Expected Recovery Time in DNA-based Distributed Storage Systems

Adi Levy* Roni Con* Eitan Yaakobi* Han Mao Kiah†

Abstract

We initiate the study of *DNA-based distributed storage systems*, where information is encoded across multiple DNA data storage containers to achieve robustness against container failures. In this setting, data are distributed over M containers, and the objective is to guarantee that the contents of any failed container can be reliably reconstructed from the surviving ones. Unlike classical distributed storage systems, DNA data storage containers are fundamentally constrained by sequencing technology, since each read operation yields the content of a uniformly random sampled strand from the container. Within this framework, we consider several erasure-correcting codes and analyze the expected recovery time of the data stored in a failed container. Our results are obtained by analyzing generalized versions of the classical Coupon Collector's Problem, which may be of independent interest.

Contents

1 Introduction **2**
1.1 Distributed DNA Storage and Coupon Collector's Problem 3

2 Our Results **4**
2.1 The *Scalar* MDS Code Case 4
2.2 The MDS *Array* Code Case 5

3 Preliminaries **7**

4 Warmup: Proof of Theorem 2 **8**

5 Proof of Theorem 3 **10**

6 Proof of Theorem 1 **15**

7 Proof of Theorem 4 **20**

*Department of Computer Science, Technion—Israel Institute of Technology, Haifa 3200003, Israel, emails: {levyadi, roni.con, yaakobi}@cs.technion.ac.il

†School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, email: hmkiah@ntu.edu.sg

A.L., R.C., and E.Y. are supported by the European Union (DiDAX, 101115134). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

1 Introduction

Recent advances in DNA-based data storage technologies have positioned DNA as a compelling substrate for long-term digital archiving, offering density, durability, and longevity far beyond the capabilities of conventional media. As global data generation accelerates at an unprecedented pace, existing storage technologies such as magnetic disks, solid-state drives, and tapes struggle to meet the demands of scalability, reliability, and archival lifetime. DNA, by contrast, provides an ultra-dense and chemically stable medium whose information can persist for thousands of years under modest conditions. These features position DNA-based storage as a promising solution for the growing archival storage crisis [CGK12,GBC+13,YKGR+16,OAC+18].

A conventional DNA-based data storage system is composed of three main elements: DNA synthesis, storage containers, and DNA sequencing. First, artificial DNA strands, called oligos, are produced to encode the user's data. These strands are then stored in an unordered manner within a storage container, with each strand present in millions of copies. Finally, DNA sequencing is used to read the stored strands and transform them into digital sequences, known as reads, which are then decoded to reconstruct the original user information. In each of these three components, errors might occur, and thus, much attention was given to designing error-correcting codes for DNA-based storage systems. We refer the reader to the following survey [SKSY24] that presents the coding challenges in such systems.

In this work, we initiate the study of *coding for DNA-based distributed storage systems*. We consider a setting with M DNA storage containers, and our objective is to encode data across these M containers so that, even if one or more containers fail, the lost information can still be reconstructed from the remaining containers. The formulation of this problem matches that of standard distributed storage systems. However, there are a few important differences between our notion of repair and the one adopted in the classical storage setting.

In a classical distributed storage system, a file is divided into k data fragments and then encoded using an erasure-correcting code, producing $M > k$ fragments in total, with each fragment stored on a different one of the M nodes. Since an erasure code is used, repairing a failed node reduces to decoding a single erasure of a codeword. A central performance metric in traditional distributed storage systems is the *repair bandwidth*, which denotes the amount of information downloaded from the remaining nodes to recover the failed one. The problem of minimizing the repair bandwidth was introduced in the seminal work of Dimakis et al. [DGW+10] and has since attracted extensive attention, leading to ingenious constructions that achieve optimal tradeoffs (for a detailed overview, the reader is referred to the survey [RBS+22]). We emphasize that in this model, nodes are allowed to perform arbitrary computations on their stored data and transmit the resulting outputs.

In DNA-based distributed storage, the process of downloading information from containers is fundamentally different. First, no computation can be performed within a container. Second, reading is carried out by a sequencer, which we assume samples a strand uniformly at random from the container in each read. Because each of the n strands is present in the tube in about a million identical copies, all in equal numbers, drawing a strand uniformly at random from the container is effectively the same as choosing one of the n distinct strands uniformly at random.

It was observed in [BLSGY24], that this sampling model is deeply connected with the classical coupon collector's problem (CCP, for short). In the CCP, there are n distinct coupons, and at each point of time, the collector picks a coupon uniformly at random from the n coupons, and then returns it. The task of the collector is to see all the n distinct coupons, and the question is how many rounds it takes. This problem has been extensively investigated in combinatorics and probability theory; see, for instance, [ER61, BH97, AR01, KJV07, AD21]. However, the classical problem received many new twists with the new connection to DNA-based storage systems [BLSGY24, AGY24, GBLRY24, GMZ24, BEG+25].

In this paper, we introduce a DNA-based distributed storage system model. We establish a connection between this model and a generalized coupon collector's problem involving multiple independent collectors, each with its own set of coupons to be collected. To the best of our knowledge, this generalized setting has not been studied previously and may be of independent interest.

1.1 Distributed DNA Storage and Coupon Collector's Problem

For a positive integer n, let $[n] \triangleq \{1, \ldots, n\}$. We consider a distributed storage system with M containers, each storing n distinct strands. Each container is equipped with its own sequencer, and reads are performed in parallel: at each time step, one strand is sampled uniformly at random from each functioning container. When one container fails, our goal is to reconstruct its data from the remaining $M-1$ containers, and we study the expected time required for this recovery.

There could be several reasons for a container to become unavailable. These include physical degradation or loss of DNA molecules due to chemical decay or handling errors [CGK12, GBC+13, YKGR+16], as well as failures in the sequencing or library preparation pipeline that render the contents of a container unreadable despite the DNA remaining intact [OAC+18, SKSY24]. At scale, additional failure modes include contamination or mislabeling events that prevent reliable identification of a container's data [YKGR+16, OAC+18]. From a coding-theoretic perspective, such events are naturally modeled as container-level erasures, in direct analogy to node failures in classical distributed storage systems [DGW+10, RBS+22].

Remark 1. *We emphasize that, in practice, each strand is stored in millions of identical copies. For simplicity, and as an approximation of this setting, we assume that at any time, each container contains exactly n distinct strands, from which one strand is sampled uniformly at random, read, and returned.*

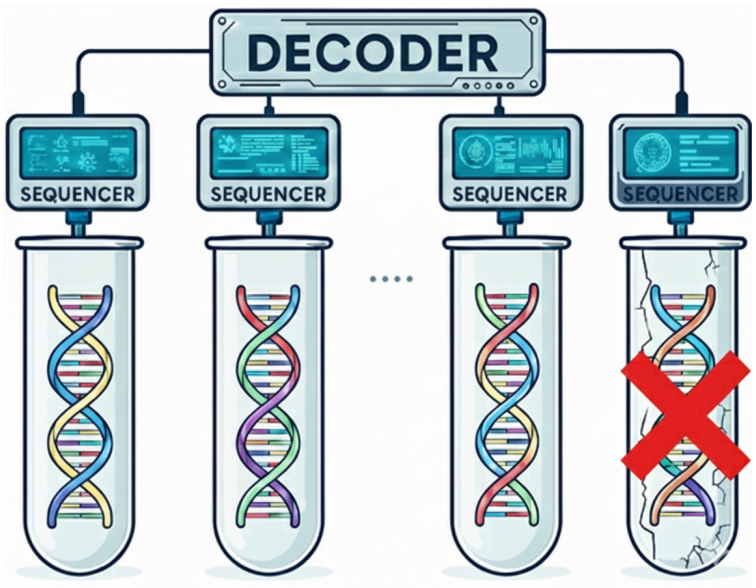

Figure 1: An illustration of a DNA-DSS

We begin by formally defining a DNA-based distributed storage system, and then introduce the main parameter of interest, namely the expected recovery time of a container.

Definition 1 ($(n, M, k, |\Sigma|)$ DNA-based distributed storage system (DNA-DSS))**.** *Let n, M, and k be positive integers and let Σ be an alphabet. An $(n, M, k, |\Sigma|)$* DNA-based distributed storage system (DNA-DSS) *is defined by an injective encoder $\mathcal{E} : \Sigma^k \to \Sigma^{n \times M}$ such that, for every $\boldsymbol{v} \in \Sigma^k$ and every $j \in [M]$, the following condition is satisfied: The j-th column of $S \triangleq \mathcal{E}(\boldsymbol{v})$ can be reconstructed from the collection of columns of S indexed by $[M] \setminus \{j\}$.*

Note that this definition is purely mathematical and does not yet refer to DNA-based storage systems. The relation to DNA-based storage is as follows. Suppose $\boldsymbol{v} \in \Sigma^k$ represents the data we wish to store and set $S = \mathcal{E}(v)$. For each $j \in [M]$, the j-th container stores the n entries of the j-th column of S. Thus, every container holds n values, corresponding to the n strands within that container. The requirement in Definition 1 can be rephrased as follows: the data in the failed j-th container can be reconstructed by reading all the data from the remaining $M-1$ containers.

Definition 2 (The expected recovery time of a container)**.** *For every $j \in [M]$, we define T_j as the random variable representing the time required to recover the j-th container under this model. The* expected recovery time of the j-th container *will be denoted by $\mathbb{E}[T_j]$.*

In this paper, we design DNA-DSS using *MDS (maximum distance separable) codes* and analyze the expected recovery time of a failed container. An MDS code of length M and redundancy r is an injective encoder $\mathcal{E} : \Sigma^{M-r} \to \Sigma^M$ such that any $\mathbf{v} \in \Sigma^{M-r}$ can be uniquely recovered from any $M - r$ entries of the *codeword* $\mathcal{E}(\mathbf{v})$. Before presenting our main results, we briefly revisit the classical coupon collector problem. We shall see that several of our results can be viewed as generalizations of this problem.

The Coupon Collector's Problem. We give a formal definition of the classical coupon collector distribution, as well as of the distribution corresponding to the version with ℓ copies.

Definition 3 (Coupon Collector's distribution)**.** *Consider n distinct coupon types, and draw one coupon at each trial independently and uniformly from these n types. Let $X^{(n)}$ be the number of draws needed to obtain at least one of each type, and let $X^{(n,\ell)}$ be the number of draws needed to obtain at least ℓ of each type. We write $X^{(n)} \sim CCP(n)$ and $X^{(n,\ell)} \sim CCP(n, \ell)$.*

It is a well-known fact that $\mathbb{E}[X^{(n)}] = n(\ln n + \gamma) + O(1)$, where $\gamma = 0.577\ldots$ is the Euler–Mascheroni constant. (e.g., [MU05, Section 2.4.1]). In [New60, Theorem 2], it was shown that when ℓ is fixed, we have $\mathbb{E}\left[X^{(n,\ell)}\right] = n \ln n + (\ell-1)n \ln\ln n + n \cdot C_\ell + o(n)$, and later, in [ER61], it was shown that $C_\ell = \gamma - \ln((\ell-1)!)$.

2 Our Results

MDS codes have the best known rate-erasure correcting tradeoff and are therefore widely studied in classical distributed storage systems. We likewise study the performance of MDS codes in the DNA-DSS setting. Our results follow the classical split: first, we consider *scalar MDS codes*, and then *array regenerating MDS codes*. We note, however, that MDS and regenerating codes may not be optimal for minimizing the expected container recovery time, and we leave this question for future work.

Throughout this paper, we work under the assumption that the number of containers, denoted by M, is fixed, while the number of distinct strands n tends to infinity. In addition, whenever we use the notation $o(1)$, it represents a term that goes to 0 as n goes to infinity. Finally, for numbers a, b and c, we will denote by $a \pm b$ the interval $[a - b, a + b]$ and $c(a \pm b) = ca \pm cb$.

