AofA'12 DMTCS proc. **AQ**, 2012, 243–264

The weighted words collector

Jérémie du Boisberranger[1], Danièle Gardy[1] and Yann Ponty[2†]

[1] *Université de Versailles, PRISM/UMR 8144, Versailles, France*
[2] *CNRS/Ecole Polytechnique/INRIA AMIB, LIX/UMR 7161 X-CNRS, Palaiseau, France*

We consider the word collector problem, i.e. the expected number of calls to a random weighted generator before all the words of a given length in a language are generated. 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 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 four exemplary languages, whose analyses reveal a large diversity of asymptotic waiting times, generally expressible as $\kappa \cdot m^p \cdot (\log m)^q \cdot (\log\log m)^r$, for m the number of words, and p, q, r some positive real numbers.

Keywords: Coupon Collector Problem; Waiting Time; Random Generation; Weighted Context-free Languages

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* [5] studied **weighted languages**, a natural generalization of context-free languages [10] 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].

These distributions, and their associated random generation algorithms, can also be found in bioinformatics, where 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 (A, C, G or 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

†Email: `yann.ponty@lix.polytechnique.fr`

a combinatorial perspective [19, 18]. A statistical sampling approach proposed by Ding and Lawrence [6] 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**, which draws 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 formal language 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 C_m of m items produced by a random source, studies the expected waiting time $E[C_m]$ of the full collection C_m, i.e. the expected number of generations before each item in C_m is present in the generated set. This problem naturally arises in a large variety of contexts, including the analysis of database [2] and network [11] 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 [8] 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 items 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. Weighted distributions over languages can be seen as highly specialized non-uniform coupon collections, whose major specificity is that many items may share the same probability or, in other words, some probability may appear with **large multiplicity**. Unfortunately, previous results either fail to apply to classes of coupons of high multiplicity, lead to bounds on the asymptotic behavior that are not tight [12], or require extensive *a priori* knowledge on the distribution, motivating further studies in the context of languages.

Intuitively, the waiting time of a non-uniform instance of the Coupon Collector problem is dominated by the generation of a subset composed of the least probable items. Indeed, some subset of items can be so improbable that it is typically fully obtained only after all the other items in the collection are generated. In such cases, a lower bound on the waiting time can be obtained by isolating the subset and analyzing its waiting-time as a uniform coupon collector problem. However, deciding which subset to study can be rather challenging, as the waiting time usually arises as a subtle tradeoff between the probability and the multiplicity. In the case of weighted languages, the presence of coupons having, simultaneously, large multiplicities and equally large discrepancy in their probabilities gives rises to a rich variety of asymptotic behaviors, and calls for a sophisticated – arguably technically involved – analysis.

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 by summarizing the contribution and describing future developments.

2 A general theorem for coupons of large multiplicities

2.1 Definitions and notations

Given a sequence $\mathbf{w} = \{w_i\}_{i=1}^m$ of positive numbers, or **weights**, associated with a collection C_m of items, one defines a **weighted probability distribution** $\{p_i\}_{i=1}^m$ over C_m 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 multiple items may share the same weight/probability. Let us then introduce $\mathbf{W}_m = \{W_{m,i}\}_i$ the increasingly-ordered, finite, sequence of all **distinct weights** in $\mathbf{w}$. Furthermore, for each $i \in [1, |\mathbf{W}_m|]$, let us denote by $M_{m,i}$ the **multiplicity** of the weight $W_{m,i}$, i.e. the number of occurrences of $W_{m,i}$ in $\mathbf{w}$. We observe that:

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

2.2 Main result

We describe a first-order asymptotical expression for the expected time of the full collection, assuming a large number m of items. Accessible weights $\mathbf{W}_m$ and their multiplicities $\mathbf{M}_m$ may in principle vary for different values of m, leading, in the extreme case, to the absence of a limit expression for the waiting time. Therefore we restrict the scope of our main theorem to distributions that obey three, essentially technical, conditions.

$\mathsf{H_1}$ - The number m of coupons and the weight rank i may interact only in a simple way within the multiplicity of a given weight. Thus we require that:

- There exists functions $f_1, \ldots, f_p$, $g_1, \ldots, g_p$, h and H, such that

$$M_{m,i} \underset{m\to\infty}{\sim} \frac{e^{\sum_{j=1}^{p} f_j(i) g_j(m)}}{h(i)}, \quad \text{and} \quad M_{m,i} \leq \frac{e^{\sum_{j=1}^{p} f_j(i) g_j(m)}}{H(i)},\ \forall m \geq 1, \forall i \leq |\mathbf{W}_m|.$$

- The functions f_1 and g_1 must effectively determine the growth of $M_{m,i}$, therefore one requires that: f_1 is positive and non-zero everywhere, $g_j(m) = o(g_1(m)), \forall j \in [2,p]$, and $g_1(m) \to +\infty$.
- Finally, $\sum_{i \in [1,|\mathbf{W}_m|]} \frac{1}{H(i)}$ must converge, to prevent H from capturing the growth of $M_{m,i}$.

