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 $f : [0,1] \rightarrow [0,1]$ 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 $\mathbf{Pr}\left[x \in B_i\right] = \frac{1}{n}$ for any bucket.

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

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

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