** require good approximations of $|\Lambda_n|$ and S_n. Such approximations strongly depend on the language, therefore we present several examples to illustrate the method.

4.3. **Application to specific languages.**

4.3.1. *The unconstrained language Σ^*.* Let us consider the language $\mathcal{L} = \Sigma^*$, where $\Sigma = (a_1, ..., a_k)$. Let us remark that the weighted distribution is stable upon multiplying each weight by a constant factor, therefore let us assume without loss of generality that $1 = \pi_{a_1} = ... = \pi_{a_l}$ for some $l \geq 1$, and $1 < \pi_{a_{l+1}} \leq ... \leq \pi_{a_k}$. For the sake of simplicity, let us further assume that the non-unit weights are pairwise incommensurable, which implies that sub-compositions can be bijectively associated with weights.

Under these assumptions, one has $\Gamma_n = \{(\mathbf{x}', |\mathbf{x}'| \leq n)\}$. The function $\phi_n(\mathbf{x})$ counts the number of sub-compositions under $H(\mathbf{x})$ which belong to Γ_n. Notice that, for sufficiently large values of

n, any sub-composition $\mathbf{x}'$ belongs to Γ_n. It follows that there exists a function ϕ such that, for all sub-composition $\mathbf{x}$ and for n sufficiently large, one has $\phi_n(\mathbf{x}) = \phi(\mathbf{x})$. From Equation (4.3), one has $W_{n,i} = \pi_{a_{l+1}}^{\tilde{\boldsymbol{\phi}}_{n,1}(i)} ... \pi_{a_k}^{\tilde{\boldsymbol{\phi}}_{n,k-l}(i)}$, and it follows that, for sufficiently large values of n, one has $W_{n,i} = \pi_{a_{l+1}}^{\tilde{\boldsymbol{\phi}}_1(i)} ... \pi_{a_k}^{\tilde{\boldsymbol{\phi}}_{k-l}(i)}$. Consequently, Condition **H2** is verified with

$$c(i) = \pi_{a_{l+1}}^{\tilde{\boldsymbol{\phi}}_1(i)} ... \pi_{a_k}^{\tilde{\boldsymbol{\phi}}_{k-l}(i)} \qquad \text{and} \qquad \omega(n) = 1.$$

In $\mathcal{L}$, the number of words of composition $(x_1, ..., x_k)$ is $M(x_1, ..., x_k) = \binom{n}{x_1, ..., x_k}$, thus the number of words of sub-composition $(x_{l+1}, ..., x_k)$ is $M(x_{l+1}, ..., x_k) = l^{n-x_{l+1}-...-x_k}\binom{n}{x_{l+1}, ..., x_k}$. Since there exists only one sub-composition $\mathbf{x}$ such that $\phi(\mathbf{x}) = i$, then it follows from Equation (4.3) that $M_{n,i} = l^{n-|\tilde{\boldsymbol{\phi}}_n(i)|}\binom{n}{\tilde{\boldsymbol{\phi}}_n(i)}$, where $\binom{n}{\mathbf{a}}$ is the multinomial coefficient $\binom{n}{a_1, ..., a_k}$. Since $\phi_n = \phi$ for sufficiently large values of n, one has

$$M_{n,i} \underset{n\to\infty}{\sim} l^{n-|\tilde{\boldsymbol{\phi}}(i)|}\binom{n}{\tilde{\boldsymbol{\phi}}(i)} \underset{n\to\infty}{\sim} \frac{l^{n-|\tilde{\boldsymbol{\phi}}(i)|} n^{|\tilde{\boldsymbol{\phi}}(i)|}}{|\tilde{\boldsymbol{\phi}}(i)|!}. \tag{4.5}$$

Let us now give some properties of the functions ϕ_n and ϕ.

Lemma 4.5. *Let $S := \sum_{j=l+1}^{k} \log \pi_{a_j}$ and $P := \prod_{j=l+1}^{k} \log \pi_{a_j}$.*

i) *For any sub-composition $\boldsymbol{x}$,*

$$\frac{|\boldsymbol{x}|_\pi^{k-l}}{(k-l)!P} \le \phi(\boldsymbol{x}) \le \frac{(|\boldsymbol{x}|_\pi + S)^{k-l}}{(k-l)!P} \tag{4.6}$$

where $|\boldsymbol{x}|_\pi = x_{l+1} \log \pi_{a_{l+1}} + ... + x_k \log \pi_{a_k}$.

ii) *For all $i > 0$, one has*

$$\begin{aligned} \sqrt[k-l]{i(k-l)!P} - S \le \quad |\tilde{\boldsymbol{\phi}}(i)|_\pi \quad &\le \sqrt[k-l]{i(k-l)!P} \\ \sqrt[k-l]{i(k-l)!\frac{P}{(\log \pi_{a_k})^{k-l}}} - \frac{S}{\log \pi_{a_k}} \le \quad |\tilde{\boldsymbol{\phi}}(i)| \quad &\le \sqrt[k-l]{i(k-l)!\frac{P}{\log \pi_{a_{l+1}}}}. \end{aligned} \tag{4.7}$$

iii) *For all $\boldsymbol{x}$ and $n > 0$, one has*

$$\phi_n(\boldsymbol{x}) \le \phi(\boldsymbol{x}). \tag{4.8}$$

iv) *For all $n > 0$ and $i \ge 1$, one has*

$$\frac{\log \pi_{a_{l+1}}}{\log \pi_{a_k}} |\tilde{\boldsymbol{\phi}}(i)| \le |\tilde{\boldsymbol{\phi}}_n(i)| \le |\tilde{\boldsymbol{\phi}}(i)|. \tag{4.9}$$

Proof.

i) $\phi(\mathbf{x})$ counts the number of points which are under the $(k-l-1)$-plane $H(\mathbf{x})$. Equation (4.6) just consists in bounding ϕ by the volume of the $(k-l-1)$-pyramid under $H(x_{l+1}, ..., x_k)$ and the $(k-l-1)$-pyramid under $H(x_{l+1}+1, ..., x_k+1)$.