$\mathsf{H_2}$ - Similarly, we restrict the possible interactions of the weight rank i and the number m of items within the i-th weight $W_{m,i}$, by requiring the existence of functions $\nu(i) > 0$ and $\omega(m) > 0$ such that

$$W_{m,i} \geq \nu(i) \cdot \omega(m), \ \forall m \geq 1, \forall i \geq 1,$$

and such that any weight at rank i, beyond some value of m, remains constant:

$$\forall k > 0, \exists m_k > 0 \text{ such that } W_{m,i} = \nu(i) \cdot \omega(m), \ \forall m \geq m_k, \forall i \leq k.$$

$\mathsf{H_3}$ - The multiplicity $M_{m,i}$ must not grow too quickly in comparison with the weight $W_{m,i}$. More precisely, if $|\mathbf{W}_m| \underset{m\to\infty}{\to} \infty$, then one must have

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

The conditions are sufficient (yet not always necessary) to obtain the asymptotic behavior of the waiting time, and hold for a large class of weighted languages.

Theorem 2.1 *Assume that, for all $m > 0$, the weights $\mathbf{W}_m$ and multiplicities $\boldsymbol{M}_m$ of the coupon collection satisfy the conditions $\mathsf{H_1}$, $\mathsf{H_2}$, and $\mathsf{H_3}$. Then, as $m \to \infty$, one has*

$$E[C_m] = t^*(F,\nu) \cdot G(m) \cdot \frac{\mu(m)}{\omega(m)} \cdot (1 + o(1)), \tag{2.1}$$

where:

- $\mu(m)$ *is the total weight of all coupons;*
- $F \equiv f_1$ *and* $G \equiv g_1$*, defined in* $\mathsf{H_1}$*, drive the leading term of the growth of* $M_{m,i}$ *as* $m \to \infty$*;*
- $\omega(m) \cdot \nu(1)$ *is the smallest weight within the collection of cardinality* m *(see* $\mathsf{H_2}$*);*
- $t^*(F,\nu)$ *is the largest value of* t *such that there exists* $x \in \mathbb{N}$ *such that* $F(x) - t \cdot \nu(x) > 0$.

Sketch of proof.

We give here a brief description on our proof, whose details can be found in Appendix A.

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

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

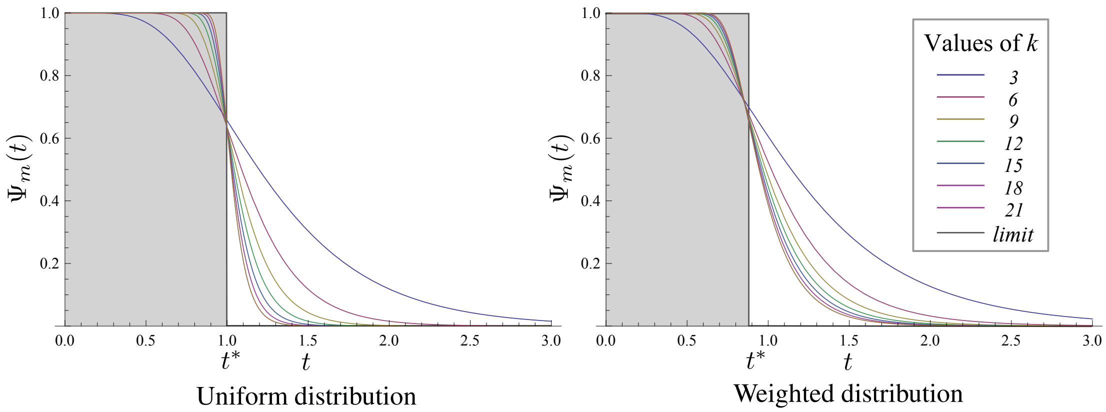

Fig. 1: Plots of the $\Psi_m(t)$ functions as appearing for a uniform (Left) and weighted (Right, $\pi(a)/\pi(b) = 2/3$) distribution over the $(a+b)^*$ language. We consider $m = 2^k$ coupons/words, for several values of $k \in \{3, 6, 9, 12, 15, 18, 21\}$. The convergence of $\Psi_m(t)$ towards a step 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.

Focusing on this expression, one shows that the integral of $\Psi_m(t)$ converges towards some constant. Indeed, numerical computations, as illustrated by Figure 2.2, suggest that Ψ_m converges toward a step function when $m \to \infty$. This can be rigorously proved under the conditions $\mathsf{H_1}$, $\mathsf{H_2}$ and $\mathsf{H_3}$, and the integral from 0 to $t^*(f_1, \nu)$ converges to $t^*(f_1, \nu)$, while the remaining integral converges to 0. $\square$

Remark 1 Theorem 2.1 applies in the special case of the uniform distribution. Indeed, considering the weight collection is $\{w_i\}_{i=1}^m = \{1\}_{i=1}^m$, one has $p_i = 1/m$ and $\mu_m = m$. The set of weights is then reduced to the singleton $\mathbf{W}_m = (1)$, which has multiplicity $M_{m,1} = m = e^{\log m}$.