2.1 The *Scalar* MDS Code Case

We design an $(n, M, n(M-r), q)$ DNA-DSS based on an MDS code of length M and redundancy r as follows. Let $S' \in \Sigma^{n\times(M-r)}$ denote the information to be stored (in the notation of Definition 1, this corresponds to $|\Sigma| = q$ and $k = n(M - r)$). Each row vector of S' is encoded using the encoder of the MDS code into a row vector $\boldsymbol{v}$ of length M. Let $S \in \Sigma^{n\times M}$ denote the matrix whose rows are these encoded vectors, and store each column of S in a distinct container. Once a container fails, the MDS property guarantees that to recover its content, it is enough to know any $M - r$ symbols from each row of S.

We therefore inquire for the expected time it takes to read this data from the functional containers. We will formulate this question, and obtain its answer, in terms of the following random process.

Theorem 1. *Let $A^{(0)} = \mathbf{0}^{n\times(m+\rho)}, A^{(1)}, A^{(2)}, \ldots$ be a sequence of matrices constructed as follows: for each $t \in \mathbb{N}$, we draw $(v_1, \ldots, v_{m+\rho}) \sim \mathrm{Unif}([n]^{m+\rho})$ and set*

$$A^{(t)} = A^{(t-1)} + \left[\boldsymbol{e}_{v_1}, \ldots, \boldsymbol{e}_{v_{m+\rho}}\right] ,$$

where $\boldsymbol{e}_a$ denotes the column vector with a 1 in the a-th position and 0 in all other entries. Define $T_{n,m,\rho}$ as the random variable that corresponds to the smallest t such that every row of $A^{(t)}$ contains at least m nonzero entries. Then,

$$\mathbb{E}\left[T_{n,m,\rho}\right] = \frac{n}{\rho+1}\left(\ln n + \ln\binom{m+\rho}{m-1} + \gamma \pm o(1)\right).$$

The next corollary states the associated DNA-DSS, derived from Theorem 1 by instantiating it with a standard MDS code having redundancy $r = \rho + 1$ and length $M = m + \rho + 1$. As concrete choices, one may use a *doubly extended Reed–Solomon code* over an alphabet of prime-power size [Rot06, Section 5, Problem 5.2], or, in the special case where a single container acts as the redundancy container, a simple parity code suffices.

Corollary 1. *Let $n, M, r \in \mathbb{N}$. If $r = 1$, choose any positive integer $q \geq 2$. If $r \geq 2$, choose q to be a prime power satisfying $q \geq M - 1$. Then there exists an $(n, M, n(M-r), q)$ DNA-DSS such that the expected recovery time, T_j, of every container $j \in [M]$ is*

$$\mathbb{E}[T_j] = \frac{n}{r} \ln n + \frac{n}{r} \ln \binom{M-1}{r} + \frac{\gamma n}{r} \pm o(n) \ .$$

A direct theorem for $r = 1$. We observe that Theorem 1 and Corollary 1 already handle the situation in which exactly one container stores redundant data. Nonetheless, to build intuition for the proof of Theorem 1, we present here and prove in Section 4a simpler statement, which can be viewed as a straightforward extension of the classical coupon collector problem. In the special case $r = 1$, whenever a container fails, recovering it necessitates reading *all* strands from each of the remaining containers. This is equivalent to asking for the expected time until m independent coupon collectors (each with its own set of n coupons) collect all their coupons. Formally, we show the following.

Theorem 2. *Let $n \in \mathbb{N}$ and let m be a constant positive integer with respect to n. Let $X_1^{(n)}, \ldots, X_m^{(n)}$ be i.i.d. random variables with $X_j^{(n)} \sim CCP(n)$. Then,*

$$\mathbb{E}\left[\max_{j \in [m]} X_j^{(n)}\right] = n\left(\ln(mn) + \gamma \pm o(1)\right) \ .$$

In practice, each strand is stored in many identical copies, so sequencing typically yields multiple reads of the same strand. When errors are present, having several corrupted versions of each strand can enhance the tradeoff between rate and reliability. In this work, however, we do not address errors. Instead, we generalize Theorem 2 to the setting where ℓ copies of each coupon are required, laying the groundwork for future extensions that incorporate also storage errors.

Theorem 3. *Let m and ℓ be constant positive integers. Let $X_1^{(n,\ell)}, \ldots, X_m^{(n,\ell)}$ be i.i.d. random variables with $X_j^{(n,\ell)} \sim CCP(n, \ell)$. Then,*

$$\mathbb{E}\left[\max_{j \in [m]} X_j^{(n,\ell)}\right] = n\left(\ln(nm) + (\ell - 1)\ln\ln n + C_\ell \pm o(1)\right)$$

where $C_\ell = \gamma - \ln((\ell - 1)!)$.

2.2 The MDS *Array* Code Case

Our next result can be viewed as a generalization of Theorem 1, motivated by insights from the theory of regenerating codes in classical distributed storage. Let $S' \in \Sigma^{n \times (M-r)}$ represent the information to be stored. In the DNA-DSS given in Corollary 1, we encoded each row of S' independently to obtain the encoded matrix S. Consequently, to reconstruct any column of S, it was necessary to read at least a fixed number of symbols from every row of S.

We now proceed with a more general construction. Partition S' into n/b consecutive submatrices $S'_1, \ldots, S'_{n/b}$, each of dimension $b \times (M - r)$. Concretely, the first submatrix is formed by the top b rows, the second by the next b rows, and so forth. Each submatrix S'_i is then encoded into a new submatrix S_i of size $b \times M$ using an encoder $\mathcal{E} : \Sigma^{b \times (M-r)} \to \Sigma^{b \times M}$ of an MDS array code. The resulting matrix is obtained by stacking $S_1, \ldots, S_{n/b}$ vertically. Note that here, the MDS property states that knowing the values of any $M - r$ columns of S suffice to recover S.

Assume now that the p-th container has failed and needs to be reconstructed, and focus, for the moment, only on the first block S_1. Let $M_p \triangleq [M] \setminus \{p\}$ denote the set of all remaining operational containers.

The encoding function $\mathcal{E}$, together with the failed index p, defines a *bad blocks configurations family* $\mathcal{B}_p \subseteq \mathcal{P}([b] \times M_p)$ (here, $\mathcal{P}$ denotes the power set) consisting of all index sets that cannot be used to reconstruct the p-th column. Namely, $B \in \mathcal{B}_p$ if and only if the p-th column of S_1 cannot be reconstructed from the collection of symbols $\{(S_1)_{i,j} \mid (i,j) \in B\}$. Formally,

Definition 4. (Bad blocks configuration with respect to $\mathcal{E}$ and p) *Let $\mathcal{E} : \Sigma^{b\times(M-r)} \to \Sigma^{b\times M}$ and fix an index $p \in [M]$. Define $M_p \triangleq [M] \setminus \{p\}$ and let $\mathcal{B}_p \subseteq \mathcal{P}([b] \times M_p)$ be a collection of subsets with the following property: a set B belongs to $\mathcal{B}_p$ if and only if there exist two distinct encoded matrices $S, S' \in \mathcal{E}(\Sigma^{b\times(M-r)})$ such that $S_{i,j} = S'_{i,j}$ for every $(i,j) \in B$, while the p-th column of S is* not equal *to the p-th column of S'. Equivalently, even after observing all entries $S_{i,j}$ with $(i,j) \in B$, one still cannot uniquely determine the p-th column of S. Moreover, define*

$$\alpha_p^* \triangleq (M-1)b - \max_{B\in\mathcal{B}_p} |B| \ , \quad \textit{and}$$
$$\beta_p^* \triangleq |\{B' \in \mathcal{B}_p \mid |B'| = \max_{B\in\mathcal{B}_p} |B|\}| \ .$$

Namely, α_p^* is the smallest number of missing entries of S for which the p-th column cannot be recovered and β_p^* is the be number of bad recovering sets of maximal size in $\mathcal{B}_p$.

Note that since the same encoder $\mathcal{E}$ is applied to every block, the family $\mathcal{B}_p$ is the same for all matrices $S_1, \ldots, S_{n/b}$. Thus, the question we consider is the expected time until, for each block, we have gathered a set of indices that *does not* belong to $\mathcal{B}_p$. Similarly to Theorem 1, the next theorem defines the random process and the stopping criterion that characterize recovery, given a bad blocks configuration.

Theorem 4. *Let $A^{(0)} = \mathbf{0}^{n\times m}, A^{(1)}, A^{(2)}, \ldots$ be a sequence of matrices constructed as follows: for each $t \in \mathbb{N}$, we draw $(v_1, \ldots, v_m) \sim \mathrm{Unif}([n]^m)$ and set*

$$A^{(t)} = A^{(t-1)} + [\boldsymbol{e}_{v_1}, \ldots, \boldsymbol{e}_{v_m}] \ .$$

Let $b \in \mathbb{N}$ that divides n and let $\mathcal{B} \subseteq \mathcal{P}([b] \times [m])$. As Definition 4, define

$$\alpha^* \triangleq mb - \max_{B\in\mathcal{B}} |B| \qquad \beta^* \triangleq |\{B' \in \mathcal{B} \mid |B'| = \max_{B\in\mathcal{B}} |B|\}|$$

Define T as the random variable that corresponds to the minimal t such that for all $a \in [n/b]$, we have that

$$\left\{(i,j) \in [b] \times [m] \mid A^{(t)}_{i+(a-1)b,j} \neq 0\right\} \notin \mathcal{B} \ .$$

Then,

$$\mathbb{E}[T] \leq \frac{n}{\alpha^*} \ln n + \frac{\beta^*}{b\alpha^*} \cdot n + o(n) \ ,$$

Theorem 4, combined with Definition 4, yields a framework for establishing an upper bound on the expected recovery time of each container in a DNA-DSS.

Corollary 2. *Let $M, b, r, n \in \mathbb{N}$ where b divides n. An injective encoding function $\mathcal{E} : \Sigma^{b\times(M-r)} \times \Sigma^{b\times M}$ yields an $(n, M, n(M-r), |\Sigma|)$ DNA-DSS code such that for every $p \in [M]$, we have*

$$\mathbb{E}[T_p] \leq \frac{n}{\alpha_p^*} \ln n + \frac{\beta_p^*}{b\alpha_p^*} n + o(n) \ ,$$

where β_p^* and α_p^* are defined in Definition 4.

Example 1. *We specialize this framework to one of the simplest examples of an MDS regenerating code [TWB13, Fig. 1], defined over $\mathbb{F}_3$. Let n be an even number. Consider the $(n, 4, 2n, 3)$ DNA-DSS defined by the encoding function $\mathcal{E} : \mathbb{F}_3^{2\times 2} \to \mathbb{F}_3^{2\times 4}$ defined by*

$$\mathcal{E}\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = \begin{pmatrix} a & b & a+b & a+2d \\ c & d & c+d & c+b \end{pmatrix} \ ,$$

where arithmetic operations are done over $\mathbb{F}_3$. Assume that $p = 1$ and observe that there are exactly two sets of size 3 that recover the first column: $\{(1,2),(1,3),(2,4)\}$ and $\{(2,2),(2,3),(1,4)\}$. Further, there are only two sets of size 4 that do not *recover the first column: $\{(1,2),(1,3),(1,4),(2,2)\}$ and $\{(1,2),(2,2),(2,3),(2,4)\}$. Thus,*

δ	0	1	2	3	4	5	6
b_δ	1	6	$\binom{6}{2} = 15$	$\binom{6}{3} - 2 = 18$	2	0	0

where $b_\delta = |\{B \in \mathcal{B}_1 \mid |B| = \delta\}|$. Since $M = 4$ and $b = 2$, we get, $\alpha_1^* = 3 \cdot 2 - 4 = 2$ and $\beta_1^* = 2$. In total, we get

$$\mathbb{E}[T_1] \leq \frac{n}{2} \ln n + \frac{n}{2} + o(n) \ .$$

Note that, setting $M = 4$ and $r = 2$ in the scalar MDS DNA-DSS construction of Corollary 1, we obtain $\mathbb{E}[T_1] \approx \frac{n}{2} \ln n + 0.838n \pm o(n)$, which shows that the regenerating code provides an improvement of at least $0.338n$.

