Space efficient quantile selection

Input: stream \( a_1, \ldots, a_n \in U \) where \( U \) has order < 

\[
\begin{array}{cccccccc}
2 & 5 & 7 & 1 & 4 & 11 & 13 & \ldots
\end{array}
\]

e.g. numerical data
- names w/ alphabetic order
- grades

allowed multiple passes

Goal: return the median w/ minimum:

(a) \# passes  (b) space

more generally: "quantile queries"

select rank \( k \) element

(kth largest)
1 pass: Input
$O(1)$ space
<table>
<thead>
<tr>
<th>Passes</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>$\alpha(\log n)$</td>
<td>$O(1)$</td>
</tr>
</tbody>
</table>

$P$ \(\tilde{O}(n^{1/p})\) sort and select quickselect (random pivot)

\[2 \sqrt{n}\]

Munro Paterson 1981

quantile summaries
Approximations

given rank $k \in [n]$ and param $\varepsilon > 0$,
return element w/ rank $k = \varepsilon n$

Sampling:

for median:
    sample $l = O\left(\frac{\log \frac{\sqrt{6}}{\varepsilon}}{\varepsilon^2}\right)$ elements
    return median of sample

for rank $k = d n$
    sample $l$
    return rank $d l$ out of the sample

Deterministic?
Quantiles

space-efficient
mergable
answer $\varepsilon$-approximate quantile queries

let elements $q_1 < q_2 < \ldots < q_\ell \in S$

along w/ intervals $I(q_i)$ rank($q_i$) $\in I(q_i)$

need $\geq \frac{1}{\varepsilon}$ points

By tracking min/max specially, assume

rank($q_1$) = 1, rank($q_\ell$) = $n$. 

$0 \leq q_i \leq n$
how to ensure such \( q_i \) exists if \( k < k' \)?

Then return \( q_i \).

If \( I(a_i) \in [k, \text{en}, k + \text{en}] \) for some \( q_i \),

then return \( q_i \).
Lemma

Then every query $k$ contains an interval $I(q_i)$. 

Proof two cases:

$k \in I(q_i)$ for some $q_i$ 

$k \notin I(q_i)$ \forall q_i
Suppose $k \in I(q_i)$ for some $i$.

If $I(q_i) \subseteq [k-\varepsilon n, k+\varepsilon n]$ then done.

Else look at $q_{i-1}$. 

$k-\varepsilon n < k < k+\varepsilon n$.
suppose $k \not\in \mathcal{I}(q_i) \cup i$

the "combined intervals" cover $[n]$
pick 1 covering $k$.

one of the intervals must lie inside
Key invariant: any two consecutive intervals have width $\leq 2\varepsilon n$.

"$\varepsilon$-APX quantile summary"
Merging, given two \( \varepsilon \)-APX quantile summaries over 2 streams, want \( \varepsilon \)-APX summary over combined stream

\[ Q = \]

\[ Q' = \]

\[ Q'' = \]

want to combine \( Q, Q'' \) to get summary of

\[ s_1 \]

\[ + s_2 \]

\"\[ Q' + Q'' = \{ q_1'', ..., q_c'', I''(q_1''), ..., I''(q_c'') \} \]\"
denote: \[ Q = \{ q_1', \ldots, q_k', I(q_1'), \ldots, I(q_k') \} \]

\[ Q'' = \{ q_1'', \ldots, q_m'', I''(q_1''), \ldots, I''(q_m'') \} \]

let \( q_j'' \in Q'' \). \( I''(q_j) \) bounds rank \( q_j'' \) wrt \( S_2 \)

goal: bound rank \( q_j'' \) wrt \( S_1+S_2 \).

\[
\begin{align*}
\text{rank of } q_j'' \text{ in } S_1 \{ & \geq \min I'(q_i') \\
& \leq \max I'(q_{i+1}') \}
\end{align*}
\]

so \( \min I'(q_i') + \min I''(q_j'') \)

\[
\leq \text{rank } (q_j'' \text{ in } S_1+S_2) \\
\leq \max I''(q_{i+1}'') + \max I''(q_j'')
\]

set \( I''(q_j'') = \left[ \min I'(q_i') + \min I''(q_j''), \max I''(q_{i+1}'') + \max I''(q_j'') \right] \)
$Q'' = \{ q_1, \ldots, q_e, q_i, \ldots, q_m \}$, w/ intervals $I''$.

To show $Q''$ is $\varepsilon$-APX, need to show "2\varepsilon n width" property.