- $\mathsf{H_1}$ is satisfied upon taking

$$F(i) := f_1(i) = 1, \qquad G(m) := g_1(m) = \log m, \qquad h(1) = 1 \qquad \text{and} \qquad H(1) = 1,$$

  noticing that $\sum_{1 \le i \le |\mathbf{W}_m|} 1/H(i) = 1/H(1)$, which obviously converges.

- Since one has $W_{m,1} = 1$, then $\mathsf{H_2}$ is satisfied with $\nu(1) = 1$ and $\omega(m) = 1$.

- Since $\mathbf{W}_m$ is finite, the limit condition of $\mathsf{H_3}$ does not need to be verified.

One then easily verifies that $t^*(f_1, \nu) = t^*(1, 1) = 1$, and applying Theorem 2.1 unsurprisingly gives $E[C_m] \sim m \log m$, which is 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 remind the reader that m is the number of coupons/words, which typically grows exponentially along 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 [16], Berenbrink and Sauerwald [2] consider the two approximations

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

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

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

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 knowledge over the multiplicity of coupons comparable to the one required 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 asymptotics 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.

Their first result requires the existence of a $\xi \in]0,1]$ such that $S := \sum_{k=1}^{\infty} \xi^{a_k} < \infty$. However, under the hypotheses of our main Theorem 2.1, there always exists a weight of unbounded multiplicity as m goes to the infinity, and S therefore diverges for any value of ξ.

Their second result is based on the assumption of a decreasing sequence $\boldsymbol{\alpha}$. However, many weighted distributions that satisfy hypothesis $\mathsf{H_1}$ to $\mathsf{H_3}$ cannot be defined by truncating a fixed decreasing sequence. For instance, suppose that for all m, the accessible weights are $\{\frac{2k-1}{k}\}_{k=1}^{m} \cup \{2\}$, each appearing with multiplicity m. It is easily checked that such a set of weights cannot be ordered into a decreasing sequence whose truncations include the families of coupons weights.

Conversely, distributions with low multiplicity satisfying their conditions are not covered by our Theorem 2.1. Therefore, their results and ours are complementary, and seldom overlapping.

Neal [14]: The limiting distribution

Neal studied the distribution of the waiting time. Although the results described in the article can in principle be used to assess the expectation of the waiting time, checking the prerequisites of its main theorem turns out to be considerably more involved than checking those of Theorem 2.1. In particular, one has to figure out suitable sequences, 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 to.

4 Applications to languages: the word collector

4.1 Weighted Languages

Let us remind some definitions introduced by Denise *et al* [5]. 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 naturally defines a weighted 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 a n, all the characteristics of the weight distribution, such as $\mathbf{W}_m$, will be indexed by n instead of m.

4.2 Verifying preconditions $\mathsf{H_1}$, $\mathsf{H_3}$ and $\mathsf{H_3}$ in the context of weighted languages

Let us outline a systematic method to verify the preconditions $\mathsf{H_1}$, $\mathsf{H_2}$ and $\mathsf{H_3}$ for a language $\mathcal{L}$ defined over an alphabet $\Sigma = (a_1, \dots, a_k)$. The idea is, firstly, to classify the words of the language according to their weights 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 this is done, the hypotheses of Theorem 2.1 are usually easily verified.

- **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 $(x_1, \dots, x_k)$. Suppose that $1 = \pi_{a_1} = \cdots = \pi_{a_l}$ for some l, then the sub-composition of a word of composition $(x_1, \dots, x_k)$ is the vector $(x_{l+1}, \dots, x_k)$, 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.*

  Firstly, let us define an ordering function over $\mathcal{L}_n$, which 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 $\mathbf{W}_n$. One has*

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

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

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

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

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

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

  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 $\mathsf{H_3}$ directly follows from steps 1 and 2. However, conditions $\mathsf{H_1}$ and $\mathsf{H_2}$ require good approximations of $|\Lambda_n|$ and S_n. Such approximations strongly depend on the language $\mathcal{L}$ of interest, therefore we present several examples to illustrate the method.

4.3 Application to specific languages

In this part, we shall denote by D_n the collection of all words of length n, and assume that pairs of non-unit weights are incommensurable, which implies that sub-compositions can be bijectively associated with weights.

4.3.1 *The unconstrained language* Σ^*

Let us consider the language $\mathcal{L} = \Sigma^*$, where $\Sigma = (a_1, \ldots, a_k)$. It is worth noticing that the weighted distribution is stable upon multiplying each weight by a constant factor, therefore we assume without loss of generality that $1 = \pi_{a_1} = \cdots = \pi_{a_l}$ for some $l \geq 1$, and $1 < \pi_{a_{l+1}} \leq \cdots \leq \pi_{a_k}$.