3 Preliminaries

We shall use the following limit theorem about the coupon collector distribution.

Theorem 5 (Theorem 5.13 [MU05])**.** *Let $X \sim CCP(n)$. Then, for any constant c,*

$$\lim_{n\to\infty} \Pr\left[X > n \ln n + cn\right] = 1 - e^{-e^{-c}}$$

This limit theorem is exactly a special case of the Gumbel distribution whose definition is given next.

Definition 5 (Gumbel distribution)**.** *The cumulative distribution function of Gumbel distribution with parameters μ, β is*

$$F(x; \mu, \beta) = e^{-e^{\frac{-(x-\mu)}{\beta}}} .$$

If X is a random variable with Gumbel distribution with parameters μ, β, we denote $X \sim Gumbel(\mu, \beta)$.

Definition 6 (Convergence in distribution)**.** *A sequence $\mathcal{X} = X^{(1)}, X^{(2)}, \ldots$ of real-valued random variables with cumulative distribution functions $F_1, F_2, \ldots$ is said to converge in distribution to a random variable X with cumulative distribution function F if $\lim_{n\to\infty} F_n(x) = F(x)$, for every $x \in \mathbb{R}$ at which F is continuous. We denote convergence in distribution by $\mathcal{X} \xrightarrow{\mathcal{D}} X$.*

Lemma 1. *Let $\mathcal{X}_1, \ldots, \mathcal{X}_m$ be m sequences of random variables such that for every $n \in \mathbb{N}$, $X_1^{(n)}, X_2^{(n)}, \ldots, X_m^{(n)}$ are i.i.d. and for all $j \in [m]$, we have $\mathcal{X}_j \xrightarrow{\mathcal{D}} Gumbel(\mu, \beta)$. For every $n \in \mathbb{N}$, denote $X_{\max}^{(n)} \triangleq \max_j X_j^{(n)}$ and let $\mathcal{X}_{\max} \triangleq X_{\max}^{(1)}, X_{\max}^{(2)}, \ldots$. It holds that*

$$\mathcal{X}_{\max} \xrightarrow{\mathcal{D}} \text{Gumbel}(\beta \ln(m) + \mu, \beta).$$

Proof. By the definition of convergence in distribution (Definition 6) and the Gumbel distribution (Definition 5), for every t and $j \in [m]$,

$$\lim_{n\to\infty} \Pr\left[X_j^{(n)} \leq t\right] = e^{-e^{-\frac{(t-\mu)}{\beta}}} .$$

For every fixed $n \in \mathbb{N}$, since $X_1^{(n)}, \ldots, X_m^{(n)}$ are i.i.d., for every fixed t and m, we have

$$\Pr\left[X_{\max}^{(n)} \leq t\right] = \Pr\left[\forall j \in [m], X_j^{(n)} \leq t\right] = \Pr\left[X_1^{(n)} \leq t\right]^m .$$

Therefore, for any fixed t, m,

$$\lim_{n\to\infty} \Pr\left[X_{\max}^{(n)} \leq t\right] = \lim_{n\to\infty} \Pr\left[X_1^{(n)} \leq t\right]^m = \exp\left(-e^{-\frac{(t-\mu)}{\beta}}\right)^m = \exp\left(-e^{-\frac{(t-(\mu+\beta\ln(m))}{\beta}}\right).$$

By Definition 5 and Definition 6, we conclude that $\mathcal{X}_{\max} \xrightarrow{\mathcal{D}} \text{Gumbel}(\beta \ln(m) + \mu, \beta)$ as required. □

For the following definition we shall need the following notation of an indicator. Let E be any event. Then, define

$$\mathbb{1}_E = \begin{cases} 1 & \text{if } E \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Definition 7 (Uniformly integrable, Definition 4.1 [Gut12])**.** *A sequence $\mathcal{X} = X^{(1)}, X^{(2)}, \ldots$ of random variables with finite expectation is called uniformly integrable if for any $n \in \mathbb{N}$, it holds that $\mathbb{E}[|X^{(n)}| \cdot \mathbb{1}_{|X^{(n)}|>K}] \to 0$ as $K \to \infty$, uniformly in n.*

For a sequence of uniformly integrable variables, we have the following.

Theorem 6 (Theorem 3.5, [Bil13])**.** *Let $\mathcal{X} = X^{(1)}, X^{(2)}, \ldots$ be a uniformly integrable sequence such that $\mathcal{X} \xrightarrow{\mathcal{D}} X$ to some random variable X. Then X has finite expectation and $\lim_{n\to\infty} \mathbb{E}[X^{(n)}] = \mathbb{E}[X]$.*

Throughout the paper, we shall want to prove that a sequence of random variables is uniformly integrable. The following lemma gives a sufficient condition for that.

Lemma 2. *Let $\mathcal{X} = X^{(1)}, X^{(2)}, \ldots$ be a sequence of random variables and assume there exist $C, c > 0$ such that for every $t \geq 0$ and any $n \in \mathbb{N}$, it holds that $\Pr[|X^{(n)}| > t] \leq Ce^{-ct}$. Then, $\mathcal{X}$ is uniformly integrable.*

Proof. We will verify that the two conditions specified in Definition 7 hold.

1. For every $n \in \mathbb{N}$, we have that

$$\begin{aligned} \mathbb{E}[|X^{(n)}|] &= \int_0^\infty \Pr\left[|X^{(n)}| > t\right] dt \\ &\leq C \int_0^\infty e^{-ct} dt = \frac{C}{c} < \infty \,. \end{aligned}$$

   Thus, $X^{(n)}$ has finite expectation.

2. For any $n \in \mathbb{N}$ and positive K,

$$\begin{aligned} 0 \leq \mathbb{E}[|X^{(n)}| \cdot \mathbb{1}_{|X^{(n)}|\geq K}] &= \int_K^\infty \Pr\left[|X^{(n)}| > t\right] dt \\ &\leq C \int_K^\infty e^{-ct} dt = \frac{C}{c} e^{-cK} \,. \end{aligned}$$

   Since $\lim_{K\to\infty} \frac{C}{c} e^{-cK} = 0$ and is independent of n, $\lim_{K\to\infty} \mathbb{E}[|X^{(n)}| \cdot \mathbb{1}_{|X^{(n)}|\geq K}] = 0$ uniformly in n.

We conclude that $\mathcal{X}$ is uniformly integrable. □

4 Warmup: Proof of Theorem 2

We start by introducing the notations of this section. Let m be a constant positive integer. For all $j \in [m]$, let $\mathcal{X}_j = \{X_j^{(n)}\}_{n\geq 2}$ be a sequence of random variables where $X_j^{(n)} \sim \mathrm{CCP}(n)$. Furthermore, we assume that for every fixed $n \in \mathbb{N}$, the variables $X_1^{(n)}, \ldots, X_m^{(n)}$ are i.i.d. For every $n \in \mathbb{N}$, let $X_{\max}^{(n)} \triangleq \max_{j\in[m]} X_j^{(n)}$.

We will be interested in the following normalized versions of $\mathcal{X}_1, \ldots, \mathcal{X}_m$. For every $j \in [m]$ and $n \in \mathbb{N}$, define $Y_j^{(n)} \triangleq \frac{X_j^{(n)}}{n} - \ln n$ and $Y_{\max}^{(n)} \triangleq \max_{j\in[m]} Y_j^{(n)}$. Finally, for every $j \in [m]$, we set $\mathcal{Y}_j \triangleq \{Y_j^{(n)}\}_{n\geq 2}$ and $\mathcal{Y}_{\max} \triangleq \{Y_{\max}^{(n)}\}_{n\geq 2}$.

The proof of Theorem 2 consists of the following steps. First, in Claim 1, we show that $\mathcal{Y}_{\max}$ converges in distribution to the Gumbel distribution with $\mu = \ln m$ and $\beta = 1$. Then, in Claim 2, we bound the tails of $Y_{\max}^{(n)}$ which is needed to show in Corollary 3, that $\mathcal{Y}_{\max}$ is uniformly integrable. Finally, to prove Theorem 2 we will apply Theorem 6.

Claim 1. *It holds that*

$$\mathcal{Y}_{\max} \xrightarrow{\mathcal{D}} \text{Gumbel}(\ln(m), 1).$$

Proof. Note that for all $n \in \mathbb{N}$, and every $j \in [m]$ we have that $nY_j^{(n)} + n\ln n \sim \text{CCP}(n)$. By Theorem 5,

$$\lim_{n\to\infty} \Pr\left[nY_j^{(n)} + n\ln n > n\ln n + cn\right] = 1 - e^{-e^{-c}} ,$$

which implies that

$$\lim_{n\to\infty} \Pr[Y_j^{(n)} \le c] = e^{-e^{-c}} ,$$

and thus, by Definition 6, the sequence $\mathcal{Y}_j = Y_j^{(1)}, Y_j^{(2)}, \ldots$ converges in distribution to Gumbel$(0, 1)$. Applying Lemma 1 with $\mathcal{Y}_1, \ldots, \mathcal{Y}_m$ proves the claim. □

Claim 2. *For any $n \ge 2$ and any $x \ge 0$,*

$$\Pr\left[\left|Y_{\max}^{(n)}\right| > x\right] \le (2 + m) \cdot e^{-x}.$$

Proof. We first show that $\Pr[Y_{\max}^{(n)} > x] \le m \cdot e^{-x}$. Indeed, By the definition of $Y_{\max}^{(n)}$, we have

$$\begin{aligned}
\Pr\left[Y_{\max}^{(n)} > x\right] &= \Pr\left[X_{\max}^{(n)} > n(\ln n + x)\right] \\
&= \Pr\left[\text{some coupon was } \textbf{not} \text{ collected at time } n(\ln n + x)\right] \\
&\le nm\left(1 - \frac{1}{n}\right)^{n(\ln n + x)} \\
&\le me^{-x} ,
\end{aligned}$$

where the first inequality is by a union bound.

Now, we show that $\Pr[Y_{\max}^{(n)} < -x] \le 2e^{-x}$. First observe that

$$\Pr\left[Y_{\max}^{(n)} < -x\right] \le \Pr\left[Y_{\max}^{(n)} \le -x\right] = \Pr\left[Y_1^{(n)} \le -x\right]^m = \Pr\left[X_1^{(n)} \le n(\ln n - x)\right]^m .$$

Next, we shall bound $\Pr[X_1^{(n)} < n(\ln n - x)]$. Fix x, set $t \triangleq n(\ln n - x)$, and define $W^{(t)}$ to be the number of missing coupons after t draws were drawn for the first collector. Note that $W^{(t)} = \sum_{j=1}^n I_j^{(t)}$ where $I_j^{(t)}$ is the indicator that j-th coupon was *not* collected in the first t draws. Clearly,

$$\Pr\left[X_1^{(n)} \le n(\ln n - x)\right] = \Pr\left[X_1^{(n)} \le t\right] = \Pr\left[W^{(t)} = 0\right] .$$

We shall bound the expectation and variance of $W^{(t)}$ in order to upper bound $\Pr[W^{(t)} = 0]$. We have that

$$\mathbb{E}\left[W^{(t)}\right] = \sum_{j=1}^n \mathbb{E}\left[I_j^{(t)}\right] = n \cdot \left(1 - \frac{1}{n}\right)^{n(\ln n - x)} \ge n \cdot e^{-\frac{n}{n-1}\cdot(\ln n - x)} \ge n^{-\frac{1}{n-1}} \cdot e^x \ge \frac{1}{2}e^x ,$$

where the first inequality follows since $(1 - 1/n)^n \ge e^{-n/(n-1)}$ for all $n \ge 2$ and last inequality follows since $n^{-1/(n-1)}$ is monotonically increasing. For the variance of $W^{(t)}$, we first observe that

$$\text{Var}\left(W^{(t)}\right) = \sum_{i\in[n]} \text{Var}\left(I_i^{(t)}\right) + 2\sum_{1\le i<j\le n} \text{Cov}\left(I_i^{(t)}, I_j^{(t)}\right) .$$

Now, since $I_i^{(t)}$ is an indicator, then $\text{Var}(I_i^{(t)}) \le \mathbb{E}[I_i^{(t)}]$. Moreover,

$$\text{Cov}\left(I_i^{(t)}, I_j^{(t)}\right) = \mathbb{E}\left[I_i^{(t)} \cdot I_j^{(t)}\right] - \mathbb{E}\left[I_i^{(t)}\right] \cdot \mathbb{E}\left[I_j^{(t)}\right] = \left(1 - \frac{2}{n}\right)^t - \left(1 - \frac{1}{n}\right)^{2t} < 0,$$

where the inequality follows since $(1-2/n) < (1-1/n)^2$. Combining, we have that $\mathrm{Var}(W^{(t)}) < \mathbb{E}[W^{(t)}]$.

