Bucket Sort When You Know The Distribution.

David Ponarovsky

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 $\Theta\left(n^{1-\varepsilon}\right)$ for any $\varepsilon > 0$.

The problem. Let $f : [0,1] \to [0,1]$ a fixed distribution function. Write an algorithm that sort n draws $x_1 ... x_n$ 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 $\mathbf{Pr}[x \in B_i] = \frac{1}{n}$ for any bucket.

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

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

$$\begin{aligned}
\mathbf{E}\left[B_i^k\right] &= \mathbf{E}\left[\left(\sum_j X_{ij}\right)^k\right] = \mathbf{E}\left[\sum_{J \in [n]^k} \prod_{l \in [k]} X_{iJ_l}\right] \\
&= \mathbf{E}\left[\sum_{l \in [k]} \sum_{\substack{J \subset [n] \\ |J| = l}} \prod_{j \in J} X_{ij}\right] \\
&= \sum_{l \in [k]} \binom{n}{l} \frac{l!}{n^l}
\end{aligned}$$

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