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 $W_{n,i} = \pi_{a_{l+1}}^{\tilde{\boldsymbol{\phi}}_{n,1}(i)} \cdots \pi_{a_k}^{\tilde{\boldsymbol{\phi}}_{n,k-l}(i)}$, and it follows that, for sufficiently large values of n, one has $W_{n,i} = \pi_{a_{l+1}}^{\tilde{\boldsymbol{\phi}}_1(i)} \cdots \pi_{a_k}^{\tilde{\boldsymbol{\phi}}_{k-l}(i)}$. Consequently, Condition $\mathsf{H_2}$ is verified with

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

In $\mathcal{L}_n$, the number of words of composition $(x_1, \cdots, x_k)$ is $M(x_1, \cdots, x_k) = \binom{n}{x_1, \cdots, x_k}$, thus the number of words of sub-composition $(x_{l+1}, \cdots, x_k)$ is $M(x_{l+1}, \cdots, x_k) = l^{n-x_{l+1}-\cdots-x_k}\binom{n}{x_{l+1}, \ldots, x_k}$. 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 $\binom{n}{a_1, \ldots, a_k}$. 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 $S := \sum_{j=l+1}^{k} \log \pi_{a_j}$, $P := \prod_{j=l+1}^{k} \log \pi_{a_j}$, and let introduce a notation*

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

Then the following inequalities hold:

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

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

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

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

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

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

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

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

Proof.

i) Remind that $\phi(\mathbf{x})$ 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 $H(x_{l+1}, \ldots, x_k)$ and the $(k-l-1)$-pyramid under $H(x_{l+1}+1, \ldots, x_k+1)$.

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

iii) The function $\phi_n(\mathbf{x})$ 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 $|\tilde{\phi}_n(i)| \leq n$. 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 $|\tilde{\phi}(i)| \geq n$ and one finally gets $|\tilde{\phi}_n(i)| \leq |\tilde{\phi}(i)|, \forall n > 0, \forall i \geq 1$.

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

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

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

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

The convergence of $\sum_i 1/\left(\frac{\log \pi_{a_{l+1}}}{\log \pi_{a_k}} |\tilde{\phi}(i)|\right)!$ follows from Equation (4.7). Therefore, $\mathsf{H_1}$ is satisfied for the following choice of functions

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

where $z = \log \pi_{a_{l+1}} / \log \pi_{a_k}$. Furthermore it can be verified that $\mathsf{H_3}$ is satisfied, since Equation (4.7) gives a lower bound for $\nu(i)/F(i)$. Consequently, Theorem 2.1 applies to the weighted distribution on Σ^*, and we get.

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

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

where l is the number of letters of lowest weight, $\mu(n) = \left(l + \sum_{j=l+1}^{k} \pi_{a_j}\right)^n$ is the total weight, $\kappa_1 = t^*\left(|\tilde{\phi}(i)|, \lambda\right)$, and $\kappa_2 = t^*(\log l, \lambda)$ with $\lambda = \pi_{a_{l+1}}^{\tilde{\phi}_1(i)} \cdots \pi_{a_k}^{\tilde{\phi}_{k-l}(i)}$.

Corollary 4.7 *Define $p = \log(\pi_{a_1} + \cdots + \pi_{a_k})/\log k$, noting that $p \geq 1$ and $p = 1$ only in the uniform case. The expected waiting time $E[C_m]$ for obtaining the $m = k^n$ words of length n in $\mathcal{L} = \Sigma^*$ is asymptotically equivalent to*

- $\kappa_1 \cdot m^p \cdot \log\log m$, *if there is a single lettre of smallest weight;*
- $(\kappa_2/\log k) \cdot m^p \cdot \log m$, *if there are at least two letters of smallest weight.*

4.3.2 Motzkin words

Motzkin words are well-parenthesized expressions featuring any number of dot characters $\bullet$. 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 to generate all Motzkin words of even length n. For the sake of readability, we replace the characters (,) and $\bullet$ by letters a, $\bar{a}$ and b respectively. Since parentheses come in pairs, any word has equal number of occurrences of a and $\bar{a}$, and the parity of the number of occurrences of b is the parity of the word length. Consequently, accessible compositions for words of length n are triplets $(x_a, x_{\bar{a}}, x_b)$ of the form $(k, k, n-2k)$, 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 shows two types of behavior depending on whether a or $\bar{a}$ have the smallest weight. To give a flavor of our result and illustrate its proof strategy, we explicitly derive two exemplary results for the cases where $1 = \pi_b < \pi_a < \pi_{\bar{a}}$ and $1 = \pi_a = \pi_{\bar{a}} < \pi_b$, and give the general form of the asymptotic equivalent for the weighted Coupon Collector.

First case: $(1 = \pi_b < \pi_a < \pi_{\bar{a}})$. Here, the sub-compositions $(x_a, x_{\bar{a}})$ 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 $\mathsf{H_2}$ is satisfied with $\nu(i) = \pi_a^{i-1}\pi_{\bar{a}}^{i-1}$ 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 $\mathsf{H_1}$ is satisfied with

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

coupled with the observation that $\sum_i 1/i(i-1)!^2$ converges. The verification of $\mathsf{H_3}$ is immediate, and applying Theorem 2.1 readily gives the following result.