Thus, by plugging these bounds into the Chebyshev inequality,

$$\Pr\left[W^{(t)}=0\right] \le \Pr\left[\left|W^{(t)}-\mathbb{E}\left[W^{(t)}\right]\right| \ge \mathbb{E}\left[W^{(t)}\right]\right] \le \frac{\mathrm{Var}\left(W^{(t)}\right)}{\mathbb{E}\left[W^{(t)}\right]^2} \le \frac{1}{\mathbb{E}\left[W^{(t)}\right]} \le 2e^{-x}.$$

We conclude that

$$\Pr\left[Y_{\max}^{(n)} < -x\right] = \left(\Pr\left[W^{(t)}=0\right]\right)^m \le \Pr\left[W^{(t)}=0\right] \le 2e^{-x} .$$

□

Corollary 3. *The sequence $\mathcal{Y}_{\max}$ is uniformly integrable.*

Proof. Follows directly from Lemma 2 and Claim 2. □

We are now ready to prove Theorem 2 which is restated for convenience.

Theorem 2. *Let $n \in \mathbb{N}$ and let m be a constant positive integer with respect to n. Let $X_1^{(n)}, \ldots, X_m^{(n)}$ be i.i.d. random variables with $X_j^{(n)} \sim CCP(n)$. Then,*

$$\mathbb{E}\left[\max_{j\in[m]} X_j^{(n)}\right] = n\left(\ln(mn) + \gamma \pm o(1)\right) .$$

Proof. According to Corollary 3, the sequence $\mathcal{Y}_{\max}$ is uniformly integrable, and converges in distribution to $\mathrm{Gumbel}(\ln(m), 1)$ and so by Theorem 6, we have that

$$\lim_{n\to\infty} \mathbb{E}\left[Y_{\max}^{(n)}\right] = \mathbb{E}\left[\mathrm{Gumbel}(\ln(m), 1)\right] = \ln(m) + \gamma ,$$

Therefore, we can write

$$\mathbb{E}\left[Y_{\max}^{(n)}\right] = \ln(m) + \gamma \pm o(1) .$$

Remembering that $X_{\max}^{(n)} = n\ln n + nY_{\max}^{(n)}$, we conclude that

$$\mathbb{E}[X_{\max}^{(n)}] = n\ln n + n\mathbb{E}[Y_{\max}^{(n)}] = n\left(\ln(mn) + \gamma \pm o(1)\right) .$$

□

5 Proof of Theorem 3

The proof of Theorem 3 follows the same general strategy as that of Theorem 2. Specifically, we combine convergence in distribution with uniform integrability to establish the convergence of expectations of a sequence of normalized random variables and then infer the expectation of the target random variable. However, the analysis in this case is more involved.

We shall first establish notations. Fix $\ell \in \mathbb{N}$. For every $j \in [m]$, let $\mathcal{X}_j = \{X_j^{(n,\ell)}\}_{n\ge 2}$ be a sequence of random variables where $X_j^{(n,\ell)} \sim \mathrm{CCP}(n,\ell)$. As before, we assume that for any fixed $n \in \mathbb{N}$, we have that $X_1^{(n,\ell)}, X_2^{(n,\ell)}, \ldots, X_m^{(n,\ell)}$ are i.i.d., and we also denote $X_{\max}^{(n,\ell)} \triangleq \max_{j\in[m]} X_j^{(n,\ell)}$. For every $j \in [m]$, define $Y_j^{(n,\ell)} \triangleq \frac{X_j^{(n,\ell)}}{n} - \ln n - (\ell-1)\ln\ln n$ and, as before, $\mathcal{Y}_j \triangleq \{Y_j^{(n,\ell)}\}_{n\ge 2}$. Finally, define $\mathcal{Y}_{\max} = \{Y_{\max}^{(n,\ell)}\}_{n\ge 2}$ where $Y_{\max}^{(n,\ell)} = \max_{j\in[m]} Y_j^{(n,\ell)}$. Recall that we assume that m and ℓ are constants with respect to n.

The next theorem, due to Erdős and Rényi [ER61], extends the limit theorem for the classical coupon collector problem (given in Theorem 5), to the version of the coupon collector problem where one seeks ℓ copies of each coupon.

Theorem 7 ([ER61]). *Let $X^{(n,\ell)} \sim CCP(n,\ell)$, then*

$$\lim_{n\to\infty} \Pr\left[\frac{X^{(n,\ell)}}{n} < \ln n + (\ell-1)\ln\ln n + x\right] = \exp\left(-\frac{e^{-x}}{(\ell-1)!}\right) .$$

Equivalently, for $Y^{(n,\ell)} \triangleq \frac{X^{(n,\ell)}}{n} - \ln n - (\ell-1)\ln\ln n$, we have that $Y^{(n,\ell)} \xrightarrow{\mathcal{D}} Gumbel(-\ln((\ell-1)!),1)$.

Theorem 7, alongside Lemma 1, lead to the following corollary.

Corollary 4. *It holds that*

$$\mathcal{Y}_{\max} \xrightarrow{\mathcal{D}} \text{Gumbel}\left(\ln\left(\frac{m}{(\ell-1)!}\right),1\right).$$

Proof. Note that for all $n \in \mathbb{N}$, and every $j \in [m]$ we have that $nY_j^{(n,\ell)} + n\ln n + n(\ell-1)\ln\ln n \sim \text{CCP}(n,\ell)$. By Theorem 7 and Definition 6, for any $j \in [m]$, $\mathcal{Y}_j \xrightarrow{\mathcal{D}} \text{Gumbel}(-\ln((\ell-1)!),1)$. Applying Lemma 1 with $\mathcal{Y}_1,\ldots,\mathcal{Y}_m$ proves the claim. □

As in the previous section, in order to prove that $\mathcal{Y}_{\max}$ is uniformly integrable, we will show that $\Pr\left[|Y_{\max}^{(n,\ell)}| > x\right]$ can be exponentially bounded and that the bound does not depend on n.

Claim 3. *There exists $C_0 = C_0(\ell,m)$, such that for any $x \geq 0$,*

$$\Pr\left[Y_{\max}^{(n,\ell)} > x\right] \leq C_0 e^{-\frac{x}{2}}.$$

Proof. Denote $t \triangleq n(\ln n + (\ell-1)\ln\ln n + x)$. We have that

$$\begin{aligned}
\Pr\left[Y_{\max}^{(n,\ell)} > x\right] &= \Pr\left[X_{\max}^{(n,\ell)} > t\right] \\
&= \Pr\left[\text{some coupon was collected } \textbf{less} \text{ than } \ell \text{ times at time } t\right] \\
&\leq nm \cdot \Pr\left[\text{a specific coupon was collected } \textbf{less} \text{ than } \ell \text{ times at time } t\right] \\
&= nm \cdot \Pr\left[\text{Bin}\left(t,\frac{1}{n}\right) \leq \ell-1\right] \\
&= nm \cdot \sum_{r=0}^{\ell-1} \binom{t}{r}\left(\frac{1}{n}\right)^r\left(1-\frac{1}{n}\right)^{t-r} \\
&= nm\left(1-\frac{1}{n}\right)^t \cdot \sum_{r=0}^{\ell-1} \binom{t}{r}\left(\frac{1}{n}\right)^r\left(1-\frac{1}{n}\right)^{-r} \\
&= nm\left(1-\frac{1}{n}\right)^t \cdot \sum_{r=0}^{\ell-1} \binom{t}{r}\left(\frac{1}{n-1}\right)^r \\
&\leq nm\left(1-\frac{1}{n}\right)^t \cdot \sum_{r=0}^{\ell-1} \left(\frac{t}{n-1}\right)^r \\
&\leq 2^\ell m\ell \cdot ne^{-\frac{t}{n}}\left(\frac{t}{n}\right)^{\ell-1},
\end{aligned}$$

where the first inequality is obtained via a union bound, the second inequality applies the bound $\binom{a}{b} < a^b$, and the third inequality uses the bound $(1-1/n)^n \leq e^{-1}$ together with the fact that the final term in the sum is the largest one. For any ℓ, there exists $x_0(\ell) \geq \ell$ such that for any $x > x_0(\ell)$, $x^{\ell-1} \leq e^{\frac{x}{2}}$. Hence, for any $x > x_0(\ell)$,

$$ne^{-\frac{t}{n}}\left(\frac{t}{n}\right)^{\ell-1} = e^{-x}\left(\frac{t}{n\ln n}\right)^{\ell-1} \leq e^{-x}(\ell+2x)^{\ell-1} \leq 3^{\ell-1}e^{-\frac{x}{2}},$$

where the first inequality holds since $\frac{t}{n\ln n} = 1 + \frac{(\ell-1)\ln\ln n}{\ln n} + \frac{x}{\ln n} < \ell + 2x$, and the second inequality holds by the requirements on $x_0(\ell)$. Put it all together,

$$\Pr\left[Y_{\max}^{(n,\ell)} > x\right] \le 2^\ell m\ell \cdot ne^{-\frac{t}{n}} \left(\frac{t}{n}\right)^{\ell-1} \le 2^\ell 3^{\ell-1} m\ell e^{-\frac{x}{2}}.$$

In addition, for any $0 \le x \le x_0(\ell)$, $\Pr[Y_{\max}^{(n,\ell)} > x] \le 1 \le e^{x_0(\ell)} e^{-\frac{x}{2}}$. Denote $C_0(\ell, m) \triangleq \max\left\{e^{x_0(\ell)}, 3^{2\ell-1} m\ell\right\}$. Then, $\Pr[Y_{\max}^{(n,\ell)} > x] \le C_0(\ell, m) e^{-\frac{x}{2}}$ for any $x \ge 0$ as required. □

To bound the left tail, namely, to bound $\Pr\left[Y_{\max}^{(n,\ell)} < -x\right]$, we will need the notion of negatively associated random variables.

Definition 8 (Negative association)**.** *A collection $X_1, X_2, \ldots, X_n$ of random variables is called* negatively associated *(NA) if for every two disjoint index sets, $I, J \subseteq [n]$, and every $f : \mathbb{R}^{|I|} \to \mathbb{R}$ and $g : \mathbb{R}^{|J|} \to \mathbb{R}$ that are both non-decreasing or both non-increasing, we have*

$$\mathbb{E}\left[f\left(X_i, i \in I\right) g\left(X_j, j \in J\right)\right] \le \mathbb{E}\left[f\left(X_i, i \in I\right)\right] \mathbb{E}\left[g\left(X_j, j \in J\right)\right] .$$

Equivalently, $\mathrm{Cov}(f(X_i, i \in I), g(X_j, j \in J)) \le 0$.

