XGBoost: A Scalable Tree Boosting System Supplementary Material

Tianqi Chen
University of Washington
tqchen@cs.washington.edu

Carlos Guestrin
University of Washington
guestrin@cs.washington.edu

CCS Concepts

•**Methodologies** → **Machine learning;** •**Information systems** → **Data mining;**

Keywords

Large-scale Machine Learning

1. WEIGHTED QUANTILE SKETCH

In this section, we introduce the weighted quantile sketch algorithm. Approximate answer of quantile queries is for many real-world applications. One classical approach to this problem is GK algorithm [1] and extensions based on the GK framework [2]. The main component of these algorithms is a data structure called quantile summary, that is able to answer quantile queries with relative accuracy of ϵ. Two operations are defined for a quantile summary:

- A merge operation that combines two summaries with approximation error ϵ_1 and ϵ_2 together and create a merged summary with approximation error $\max(\epsilon_1, \epsilon_2)$.
- A prune operation that reduces the number of elements in the summary to $b+1$ and changes approximation error from ϵ to $\epsilon + \frac{1}{b}$.

A quantile summary with merge and prune operations forms basic building blocks of the distributed and streaming quantile computing algorithms [2].

In order to use quantile computation for approximate tree boosting, we need to find quantiles on weighted data. This more general problem is not supported by any of the existing algorithm. In this section, we describe a non-trivial weighted quantile summary structure to solve this problem. Importantly, the new algorithm contains merge and prune operations with *the same guarantee* as GK summary. This allows our summary to be plugged into all the frameworks used GK summary as building block and answer quantile queries over weighted data efficiently.

DOI: 10.475/123_4

1.1 Formalization and Definitions

Given an input multi-set $\mathcal{D} = \{(x_1, w_1), (x_2, w_2) \cdots (x_n, w_n)\}$ such that $w_i \in [0, +\infty), x_i \in \mathcal{X}$. Each x_i corresponds to a position of the point and w_i is the weight of the point. Assume we have a total order $<$ defined on $\mathcal{X}$. Let us define two rank functions $r^-_{\mathcal{D}}, r^+_{\mathcal{D}} : \mathcal{X} \to [0, +\infty)$

$$r^-_{\mathcal{D}}(y) = \sum_{(x,w)\in\mathcal{D}, x<y} w \tag{1}$$

$$r^+_{\mathcal{D}}(y) = \sum_{(x,w)\in\mathcal{D}, x\leq y} w \tag{2}$$

We should note that since $\mathcal{D}$ is defined to be a *multiset* of the points. It can contain multiple record with exactly same position x and weight w. We also define another weight function $\omega_{\mathcal{D}} : \mathcal{X} \to [0, +\infty)$ as

$$\omega_{\mathcal{D}}(y) = r^+_{\mathcal{D}}(y) - r^-_{\mathcal{D}}(y) = \sum_{(x,w)\in\mathcal{D}, x=y} w. \tag{3}$$

Finally, we also define the weight of multi-set $\mathcal{D}$ to be the sum of weights of all the points in the set

$$\omega(\mathcal{D}) = \sum_{(x,w)\in\mathcal{D}} w \tag{4}$$

Our task is given a series of input $\mathcal{D}$, to estimate $r^+(y)$ and $r^-(y)$ for $y \in \mathcal{X}$ as well as finding points with specific rank. Given these notations, we define quantile summary of weighted examples as follows:

DEFINITION 1.1. *Quantile Summary of Weighted Data*
A quantile summary for $\mathcal{D}$ is defined to be tuple $Q(\mathcal{D}) = (S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$, where $S = \{x_1, x_2, \cdots, x_k\}$ is selected from the points in $\mathcal{D}$ (i.e. $x_i \in \{x|(x,w) \in \mathcal{D}\}$) with the following properties:

1) $x_i < x_{i+1}$ for all i, and x_1 and x_k are minimum and maximum point in $\mathcal{D}$:

$$x_1 = \min_{(x,w)\in\mathcal{D}} x, \quad x_k = \max_{(x,w)\in\mathcal{D}} x$$

2) $\tilde{r}^+_{\mathcal{D}}$, $\tilde{r}^-_{\mathcal{D}}$ and $\tilde{\omega}_{\mathcal{D}}$ are functions in $S \to [0, +\infty)$, that satisfies

$$\tilde{r}^-_{\mathcal{D}}(x_i) \leq r^-_{\mathcal{D}}(x_i), \;\; \tilde{r}^+_{\mathcal{D}}(x_i) \geq r^+_{\mathcal{D}}(x_i), \;\; \tilde{\omega}_{\mathcal{D}}(x_i) \leq \omega_{\mathcal{D}}(x_i), \tag{5}$$

the equality sign holds for maximum and minimum point ($\tilde{r}^-_{\mathcal{D}}(x_i) = r^-_{\mathcal{D}}(x_i)$, $\tilde{r}^+_{\mathcal{D}}(x_i) = r^+_{\mathcal{D}}(x_i)$ and $\tilde{\omega}_{\mathcal{D}}(x_i) = \omega_{\mathcal{D}}(x_i)$ for $i \in \{1, k\}$).