ii) The first equation is obtained from equation (4.6), taking $\mathbf{x} = \tilde{\boldsymbol{\phi}}(i)$. For the second equation, one uses the fact that $|\mathbf{x}| \cdot \log \pi_{a_{l+1}} \le |\mathbf{x}|_\pi \le |\mathbf{x}| \cdot \log \pi_{a_k}$.

iii) $\phi_n(\mathbf{x})$ counts the number of sub-compositions which are both under $H(\mathbf{x})$ and belong to Γ_n, whereas $\phi(\mathbf{x})$ counts the number of sub-compositions which are under $H(\mathbf{x})$.

iv) For a given length $n > 0$, any sub-composition is found below the $|\mathbf{x}| = n$ hyperplane and, in particular, one has $|\tilde{\boldsymbol{\phi}}_n(i)| \le n$. For some sufficiently large value of $n' > n$, the sub-composition of i-th weight becomes fixed and is necessarily a sub-composition of $\Gamma_{n'}$ that did not belong to Γ_n. Consequently, this sub-composition is above the $|\mathbf{x}| = n$ hyperplane, one has $|\tilde{\boldsymbol{\phi}}(i)| \ge n$ and one finally gets $|\tilde{\boldsymbol{\phi}}_n(i)| \le |\tilde{\boldsymbol{\phi}}(i)|, \forall n > 0, \forall i \ge 1$.

On the other hand, the sub-composition $\tilde{\boldsymbol{\phi}}(i)$ must be below the hyperplane $|\mathbf{x}| = |\tilde{\boldsymbol{\phi}}_n(i)|_\pi$ otherwise its weight would be larger than the one of $\tilde{\boldsymbol{\phi}}_n(i)$. This gives $|\tilde{\boldsymbol{\phi}}_n(i)|_\pi \ge |\tilde{\boldsymbol{\phi}}(i)|_\pi$.

Since any sub-composition obeys $\frac{|\mathbf{x}|_\pi}{\log \pi_{a_{l+1}}} \geq |\mathbf{x}| \geq \frac{|\mathbf{x}|_\pi}{\log \pi_{a_k}}$, one has

$$|\tilde{\boldsymbol{\phi}}_n(i)| \geq \frac{|\tilde{\boldsymbol{\phi}}_n(i)|_\pi}{\log \pi_{a_k}} \geq \frac{|\tilde{\boldsymbol{\phi}}(i)|_\pi}{\log \pi_{a_k}} \geq \frac{\log \pi_{a_{l+1}}}{\log \pi_{a_k}} |\tilde{\boldsymbol{\phi}}(i)|,$$

which concludes the proof.

□

Combining Equations (4.5) and (4.8), one obtains bounds for the leading term of $M_{n,i}$, for all i and as $n \to \infty$, such that

$$l^{n-|\tilde{\boldsymbol{\phi}}_n(i)|} \binom{n}{\tilde{\boldsymbol{\phi}}_n(i)} \leq l^{n-|\tilde{\boldsymbol{\phi}}_n(i)|} \frac{n^{|\tilde{\boldsymbol{\phi}}_n(i)|}}{|\tilde{\boldsymbol{\phi}}_n(i)|!} \leq l^{n-\frac{\log \pi_{a_{l+1}}}{\log \pi_{a_k}}|\tilde{\boldsymbol{\phi}}(i)|} \frac{n^{|\tilde{\boldsymbol{\phi}}(i)|}}{\left(\frac{\log \pi_{a_{l+1}}}{\log \pi_{a_k}}|\tilde{\boldsymbol{\phi}}(i)|\right)!}.$$

The convergence of $\sum_i 1/\left(\frac{\log \pi_{a_{l+1}}}{\log \pi_{a_k}}|\tilde{\boldsymbol{\phi}}(i)|\right)!$ follows from Equation (4.7). Therefore, **H1** is satisfied for the following choice of functions

	$F(i) := f_1(i)$	$f_2(i)$	$G(i) := g_1(n)$	$g_2(n)$	$h(i)$	$H(i)$
$l = 1$	$\lvert\tilde{\boldsymbol{\phi}}(i)\rvert$		$\log n$		$\lvert\tilde{\boldsymbol{\phi}}(i)\rvert!$	$\left(z\lvert\tilde{\boldsymbol{\phi}}(i)\rvert\right)!$
$l > 1$	$\log l$	$\lvert\tilde{\boldsymbol{\phi}}(i)\rvert$	n	$\log n$	$l^{\lvert\tilde{\boldsymbol{\phi}}(i)\rvert}\lvert\tilde{\boldsymbol{\phi}}(i)\rvert!$	$l^{\left(z\lvert\tilde{\boldsymbol{\phi}}(i)\rvert\right)}\left(z\lvert\tilde{\boldsymbol{\phi}}(i)\rvert\right)!$

where $z = \log \pi_{a_{l+1}} / \log \pi_{a_k}$. Furthermore it can be verified that **H3** is satisfied, since Equation (4.7) gives a lower bound for $c(i)/F(i)$. Consequently, Theorem 2.1 applies to the weighted distribution on Σ^*, and gives the following asymptotic behavior.