Theorem 8 (Theorem 13, [DR96])**.** *Suppose we throw m balls into n bins independently at random. For $i \in [n]$, let B_i be the random variable denoting the number of balls in the ith bin. Then, the variables $B_1, \ldots, B_n$ are NA.*

Claim 4. *There exists $C_1 = C_1(\ell)$, such that for any $x \ge 0$,*

$$\Pr\left[Y_{\max}^{(n,\ell)} \le -x\right] \le C_1 e^{-\frac{x}{2}}.$$

Proof. Similarly to the proof of Claim 2,

$$\Pr\left[Y_{\max}^{(n,\ell)} \le -x\right] = \Pr\left[Y_1^{(n,\ell)} \le -x\right]^m \le \Pr\left[Y_1^{(n,\ell)} \le -x\right] = \Pr\left[X_1^{(n,\ell)} \le n\left(\ln n + (\ell-1)\ln\ln n - x\right)\right].$$

Fix x, and define $t \triangleq n(\ln n + (\ell-1)\ln\ln n - x)$. If $x > \ln n + (\ell-1)\ln\ln n - \ell$, then $t < n\ell$. Hence $\Pr[X_1^{(n,\ell)} < t] = 0$, and in particular $\Pr[Y_{\max}^{(n,\ell)} < -x] = 0 \le Ce^{-cx}$ for any $C, c > 0$. Therefore, for the rest of the proof, we assume that $0 \le x \le \ln n + (\ell-1)\ln\ln n - \ell$.

Let $B_j^{(t)}$ be the number of copies of coupon j after the first t draws of one collector, and define $I_j^{(t)} \triangleq \mathbb{1}_{B_j^{(t)} < \ell}$. Then $I_j^{(t)}$ is the indicator that coupon j has less than ℓ copies after the first t draws, and $W^{(t)} \triangleq \sum_{j=1}^n I_j^{(t)}$ is the number of such coupons. Hence,

$$\Pr\left[X_1^{(n,\ell)} \le n\left(\ln n + (\ell-1)\ln\ln n - x\right)\right] = \Pr\left[X_1^{(n,\ell)} \le t\right] = \Pr\left[W^{(t)} = 0\right].$$

Analogously to the proof of Claim 2, we now derive bounds on the expectation and variance of $W^{(t)}$ in order to obtain an upper bound on $\Pr[W^{(t)} = 0]$. Note that the variables $B_1^{(t)}, \ldots, B_n^{(t)}$ correspond exactly to the balls and bins paradigm (with t balls and n bins), and thus, they are NA according to Theorem 8. Moreover, observe that the indicator function $f(B) = \mathbb{1}_{B<\ell}$ is non-increasing, thus for any $i \ne j$, $\mathrm{Cov}(I_i^{(t)}, I_j^{(t)}) \le 0$. Therefore,

$$\begin{aligned}\mathrm{Var}\left(W^{(t)}\right) &= \sum_{j\in[n]} \mathrm{Var}\left(I_j^{(t)}\right) + 2\sum_{1\le j<i\le n} \mathrm{Cov}\left(I_j^{(t)}, I_i^{(t)}\right) \\ &\le \sum_{j\in[n]} \mathrm{Var}\left(I_j^{(t)}\right) \le \sum_{j\in[n]} \mathbb{E}\left[I_j^{(t)}\right] = \mathbb{E}\left[W^{(t)}\right],\end{aligned}$$

where the last inequality holds again since for any indicator I, $\mathrm{Var}(I) \leq \mathbb{E}[I]$. Moreover,

$$\mathbb{E}\left[W^{(t)}\right] = \sum_{j\in[n]} \mathbb{E}\left[I_j^{(t)}\right] = n\sum_{r=0}^{\ell-1}\binom{t}{r}\left(\frac{1}{n}\right)^r\left(1-\frac{1}{n}\right)^{t-r} > 0.$$

Hence, by the Chebyshev inequality,

$$\Pr\left[W^{(t)}=0\right] \leq \Pr\left[\left|W^{(t)} - \mathbb{E}\left[W^{(t)}\right]\right| \geq \mathbb{E}\left[W^{(t)}\right]\right] \leq \frac{\mathrm{Var}\left(W^{(t)}\right)}{\mathbb{E}\left[W^{(t)}\right]^2} \leq \frac{1}{\mathbb{E}\left[W^{(t)}\right]}.$$

In addition,

$$\begin{aligned}
\mathbb{E}\left[W^{(t)}\right] &= n\sum_{r=0}^{\ell-1}\binom{t}{r}\left(\frac{1}{n}\right)^r\left(1-\frac{1}{n}\right)^{t-r} \\
&= n\left(1-\frac{1}{n}\right)^t\sum_{r=0}^{\ell-1}\binom{t}{r}\left(\frac{1}{n-1}\right)^r \\
&\geq ne^{-\frac{t}{n-1}}\sum_{r=0}^{\ell-1}\binom{t}{r}\left(\frac{1}{n-1}\right)^r.
\end{aligned}$$

Divide into cases:

1. $2(\ell-1)\ln\ln n \leq x$:

$$ne^{-\frac{t}{n-1}}\sum_{r=0}^{\ell-1}\binom{t}{r}\left(\frac{1}{n-1}\right)^r \geq ne^{-\frac{t}{n-1}} = n^{-\frac{1}{n-1}}e^{\frac{n}{n-1}(x-(\ell-1)\ln\ln n)} \geq \frac{1}{2}e^{\frac{n}{n-1}\frac{x}{2}} \geq \frac{1}{2}e^{\frac{x}{2}},$$

   where the second inequality holds by the requirements on x and the monotonicity of $n^{-1/(n-1)}$.

2. $0 \leq x \leq 2(\ell-1)\ln\ln n$: Let $n_1 = n_1(\ell)$ be such that for any $n > n_1$:

   (a) $\ln n \geq 2(\ell-1)\ln\ln n$

   (b) $(\ln n)^{-\frac{\ell-1}{n-1}} > \frac{1}{2}$

   In this case, we have

$$\begin{aligned}
\sum_{r=0}^{\ell-1}\binom{t}{r}\left(\frac{1}{n-1}\right)^r &\geq \binom{t}{\ell-1}\left(\frac{1}{n-1}\right)^{\ell-1} \\
&\geq \frac{1}{(\ell-1)^{\ell-1}}\left(\frac{t}{n-1}\right)^{\ell-1} \\
&= \frac{1}{(\ell-1)^{\ell-1}}\left(\frac{n(\ln n+(\ell-1)\ln\ln n-x)}{n-1}\right)^{\ell-1} \\
&\geq \frac{1}{(\ell-1)^{\ell-1}}\left(\ln n-(\ell-1)\ln\ln n\right)^{\ell-1} \\
&\geq \frac{1}{(\ell-1)^{\ell-1}2^{\ell-1}}\left(\ln n\right)^{\ell-1},
\end{aligned}$$

where the first inequality just considers the last summand, the second inequality applies the bound $\binom{a}{b} \geq a^a/b^b$, and the last inequality is due to our assumption 2a on n. Thus,

$$\begin{aligned}
ne^{-\frac{t}{n-1}} \sum_{r=0}^{\ell-1} \binom{t}{r} \left(\frac{1}{n-1}\right)^r &\geq ne^{-\frac{t}{n-1}} \frac{1}{(\ell-1)^{\ell-1}2^{\ell-1}} (\ln n)^{\ell-1} \\
&= n^{-\frac{1}{n-1}} e^{-\frac{n}{n-1}((\ell-1)\ln\ln n - x)} \frac{1}{(\ell-1)^{\ell-1}2^{\ell-1}} (\ln n)^{\ell-1} \\
&\geq \frac{1}{(\ell-1)^{\ell-1}2^{\ell}} e^x (\ln n)^{-\frac{\ell-1}{n-1}} \\
&\geq \frac{1}{(\ell-1)^{\ell-1}2^{\ell+1}} e^x,
\end{aligned}$$

where the last inequality holds due to assumption 2b on n.

Therefore, for the cases where either $n > n_1$ or $n \leq n_1$ and $2(\ell-1)\ln\ln n \leq x$,

$$\begin{aligned}
\mathbb{E}\left[W^{(t)}\right] &\geq ne^{-\frac{t}{n-1}} \sum_{r=0}^{\ell-1} \binom{t}{r} \left(\frac{1}{n-1}\right)^r \\
&\geq \min\left\{\frac{1}{2}e^{\frac{x}{2}}, \frac{1}{(\ell-1)^{\ell-1}2^{\ell+1}} e^x\right\} \\
&\geq \frac{1}{(\ell-1)^{\ell-1}2^{\ell+1}} e^{\frac{x}{2}} ,
\end{aligned}$$

and therefore,

$$\begin{aligned}
\Pr\left[Y_{\max}^{(n,\ell)} \leq -x\right] &\leq \Pr\left[W^{(t)} = 0\right] \\
&\leq \frac{1}{\mathbb{E}\left[W^{(t)}\right]} \\
&\leq (\ell-1)^{\ell-1}2^{\ell+1}e^{-\frac{x}{2}}.
\end{aligned}$$

Lastly, for the case where $x \leq 2(\ell-1)\ln\ln n$ and $2 \leq n \leq n_1$, it holds that $x \leq 2(\ell-1)\ln\ln n \leq 2(\ell-1)\ln\ln n_1$, and

$$\Pr\left[Y_{\max}^{(n,\ell)} \leq -x\right] \leq 1 = (\ln n_1)^{(\ell-1)} e^{-(\ell-1)\ln\ln n_1} \leq (\ln n_1)^{(\ell-1)} e^{-\frac{x}{2}}.$$

We conclude the proof of the lemma by noting that for $C_1(\ell) \triangleq \max\{(\ell-1)^{\ell-1}2^{\ell+1}, (\ln n_1)^{(\ell-1)}\}$ and any $x \geq 0, n \geq 2$, it holds that $\Pr\left[Y_{\max}^{(n,\ell)} < -x\right] \leq C_1(\ell)e^{-\frac{x}{2}}$ as required. □

Corollary 5. *The sequence* $\left\{Y_{\max}^{(n,\ell)}\right\}_{n\geq 2}$ *is uniformly integrable.*

Proof. By Claim 3 and Claim 4, we have that,

$$\Pr\left[\left|Y_{\max}^{(n,\ell)}\right| > x\right] \leq \Pr\left[Y_{\max}^{(n,\ell)} > x\right] + \Pr\left[Y_{\max}^{(n,\ell)} \leq -x\right] \leq (C_0 + C_1)\, e^{-\frac{x}{2}} .$$

The claim follows according to Lemma 2. □

Now we will prove Theorem 3 which is restated for convenience.

Theorem 3. *Let m and ℓ be constant positive integers. Let $X_1^{(n,\ell)}, \ldots, X_m^{(n,\ell)}$ be i.i.d. random variables with $X_j^{(n,\ell)} \sim CCP(n,\ell)$. Then,*

$$\mathbb{E}\left[\max_{j\in[m]} X_j^{(n,\ell)}\right] = n\left(\ln(nm) + (\ell-1)\ln\ln n + C_\ell \pm o(1)\right)$$

where $C_\ell = \gamma - \ln((\ell-1)!)$.