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

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

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 case: $(1 = \pi_a = \pi_{\bar{a}} < \pi_b)$. In this second case, the sub-compositions (x_b) are of the form $(n-2k)$, for $0 \le k \le n/2$, and the weight of a word increases with the number of occurrences of b. Consequently, one has $W_{n,i} = \pi_b^{2(i-1)}$, and $\mathsf{H_2}$ is satisfied with $\nu(i) = \pi_b^{2(i-1)}$ 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

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

Finally, one has $M_{n,i} \le 2^n \frac{n^{2(i-1)-\frac{3}{2}}}{\sqrt{\pi}2^{2(i-1)-\frac{3}{2}}(2(i-1))!}$, for $i \le n/2$, and $\mathsf{H_1}$ is satisfied with

$f_1(i)$	$f_2(i)$	$g_1(n)$	$g_2(n)$	$h(i)$	$H(i)$
$\log 2$	$2(i-1)-\frac{3}{2}$	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 $\mathsf{H_3}$ is immediate.

Proposition 4.9 *The expected waiting time of the full collection of weighted Motzkin words of even length n, under the configuration $1 = \pi_a = \pi_{\bar{a}} < \pi_b$, admits the following asymptotic behavior:*

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

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

This approach can be extended to any relative positioning of π_a, $\pi_{\bar{a}}$ and π_b. The symmetrical roles played by the letters a and $\bar{a}$, allow for a restriction, without loss of generality, to cases where $\pi_{\bar{a}} \le \pi_a$. Also, singularity analysis can be applied to the generating function of the weighted Motzkin language, giving $\mu_n \sim \kappa \cdot \rho^{-n} \cdot n^{-3/2}$, with $\rho = (\pi_b + 2\sqrt{\pi_a}\pi_{\bar{a}})^{-1}$.

Proposition 4.10 *The expected waiting time for generating all Motzkin words of length n obeys:*

$$E[D_n] \sim \begin{cases} \kappa\frac{\rho^{-n}}{\sqrt{n}} & \text{with } \rho = \frac{\sqrt{\pi_a}}{\pi_b+2\sqrt{\pi_a\pi_{\bar{a}}}} \quad \text{if } \pi_b^2 > \pi_a\pi_{\bar{a}}, \\ \kappa'\frac{\rho'^{-n}}{n\sqrt{n}}\log n & \text{with } \rho' = \frac{\pi_b}{\pi_b+2\sqrt{\pi_a\pi_{\bar{a}}}} \quad \text{if } \pi_b^2 < \pi_a\pi_{\bar{a}}, \end{cases}$$

where κ and κ' are constants of n which can be explicitly computed (and depends on the relative positions of the weights).

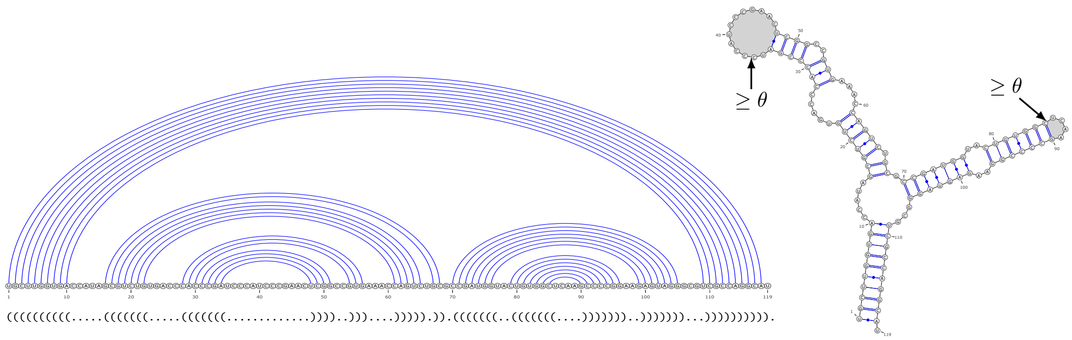

Fig. 2: Secondary structure of a 5s ribosomal RNA. A well-parenthesized expression (lower-left) unambiguously defines a set of matching position (upper-left) which *folds* into a projection of a three-dimensional conformation of the molecule. The latter representation illustrates the relationship between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

Corollary 4.11 *Let m be the number of Motzkin words of length n ($m \sim 3(\sqrt{3}/2\sqrt{\pi})3^n n^{-3/2}$). The expected waiting time for generating the complete collection of m words obeys*

$$E[C_n] \sim \begin{cases} \kappa \cdot m^p \cdot \log(m)^{\frac{3p-1}{2}} & \textit{with } p = \frac{\log(\pi_b + 2\sqrt{\pi_a \pi_{\bar{a}}}) - \log\sqrt{\pi_a}}{\log 3} \quad \textit{if } \pi_b^2 > \pi_a \pi_{\bar{a}}, \\ \kappa' \cdot m^{p'} \cdot \log(m)^{\frac{3(p'-1)}{2}} \cdot \log\log m & \textit{with } p' = \frac{\log(\pi_b + 2\sqrt{\pi_a \pi_{\bar{a}}}) - \log \pi_b}{\log 3} \quad \textit{if } \pi_b^2 < \pi_a \pi_{\bar{a}}, \end{cases}$$

for constants κ and κ' that can be explicitly computed (and depend on the relative positions of the weights).

