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 $\mathbf{Pr}[x \in B_i] = \frac{1}{n}$ for any bucket.

Claim. The probability that the size of the ith bucket excceds $t \in \mathbb{N}$ is bounded by: $\mathbf{Pr}[B_i \geq t] \leq kt^{-k}$ 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:

$$\mathbf{Pr}\left[B_i \geq (1+\delta)\mu\right] \leq e^{-2\frac{\delta^2\mu^2}{n}}$$

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

And noitce that quantinue of sequance elements in summation is bounded by:

$$\binom{n}{l+1}\frac{(l+1)!}{n^{l+1}} / \binom{n}{l}\frac{l!}{n^l} = \frac{n-l}{n} = 1 - \frac{l}{n} \leq 1$$

Hence the sum over k elements is lower than k. Unsing the markov inequality we have that:

$$\mathbf{Pr}\left[B_i \geq t\right] \leq \frac{\mathbf{E}\left[B_i^k\right]}{t^k} \leq kt^{-k}$$

$\square$

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 expecion of the total work is linear. It lefts to show that knowing the distribution enables to compute efficently the buckets.

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