Proposition 4.6. *The expected waiting time for obtaining all words of length n in $\mathcal{L} = \Sigma^*$ admits the following asymptotic behavior:*

$$E[C_n] \sim \begin{cases} \kappa_1 \cdot \mu(n) \cdot \log n & \text{if } l = 1, \\ \kappa_2 \cdot \mu(n) \cdot n & \text{otherwise,} \end{cases}$$

where l is the number of letters of lowest weights,

$$(\kappa_1, \kappa_2) = \left(t^* \left(|\tilde{\boldsymbol{\phi}}(i)|, \pi_{a_{l+1}}^{\tilde{\boldsymbol{\phi}}_1(i)} ... \pi_{a_k}^{\tilde{\boldsymbol{\phi}}_{k-l}(i)} \right), t^* \left(\log l, \pi_{a_{l+1}}^{\tilde{\boldsymbol{\phi}}_1(i)} ... \pi_{a_k}^{\tilde{\boldsymbol{\phi}}_{k-l}(i)} \right) \right)$$

and

$$\text{and } \mu(n) = \sum_{0 \leq i_1 + ... + i_k \leq n} \binom{n}{i_1, ..., i_k} \pi_{a_1}^{i_1} ... \pi_{a_k}^{i_k} = \left(l + \sum_{i=l+1}^{k} \pi_{a_j} \right)^n.$$

4.3.2. *Motzkin words.* Motzkin words are well-parenthesized words featuring any number of dots characters •. This language, denoted by $\mathcal{L}^{(m)}$, is generated by the context-free grammar

$$S \to (\ S\ )\ S \mid \bullet\ S \mid \varepsilon.$$

Here we study the expected waiting time of Motzkin words of even length n. For the sake of readability, we replace the characters (,) and • respectively by the letters a, $\bar{a}$ and b. Since parentheses come in pairs, any word has equal number of occurrences of a and $\bar{a}$, and the parity of the number of b is the parity of the word length. Consequently, accessible compositions for words of length n are of the form $(k, k, n-2k)$ occurrences of the letters a, $\bar{a}$ and b respectively, with $0 \leq k \leq n/2$. The number of words of size n is then given by

$$M(k, k, n-2k) = \frac{1}{k+1} \binom{2k}{k} \binom{n}{2k}.$$

The expected waiting time presents two different behaviors depending on whether a and $\bar{a}$ have the smallest weight. We derive two examples, corresponding to the $1 = \pi_b < \pi_a < \pi_{\bar{a}}$ and $1 = \pi_a = \pi_{\bar{a}} < \pi_b$ cases respectively.

First scenario ($1 = \pi_b < \pi_a < \pi_{\bar{a}}$). In this first case, the sub-compositions are of the form (k, k), $0 \leq k \leq n/2$, and the associated weights are of the form $\pi_a^k \pi_{\bar{a}}^k$, increasing with k. Therefore one has

$W_{n,i} = \pi_a^{i-1}\pi_{\bar{a}}^{i-1}$, and **H2** is satisfied with $c(i) = \pi_a^{i-1}\pi_{\bar{a}}^{i-1}$ and $\omega(n) = 1$. The number of words having weight $W_{n,i}$, or equivalently of sub-composition $(i-1, i-1)$, is given by

$$M_{n,i} = \frac{1}{i}\binom{2i-2}{i-1}\binom{n}{2i-2} \underset{n\to\infty}{\sim} \frac{n^{2i-2}}{i(2i-2)!}\binom{2i-2}{i-1} = \frac{n^{2i-2}}{i(i-1)!^2}.$$

Moreover, for all $i \leq n/2$, one has $M_{n,i} \leq \frac{n^{2i-2}}{i(2i-2)!}\binom{2i-2}{i-1}$, and **H1** is satisfied with

$F(i) := f_1(i)$	$G(n) := g_1(n)$	$h(i)$	$H(i)$
$2i-2$	$\log n$	$\frac{1}{i(i-1)!^2}$	$\frac{1}{i(i-1)!^2}$

noticing that $\sum_i 1/i(i-1)!^2$ converges. The verification of **H3** is immediate, and applying Theorem 2.1 readily gives the following result.

Proposition 4.7. *The expected waiting time for obtaining all weighted Motzkin words of even length n, under the configuration $1 = \pi_b < \pi_a < \pi_{\bar{a}}$, admits the following asymptotic behavior:*

$$E[C_n] \sim \kappa \cdot \mu(n) \cdot \log n$$

where $\kappa = t^*\left(2i-2, \pi_a^{i-1}\pi_{\bar{a}}^{i-1}\right)$ and $\mu(n) = \sum_{k=0}^{n/2} \frac{1}{k+1}\binom{2k}{k}\binom{n}{2k}\pi_a^k\pi_{\bar{a}}^k$.

Second scenario $(1 = \pi_a = \pi_{\bar{a}} < \pi_b)$. In this second case, the sub-compositions are of the form $(n-2k)$, for $0 \leq k \leq n/2$, and the weight of a word increases with the number of b. Consequently, one has $W_{n,i} = \pi_b^{2(i-1)}$, and **H2** is satisfied with $c(i) = \pi_b^{2(i-1)}$ and $\omega(n) = 1$. Furthermore, if $(n-2k)$ is the sub-composition of the i-th weight, then $n - 2k = 2(i-1)$, leading to $k = n/2 - (i-1)$ and one finally has

$$M_{n,i} = \frac{1}{\frac{n}{2}-(i-1)+1}\binom{n-2(i-1)}{\frac{n}{2}-(i-1)}\binom{n}{n-2(i-1)} \underset{n\to\infty}{\sim} 2^n \frac{n^{2(i-1)-\frac{3}{2}}}{\sqrt{\pi}2^{2(i-1)-\frac{3}{2}}(2(i-1))!}.$$

Furthermore, one has $M_{n,i} \leq 2^n \frac{n^{2(i-1)-\frac{3}{2}}}{\sqrt{\pi}2^{2(i-1)-\frac{3}{2}}(2(i-1))!}$, for $i \leq n/2$, and **H1** is satisfied with

$F(i) := f_1(i)$	$f_2(i)$	$G(n) := g_1(n)$	$g_2(n)$	$h(i)$	$H(i)$
$\log 2$	$2(i-1)-\frac{3}{2}$	n	$\log n$	$\sqrt{\pi}2^{2(i-1)-\frac{3}{2}}(2(i-1))!$	$\sqrt{\pi}2^{2(i-1)-\frac{3}{2}}(2(i-1))!$

since $\sum_i 1/2^{2i}(2(i-1))!$ converges. Again, verifying **H3** is immediate.