4.3.3 RNA secondary structures

Through an adaptation of Viennot *et al* [17], secondary structures can be generated by a grammar:

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

where θ is the minimal distance between matching parenthesis, enforcing steric constraints. The connection between the secondary structure and the conformations of an RNA sequence is illustrated by Figure 2: Matching parentheses represent base-pairs, or interacting pairs of nucleotides mediated by hydrogen bonds. Such base-pairs are known to stabilize a secondary structure, thus decreasing its free-energy. In this model, we consider a simple free-energy model proposed by Nussinov[15] which assigns a -1 kcal/mol contribution to each base-pair. The free-energy $E(S)$ of a secondary structure S is then inherited additively by summing the individual contributions of its base-pairs.

One can assume a Boltzmann distribution on the set of secondary structures, where the probability of any secondary structure S of length n is proportional to its Boltzmann factor $e^{-E(S)/RT}$, with R the gas constant and T the temperature. Such a non-deterministic perspective over the RNA folding process is fundamental to a recent paradigm shift in RNA structure prediction [6] based on random generation. In the worst-case scenario, the complexity of this algorithm is equivalent to a coupon collector for Boltzmann weighted secondary structures. It is then worth noticing that the Boltzmann distribution is just a special case of a weighted distribution, where a neutral weight 1 is assigned to unpaired positions, and a weight $e^{1/RT}$ to each pair of matching parentheses.

Again in this example, we replace the characters (,) and • by letters a, $\bar{a}$ and b respectively. Let us denote by $\mathcal{L}^{(rna)}$ the language of RNA secondary structure. For the sake of simplicity, let us assume, without loss of generality, that $1 = \pi_b < \pi_a < \pi_{\bar{a}}$, with $\pi_a \cdot \pi_{\bar{a}} = e^{1/RT}$. The compositions are triplets $(x_a, x_{\bar{a}}, x_b)$ of the form $(k, k, n-2k)$, $0 \leq k \leq n/2$. The 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 the 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 $\bar{a}$. It follows that $W_{n,i} = \pi_a^{(i-1)}\pi_{\bar{a}}^{(i-1)}$, and $\mathsf{H_2}$ is satisfied with $\nu(i) = \pi_a^{(i-1)}\pi_{\bar{a}}^{(i-1)}$ and $\omega(n) = 1$.

Moreover, the multiplicity of the weight $W_{n,i}$ is the number of words having sub-composition $(x_a, x_{\bar{a}})$ of the form $(i-1, i-1)$, and is given by

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

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 $p > (i-1)$ has null contribution.

One also has $M_{n,i} \leq 2\frac{n^{2(i-1)}}{i(i-1)!^2}$, for all i, thus $\mathsf{H_1}$ 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 $\mathsf{H_3}$ is immediate. Setting

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

one verifies, e.g. from the strong-connectivity of the grammar [7], that

$$\mu_n \sim c \cdot \left(\frac{1}{\rho^\theta}\right)^n n^{-3/2}$$

where c is a constant, and ρ^θ is the dominant singularity of $\sum_{n\geq 0} \mu(n) \cdot z^n$.

Proposition 4.12 *The expected waiting time for the collection of Boltzmann-factor weighted RNA secondary structures of length n, assuming $1 = \pi_b < \pi_a < \pi_{\bar{a}}$, admits the following asymptotic behavior:*

$$E[D_n] \sim \kappa \cdot \frac{\rho_\theta^{-n}}{n\sqrt{n}} \cdot \log n, \ \forall \theta \in \mathbb{N}^+$$

where $\kappa = t^*\left(2(i-1), (\pi_a\pi_{\bar a})^{i-1}\right)$ *and, setting* $q = \pi_a\pi_{\bar a}$*,* ρ_θ *is the smallest positive real solution of*

$$1-4z+(6-2q)z^2+4(q-1)z^3+(1-2q)z^4-2qz^{\theta+2}+4qz^{\theta+3}-2q(1+q)z^{\theta+4}+q^2z^{2\theta+4}=0.$$

Corollary 4.13 *Define* η_θ *as the smallest positive solution of the equation*

$$1-4z+4z^2-z^4-2z^{\theta+2}+4z^{\theta+3}-4z^{\theta+4}+z^{2\theta+4}=0.$$

Then the number m *of RNA structures of length* n *is asymptotically equal to* $\lambda\cdot\eta_\theta^n\cdot n^{-3/2}$*, and the asymptotic waiting time of the full collection is given by*

$$E[C_m]\sim\kappa\cdot m^p\cdot(\log m)^{3p/2}\cdot\log\log m,$$

where $p=-\frac{\log\rho_\theta}{\log\eta_\theta}$*, and* λ *and* κ *are constants that can be fully specified.*

4.3.4 A non strongly-connected language

Let us finally consider the language $\mathcal{L}^{(nc)}$ over an alphabet $\{a,\bar a,b\}$ and generated by the grammar

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

It is worth noticing that this grammar is not strongly connected, and the distributions of letters may therefore be untypical (non-normal and/or expectation/variance not in $O(n)$ [7]). Here, this grammar models binary trees, whose leftward edges along the leftmost branch are marked by a dedicated letter $\bar a$, and any other leftward (resp. rightward) edge is marked by a (resp. b).