Proof. By Corollary 4 and Corollary 5, $\mathcal{Y}_{\max} = \left\{Y_{\max}^{(n,\ell)}\right\}_{n\geq 2}$ is uniformly integrable and converges in distribution to Gumbel $\left(\ln\left(\frac{m}{(\ell-1)!}\right), 1\right)$. Therefore, by Theorem 6,

$$\lim_{n\to\infty} \mathbb{E}\left[Y_{\max}^{(n,\ell)}\right] = \mathbb{E}\left[\text{Gumbel}\left(\ln\left(\frac{m}{(\ell-1)!}\right), 1\right)\right] = \ln\left(\frac{m}{(\ell-1)!}\right) + \gamma.$$

We note that $X_{\max}^{(n,\ell)} = n\ln n + n(\ell-1)\ln\ln n + nY_{\max}^{(n,\ell)}$. Hence,

$$\begin{aligned}\mathbb{E}[X_{\max}^{(n,\ell)}] &= n\ln n + n(\ell-1)\ln\ln n + n\mathbb{E}\left[Y_{\max}^{(n,\ell)}\right] \\ &= n\left(\ln(nm) + (\ell-1)\ln\ln n + C_\ell \pm o(1)\right)\end{aligned}$$

where $C_\ell = \gamma - \ln((\ell-1)!)$. □

6 Proof of Theorem 1

The proof of Theorem 1 is based on the same general framework as the proofs of Theorem 2 and Theorem 3, namely, analyzing normalized versions of the relevant random variables and establishing their convergence in distribution and uniform integrability. Having said that, unlike the previous theorems, there is no well known normalization related to the problem. Consequently we had to find it and prove its limiting distribution.

Throughout this section, we adopt the following notation. Let $A^{(0)} = \mathbf{0}^{n\times(m+\rho)}, A^{(1)}, A^{(2)}, \ldots$ be a sequence of matrices constructed as follows: for each $t \in \mathbb{N}$, we draw $(v_1, \ldots, v_{m+\rho}) \sim \text{Unif}([n]^{m+\rho})$ and set

$$A^{(t)} = A^{(t-1)} + \left[\boldsymbol{e}_{v_1}, \ldots, \boldsymbol{e}_{v_{m+\rho}}\right] ,$$

where $\boldsymbol{e}_a$ denotes the column vector with a 1 in the a-th position and 0 in all other entries. Define $T_{n,m,\rho}$ as the random variable that corresponds to the smallest t such that every row of $A^{(t)}$ contains at least m nonzero entries. For each row $i \in [n]$ and time $t \geq 0$, let $X_i(t)$ denote the *number of nonzero entries in row i of $A^{(t)}$*, and by $E_i(t)$ the event that row i has less than m nonzero entries at time t, i.e., $E_i(t) = \{(X_1(t), \ldots, X_n(t)) \in ([m+\rho] \cup \{0\})^n \mid X_i(t) < m\}$. In addition, denote $Z_{n,m,\rho} \triangleq \frac{\rho+1}{n} T_{n,m,\rho} - \ln n$.

The proof of Theorem 1 follows a three-step argument. We begin by showing in Theorem 9 that $\{Z_{n,m,\rho}\}_{n\geq 2}$ converges in distribution to a Gumbel law with parameters $\mu = \ln\binom{m+\rho}{m-1}$ and $\beta = 1$. We then establish in Claim 5 and in Claim 6 suitable tail bounds of $\{Z_{n,m,\rho}\}_{n\geq 2}$, which yield uniform integrability in Corollary 6. Combining these ingredients, we conclude the proof of Theorem 1 via Theorem 6.

We begin with some estimations which will be used in the proof of Theorem 9.

Lemma 3. *Let $m, \rho, k \in \mathbb{N}, x \in \mathbb{R}$ be fixed constants with respect to $n \in \mathbb{N}$, and set*

$$\psi(x) = \binom{m+\rho}{m-1} e^{-x}.$$

If $t = \frac{n}{\rho+1}(\ln n + x)$, then[1]

$$\Pr\left[E_1(t) \cap \cdots \cap E_k(t)\right] = \frac{\psi(x)^k}{n^k}(1 \pm o(1)). \tag{1}$$

Consequently, if

$$U_k(n) \triangleq \sum_{1\leq i_1<\cdots<i_k\leq n} \Pr\Big(\bigcap_{j=1}^{k} E_{i_j}(t)\Big),$$

then

$$U_k(n) = \frac{\psi(x)^k}{k!}(1 \pm o(1)) . \tag{2}$$

[1]Since $E_i(t) = \{(X_1(t), \ldots, X_n(t)) \in ([m+\rho] \cup \{0\})^n \mid X_i(t) < m\}$, the event $E_1(t) \cap \cdots \cap E_k(t)$ is the set $\{(X_1(t), \ldots, X_n(t)) \in ([m+\rho] \cup \{0\})^n \mid \forall i \in [k] : X_i(t) < m\}$.

Proof. Fix $k \geq 1$ and consider the event $E_1(t) \cap \cdots \cap E_k(t)$ which is the event where at time t, every row $i \in [k]$, contains less than m nonzero entries. In order to calculate $\Pr\Big(E_1(t) \cap \cdots \cap E_k(t)\Big)$, we will estimate the probabilities of different possible patterns of $A^{(t)}$. For each $j \in [m+\rho]$, denote by $\xi_j \subseteq [k]$ the set of **rows** with zero entries at column j, and call $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{m+\rho})$ the configuration at time t. A configuration $\boldsymbol{\xi}$ for which the event $E_1(t) \cap \cdots \cap E_k(t)$ occurs, it called a *valid configuration*.

For $t = \frac{n}{\rho+1}(\ln n + x)$, and each fixed $\xi \subseteq [k], j \in [m+\rho]$,

$$\begin{aligned}
\Pr\left[\text{At time } t,\, \xi_j = \xi\right] =& \Pr\left[\forall i \in \xi : A_{i,j}^{(t)} = 0\right] \Pr\left[\forall i \in [k] \setminus \xi : A_{i,j}^{(t)} \geq 1 \mid \forall i \in \xi : A_{i,j}^{(t)} = 0\right] \\
=& \Big(1 - \frac{|\xi|}{n}\Big)^t \sum_{a=0}^{k-|\xi|} (-1)^a \binom{k-|\xi|}{a} \left(1 - \frac{a}{n-|\xi|}\right)^t \\
=& \sum_{a=0}^{k-|\xi|} (-1)^a \binom{k-|\xi|}{a} \left(1 - \frac{|\xi|+a}{n}\right)^t \\
=& \sum_{a=0}^{k-|\xi|} (-1)^a \binom{k-|\xi|}{a} \left(1 - \frac{|\xi|+a}{n}\right)^{\frac{n}{\rho+1}(\ln n + x)} \\
=& (n^{-1}e^{-x})^{|\xi|/(\rho+1)}(1 \pm o(1)) \,.
\end{aligned}$$

Here, the last equality holds since $\left(1 - \frac{|\xi|+a}{n}\right)^{\frac{n}{\rho+1}(\ln n + x)} = (n^{-1}e^{-x})^{(|\xi|+a)/(\rho+1)}(1 \pm o(1))$, the dominant term in the sum is obtained when $a = 0$, namely, $(n^{-1}e^{-x})^{|\xi|/(\rho+1)}(1 \pm o(1))$, while all other terms (in absolute value) are $o((n^{-1}e^{-x})^{|\xi|/(\rho+1)})$.

As the columns are independent, the probability that at time t, the obtained configuration is $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{m+\rho})$ is

$$\prod_{j=1}^{m+\rho} \left(n^{-1}e^{-x}\right)^{|\xi_j|/(\rho+1)}(1 \pm o(1)) \;=\; \left(n^{-1}e^{-x}\right)^{(\sum_{j=1}^{m+\rho}|\xi_j|)/(\rho+1)}(1 \pm o(1)) \,. \tag{3}$$

Now, note that for the event $E_1(t) \cap \cdots \cap E_k(t)$ to occur, for any $i \in [k]$, we must have $\sum_{j=1}^{m+\rho} \mathbb{1}_{i \in \xi_j} \geq \rho+1$. Therefore, it holds that $\sum_{i=1}^{k}\sum_{j=1}^{m+\rho} \mathbb{1}_{i \in \xi_j} \geq \sum_{i=1}^{k}(\rho+1) = k(\rho+1)$. However, by changing the order of summation, we get

$$k(\rho+1) \leq \sum_{i=1}^{k}\sum_{j=1}^{m+\rho} \mathbb{1}_{i \in \xi_j} = \sum_{j=1}^{m+\rho}\sum_{i=1}^{k} \mathbb{1}_{i \in \xi_j} = \sum_{j=1}^{m+\rho} |\xi_j|.$$

If $\sum_{j=1}^{m+\rho}|\xi_j| = k(\rho+1)$, then (3) becomes $n^{-k}e^{-kx}(1 \pm o(1))$. Also observe that $\boldsymbol{\xi}$ is valid and $\sum_{j=1}^{m+\rho}|\xi_j| = k(\rho+1)$ if and only if for all $i \in [k]$, it holds that $\sum_{j=1}^{m+\rho} \mathbb{1}_{i \in \xi_j} = \rho+1$. Thus, there are exacly $\binom{m+\rho}{\rho+1}^k = \binom{m+\rho}{m-1}^k$ valid configurations $\boldsymbol{\xi}$ for which $\sum_{j=1}^{m+\rho}|\xi_j| = k(\rho+1)$, and we have get that at time t,

$$\Pr\left[\boldsymbol{\xi} \text{ is valid and } \sum_{j=1}^{m+\rho}|\xi_j| = k(\rho+1)\right] = \binom{m+\rho}{m-1}^k n^{-k}e^{-kx}(1 \pm o(1)) = \frac{\psi(x)^k}{n^k}(1 \pm o(1)) \,.$$

where $\psi(x) = \binom{m+\rho}{m-1}e^{-x}$.

If $\sum_{j=1}^{m+\rho}|\xi_j| > k(\rho+1)$, then (3) is $o(n^{-k})$. As k, m, and ρ are constants and the number of total configurations is $2^{k(m+\rho)}$, we get that

$$\Pr\left[\boldsymbol{\xi} \text{ is valid and } \sum_{j=1}^{m+\rho}|\xi_j| > k(\rho+1)\right] \leq 2^{k(m+\rho)} \cdot o(n^{-k}) = o(n^{-k}) \,.$$

Thus, since we cannot have a valid configuration $\boldsymbol{\xi}$ with $\sum_{j=1}^{m+\rho} |\xi_j| < k(\rho+1)$, we conclude that

$$\Pr\left[E_1(t) \cap \cdots \cap E_k(t)\right] = \frac{\psi(x)^k}{n^k}(1 \pm o(1)),$$

as claimed in Lemma 3. Lastly, since $\Pr\left[\bigcap_{j=1}^k E_{i_j}(t)\right] = \Pr\left[\bigcap_{j=1}^k E_j(t)\right]$ is the same for any set $\{i_1, \ldots, i_k\} \subset [n]$, if $U_k(n) \triangleq \sum_{1 \le i_1 < \cdots < i_k \le n} \Pr\Big(\bigcap_{j=1}^k E_{i_j}(t)\Big)$, then $U_k(n) = \binom{n}{k}\frac{\psi(x)^k}{n^k}(1 \pm o(1)) = \frac{\psi(x)^k}{k!}(1 \pm o(1))$ as stated. □

The proof of Theorem 9 uses the asymptotic evaluation we analyzed in Lemma 3 and Bonferroni's inequalities to prove the convergence in distribution of $\{Z_{n,m,\rho}\}_{n \ge 2}$.