Proposition 4.8. *The expected waiting time for obtaining all weighted Motzkin words of even length n, under the configuration $1 = \pi_a = \pi_{\bar{a}} < \pi_b$, admits the following asymptotic behavior:*

$$E[C_n] \sim \kappa \cdot \mu(n) \cdot n$$

where $\kappa = t^*\left(\log 2, \pi_b^{2(i-1)}\right)$ and $\mu(n) = \sum_{k=0}^{n/2} \frac{1}{k+1}\binom{2k}{k}\binom{n}{2k}\pi_b^{n-2k}$.

4.3.3. *RNA secondary structures.* Through an adaptation of Viennot et al [16], secondary structures are generated by the following grammar:

$$S \to (\, S_{\geq\theta}\,)\, S \mid \bullet\, S \mid \varepsilon \qquad \text{and} \qquad S_{\geq\theta} \to (\, S_{\geq\theta}\,)\, S \mid \bullet\, S_{\geq\theta} \mid \bullet^{\theta}.$$

The connection between this language and the conformations of RNA is illustrated by Figure 4.3.3. Matching parentheses represent base-pairs, whose contribution to the free-energy can be set to -1 kcal.mol^{-1} and extended additively on secondary structures. Under this setting, we are interested in weighting each secondary structure S with its Boltzmann factor $e^{\#bp(S)/RT}$ where R is usually the gas constant and T the temperature (K). This can very simply be done by assigning a weight $e^{1/RT}$ to each pair of matching parentheses, and 1 to unpaired positions. As stated in the introduction, the study of the coupon collector for Boltzmann weighted secondary structures is closely related to the worst-case complexity of a method based on a redundant statistical sampling algorithm [7].

Again, let us replace the characters (,) and $\bullet$ by letters a, $\bar{a}$ and b respectively. Let us denote by $\mathcal{L}^{(rna)}$ the language of RNA secondary structure, and study the expected waiting time in the only relevant case $1 = \pi_b < \pi_a < \pi_{\bar{a}}$. The compositions are of the form $(k, k, n-2k)$, for $0 \leq k \leq n/2$. The

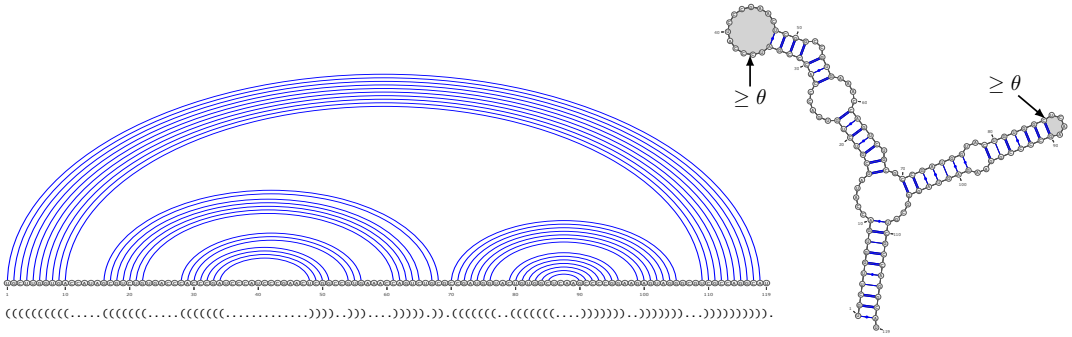


FIGURE 2. Secondary structure of a 5s ribosomal RNA. A well-parenthesized expression (lower-left) unambiguously defines a set of matching position (upper-left) which *folds* into a projection of a three-dimensional conformation of the molecule. The latter representation illustrates the relationship between the $\geq \theta$ steric constraint and the absence of sharp turns.

number of words of size n having p *plateaux* and k occurrences of a is given by 1 if $(p,k)=(0,0)$, and $s_{n,k,p,\theta} = \frac{1}{k}\binom{k}{p}\binom{k}{p-1}\binom{n-\theta p}{2k}$ otherwise. Consequently, the number of words having a given composition $(k,k,n-2k)$ is such that

$$M(k,k,n-2k) = \delta_{k,0} + \sum_{p=1}^{\lfloor \frac{n-2k}{\theta} \rfloor} s_{n,k,p,\theta} = \delta_{k,0} + \sum_{p=1}^{\lfloor \frac{n-2k}{\theta} \rfloor} \frac{1}{k}\binom{k}{p}\binom{k}{p-1}\binom{n-\theta p}{2k}$$

where δ is Kronecker symbol ($\delta_{a,b}=1$ if $a=b$ and 0 otherwise). Since $1=\pi_b<\pi_a<\pi_{\bar{a}}$, the weights of words are increasing with the number of $\bar{a}$, one has $W_{n,i} = \pi_a^{(i-1)}\pi_{\bar{a}}^{(i-1)}$, and **H2** is satisfied with $c(i) = \pi_a^{(i-1)}\pi_{\bar{a}}^{(i-1)}$ and $\omega(n)=1$. Moreover, the number of words having weight $W_{n,i}$ is the number of words of sub-composition $(i-1,i-1)$, and it follows that

$$M_{n,i} = \delta_{i-1,0} + \sum_{p=1}^{\lfloor \frac{n-2(i-1)}{\theta} \rfloor} \frac{1}{(i-1)}\binom{i-1}{p}\binom{i-1}{p-1}\binom{n-\theta p}{2(i-1)} \underset{n\to\infty}{\sim} \frac{n^{2(i-1)}}{i(i-1)!^2}.$$

Indeed, for large values of n, the scope of the sum above can be limited to $p \in [1, i-1]$ since any term such that $p>(i-1)$ will cancel. Therefore one is left with the behavior of Motzkin words in the $1=\pi_b<\pi_a<\pi_{\bar{a}}$ case. One also has $M_{n,i} \leq 2\frac{n^{2(i-1)}}{i(i-1)!^2}$, for all i, thus **H1** is satisfied with

$f_1(i)$	$g_1(n)$	$h(i)$	$H(i)$
$2(i-1)$	$\log n$	$i(i-1)!^2$	$\frac{i}{2}(i-1)!^2$

where $\sum_i 1/H(i)$ obviously converges, and the verification of **H3** is immediate.

