Bucket Sort When You Know The Distribution.

David Ponarovsky

January 21, 2023

Abstract

None.

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 $\mathbf{Pr}[x \in B_i] = \frac{1}{n}$ 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:

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

It follows that for any function $t : \mathbb{N} \rightarrow \mathbb{R}$, such that $n = o(t)$, sorting quaderic each bucket at turn would last almost surly less than $t(n)$. It lefts to show that knowing the distribution enables to compute efficently the buckets. Ensuring the uniform partitonized prtoperty leads for the follow requrisive relation:

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

Hence, if f can be computed in sublinear time then we obtained am expected linear time algortihm for sorting $\square$