Finally, the function value must also satisfy the following constraints

$$\tilde{r}^-_{\mathcal{D}}(x_i)+\tilde{\omega}_{\mathcal{D}}(x_i) \leq \tilde{r}^-_{\mathcal{D}}(x_{i+1}),\ \ \tilde{r}^+_{\mathcal{D}}(x_i) \leq \tilde{r}^+_{\mathcal{D}}(x_{i+1})-\tilde{\omega}_{\mathcal{D}}(x_{i+1}) \tag{6}$$

Since these functions are only defined on S, it is suffice to use $4k$ record to store the summary. Specifically, we need to remember each x_i and the corresponding function values of each x_i.

DEFINITION 1.2. *Extension of Function Domains*
Given a quantile summary $Q(\mathcal{D}) = (S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$ defined in Definition 1.1, the domain of $\tilde{r}^+_{\mathcal{D}}$, $\tilde{r}^-_{\mathcal{D}}$ and $\tilde{\omega}_{\mathcal{D}}$ were defined only in S. We extend the definition of these functions to $\mathcal{X} \to [0, +\infty)$ as follows
When $y < x_1$:

$$\tilde{r}^-_{\mathcal{D}}(y) = 0,\ \tilde{r}^+_{\mathcal{D}}(y) = 0,\ \tilde{\omega}_{\mathcal{D}}(y) = 0 \tag{7}$$

When $y > x_k$:

$$\tilde{r}^-_{\mathcal{D}}(y) = \tilde{r}^+_{\mathcal{D}}(x_k),\ \tilde{r}^+_{\mathcal{D}}(y) = \tilde{r}^+_{\mathcal{D}}(x_k),\ \tilde{\omega}_{\mathcal{D}}(y) = 0 \tag{8}$$

When $y \in (x_i, x_{i+1})$ for some i:

$$\begin{aligned} \tilde{r}^-_{\mathcal{D}}(y) &= \tilde{r}^-_{\mathcal{D}}(x_i) + \tilde{\omega}_{\mathcal{D}}(x_i), \\ \tilde{r}^+_{\mathcal{D}}(y) &= \tilde{r}^+_{\mathcal{D}}(x_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x_{i+1}), \\ \tilde{\omega}_{\mathcal{D}}(y) &= 0 \end{aligned} \tag{9}$$

LEMMA 1.1. *Extended Constraint*
The extended definition of $\tilde{r}^-_{\mathcal{D}}$, $\tilde{r}^+_{\mathcal{D}}$, $\tilde{\omega}_{\mathcal{D}}$ satisfies the following constraints

$$\tilde{r}^-_{\mathcal{D}}(y) \leq r^-_{\mathcal{D}}(y),\ \ \tilde{r}^+_{\mathcal{D}}(y) \geq r^+_{\mathcal{D}}(y),\ \ \tilde{\omega}_{\mathcal{D}}(y) \leq \omega_{\mathcal{D}}(y) \tag{10}$$

$$\tilde{r}^-_{\mathcal{D}}(y)+\tilde{\omega}_{\mathcal{D}}(y) \leq \tilde{r}^-_{\mathcal{D}}(x),\ \ \tilde{r}^+_{\mathcal{D}}(y) \leq \tilde{r}^+_{\mathcal{D}}(x)-\tilde{\omega}_{\mathcal{D}}(x),\ \textit{for all } y < x \tag{11}$$

PROOF. The only non-trivial part is to prove the case when $y \in (x_i, x_{i+1})$:

$$\tilde{r}^-_{\mathcal{D}}(y) = \tilde{r}^-_{\mathcal{D}}(x_i) + \tilde{\omega}_{\mathcal{D}}(x_i) \leq r^-_{\mathcal{D}}(x_i) + \omega_{\mathcal{D}}(x_i) \leq r^-_{\mathcal{D}}(y)$$

$$\tilde{r}^+_{\mathcal{D}}(y) = \tilde{r}^+_{\mathcal{D}}(x_{i+1})-\tilde{\omega}_{\mathcal{D}}(x_{i+1}) \geq r^+_{\mathcal{D}}(x_{i+1})-\omega_{\mathcal{D}}(x_{i+1}) \geq r^+_{\mathcal{D}}(y)$$