Proposition 4.9. *The expected waiting time for obtaining all Boltzmann-factor weighted RNA secondary structures of length n, under the configuration $1=\pi_b<\pi_a<\pi_{\bar{a}}$, admits the following asymptotic behavior:*

$$E[C_n] \sim \kappa \cdot \mu(n) \cdot \log n$$

where $\kappa = t^*\left(2(i-1), \pi_a^{(i-1)}\pi_{\bar{a}}^{(i-1)}\right)$ *and* $\mu(n) = \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor}\left[\delta_{k,0} + \sum_{p=1}^{\lfloor \frac{n-2k}{\theta} \rfloor} \frac{1}{k}\binom{k}{p}\binom{k}{p-1}\binom{n-\theta p}{2k}\right] a^k\bar{a}^k.$

4.3.4. *A non strongly-connected language.* Let us finally consider the language $\mathcal{L}^{(nc)}$ on an alphabet $(a,\bar{a},b)$ generated by the grammar

$$S \to \bar{a}\,S\,b\,U \mid \varepsilon \qquad \text{and} \qquad U \to a\,U\,b\,U \mid \varepsilon.$$

The restriction of $\mathcal{L}^{(nc)}$ to words of odd length is empty, thus we only study the word collector on even sizes. The structure of this grammar is such that each word of size n has exactly $n/2$ occurrences of the letter b and any composition is of the form $(n/2-k, k, n/2)$, for $1 \leq k \leq n/2$. An elementary calculus shows that the number words of a given composition is

$$M(n/2-k, k, n/2) = \binom{n-k-1}{n/2-1} - \binom{n-k-1}{n/2}.$$

The expected waiting time depends on the relative position of the weights associated with letters (9 distinct configurations), but they all lead to the same behavior. We illustrate one of the 9 configurations: $1 = \pi_b < \pi_a < \pi_{\bar{a}}$.

In this case, the sub-compositions are of the form $(n/2-k, k)$, for $1 \leq k \leq n/2$, and the weight of the word increases with the number of $\bar{a}$, thus $W_{n,i} = \pi_a^{n/2-i}\pi_{\bar{a}}^i$. Then **H2** is satisfied with $c(i) = \left(\frac{\pi_{\bar{a}}}{\pi_a}\right)^i$ and $\omega(n) = \pi_a^{\frac{n}{2}}$.

Remark 4.10. *The difference with the other configurations only reflects in the letters involved in the functions c and ω. The function ω may be constant (1) when $\pi_b = \pi_a = 1$ or $\pi_b = \pi_{\bar{a}} = 1$.*

Moreover, the number of words having i-th weight, or sub-composition $(i, n/2-i)$, is given by

$$M_{n,i} = \binom{n-i-1}{n/2-1} - \binom{n-i-1}{n/2} \underset{n\to\infty}{\sim} 2^n n^{-\frac{3}{2}} \sqrt{\frac{2}{\pi}} \frac{i}{2^i}$$

Since $M_{n,i} \leq 2^n n^{-\frac{3}{2}} 2\sqrt{\frac{2}{\pi}} \frac{i}{2^i}$, for all $1 \leq i \leq n/2$, and $\sum_i i/2^i$ converges, then **H1** is satisfied with

$F(i) := f_1(i)$	$f_2(i)$	$G(n) := g_1(n)$	$g_2(n)$	$h(i)$	$H(i)$
$\log 2$	$-\frac{3}{2}$	n	$\log n$	$\frac{2^i}{i}\sqrt{\frac{\pi}{2}}$	$\frac{2^{i-1}}{i}\sqrt{\frac{\pi}{2}}$

The verification of **H3** is immediate.

Proposition 4.11. *The expected waiting time for obtaining all words in $\mathcal{L}^{(nc)}$ of even length n, under the configuration $1 = \pi_b < \pi_a < \pi_{\bar{a}}$, admits the following asymptotic behavior:*

$$E[C_n] \sim \kappa \cdot \frac{\mu(n)}{\pi_a^{\frac{n}{2}}} \cdot n$$

where $\kappa = t^*\left(\log 2, \left(\frac{\pi_{\bar{a}}}{\pi_a}\right)^i\right)$ and $\mu(n) = \sum_{k=1}^{n/2} \left[\binom{n-k-1}{n/2-1} - \binom{n-k-1}{n/2}\right] \pi_a^{\frac{n}{2}-k}\pi_{\bar{a}}^k$.

5. Conclusion

In this extended abstract, we studied a language generalization of the ubiquitous Coupon Collector Problem. Focusing on weighted coupons with high multiplicities, we contributed a new theorem that relates the asymptotic waiting time to the growth of the multiplicity of coupons of a given weight. The novelty of the contribution was discussed against pre-existing work on the subject. The application of this theorem to weighted languages was illustrated, revealing asymptotic behaviors in $\Theta(\mu(n) \cdot n)$ and $\Theta(\mu(n) \cdot \log n)$ where $\mu(n)$ is the total weight of all words of length n.

Perhaps the main limitation of our work lies in the prerequisites of Theorem 2.1. As shown in Section 4, verifying these – technically involved – conditions is already made easier in the context of languages. However, one could imagine characterizing broad classes of languages that automatically verify these conditions. For instance, conditions of aperiodicity (a.k.a. lattice-type [10]) and strong-connectivity of a context-free grammar are known to ensure typical asymptotic growths, both for the total number of words, their cumulated weight and the total number of words of a given composition [8]. We hope that such conditions, possibly in addition to other easily-checkable properties, could provide a sufficient set of conditions for a given regime.

Another natural extension may generalize the results to multi-parameterized combinatorial classes, as generated by decomposable combinatorial classes [11]. The main difficulties behind such an extension is the variety of asymptotic growths that may appear, e.g. for the substitution construct,

in addition to an increased level of difficulty for determining the number of words of a given composition/weight. This both motivates a further relaxation of the – sufficient but not necessary – conditions of Theorem 2.1, along with a study of accessible asymptotics for the growth of coefficients in multivariate generating functions.

