

# Bucket Sort When You Know The Distribution.

David Ponnarovsky

January 20, 2023

## Abstract

We propose a new simple construction based on Tanner Codes, which yields a good LDPC code with testability query complexity of  $\Theta(n^{1-\varepsilon})$  for any  $\varepsilon > 0$ .

**The problem.** Let  $f : [0, 1] \rightarrow [0, 1]$  a fixed distribution function. Write an algorithm that sort  $n$  draws  $x_1 \dots x_n$  at linear expectation time.

**Solution.** We will define a partition of the input into a seira of  $n$  buckets  $\mathcal{B} = \{B_k = [t_k, t_{k+1}] : k \in [n]\}$  such that  $\Pr[x \in B_i] = \frac{1}{n}$  for any bucket.

**Claim.** The probability that the size of the  $i$ th bucket exceeds  $t \in \mathbb{N}$  is bounded by:  $\Pr[B_i \geq t] \leq \frac{e}{t^k}$  for every intrger  $k \leq n$ .

**Proof.** Let the  $X_{ij}$  be the indecator of the event that  $x_j$  belongs to  $B_i$ . Then we have:

$$\begin{aligned}
 \mathbf{E}[B_i^k] &= \mathbf{E}\left[\left(\sum_j X_{ij}\right)^k\right] = \mathbf{E}\left[\sum_{J \in [n]^k} \prod_{l \in [k]} X_{iJ_l}\right] \\
 &= \mathbf{E}\left[\sum_{\substack{l \in [k] \\ |J|=l}} \sum_{J \subseteq [n]} \prod_{j \in J} \binom{n}{l} \frac{(n-l)^{k-l}}{n^l}\right] \\
 &= \sum_{j \neq j'} \mathbf{E}[X_{ij}] \mathbf{E}[X_{ij'}] + \sum_j \mathbf{E}[X_{ij}] \\
 &= \sum_{l \in [4]} \frac{1}{n^l} \binom{n}{l} = O(1) \\
 \mathbf{V}[B_i^2] &= \sum_{l \in [4]} \binom{n}{l} \left(\frac{1}{n^l} - \frac{1}{n^4}\right) \leq e \\
 \mathbf{E}[(B_i^2)^k] &\leq \left(1 + \frac{1}{n}\right)^n \leq e \\
 \Pr[B_i \geq t] &\leq \frac{e}{t^k}
 \end{aligned}$$

$$\begin{aligned}
 \frac{1}{n} &= \Pr[x \in B_k] = f(t_{k+1}) - f(t_k) \\
 &\Rightarrow t_{k+1} \leftarrow f^{-1}\left(\frac{1}{n} + f(t_k)\right)
 \end{aligned}$$