This proves Eq. (10). Furthermore, we can verify that

$$\tilde{r}^+_{\mathcal{D}}(x_i) \leq \tilde{r}^+_{\mathcal{D}}(x_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x_{i+1}) = \tilde{r}^+_{\mathcal{D}}(y) - \tilde{\omega}_{\mathcal{D}}(y)$$

$$\tilde{r}^-_{\mathcal{D}}(y) + \tilde{\omega}_{\mathcal{D}}(y) = \tilde{r}^-_{\mathcal{D}}(x_i) + \tilde{\omega}_{\mathcal{D}}(x_i) + 0 \leq \tilde{r}^-_{\mathcal{D}}(x_{i+1})$$

$$\tilde{r}^+_{\mathcal{D}}(y) = \tilde{r}^+_{\mathcal{D}}(x_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x_{i+1})$$

Using these facts and transitivity of $<$ relation, we can prove Eq. (11) □

We should note that the extension is based on the ground case defined in S, and we do not require extra space to store the summary in order to use the extended definition. We are now ready to introduce the definition of ϵ-approximate quantile summary.

DEFINITION 1.3. *ϵ-Approximate Quantile Summary*
Given a quantile summary $Q(\mathcal{D}) = (S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$, we call it is ϵ-approximate summary if for any $y \in \mathcal{X}$

$$\tilde{r}^+_{\mathcal{D}}(y) - \tilde{r}^-_{\mathcal{D}}(y) - \tilde{\omega}_{\mathcal{D}}(y) \leq \epsilon\omega(\mathcal{D}) \tag{12}$$

We use this definition since we know that $r^-(y) \in [\tilde{r}^-_{\mathcal{D}}(y), \tilde{r}^+_{\mathcal{D}}(y)-\tilde{\omega}_{\mathcal{D}}(y)]$ and $r^+(y) \in [\tilde{r}^-_{\mathcal{D}}(y) + \tilde{\omega}_{\mathcal{D}}(y), \tilde{r}^+_{\mathcal{D}}(y)]$. Eq. (12) means the we can get estimation of $r^+(y)$ and $r^-(y)$ by error of at most $\epsilon\omega(\mathcal{D})$.

LEMMA 1.2. *Quantile summary $Q(\mathcal{D}) = (S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$ is an ϵ-approximate summary if and only if the following two condition holds*

$$\tilde{r}^+_{\mathcal{D}}(x_i) - \tilde{r}^-_{\mathcal{D}}(x_i) - \tilde{\omega}_{\mathcal{D}}(x_i) \leq \epsilon\omega(\mathcal{D}) \tag{13}$$

$$\tilde{r}^+_{\mathcal{D}}(x_{i+1}) - \tilde{r}^-_{\mathcal{D}}(x_i) - \tilde{\omega}_{\mathcal{D}}(x_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x_i) \leq \epsilon\omega(\mathcal{D}) \tag{14}$$

PROOF. *The key is again consider $y \in (x_i, x_{i+1})$*

$$\tilde{r}^+_{\mathcal{D}}(y)-\tilde{r}^-_{\mathcal{D}}(y)-\tilde{\omega}_{\mathcal{D}}(y) = [\tilde{r}^+_{\mathcal{D}}(x_{i+1})-\tilde{\omega}_{\mathcal{D}}(x_{i+1})]-[\tilde{r}^+_{\mathcal{D}}(x_i)+\tilde{\omega}_{\mathcal{D}}(x_i)]-0$$

This means the condition in Eq. (14) plus Eq.(13) can give us Eq. (12) □

Property of Extended Function In this section, we have introduced the extension of function $\tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}}$ to $\mathcal{X} \to [0, +\infty)$. The key theme discussed in this section is the relation of *constraints on the original function and constraints on the extended function.* Lemma 1.1 and 1.2 show that the constraints on the original function can lead to in more general constraints on the extended function. This is a very useful property which will be used in the proofs in later sections.

1.2 Construction of Initial Summary

Given a small multi-set $\mathcal{D} = \{(x_1, w_1), (x_2, w_2), \cdots, (x_n, w_n)\}$, we can construct initial summary $Q(\mathcal{D}) = \{S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}}\}$, with S to the set of all values in $\mathcal{D}$ ($S = \{x|(x, w) \in \mathcal{D}\}$), and $\tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}}$ defined to be

$$\tilde{r}^+_{\mathcal{D}}(x) = r^+_{\mathcal{D}}(x),\ \ \tilde{r}^-_{\mathcal{D}}(x) = r^-_{\mathcal{D}}(x),\ \ \tilde{\omega}_{\mathcal{D}}(x) = \omega_{\mathcal{D}}(x) \text{ for } x \in S \tag{15}$$

The constructed summary is 0-approximate summary, since it can answer all the queries accurately. The constructed summary can be feed into future operations described in the latter sections.