Acknowledgements

This work was supported by the French *Agence Nationale de la Recherche* through the BOOLE `ANR 09 BLAN 0011` (JDB and DG) and MAGNUM `ANR 2010 BLAN 0204` (YP) grants.

References

1. I. Adler, S. Oren, and S. Ross, *The coupon collector's problem revisited*, Journal of Applied Probability **40** (2003), no. 2, 513–518.
2. P. Berenbrink and T. Sauerwald, *The weighted coupon collector's problem and applications*, 15th International Computing and Combinatorics Conference (COCOON'10), 2009.
3. O. Bodini and Y. Ponty, *Multi-dimensional Boltzmann sampling of languages*, Proceedings of AOFA'10 (Vienna), June 2010.
4. Shahar Boneh and Vassilis G. Papanicolaou, *General asymptotic estimates for the coupon collector problem*, J. Comput. Appl. Math. **67** (1996), no. 2, 277–289.
5. A. Denise, Y. Ponty, and M. Termier, *Controlled non-uniform random generation of decomposable structures*, Theoretical Computer Science **411** (2010), no. 40-42, 3527 – 3552.
6. A. Denise, O. Roques, and M. Termier, *Random generation of words of context-free languagesaccording to the frequencies of letters*, Mathematics and Computer Science: Algorithms,Trees, Combinatorics and probabilities (D. Gardy and A. Mokkadem, eds.), Trends in Mathematics, Birkhaüser, 2000, pp. 113–125.
7. Y. Ding and E. Lawrence, *A statistical sampling algorithm for RNA secondary structure prediction*, Nucleic Acids Research **31** (2003), no. 24, 7280–7301.
8. M. Drmota, *Systems of functional equations*, Random Struct. Alg. **10** (1997), 103–124.
9. P. Flajolet, D. Gardy, and L. Thimonier, *Birthday paradox, coupon collectors, caching algorithms and self-organizing search*, Discrete Appl. Math. **39** (1992), no. 3, 207–229.
10. P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge University Press, 2009.
11. P. Flajolet, P. Zimmermann, and B. Van Cutsem, *Calculus for the random generation of labelled combinatorial structures*, Theoretical Computer Science **132** (1994), 1–35.
12. D. Gardy, *Occupancy urn models in the analysis of algorithms*, Journal of Statistical Planning and Inference **101** (2002), no. 1-2, 95 – 105.
13. J.S. McCaskill, *The equilibrium partition function and base pair binding probabilities for RNA secondary structure*, Biopolymers **29** (1990), 1105–1119.
14. Peter Neal, *The generalised coupon collector problem*, Journal of Applied Probabilities **45** (2008), no. 3, 621–629.
15. S.M. Ross, *Introduction to probability models*, 10th ed., Elsevier Science, 2009.
16. M. Vauchaussade de Chaumont and G. Viennot, *Polynômes orthogonaux et problèmes d'énumération en biologie moléculaire*, Séminaire Lotharingien de Combinatoire (1983).
17. M. Vauchaussade de Chaumont and X.G. Viennot, *Enumeration of RNA's secondary structures by complexity*, Mathematics in Medecine and Biology (V. Capasso, E. Grosso, and S.L. Paven-Fontana, eds.), Lecture Notes in Biomathematics, vol. 57, 1985, pp. 360–365.
18. M. S. Waterman, *Secondary structure of single stranded nucleic acids*, Advances in Mathematics Supplementary Studies **1** (1978), no. 1, 167–212.

Appendix A. Proof of Theorem 2.1

For the proof of the theorem, we need the following lemma.

Lemma A.1. *Let $E \subset \mathbb{N}^*$. Let f and g be two non-zero positive functions on E, such that if E is not finite, $\lim_{x\to\infty} \frac{g(x)}{f(x)} = +\infty$. Then,*
- $\exists\ t^*(f,g) > 0$ such that

(1) $\forall 0 \le t < t^*(f,g), \quad \exists\ x_0 \in E, \quad f(x_0) - tg(x_0) > 0$
(2) $\forall t > t^*(f,g), \qquad \forall\ x \in E, \qquad f(x) - tg(x) < 0$

- $\exists\ x_1 \in \mathbb{N}^*$ such that

(3) $f(x_1) - g(x_1)t = \max_{x\in E}(f(x) - tg(x))$

Proof.

Throughout the proof, $f(x) - tg(x)$ is seen as a function of x with a parameter t.
Let us define $t_x = \frac{f(x)}{g(x)}$, for all $x \in E$. $\forall\ t < t_x$, $f(x) - tg(x) > 0$ and $\forall t > t_x$, $f(x) - tg(x) < 0$. If E is finite, it is obvious that t_x reaches its maximum, i.e. there is $X \in E$ such that $t_X = \max_{x\in E}(t_x)$. This property is still true when E is not finite because $t_x \to 0$ as $x \to \infty$. Then, (1) and (2) are satisfied, taking $t^*(f,g) = t_X$.
If E is finite, it is obvious that $f(x) - tg(x)$ reaches its maximum for all $t > 0$. If E is not finite, using the fact that $\lim_{x\to\infty} \frac{g(x)}{f(x)} = +\infty$, we have $\forall t > 0$, $f(x) - tg(x) \underset{x\to\infty}{\to} -\infty$. Then $f(x) - tg(x)$ reaches its maximum, i.e. there is $x_1 \in E$ such that $f(x_1) - (x_1)t = \max_{x\in E}(f(x) - g(x)t)$, which proves (3). □

Proof of the theorem.
Let us suppose that $\mathbf{W}_m$ satisfies **H1**, **H2**, and **H3**. From equation (1.1), we have

$$E[C_m] = \int_0^\infty \left[1 - \prod_{i=1}^{|\mathbf{W}_m|} \left(1 - e^{-\frac{W_{m,i}}{\mu_m}u}\right)^{M_{m,i}}\right] du.$$

