

Bucket Sort When You Know The Distribution.

David Ponnarovsky

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 $\Theta(n^{1-\varepsilon})$ 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 serie 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. Assume that we seccused to compute the buckets efficiently. Let the X_{ij} be the indicator of the event that x_j fall to B_i . Then we have:

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

It follows that the probability that all the buckets will have at most 100 items is bounded by $n^2 (100)^{-n} \rightarrow 0$. 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 efficiently the buckets.

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