The Coded Coupons Collector's Problem

Daniella Bar-Lev
Universität Zürich
Department of Mathematics
CH-8057 Zürich, Switzerland
email: daniella.bar-lev@math.uzh.ch

***Abstract*—The classical Coupon Collector's Problem asks how many random draws (with replacement) are needed to collect all n distinct items? This fundamental problem in probability theory has applications across computer science, statistics, and network theory. This talk introduces a natural extension where information is encoded before storage and must be decoded from random samples. While the framework is mainly motivated by DNA-based storage, the mathematical problem is of independent interest and applies to any setting where coded data is accessed through probabilistic sampling.**

I. Problem Setup and Motivation

Consider k information symbols encoded into n coded symbols via an (n, k) code. Retrieval proceeds by sampling coded symbols at random with replacement. The goal is to retrieve any desired subset of the k information symbols. This talk focuses on linear codes with a generator matrix $G \in \mathbb{F}_q^{k\times n}$ and covers two extreme scenarios: *full recovery* (decode all k information symbols) and *random access* (decode a single requested information symbol). For full recovery, it was proven that MDS codes minimize the expected retrieval time. For random access, we define $T_{\max}(G)$ as the worst-case expected retrieval time over all k information symbols. Without coding, $T_{\max} = k$. The fundamental question is whether coding can reduce this baseline, and if so, what structural properties of generator matrices enable such improvements. This talk focuses primarily on random access.

While this mathematical framework applies broadly, it is originally motivated by DNA-based storage systems. In DNA-based storage, synthetic strands are created to encode digital information, stored in an unordered manner, and retrieved via sequencing that produces multiple copies of each strand without order. The efficiency bottleneck of DNA sequencers is tied to the coverage depth (ratio of sequenced reads to designed strands). Reducing coverage depth offers opportunities for improvements in latency and cost, making the mathematical problem of minimizing expected retrieval time practically significant.

II. Main Contributions

In [1], we initiated the study of the coded coupon collector problem. We established that for full recovery, the uniform sampling distribution minimizes the expected retrieval time over all possible channel distributions. Comprehensive upper and lower bounds on both the probability distribution and expected value for full recovery were also derived. The analysis proved that MDS codes are optimal for minimizing expected retrieval time when retrieving complete datasets. Surprisingly, for the random access setting, we showed that systematic MDS codes require an expected retrieval time of at least k and provide no improvement over the uncoded baseline. We established lower bounds on the maximum expected retrieval time and presented explicit code constructions that achieve expected retrieval times below k, demonstrating through analytical methods and simulations that carefully designed codes can improve retrieval efficiency.

We further extended the random access study in [2], introducing new combinatorial techniques to capture structural properties of generator matrices that enable improved performance. We derived two formulas for computing $T_{\max}(G)$: one based on column dependencies and the other using recovery set intersections. A central contribution is the notion of *recovery balanced codes*, for which we provided three testable criteria. Applying these criteria, we determined that classical families of codes, such as MDS, simplex, Hamming, and binary Reed-Muller codes, all satisfy $T_{\max}(G) = k$, offering no benefit over uncoded systems. Crucially, we showed that disrupting this balance enables $T_{\max}(G) < k$. We analyzed modified MDS constructions where information symbols are replicated alongside parity symbols, achieving substantial gains in worst-case retrieval time. The work also characterizes operations on codes that maintain or break the balanced property and provides systematic methods for designing matrices with improved random access performance.

The talk will present the mathematical framework and proof techniques, and discuss properties and constructions. Additionally, we will cover the main results by other researchers, and conclude with open problems and future directions.

Acknowledgment

I thank my co-authors Omer Sabary, Anina Gruica, Ryan Gabrys, Alberto Ravagnani, and Eitan Yaakobi for their valuable contributions to the works presented in this talk. I am also grateful to Ryan Gabrys and Ohad Elishco for our ongoing collaboration on extensions of this problem.

References

[1] D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems," in *IEEE Transactions on Information Theory*, vol. 71, no. 1, pp. 192-218, Jan. 2025, doi: 10.1109/TIT.2024.3496587.

[2] A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A Combinatorial Perspective on Random Access Efficiency for DNA Storage," in *IEEE Transactions on Information Theory*, vol. 71, no. 12, pp. 9395-9412, Dec. 2025, doi: 10.1109/TIT.2025.3623202.