The substitution $u \frac{\omega(m)}{\mu_m \sum_{j=1}^{p} g_j(m)} \to t$ gives

$$\begin{aligned} E[C_m] &= \frac{\mu_m}{\omega(m)} \sum_{j=1}^{p} g_j(m) \int_0^\infty \left[1 - \prod_{i=1}^{|\mathbf{W}_m|} \left(1 - e^{-t\frac{W_{m,i}}{\omega(m)} \sum_{j=1}^{p} g_j(m)}\right)^{M_{m,i}}\right] dt \\ &= \frac{\mu_m}{\omega(m)} \sum_{j=1}^{p} g_j(m) \int_0^\infty \left[1 - \exp\left(\sum_{i=1}^{|\mathbf{W}_m|} M_{m,i} \log\left(1 - e^{-\frac{W_{m,i}}{\omega(m)}t \sum_{j=1}^{p} g_j(m)}\right)\right)\right] dt. \end{aligned}$$

From **H1**, we have $\sum_{j=1}^{p} g_j(m) \sim g_1(m)$. To conclude, we have to show that the integral converges when m goes to infinty. First, we show that the integral from 0 to $t^*(f_1, c)$ converges to $t^*(f_1, c)$. Then, we show that the remaining integral converges to 0.

• From Lemma A.1, applied to E_m, and **H3** (if $|\mathbf{W}_m| \underset{m\to\infty}{\to} \infty$), there is $i_0 \in E_m$ such that $f_1(i_0) - c(i_0)t > 0$. Moreover, from **H2**, for m sufficiently large, $W_{m,i_0} = c(i_0)\omega(m)$. Then

$$\begin{aligned} \sum_{i=1}^{|\mathbf{W}_m|} M_{m,i} \log\left(1 - e^{-\frac{W_{m,i}}{\omega(m)}t \sum_{j=1}^{p} g_j(m)}\right) &\le -\sum_{i=1}^{|\mathbf{W}_m|} M_{m,i} e^{-\frac{W_{m,i}}{\omega(m)}t \sum_{j=1}^{p} g_j(m)} \\ &\le -M_{m,i_0} e^{-\frac{W_{m,i_0}}{\omega(m)}t \sum_{j=1}^{p} g_j(m)} = -M_{m,i_0} e^{-c(i_0)t \sum_{j=1}^{p} g_j(m)}. \end{aligned}$$

From **H1**, for m sufficiently large, $M_{m,i_0} \geq \frac{1}{2}\frac{e^{\sum_{j=1}^{p} f_j(i_0)g_j(m)}}{h(i_0)}$. Then,

$$\sum_{i=1}^{|\mathbf{W}_m|} M_{m,i} \log\left(1 - e^{-\frac{W_{m,i}}{\omega(m)} t \sum_{j=1}^{p} g_j(m)}\right) \leq -\frac{e^{\sum_{j=1}^{p}(f_j(i_0)-c(i_0)t)g_j(m)}}{2h(i_0)}.$$

As $f_1(i_0) - c(i_0)t > 0$ and $g_j(m) = o(g_1(m))$ for all $j > 1$, $\sum_{j=1}^{p} (f_j(i_0) - c(i_0)t)\, g_j(m) \underset{m\to\infty}{\to} +\infty$. Then,

$$\sum_{i=1}^{|\mathbf{W}_m|} M_{m,i} \log\left(1 - e^{-\frac{W_{m,i}}{\omega(m)} t \sum_{j=1}^{p} g_j(m)}\right) \underset{m\to\infty}{\to} -\infty,$$

and

$$\prod_{i=1}^{|\mathbf{W}_m|} \left(1 - e^{-\frac{W_{m,i}}{\omega(m)} t \sum_{j=1}^{p} g_j(m)}\right)^{M_{m,i}} \underset{m\to\infty}{\to} 0.$$

This leads to

$$\int_0^{t^*(f_1,c)} \left[1 - \prod_{i=1}^{|\mathbf{W}_m|} \left(1 - e^{-\frac{W_{m,i}}{\omega(m)} t \sum_{j=1}^{p} g_j(m)}\right)^{M_{m,i}}\right] dt \underset{m\to\infty}{\to} t^*(f_1, c). \tag{A.1}$$

By definition, $W_{m,i}/\omega(m)$ is increasing in i, and from **H2**, for m sufficiently large, $W_{m,1}/\omega(m) = c(1)$. Moreover, $\sum_{j=1}^{p} g_j(m) \sim g_1(m) \to +\infty$, from **H1**. Then, for m sufficiently large, $\forall t > t^*(f_1, c)$, $e^{-\frac{W_{m,i}}{\omega(m)} t \sum_{j=1}^{p} g_j(m)} < \frac{1}{2}$. Using $\log(1-x) \geq -2x$ for all $x \leq 1/2$, we have

$$\sum_{i=1}^{|\mathbf{W}_m|} M_{m,i} \log\left(1 - e^{-\frac{W_{m,i}}{\omega(m)} t \sum_{j=1}^{p} g_j(m)}\right) \geq -2 \sum_{i=1}^{|\mathbf{W}_m|} M_{m,i} e^{-\frac{W_{m,i}}{\omega(m)} t \sum_{j=1}^{p} g_j(m)}.$$

From **H1**, we have that for all i, $M_{m,i} \leq \frac{e^{\sum_{j=1}^{p} f_j(i)g_j(m)}}{H(i)}$. From **H2**, for all i, $W_{m,i} \geq c(i)\omega(m)$. Thus,

$$\sum_{i=1}^{|\mathbf{W}_m|} M_{m,i} \log\left(1 - e^{-\frac{W_{m,i}}{\omega(m)} t \sum_{j=1}^{p} g_j(m)}\right) \geq -2 \sum_{i=1}^{|\mathbf{W}_m|} \frac{e^{\sum_{j=1}^{p}(f_j(i)-c(i)t)g_j(m)}}{H(i)}.$$

