All Range Medians

S. Muthukrishnan

March 14, 2009

We are given an array $A[1, \ldots, n]$ of numbers. Our problem is to output, for each interval $[l, r]$,

- The *median* of items $A[l], A[l+1], \ldots, A[r]$, denoted $\text{MED}[l, r]$,
- The *left sum*, that is, $\sum_{i \in [l,r] \mid A[i] \leq \text{MED}[l,r]} (\text{MED}[l, r] - A[i])$, and
- The *right sum*, that is, $\sum_{i \in [l,r] \mid A[i] \geq \text{MED}[l,r]} (A[i] - \text{MED}[l, r])$.

The currently best known algorithm [1] takes $O(n \log n + k \log n)$ time for finding the median of each of the k intervals (we assume that this algorithm can be extended to find the left and right sums). Since we have $O(n^2)$ total intervals in all, this algorithm takes $O(n^2 \log n)$ time for our problem. We give an improved algorithm here. We will focus on finding the median and the left sum; finding right sum is analogous.

Step 1. For each interval $[l, r]$ with its endpoints at multiples of integer x, for some x to be determined later, we find sorted set $S[l, r]$ of items of *rank* $[|l - r + 1|/2 - 2x, \ldots, [|l - r + 1|/2 + 2x]$ in $A[l], \ldots, A[r]$; if the interval has fewer than $4x + 1$ items,keep them all. For each such item $i \in S[l, r]$, we keep the sum s_i of smaller items in the interval.
This step can be done using [1] to find for each interval, the item L of rank $|l - r + 1|/2 - 2x$ and U of rank $|l - r + 1|/2 + 2x$. We use a 2d range structure on points $(i, A[i])$ to determine the points in the set $S[l, r]$ (and their total sum) in time $O(\log n + x)$ for each interval with the range query $[l, r] \times [L, U]$. Further, we assume the points in $S[l, r]$ can be sorted in $O(x \log x)$ time. This also suffices to compute s_i for such items $i \in S[l, r]$. The total time for this step is $O((n/x)^2(\log n + x \log x))$.

Step 2. For each interval $[l, r]$ such that $jx \leq l < r \leq (j + 1)x$, we sort all points and keep prefix and suffix sums for all points. There are $O(x^2)$ such intervals for each j, and for each, we take $O(x \log x)$ time. In all, the time taken is $O((n/x)x^3 \log x)$.

Step 3. Now we complete the computation for the remaining intervals $[l, r]$. These intervals have the form $[l, jx - 1] : [jx, j'x] : [j'x + 1, r]$ where the middle interval $[jx, j'x]$ was processed in Step 1, and the other intervals were processed in Step 2. We have,

Lemma 1 $\text{MED}[l, r]$ *is the median of the set of elements in* $[l, jx - 1]$*, those in* $[j'x + 1, r]$*, and those in set* $S[jx, j'x]$.

Proof. Recall the definition of L and U for the interval $[jx, j'x]$. Observe that we have all the elements in $[l, r]$ between L and U in the set under consideration. Further observe that $\text{MED}[l, r]$ has to lie between L and U, because we add no more than $2x$ elements to $[jx, j'x]$. Now the lemma follows. ∎

There are $O(n^2)$ such intervals and each takes $O(x)$ time to process, for total running time $O(n^2 x)$. ∎

The complexity of the entire algorithm is $O((n/x)^2(\log n + x \log x) + (n/x)x^3 \log x + n^2 x)$. We will have $x \log x < \log n$. Hence, the dominating complexity is $O((n/x)^2 \log n + n^2 x)$ which is minimized when $x = \log^{1/3} n$, giving:

Theorem 1 *There is an $O(n^2 \log^{1/3} n)$ time algorithm for our problem.*

An improvement follows from observing that Step 3 involves finding the median of three sorted arrays of total size $O(x)$ which can be done in $O(\log x)$ time. Then the overall complexity is $O((n/x)^2(\log n + x \log x) + (n/x)x^3 \log x + n^2 \log x)$. The dominating complexity is now $O((n/x)^2(\log n + x \log x) + n^2 \log x)$ which is minimized when $x^2 \log x = \log n$, or $x \sim \sqrt{\frac{\log n}{\log \log n}}$. The overall complexity is then $O(n^2 \log \log n)$.

Theorem 2 *There is an $O(n^2 \log \log n)$ time algorithm for our problem.*

To extend this further, notice that the algorithm above in fact can be used to solve Step 1 for a subset of positions. Then, if we iterate, we should minimize complexity when $x^2 \log x = \log \log n$ which gives $O(n^2 \log \log \log n)$ and so on.

References

[1] B. Gfeller and P. Sanders. Towards Optimal Range Medians *Manuscript*, 2009.