Theorem 9. *For each fixed $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} \Pr\left[T_{n,m,\rho} \le \frac{n}{\rho+1}(\ln n + x)\right] = \exp\left(-\binom{m+\rho}{m-1}e^{-x}\right).$$

Equivalently, for $Z_{n,m,\rho} \triangleq \frac{\rho+1}{n} T_{n,m,\rho} - \ln n$, we have

$$Z_{n,m,\rho} \xrightarrow{\mathcal{D}} \text{Gumbel}\left(\ln\binom{m+\rho}{m-1}, 1\right).$$

Proof. Recall that

$$\{T_{n,m,\rho} \le t\} = \Big(\bigcup_{i=1}^n E_i(t)\Big)^c,$$

since $T_{n,m,\rho} \le t$ means that all rows of $A^{(t)}$ have at least m nonzero entries. Therefore,

$$\Pr\left[T_{n,m,\rho} \le t\right] = 1 - \Pr\left[\bigcup_{i=1}^n E_i(t)\right].$$

By Bonferroni's inequalities, for every $\frac{n}{2} \ge L \ge 1$,

$$\sum_{k=1}^{2L}(-1)^{k+1}U_k(n) \le \Pr\left[\bigcup_{i=1}^n E_i(t)\right] \le \sum_{k=1}^{2L-1}(-1)^{k+1}U_k(n),$$

where recall that $U_k(n)$ is $\sum_{1 \le i_1 < \cdots < i_k \le n} \Pr\Big[\bigcap_{j=1}^k E_{i_j}(t)\Big]$. For any fixed L, there exists $n_0(L)$ such that for all $n \ge n_0(L)$, we have

$$\left|U_k(n) - \frac{\psi(x)^k}{k!}\right| \le \frac{1}{2L^2} \qquad \text{for all } 1 \le k \le 2L.$$

Hence,

$$\sum_{k=1}^{2L}(-1)^{k+1}\frac{\psi(x)^k}{k!} - \frac{1}{L} \le \Pr\left[\bigcup_{i=1}^n E_i(t)\right] \le \sum_{k=1}^{2L-1}(-1)^{k+1}\frac{\psi(x)^k}{k!} + \frac{1}{L}.$$

As $L \to \infty$, both bounds converge to the alternating series

$$\sum_{k=1}^{\infty}(-1)^{k+1}\frac{\psi(x)^k}{k!} = 1 - e^{-\psi(x)}.$$

Therefore,

$$\lim_{n\to\infty} \Pr[T_{n,m,\rho} \le t] = 1 - \lim_{n\to\infty} \Pr\left[\bigcup_{i=1}^n E_i(t)\right] = e^{-\psi(x)},$$

which completes the proof. □

We now head to the proof that the set $\{Z_{n,m,\rho}\}_{n\geq 2}$ is uniformly integrable.

Claim 5. *For any $x \geq 0$ and any $n \in \mathbb{N}$:*

$$\Pr[Z_{n,m,\rho} > x] \leq 2^{m+\rho} e^{-x}$$

Proof. Fix $x \in \mathbb{R}$ and set $u = \ln n + x$, so that $\Pr[Z_{n,m,\rho} > x] = \Pr[T_{n,m,\rho} > \frac{n}{\rho+1}u]$. Recall that we define the random variable $X_i(t)$ to represent the number of columns that were collected in the i-th row by time t. We have that $X_i(t) \sim \mathrm{Bin}(m+\rho, p_t)$ where $p_t = 1 - \left(1 - \frac{1}{n}\right)^t$ is the probability that a specific column draws a specific row by time t.

$$\Pr[X_i(t) < m] = \sum_{k=0}^{m-1} \binom{m+\rho}{k} p_t^k (1-p_t)^{m+\rho-k} \leq (1-p_t)^{\rho+1} \sum_{k=0}^{m-1} \binom{m+\rho}{k} \leq 2^{m+\rho}(1-p_t)^{\rho+1}.$$

With $t = \frac{n}{\rho+1}\, u$ this gives

$$1 - p_t = \left(1 - \frac{1}{n}\right)^t \;\leq\; e^{-t/n} = e^{-u/(\rho+1)}.$$

Therefore,

$$\Pr\left[E_i(t)\right] = \Pr\left[X_i(t) < m\right] \leq 2^{m+\rho}(1-p_t)^{\rho+1} \leq 2^{m+\rho} e^{-u} = 2^{m+\rho} n^{-1} e^{-x}.$$

Applying the union bound over all n rows yields

$$\Pr[Z_{n,m,\rho} > x] = \Pr\left[T_{n,m,\rho} > \frac{n}{\rho+1}u\right] \leq 2^{m+\rho} e^{-x}\ .$$

□

For the proof of the left tail, we recall some properties of NA random variables,

Proposition 1 (Property P_6, [JDP83])**.** *Non-decreasing functions defined on disjoint subsets of a set of NA random variables are NA.*

Proposition 2 (Property P_7, [JDP83])**.** *The union of independent sets of NA random variables is NA.*

Claim 6. *For any $x \geq 0$ and $2 \leq n \in \mathbb{N}$:*

$$\Pr[Z_{n,m,\rho} \leq -x] \leq \frac{1}{C_2(m,\rho)} e^{-x}$$

where $C_2(m,\rho) = \frac{1}{2}\binom{m+\rho}{m-1}(1 - e^{-\frac{m}{m+\rho}})^{m-1}$.

Proof. Fix x, n, and set $t = \frac{n}{\rho+1}(\ln n - x)$. If $x > \ln n - \frac{(\rho+1)m}{m+\rho}$ then $t = \frac{n}{\rho+1}(\ln n - x) < \frac{nm}{m+\rho}$, which is the minimal number of rounds in order to have at least m nonzero entries in each row, and in particular $\Pr[Z_{n,m,\rho} \leq -x] = \Pr[T_{n,m,\rho} \leq t] = 0 \leq Ce^{-x}$ for any $C > 0$. Assume that $0 \leq x \leq \ln n - \frac{(\rho+1)m}{m+\rho}$ and let $W^{(t)}$ be the number of rows with less than m nonzero entries in $A^{(t)}$: $W^{(t)} = \sum_{i=1}^n I_i^{(t)}$ where $I_i^{(t)}$ is the indicator that row i of $A^{(t)}$ has less than m nonzero entries. We have that

$$\Pr[Z_{n,m,\rho} \leq -x] = \Pr[T_{n,m,\rho} \leq t] = \Pr[W^{(t)} = 0].$$

For every $i_1 \neq i_2$, $\mathrm{Cov}(I_{i_1}^{(t)}, I_{i_2}^{(t)}) \leq 0$: Similarly to the proof of Claim 4, for any fixed j, by Theorem 8 the set of entries $\{A_{i,j}^{(t)}\}_{i\in[n]}$ is NA. For any $j_1 \neq j_2$ the sets are independent, hence, by Proposition 2 the entire set $\{A_{i,j}^{(t)}\}_{i\in[n],j\in[m+\rho]}$ is NA. For any $i \in [n], j \in [m+\rho]$ let $I_{i,j}^{(t)}$ be the indicator that $A_{i,j}^{(t)}$ is nonzero. Since these are non-decreasing functions defined on disjoint subsets of a set of NA random variables, the set $\{I_{i,j}\}_{i\in[n],j\in[m+\rho]}$ is NA. Note that $X_i(t) = \sum_{j=1}^{m+\rho} I_{i,j}^{(t)}$, therefore by Proposition 1 $\{X_i(t)\}_{i\in[n]}$ are NA.

Finally, since the indicator function $f(X) = \mathbb{1}_{X<m}$ is non-increasing, and $I_i^{(t)} = \mathbb{1}_{X_i(t)<m}$, by the definition of NA (Definition 8), we have $\mathrm{Cov}(I_{i_1}^{(t)}, I_{i_2}^{(t)}) \leq 0$ and thus $0 < \mathrm{Var}(W^{(t)}) \leq \mathbb{E}[W^{(t)}]$. Moreover,

$$\begin{aligned}
\mathbb{E}[W^{(t)}] = \sum_{j=1}^{n} \mathbb{E}\left[I_j^{(t)}\right] &= n \cdot \Pr\left[X_i(t) < m\right] \\
&= n \cdot \sum_{k=0}^{m-1} \binom{m+\rho}{k} p_t^k \left(1-p_t\right)^{m+\rho-k} \\
&\geq n \binom{m+\rho}{m-1} p_t^{m-1} \left(1-p_t\right)^{m+\rho-(m-1)} \\
&= n \binom{m+\rho}{m-1} p_t^{m-1} \left(1-p_t\right)^{\rho+1} \quad .
\end{aligned}$$

In addition,

$$(1-p_t)^{\rho+1} = \left(1-\frac{1}{n}\right)^{(\rho+1)t} \geq e^{-\frac{(\rho+1)t}{n-1}} = e^{-\frac{n}{n-1}(\ln n - x)} \geq \frac{1}{2n} e^x,$$

and since we assume that $0 \leq x \leq \ln n - \frac{(\rho+1)m}{m+\rho}$,

$$p_t = 1 - \left(1-\frac{1}{n}\right)^t \geq 1 - e^{-\frac{t}{n}} = 1 - e^{-\frac{1}{\rho+1}(\ln n - x)} \geq 1 - e^{-\frac{m}{m+\rho}} > 0.$$

Thus,

$$\mathbb{E}[W^{(t)}] \geq n \binom{m+\rho}{m-1} p_t^{m-1} (1-p_t)^{\rho+1} \geq \frac{1}{2} \binom{m+\rho}{m-1} (1 - e^{-\frac{m}{m+\rho}})^{m-1} e^x.$$

Therefore, for $C_2(m,\rho) \triangleq \frac{1}{2}\binom{m+\rho}{m-1}(1-e^{-\frac{m}{m+\rho}})^{m-1}$ we obtain that

$$\begin{aligned}
\Pr[Z_{n,m,\rho} \leq -x] &= \Pr[W^{(t)} = 0] \\
&\leq \Pr\left[|W^{(t)} - \mathbb{E}[W^{(t)}]| \geq \mathbb{E}[W^{(t)}]\right] \\
&\leq \frac{\mathrm{Var}(W^{(t)})}{\mathbb{E}[W^{(t)}]^2} \\
&\leq \frac{1}{\mathbb{E}[W^{(t)}]} \\
&\leq \frac{1}{C_2(m,\rho)} e^{-x}.
\end{aligned}$$

as claimed. □

Corollary 6. *The sequence $\{Z_{n,m,\rho}\}_{n\geq 2}$ is uniformly integrable.*

Proof. By Claim 5 and Claim 6, for every $x \in \mathbb{R}, n \geq 2$ we have that,

$$\Pr\left[|Z_{n,m,\rho}| > x\right] \leq \Pr\left[Z_{n,m,\rho} > x\right] + \Pr\left[Z_{n,m,\rho} \leq -x\right] \leq \left(2^{m+\rho} + \frac{1}{C_2(m,\rho)}\right) e^{-x} \ .$$

The corollary follows according to Lemma 2. □

We can now complete the proof of Theorem 1, restated below for clarity.

Theorem 1. *Let $A^{(0)} = \mathbf{0}^{n\times(m+\rho)}, A^{(1)}, A^{(2)}, \dots$ be a sequence of matrices constructed as follows: for each $t \in \mathbb{N}$, we draw $(v_1, \dots, v_{m+\rho}) \sim \mathrm{Unif}([n]^{m+\rho})$ and set*

$$A^{(t)} = A^{(t-1)} + \left[\boldsymbol{e}_{v_1}, \dots, \boldsymbol{e}_{v_{m+\rho}}\right] ,$$

where $\boldsymbol{e}_a$ denotes the column vector with a 1 in the a-th position and 0 in all other entries. Define $T_{n,m,\rho}$ as the random variable that corresponds to the smallest t such that every row of $A^{(t)}$ contains at least m nonzero entries. Then,

$$\mathbb{E}\left[T_{n,m,\rho}\right] = \frac{n}{\rho+1}\left(\ln n + \ln\binom{m+\rho}{m-1} + \gamma \pm o(1)\right).$$

Proof. According to Corollary 6, and Theorem 9, the sequence $\{Z_{n,m,\rho}\}_{n\geq 2}$ is uniformly integrable and converges in distribution to $\mathrm{Gumbel}(\ln\binom{m+\rho}{m-1}, 1)$. Therefore, by Theorem 6, we have that

$$\lim_{n\to\infty}\mathbb{E}\left[Z_{n,m,\rho}\right] = \ln\binom{m+\rho}{m-1} + \gamma.$$

Recall that $T_{n,m,\rho} = \frac{n}{\rho+1}\left(Z_{n,m,\rho} + \ln n\right)$, hence,

$$\mathbb{E}\left[T_{n,m,\rho}\right] = \frac{n}{\rho+1}\ln n + \frac{n}{\rho+1}\ln\binom{m+\rho}{m-1} + \frac{n}{\rho+1}\gamma \pm o(n).$$

□

7 Proof of Theorem 4

We start by recalling the statement of the theorem.