$\forall t > t^*(f_1, c)$, we have $f_1(i) - c(i)t < 0$ for all $i \leq |\mathbf{W}_m|$. From **H3**, there exists $K > 0$ such that for all $1 < j \leq p$, for all $i \leq |\mathbf{W}_m|$ and for all $t > t^*(f_1, c)$, $(f_j(i) - c(i)t) \leq K$. Then, $\forall i \in E_m$,

$$\sum_{j=1}^{p} (f_j(i) - c(i)t)\, g_j(m) \leq K \sum_{j=2}^{p} g_j(m) + (f_1(i) - c(i)t) g_1(m).$$

For all $j \neq 1$ we have $g_j = o(g_1)$. Thus, for m sufficiently large, $\sum_{j=1}^{p} (f_j(i) - c(i)t)\, g_j(m) \leq 2(f_1(i) - c(i)t)g_1(m)$. Then,

$$\begin{aligned}
\sum_{i=1}^{|\mathbf{W}_m|} M_{m,i} \log\left(1 - e^{-\frac{W_{m,i}}{\omega} t \sum_{j=1}^{p} g_j(m)}\right) &\geq -2 \sum_{i=1}^{|\mathbf{W}_m|} \frac{e^{2(f_1(i)-c(i)t)g_1(m)}}{H(i)} \\
&\geq -2e^{2g_1(m) \max_{i\in E_m}(f_1(i)-c(i)t)} \sum_{i=1}^{|\mathbf{W}_m|} \frac{1}{H(i)}.
\end{aligned}$$

From **H1**, there is $C>0$ such that $\sum_{i=1}^{|\mathbf{W}_m|}\frac{1}{H(i)}\leq C$. Moreover, we obviously have $\max_{i\in E_m}(f_1(i)-c(i)t)\leq \max_{i\in\mathbb{N}}(f_1(i)-c(i)t)$, which leads to

$$\sum_{i=1}^{|\mathbf{W}_m|} M_{m,i}\log\left(1-e^{-\frac{W_{m,i}}{\omega(m)}t\sum_{j=1}^{p}g_j(m)}\right)\geq -2Ce^{2g_1(m)\max_{i\in\mathbb{N}}(f_1(i)-c(i)t)}.$$

Then,

$$\begin{aligned}\int_{t^*(f_1,c)}^{\infty}\left[1-\exp\left(\sum_{i=1}^{|\mathbf{W}_m|} M_{m,i}\log\left(1-e^{-\frac{W_{m,i}}{\omega(m)}t\sum_{j=1}^{p}g_j(m)}\right)\right)\right]dt &\leq \int_{t^*(f_1,c)}^{\infty}\left[1-e^{-2Ce^{2g_1(m)\max_{i\in\mathbb{N}}(f_1(i)-c(i)t)}}\right]dt\\ &\leq 2C\int_{t^*(f_1,c)}^{\infty}e^{2g_1(m)\max_{i\in\mathbb{N}}(f_1(i)-c(i)t)}dt.\end{aligned}$$

Let $t^+>t^*(f_1,c)$, fixed with no other assumption. As for all $t>t^*(f_1,c)$, $\max_{i\in\mathbb{N}}(f_1(i)-c(i)t)<0$, and $g_1(m)\to+\infty$, we have $e^{2g_1(m)\max_{i\in\mathbb{N}}(f_1(i)-c(i)t)}\underset{m\to\infty}{\to}0$. Then

$$\int_{t^*(f_1,c)}^{t^+}e^{2g_1(m)\max_{i\in\mathbb{N}}(f_1(i)-c(i)t)}dt\underset{m\to\infty}{\to}0.$$

Besides, for all $t\geq t^+$, we have $f_1(i)-c(i)t\leq f_1(i)\frac{t}{t^+}-c(i)t$, hence

$$\max_{i\in E_m}(f_1(i)-c(i)t)\leq\max_{i\in E_m}(f_1(i)\frac{t}{t^+}-c(i)t)=\frac{t}{t^+}\max_{i\in E_m}(f_1(i)-c(i)t^+).$$

From Lemma A.1 and **H3**, this last maximum, denoted $-\gamma$, is actually reached and we have $-\gamma=\max_{i\in\mathbb{N}}(f_1(i)-c(i)t^+)<0$. Then,

$$\begin{aligned}\int_{t^+}^{\infty}e^{2g_1(m)\max_{i\in\mathbb{N}}(f_1(i)-c(i)t)}dt &\leq \int_{t^+}^{\infty}e^{-2\gamma g_1(m)\frac{t}{t^+}}dt\\ &= \frac{e^{-2\gamma g_1(m)}}{2\gamma g_1(m)}t^+\underset{m\to\infty}{\to}0\end{aligned}$$

and finally,

$$\int_{t^*(f_1,c)}^{\infty}\left[1-\prod_{i=1}^{|\mathbf{W}_m|}\left(1-e^{-\frac{W_{m,i}}{\omega(m)}t\sum_{j=1}^{p}g_j(m)}\right)^{M_{m,i}}\right]dt\underset{m\to\infty}{\to}0. \tag{A.2}$$

• Equations (A.1) and (A.2) lead to

$$\int_{0}^{\infty}\left[1-\prod_{i=1}^{|\mathbf{W}_m|}\left(1-e^{-\frac{W_{m,i}}{\omega(m)}t\sum_{j=1}^{p}g_j(m)}\right)^{M_{m,i}}\right]dt\underset{m\to\infty}{\to}t^*(f_1,c). \tag{A.3}$$

And finally, using **H1** and equation (A.3),

$$E[C_m]\sim t^*(f_1,c)\mu_m\sum_{j=1}^{p}g_j(m)\sim t^*(f_1,c)\mu_m g_1(m).$$

□

† Université de Versailles, PRISM/UMR 8144, Versailles, France

★ CNRS/Ecole Polytechnique/INRIA AMIB, LIX/UMR 7161 X-CNRS, Palaiseau, France