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 compositions are therefore triplets $(x_a,x_{\bar a},x_b)$ of the form $(n/2-k,k,n/2)$, for $1\le k\le n/2$. An elementary computation shows that the number of 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, leading to different behaviors. Let us illustrate the approach on one out of the 9 possible configurations, such that $1=\pi_b<\pi_a<\pi_{\bar a}$.

In this case, the sub-compositions are pairs $(x_a,x_{\bar a})$ of the form $(n/2-k,k)$, for $1\le k\le n/2$. Moreover, since the weight of the word increases with the number of $\bar a$, then one has $W_{n,i}=\pi_a^{n/2-i}\pi_{\bar a}^i$, and $\mathsf{H_2}$ is therefore satisfied with $\nu(i)=(\pi_{\bar a}/\pi_a)^i$ and $\omega(n)=\pi_a^{n/2}$.

Remark 2 The influence of the configuration (ordering of the weights) only appears in the definition of the functions ν and ω. The function ω may become constant (equal to 1) when either $\pi_b=\pi_a=1$ or $\pi_b=\pi_{\bar a}=1$.

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

$$M_{n,i}=\binom{n-i-1}{n/2-1}-\binom{n-i-1}{n/2}\underset{n\to\infty}{\sim}2^{n-i}n^{-\frac{3}{2}}i\sqrt{\frac{2}{\pi}}.\tag{4.10}$$

Since $M_{n,i}\le 2^{n-i+1}n^{-\frac{3}{2}}i\sqrt{2/\pi}$ for all $1\le i\le n/2$ and $\sum_i i/2^i$ converges, the condition $\mathsf{H_1}$ is satisfied for the following functions:

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

The verification of $\mathsf{H_3}$ is immediate.

From (4.10), one has $\mu(n) = \sum_{k=1}^{n/2} \left[\binom{n-k-1}{n/2-1} - \binom{n-k-1}{n/2}\right] \pi_a^{\frac{n}{2}-k} \pi_{\bar{a}}^k$, whose asymptotic behaviour obeys

$$\mu(n) \sim 2\sqrt{2}\frac{\pi_a \pi_{\bar{a}}}{(2\pi_a - \pi_{\bar{a}})^2}\left(2\sqrt{\pi_a}\right)^n n^{-3/2}.$$

Proposition 4.14 *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:*

$$E[D_n] \sim \kappa \cdot \frac{2^n}{\sqrt{n}},$$

where $\kappa = t^* \left(\log 2, (\pi_{\bar{a}}/\pi_a)^i\right) \cdot \frac{2\sqrt{2}\pi_a\pi_{\bar{a}}}{(2\pi_a - \pi_{\bar{a}})^2}$.

Corollary 4.15 *Let m be the number of words of even length n in $\mathcal{L}^{(nc)}$, asymptotically equivalent to $2\sqrt{2/\pi} \cdot 2^n \cdot n^{-3/2}$. The expected waiting time of the full collection is*

$$E[C_m] \sim \kappa \cdot m \cdot (\log m)^{5/2},$$

where κ is a constant that can be explicitly computed.

Again, these results can be extended to any relative ordering of π_a, $\pi_{\bar{a}}$ and π_b, and one obtains the following result.

Proposition 4.16 *The expected waiting time for all words of even length n in $\mathcal{L}^{(nc)}$ is equivalent to*

$$E[D_n] \sim \begin{cases} \kappa \cdot \frac{2^n}{\sqrt{n}} & \text{if } \pi_a = 1, \text{ or } 1 = \pi_b \le \pi_a < \pi_{\bar{a}}, \\ \kappa' \cdot \left(\frac{\pi_a}{\pi_{\bar{a}}}\right)^{n/2} \cdot 2^n \cdot \frac{\log n}{n\sqrt{n}} & \text{otherwise,} \end{cases}$$

where κ and κ' are constants that can be explicitly computed.

Corollary 4.17 *Let m be the number of words of even length m in $\mathcal{L}^{(nc)}$. Then the expected waiting time of the complete collection is asymptotically equal to*

$$E[C_m] \sim \begin{cases} \kappa \cdot m^2 \cdot (\log m)^{5/2} & \text{if } \pi_a = 1 \text{ or } \pi_b = 1 \le \pi_a < \pi_{\bar{a}}, \\ \kappa' \cdot m^{2+q} \cdot (\log m)^{2+q/2} \cdot \log\log m & \text{otherwise, with } q = \log_2(\pi_a/\pi_{\bar{a}}) \end{cases}$$

where κ and κ' are constants that can be explicitly computed.

