Space efficient quantile selection

Input: stream $a_1, \ldots, a_n \in U$ where $U$ has order $< 5\ 9\ 3\ 11\ 1\ 4\ 6\ \ldots$

- 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

(sort)

sort and select
count # above, below pivot
recurse on appropriate half
<table>
<thead>
<tr>
<th>Passes</th>
<th>Space</th>
<th>Algorithm</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$O(n)$</td>
<td>sort and select</td>
<td></td>
</tr>
<tr>
<td>$O(\log n)$</td>
<td>$O(1)$</td>
<td>quickselect (random pivot)</td>
<td>Munro, Paterson [1980]</td>
</tr>
<tr>
<td>$p$</td>
<td>$\tilde{O}(n^{1/p})$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

This lecture...
Approximations

given rank $k \in [n]$ and param $\varepsilon > 0$,
return element w/ rank $k = 3n$

Sampling:

for median:

sample $l = O\left(\frac{\log(1/\varepsilon)}{\varepsilon^2}\right)$ elements
return median of sample

for rank $k = d n$

return rank $d l$ element of sample

Deterministic?
Quantiles

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

$\ell$ elements $q_1 < q_2 < \cdots < q_{\ell}$

divide w/ intervals $I(q_i) \subseteq [1, n]$ s.t.

$\text{rank}(q_i) \in I(q_i)$

By tracking min/max specially, assume

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

By tracking min/max specially, assume

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

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

Querying

If \( I(q_i) \leq [k-\text{len}, k+\text{len}] \) for some \( q_i \),
then return \( q_i \).

how to ensure such \( q_i \) exists for \( k \)?
Lemma

\[
\text{total width} \leq 2En
\]

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

Proof two cases:

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

\( k \not\in I(q_i) \) \( \forall q_i \)
Suppose \( k \in I(q_i) \) for some \( i \)
if \( I(q_i) \leq [k-2\epsilon, k+2\epsilon] \) then done
else look at \( q_{i-1} \)
suppose $k \notin I(q_i) + i$

the "combined intervals" cover $[n_j]$

pick 1 covering $k$.

\[ \leq 2\varepsilon n \]

\[ k \leq n \]

\[ k + \varepsilon n \]

\[ = 2\varepsilon n \]

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

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

\[
S_1, \quad S_2
\]

\[
\Rightarrow Q' = [ q'_1, q'_2, q'_3, q'_4 ]
\]

\[
S_2
\]

\[
\Rightarrow Q'' = [ q''_1, q''_2, q''_3 ]
\]

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

\[
S_1 + S_2
\]

\[
"Q'UQ''" = \{ q''_1, \ldots, q''_c, I''(q''_1), \ldots, I''(q''_c) \}
\]
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 \( q'_j \) w/r/t \( S_2 \)
goal: bound rank \( q''_j \) w/r/t \( S_1 + S_2 \).

\[ q'_j \rightarrow q'_{j+1} \]

rank of \( q'_j \) in \( S_1 \) \( \geq \min \ I'(q_i) \]
\[ \leq \max \ I'(q_{i+1}) \]

so \( \min I''(q'_j) + \min I''(q''_j) \)
\[ \leq \text{rank}(q''_j \mid S_1 + S_2) \]
\[ \leq \max I''(q''_j) + \max I''(q''_j) \]

set \( I''(q'_j) = \left[ \min I'(q'_i) + \min I''(q''_j), \right. \]
\[ \max I'(q_{i+1}) + \max I''(q''_j) \]
\[ Q'' = \{q_1, \ldots, q_e, q''_1, \ldots, q''_m\}, \text{ w/ intervals } I'' \]

To show \( Q'' \) is \( \epsilon \)-APX, need to show "2\( \epsilon \)en width" property.

Take two consecutive intervals in \( Q'' \).

\[
\begin{array}{c}
q_1' \quad q_2' \\
\overline{[\text{---}]} \\
q_3' \quad q_4'
\end{array}
\]

Two cases:

1. elements from diff sets
2. elements from same sets
1) diff sets:

\[
\begin{align*}
\text{min}(I'(q_i)) + \text{min}(I''(q_{j-1})) \quad \text{(1)} \\
\text{max}(I''(q_i) + \text{max}(I'(q_{i+1})) \quad \text{(2)} \\
\end{align*}
\]

\[
\text{(3)} - (\text{(1)}) = \text{max}(I'(q_{i+1})) - \text{min}(I''(q_i)) \leq 2\varepsilon_n, \\
\text{max}(I''(q_i)) - \text{min}(I''(q_{i-1})) \leq 2\varepsilon_{n_2}
\]

\[
\leq 2\varepsilon(n_1 + n_2)
\]
This shows that merging 2 $\varepsilon$-APX QS's gives $\varepsilon$-APX QS of combined streams.
Lemma

given \( \varepsilon \)-approximate quantile summaries \( Q_1, \ldots, Q_h \) for \( S_1, \ldots, S_n \), we can combine to make \( \varepsilon \)-approximate summary \( Q_1 U \cdots U Q_h \) of \( S_1 U \cdots U S_h \).

size?
Pruning

Input:

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

Goal: sparser summary that’s still very good
claim: resulting quantile is \((3+\frac{1}{3e})\)-APX

**Proof**

suppose we query a particular rank \(k\)

look at original summary

\[
\begin{align*}
\frac{1}{e} & \leq \frac{i}{e} \leq \frac{n}{e} \\
\frac{i}{e} & \leq k \leq \frac{n}{e}
\end{align*}
\]

\[
\text{width} \leq \frac{1}{e}
\]

total width \(\leq 2en + \frac{1}{e}\) i.e. \((3+\frac{1}{3e})\)-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 \( n=1 \)?
make a summary of just that point.

I claim that's all we need!

\[ \text{APX} \]

\[ \frac{3\varepsilon}{\log n} \]

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

\[ \frac{\varepsilon}{\log n} \]

take \( k = \frac{\log n}{\varepsilon} \)
at the root,
\[ \Delta \]
\[ \frac{3\log n}{\Delta} \]
\[ \frac{0}{n} \]
\varepsilon-approximate quantiles
Space?
each summary takes $O(\log(n)/\varepsilon)$ space.
How many summaries?

only keep "root summaries"
how many "root summaries"?

One summary per height, so $\log n$ total summaries at any time
Theorem

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

idea: mergability + dyadic intervals trick

slightly better:

```
  /\  \\
 /   \  \\
V\   V\  \\
 V\   V\  \\
```

have first level contain $\frac{1}{\varepsilon}$ points

```
          /\  \\
         /   \  \\
        /     \  \\
       /       \  \\
      /         \  \\
     /           \\
```

$\varepsilon n$ summaries at "leaf level"

$\Rightarrow \log(\varepsilon n)$ height
Theorem++

- 1-pass
- $O(\log^2(\frac{en}{\varepsilon})/\varepsilon)$ space
- deterministic
- $\varepsilon$-APX quantile over stream

Even better?

Khanna- Greenwald:

$\frac{1}{\varepsilon} \log(en)$ space

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

Fix $p=2$ for simplicity.

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

suppose we are querying rank $k$.

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

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

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

\[
\begin{align*}
\text{rank}(a) & \leq k - \sqrt{n} \\
\text{rank}(b) & \geq k + \sqrt{n}
\end{align*}
\]
2nd pass:

- take all elements between a and b

For general \( p \):

- make \( \frac{1}{n^{1/p}} \)-APX quantile summaries and filter.

After \( p \) passes, down to \( n^{1/p} \) elements sort and select.