Bucket Sort When You Know The Distribution.

David Ponarovsky

January 21, 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 serie 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. Assume that we seccsused to compute the buckets efficently. Let the X_{ij} be the indecator of the event that x_j fall to B_i. Then we have:

$$\mathbf{Pr}\left[\sum_i |B_i|^2 \geq t\right] = \mathbf{Pr}\left[\sum_i \left(\sum_j X_{ij}\right)^2 \geq t\right]$$
$$= \mathbf{Pr}\left[\sum_{i,j,j'} X_{i,j} X_{i,j'} \geq t\right] = \mathbf{Pr}\left[\sum_{i,j\neq j'} X_{i,j} X_{i,j'} \geq t - n\right]$$

It follows that the probability that all the buckets will have at most 100 items is bounded by $n^2 \left(100\right)^{-n} \rightarrow 0$. Therefore any computaion made over single bucket requires a constant time (w.h.p) and the expection of the total work is linear. It lefts to show that knowing the distribution enables to compute efficently the buckets.

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