5 Conclusion

In this extended abstract, we studied a language generalization of the ubiquitous Coupon Collector Problem. Focusing on collections of weighted coupons having large multiplicities, we contributed a new theorem that relates the asymptotic waiting time of the full to the growth of the multiplicity of coupons of a given weight. We compared the novelty of the contribution against pre-existing work on the subject. We discussed the application of our theorem to weighted languages in general, and particularly on four languages showing different properties (rational vs context-free, simple-type vs non-square-root singularities, limited vs parameterized alphabet…).

Quite interestingly, our study of four illustrative examples reveals a large variety of expressions for the waiting-time. As a function of the word length n, we observed waiting times of the form $\kappa \cdot \mu(n) \cdot n$ and $\kappa \cdot \mu(n) \cdot \log(n)$, depending essentially on the multiplicity of the smallest weights. As a function of the number of coupons m, we obtained estimates of the general form $\kappa \cdot m^p \cdot (\log m)^q \cdot (\log\log m)^\theta$, where p and q are irrational numbers and $\theta \in \{0, 1\}$. Such a diversity partly not only arises from differences regarding the nature of the asymptotical growth within the language, but also reflects subtle differences in the accumulation of the contributions of the least probable words. To our opinion, this illustrates the versatility of the method, and hints toward a significant amount of work being required, in the case of approximations [2].

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 [9]) 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 [7]. 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 [10]. The main difficulties behind such an extension are related to 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

The authors wish to thank an anonymous reviewer for suggesting a more intuitive presentation of our main result. 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), DMTCS Proceedings, June 2010, pp. 49–64.

[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] Y. Ding and E. Lawrence, *A statistical sampling algorithm for RNA secondary structure prediction*, Nucleic Acids Research **31** (2003), no. 24, 7280–7301.

[7] M. Drmota, *Systems of functional equations*, Random Struct. Alg. **10** (1997), 103–124.

[8] 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.

[9] P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge University Press, 2009.

[10] P. Flajolet, P. Zimmermann, and B. Van Cutsem, *Calculus for the random generation of labelled combinatorial structures*, Theoretical Computer Science **132** (1994), 1–35.

[11] D. Gardy, *Occupancy urn models in the analysis of algorithms*, Journal of Statistical Planning and Inference **101** (2002), no. 1-2, 95 – 105.

[12] Danièle Gardy and Yann Ponty, *Weighted random generation of context-free languages: Analysis of collisions in random urn occupancy models*, Proceedings of GASCom'10, 2010.

[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] R. Nussinov and A.B. Jacobson, *Fast algorithm for predicting the secondary structure of single-stranded rna*, Proc Natl Acad Sci U S A **77** (1980), 6903–13.

[16] S.M. Ross, *Introduction to probability models*, 10th ed., Elsevier Science, 2009.

[17] M. Vauchaussade de Chaumont and G. Viennot, *Polynômes orthogonaux et problèmes d'énumération en biologie moléculaire*, Séminaire Lotharingien de Combinatoire (1983).

[18] 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.

[19] M. S. Waterman, *Secondary structure of single stranded nucleic acids*, Advances in Mathematics Supplementary Studies **1** (1978), no. 1, 167–212.

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) $\forall t > t^*(f,g), \quad \forall\, x \in E, \quad f(x) - tg(x) < 0$

- $\exists\, x_1 \in \mathbb{N}^*$ *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 infinity. 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, and **H3** (if $|\mathbf{W}_m| \underset{m\to\infty}{\to} \infty$), there is $i_0 \in E$ such that $f_1(i_0) - \nu(i_0)t > 0$. Moreover, from **H2**, for m sufficiently large, $W_{m,i_0} = \nu(i_0)\omega(m)$. Then

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

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,

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

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

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

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

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

$\forall t > t^*(f_1, \nu)$, we have $f_1(i) - \nu(i)t < 0$ 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 $t > t^*(f_1, \nu)$, $(f_j(i) - \nu(i)t) \leq K$. Then, $\forall i \in E$,

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

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

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

From **H1**, there is $C > 0$ such that $\sum_{i=1}^{|\mathbf{W}_m|} \frac{1}{H(i)} \leq C$. Moreover, we obviously have

$$\max_{i \in E} (f_1(i) - \nu(i)t) \leq \max_{i \in \mathbb{N}} (f_1(i) - \nu(i)t)$$

, which leads to

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

Then,

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

Choose $t^+ > t^*(f_1, \nu)$, without any other assumption. As for all $t > t^*(f_1, \nu)$, $\max\limits_{i \in \mathbb{N}}(f_1(i) - \nu(i)t) < 0$, and $g_1(m) \to +\infty$, we have $e^{2g_1(m) \max\limits_{i\in\mathbb{N}}(f_1(i)-\nu(i)t)} \underset{m\to\infty}{\to} 0$. Then

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

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

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

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

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

and finally,

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

• Equations (A.1) and (A.2) lead to

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

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

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

□