Take two consecutive intervals in $Q''$.

\[
\begin{array}{c}
[ q_i ] \\
\end{array}
\quad
\begin{array}{c}
[ q_j ] \\
\end{array}
\]

Two cases:

1. elements from diff sets
2. elements from same sets
\[ \min I(q) + \min I''(q) = \text{min} \quad \text{(1)} \]

\[ \max I(q) + \max I''(q) = \text{max} \quad \text{(2)} \]

\[ \text{(1)} - \text{(2)} = \max I'(q_{i+1}) + \max I''(q_{j+1}) - \min I'(q_i) - \min I''(q_{j+1}) \]

\[ \leq \epsilon n_1 + 2\epsilon n_2 \]

\[ = 2\epsilon (n_1 + n_2) \]
same sets

\[ \min I'(q_i^j) + \min I''(q_j^i) \quad \max I'(q_{i+1}^j) + \max I''(q_{j+1}^i) \]

\[ \frac{\max I'(q_{i+1}^j) + \max I''(q_{j+1}^i) - \min I'(q_i^j) - \min I''(q_j^i)}{2 \varepsilon n_1 + 2 \varepsilon n_2 = 2 \varepsilon (n_1 + n_2)} \]
This shows that merging 2 ε-APX QS's gives ε-APX QS of combined streams

size?
Pruning.

Input:

$\epsilon$-approximate quantile summary w/ too many points

Goal: sparser summary that's still very good

& keep

$l$ queries
Claim: resulting quantile is \((\varepsilon + \frac{1}{2\varepsilon})\)-APX

Proof

Suppose we query a rank \(k\).

Look at original summary.

\[ \leq \frac{1}{2\varepsilon} n \]

\[ 2 \varepsilon n + \frac{n}{\varepsilon} \]

\((\varepsilon + \frac{1}{2\varepsilon})\)-Approx
Recap: we can combine \( \varepsilon \)-APX quantile summaries to get \( \varepsilon \)-APX quantile summary of whole thing. Sparsify \( \varepsilon \)-APX quantile summary to \((\varepsilon + \frac{1}{2k})\)-APX quantile summary w/ k points.

Remains to address:

how to make one at all??
what if $n=1$?

Take the point

I claim that's all we need!

\[ APX \]

\[ \frac{3\epsilon}{\log n} \]
\[ \frac{2\epsilon}{\log n} \]
\[ \frac{\epsilon}{\log n} \]

\[ O \]

\[ \ell \]

take $k = \frac{\log n}{2\epsilon}$

at the root,

\[ \ell \approx \frac{3\log n}{\log n} \]

$\epsilon$-approximate quantiles
Space?

only keep "root summaries"
Theorem

- 1-pass
- $O(\log^2(n)/\varepsilon)$ space
- deterministic
- $\varepsilon$-APX quantile over stream

idea: mergability + dyadic intervals trick

slightly better:
have first level contain \( \frac{1}{2} \) points

\[ \log(n) \]

"leaves" \( \leq n \)
Theorem++

- 1-pass
- \(O(\log^2(\varepsilon n)/\varepsilon)\) space
- deterministic
- \(\varepsilon\)-APX quantile over stream

Even better?

Khanna-Greenwald [2001]:

\[ \frac{1}{\varepsilon} \log(\varepsilon n) \] space

- more sophisticated quantile summary, merging
- interval trick
Finding the median (and other ranks) in $p$ passes

Fix $p=2$ for simplicity.

goal: $O(\sqrt{n \log(n)})$ space

suppose we are querying rank $k$.

1st pass: build $\varepsilon$-APX quantile summary

for $\varepsilon = \frac{1}{\sqrt{n}}$ ($\log(n)$ space w/ GK)

query $k - \sqrt{n}$, $k + \sqrt{n}$ \Rightarrow a, b

\[
\frac{b - a}{\sqrt{n}} \leq \frac{4 \sqrt{n}}{n}
\]

\[
\begin{array}{c}
0 \\
\alpha \quad \beta
\end{array}
\]

\[
k - 2\sqrt{n} \leq \text{rank}(a) \leq k \leq \text{rank}(b) \leq k + 2\sqrt{n}
\]

\[a = \text{query}(k - \sqrt{n})\]
2nd pass:

$$\text{Count}^\#$$

Take the \((k-\#(\leq a/b))\)th in the sorted set

for general \(p\):

- make \(\frac{1}{n^p}\)-APX quantile summaries and filter.
- After \(p\) passes, down to \(n^{1/p}\) elements sort and select.