Theorem 4. *Let $A^{(0)} = \mathbf{0}^{n\times m}, A^{(1)}, A^{(2)}, \dots$ be a sequence of matrices constructed as follows: for each $t \in \mathbb{N}$, we draw $(v_1, \dots, v_m) \sim \mathrm{Unif}([n]^m)$ and set*

$$A^{(t)} = A^{(t-1)} + \left[\boldsymbol{e}_{v_1}, \dots, \boldsymbol{e}_{v_m}\right] .$$

Let $b \in \mathbb{N}$ that divides n and let $\mathcal{B} \subseteq \mathcal{P}([b]\times[m])$. As Definition 4, define

$$\alpha^* \triangleq mb - \max_{B\in\mathcal{B}}|B| \qquad \beta^* \triangleq \left|\{B' \in \mathcal{B} \mid |B'| = \max_{B\in\mathcal{B}}|B|\}\right|$$

Define T as the random variable that corresponds to the minimal t such that for all $a \in [n/b]$, we have that

$$\left\{(i,j) \in [b]\times[m] \mid A^{(t)}_{i+(a-1)b,j} \neq 0\right\} \notin \mathcal{B} .$$

Then,

$$\mathbb{E}[T] \leq \frac{n}{\alpha^*}\ln n + \frac{\beta^*}{b\alpha^*}\cdot n + o(n) ,$$

We will first prove the following upper bound

Claim 7. *For every positive x, we have*

$$\Pr\left[T \geq \frac{n}{\alpha^*}\cdot(\ln n + x)\right] \leq \frac{\beta^*}{b}\cdot e^{-x}\left(1 + (mb)^2 n^{-\frac{1}{\alpha^*}}\right) .$$

Proof. Define $t \triangleq \frac{n}{\alpha^*}(\ln n + x)$. We begin by focusing only on the first block, i.e., the case $a = 1$. Our goal is to upper bound the probability that, at time t, the collection of all nonzero entries is still a bad set, that is, lies in $\mathcal{B}$. This will ensure that the block is not yet in a recovering state. More precisely, we will prove that

$$\Pr\left[\left\{(i,j) \in [b]\times[m] \mid A^{(t)}_{i,j} \neq 0\right\} \in \mathcal{B}\right] \leq \beta^*\cdot n^{-1}e^{-x}\left(1 + (mb)^2 n^{-\frac{1}{\alpha^*}}\right) .$$

Then, the claim will follow from taking a union bound over all the n/b blocks.

It holds that

$$\Pr\left[\left\{(i,j) \in [b] \times [m] \mid A_{i,j}^{(t)} \neq 0\right\} \in \mathcal{B}\right] = \sum_{B \in \mathcal{B}} \Pr\left[\left\{(i,j) \in [b] \times [m] \mid A_{i,j}^{(t)} \neq 0\right\} = B\right] .$$

Fix $B \in \mathcal{B}$ and denote by $\overline{B}$ its complement in the block, i.e., $B \cup \overline{B} = [b] \times [m]$ and $B \cap \overline{B} = \emptyset$. We have

$$\Pr\left[\left\{(i,j) \in [b] \times [m] \mid A_{i,j}^{(t)} \neq 0\right\} = B\right] = \Pr\left[\left\{(i,j) \in [b] \times [m] \mid A_{i,j}^{(t)} = 0\right\} = \overline{B}\right] ,$$

and if we denote $s_j \triangleq |\{i \in [b] \mid (i,j) \in \overline{B}\}|$, then,

$$\Pr\left[\left\{(i,j) \in [b] \times [m] \mid A_{i,j}^{(t)} = 0\right\} = \overline{B}\right] = \prod_{j=1}^{m} \left(1 - \frac{s_j}{n}\right)^t \leq \prod_{j=1}^{m} \left(1 - \frac{1}{n}\right)^{s_j t} = \left(1 - \frac{1}{n}\right)^{(mb - |B|) \cdot t} .$$

Now, observe that this upper bound depends only on the size of B. Thus,

$$\Pr\left[\left\{(i,j) \in [b] \times [m] \mid A_{i,j}^{(t)} \neq 0\right\} \in \mathcal{B}\right] \leq \sum_{B \in \mathcal{B}} \left(1 - \frac{1}{n}\right)^{(mb - |B|) \cdot t} \leq \sum_{B \in \mathcal{B}} e^{-\frac{(mb - |B|)(\ln n + x)}{\alpha^*}} .$$

Note that $mb - |B| \geq \alpha^*$. For all $s \in [0, mb]$, denote by $\beta_s \triangleq |\{B \in \mathcal{B} \mid |B| = s\}|$, i.e., the number of the sets in $\mathcal{B}$ of size exactly s. Furthermore, set $s^* \triangleq mb - \alpha^*$. Then,

$$\begin{aligned}
\sum_{B \in \mathcal{B}} e^{-\frac{(mb - |B|)(\ln n + x)}{\alpha^*}} &= \sum_{s=s^*}^{mb} \beta_s \cdot e^{-\frac{(mb - s)(\ln n + x)}{\alpha^*}} \\
&\leq \beta_{s^*} \cdot n^{-1} e^{-x} + \sum_{s=s^*+1}^{mb} \beta_s \cdot e^{-\frac{\alpha^*+1}{\alpha^*} \cdot (\ln n + x)} \\
&\leq \beta_{s^*} \cdot n^{-1} e^{-x} \left(1 + (mb)^2 n^{-\frac{1}{\alpha^*}}\right) .
\end{aligned}$$

The claim follows by a union bound over all $a \in [n/b]$. □

With this bound, we can prove Theorem 4.

Proof of Theorem 4. Define $t_0 = (n \ln n)/\alpha^*$. It holds that

$$\mathbb{E}[T] = \sum_{t \geq 1} \Pr[T \geq t] = \sum_{t=1}^{t_0 - 1} \Pr[T \geq t] + \sum_{t=t_0}^{\infty} \Pr[T \geq t] .$$

The first sum can be bounded trivially by t_0 (every term in the sum is at most 1). We focus on the second sum.

$$\begin{aligned}
\sum_{t=t_0}^{\infty} \Pr[T \geq t] &= \sum_{k=0}^{\infty} \Pr\left[T \geq \frac{n \ln n}{\alpha^*} + k\right] \\
&\leq \sum_{k=0}^{\infty} \Pr\left[T \geq \frac{n}{\alpha^*}\left(\ln n + k \cdot \frac{\alpha^*}{n}\right)\right] \\
&\leq \frac{\beta_{s^*}}{b} \cdot \left(1 + (mb)^2 n^{-\frac{1}{\alpha^*}}\right) \cdot \sum_{k=0}^{\infty} e^{-k \cdot \frac{\alpha^*}{n}} \\
&\leq \frac{\beta_{s^*}}{b} \cdot \left(1 + (mb)^2 n^{-\frac{1}{\alpha^*}}\right) \cdot \frac{1}{1 - e^{-\frac{\alpha^*}{n}}} .
\end{aligned}$$

Now, by Taylor expansion, $1 - e^{-x} = x - x^2/2 + O(x^3)$, we conclude that

$$\mathbb{E}[T] \leq \frac{n}{\alpha^*} \ln n + \frac{\beta_{s^*}}{b\alpha^*} \cdot n \cdot (1 + m^2 b n^{-\frac{1}{\alpha^*}})(1 + o(1)) ,$$

and the result follows by recalling that m and b are constant with respect to n. □

References

[AD21] Dan Alistarh and Peter Davies. Collecting coupons is faster with friends. In *International Colloquium on Structural Information and Communication Complexity*, pages 3–12. Springer, 2021.

[AGY24] Hadas Abraham, Ryan Gabrys, and Eitan Yaakobi. Covering all bases: The next inning in dna sequencing efficiency. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 464–469. IEEE, 2024.

[AR01] Ilan Adler and Sheldon M Ross. The coupon subset collection problem. *Journal of Applied Probability*, 38(3):737–746, 2001.

[BEG+25] Avital Boruchovsky, Ohad Elishco, Ryan Gabrys, Anina Gruica, Itzhak Tamo, and Eitan Yaakobi. Making it to first: The random access problem in dna storage. *arXiv preprint arXiv:2501.12274*, 2025.

[BH97] Arnon Boneh and Micha Hofri. The coupon-collector problem revisited—a survey of engineering problems and computational methods. *Stochastic Models*, 13(1):39–66, 1997.

[Bil13] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[BLSGY24] Daniella Bar-Lev, Omer Sabary, Ryan Gabrys, and Eitan Yaakobi. Cover your bases: How to minimize the sequencing coverage in dna storage systems. *IEEE Transactions on Information Theory*, 2024.

[CGK12] George M Church, Yuan Gao, and Sriram Kosuri. Next-generation digital information storage in dna. *Science*, 337(6102):1628–1628, 2012.

[DGW+10] Alexandros G Dimakis, P Brighten Godfrey, Yunnan Wu, Martin J Wainwright, and Kannan Ramchandran. Network coding for distributed storage systems. *IEEE transactions on information theory*, 56(9):4539–4551, 2010.

[DR96] Devdatt P Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *BRICS Report Series*, 3(25), 1996.

[ER61] Pál Erdős and Alfréd Rényi. On a classical problem of probability theory. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 6(1-2):215–220, 1961.

[GBC+13] Nick Goldman, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M LeProust, Botond Sipos, and Ewan Birney. Towards practical, high-capacity, low-maintenance information storage in synthesized dna. *nature*, 494(7435):77–80, 2013.

[GBLRY24] Anina Gruica, Daniella Bar-Lev, Alberto Ravagnani, and Eitan Yaakobi. A combinatorial perspective on random access efficiency for dna storage. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 675–680. IEEE, 2024.

[GMZ24] Anina Gruica, Maria Montanucci, and Ferdinando Zullo. The geometry of codes for random access in dna storage. *arXiv preprint arXiv:2411.08924*, 2024.

[Gut12] Allan Gut. *Probability: a graduate course*, volume 200. Springer New York, NY, 2012.

[JDP83] Kumar Joag-Dev and Frank Proschan. Negative association of random variables with applications. *The Annals of Statistics*, pages 286–295, 1983.

[KJV07] John E Kobza, Sheldon H Jacobson, and Diane E Vaughan. A survey of the coupon collector's problem with random sample sizes. *Methodology and Computing in Applied Probability*, 9(4):573–584, 2007.

[MU05] M Mitzenmacher and Eli Upfal. *Probability and Computing*. 01 2005.

[New60] Donald J Newman. The double dixie cup problem. *The American Mathematical Monthly*, 67(1):58–61, 1960.

[OAC+18] Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, et al. Random access in large-scale dna data storage. *Nature biotechnology*, 36(3):242–248, 2018.

[RBS+22] Vinayak Ramkumar, SB Balaji, Birenjith Sasidharan, Myna Vajha, M Nikhil Krishnan, and P Vijay Kumar. Codes for distributed storage. *Foundations and Trends in Communications and Information Theory*, 19(4):547–813, 2022.

[Rot06] Ron Roth. *Introduction to coding theory*. Cambridge University Press, 2006.

[SKSY24] Omer Sabary, Han Mao Kiah, Paul H Siegel, and Eitan Yaakobi. Survey for a decade of coding for dna storage. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, 10(2):253–271, 2024.

[TWB13] Itzhak Tamo, Zhiying Wang, and Jehoshua Bruck. Zigzag codes: MDS array codes with optimal rebuilding. *IEEE Trans. Inf. Theory*, 59(3):1597–1616, 2013.

[YKGR+16] SM Hossein Tabatabaei Yazdi, Han Mao Kiah, Eva Garcia-Ruiz, Jian Ma, Huimin Zhao, and Olgica Milenkovic. Dna-based storage: Trends and methods. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, 1(3):230–248, 2016.