1.3 Merge Operation

In this section, we define how we can merge the two summaries together. Assume we have $Q(\mathcal{D}_1) = (S_1, \tilde{r}^+_{\mathcal{D}_1}, \tilde{r}^-_{\mathcal{D}_1}, \tilde{\omega}_{\mathcal{D}_1})$ and $Q(\mathcal{D}_2) = (S_2, \tilde{r}^+_{\mathcal{D}_1}, \tilde{r}^-_{\mathcal{D}_2}, \tilde{\omega}_{\mathcal{D}_2})$ quantile summary of two dataset $\mathcal{D}_1$ and $\mathcal{D}_2$. Let $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$, and define the merged summary $Q(\mathcal{D}) = (S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$ as follows.

$$S = \{x_1, x_2 \cdots, x_k\}, x_i \in S_1 \text{ or } x_i \in S_2 \tag{16}$$

The points in S are combination of points in S_1 and S_2. And the function $\tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}}$ are defined to be

$$\tilde{r}^-_{\mathcal{D}}(x_i) = \tilde{r}^-_{\mathcal{D}_1}(x_i) + \tilde{r}^-_{\mathcal{D}_2}(x_i) \tag{17}$$

$$\tilde{r}^+_{\mathcal{D}}(x_i) = \tilde{r}^+_{\mathcal{D}_1}(x_i) + \tilde{r}^+_{\mathcal{D}_2}(x_i) \tag{18}$$

$$\tilde{\omega}_{\mathcal{D}}(x_i) = \tilde{\omega}_{\mathcal{D}_1}(x_i) + \tilde{\omega}_{\mathcal{D}_2}(x_i) \tag{19}$$

Here we use functions defined on $S \to [0, +\infty)$ on the left sides of equalities and use the extended function definitions on the right sides.

Due to additive nature of r^+, r^- and ω, which can be formally written as

$$\begin{aligned} r^-_{\mathcal{D}}(y) &= r^-_{\mathcal{D}_1}(y) + r^-_{\mathcal{D}_2}(y), \\ r^+_{\mathcal{D}}(y) &= r^+_{\mathcal{D}_1}(y) + r^+_{\mathcal{D}_2}(y), \\ \omega_{\mathcal{D}}(y) &= \omega_{\mathcal{D}_1}(y) + \omega_{\mathcal{D}_2}(y), \end{aligned} \tag{20}$$

Algorithm 1: Query Function $g(Q, d)$

Input: d: $0 \le d \le \omega(\mathcal{D})$
Input: $Q(\mathcal{D}) = (S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$ where $S = x_1, x_2, \cdots, x_k$
if $d < \frac{1}{2}[\tilde{r}^-_{\mathcal{D}}(x_1) + \tilde{r}^+_{\mathcal{D}}(x_1)]$ **then return** x_1 ;
if $d \ge \frac{1}{2}[\tilde{r}^-_{\mathcal{D}}(x_k) + \tilde{r}^+_{\mathcal{D}}(x_k)]$ **then return** x_k ;
Find i such that
$\frac{1}{2}[\tilde{r}^-_{\mathcal{D}}(x_i) + \tilde{r}^+_{\mathcal{D}}(x_i)] \le d < \frac{1}{2}[\tilde{r}^-_{\mathcal{D}}(x_{i+1}) + \tilde{r}^+_{\mathcal{D}}(x_{i+1})]$
if $2d < \tilde{r}^-_{\mathcal{D}}(x_i) + \tilde{\omega}_{\mathcal{D}}(x_i) + \tilde{r}^+_{\mathcal{D}}(x_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x_{i+1})$ **then**
  return x_i
else
  return x_{i+1}
end

and the extended constraint property in Lemma 1.1, we can verify that $Q(\mathcal{D})$ satisfies all the constraints in Definition 1.1. Therefore it is a valid quantile summary.

LEMMA 1.3. *The combined quantile summary satisfies*

$$\tilde{r}^-_{\mathcal{D}}(y) = \tilde{r}^-_{\mathcal{D}_1}(y) + \tilde{r}^-_{\mathcal{D}_2}(y) \tag{21}$$

$$\tilde{r}^+_{\mathcal{D}}(y) = \tilde{r}^+_{\mathcal{D}_1}(y) + \tilde{r}^+_{\mathcal{D}_2}(y) \tag{22}$$

$$\tilde{\omega}_{\mathcal{D}}(y) = \tilde{\omega}_{\mathcal{D}_1}(y) + \tilde{\omega}_{\mathcal{D}_2}(y) \tag{23}$$

for all $y \in \mathcal{X}$

This can be obtained by straight-forward application of Definition 1.2.

THEOREM 1.1. *If $Q(\mathcal{D}_1)$ is ϵ_1-approximate summary, and $Q(\mathcal{D}_2)$ is ϵ_2-approximate summary. Then the merged summary $Q(\mathcal{D})$ is $\max(\epsilon_1, \epsilon_2)$-approximate summary.*

PROOF. For any $y \in \mathcal{X}$, we have

$$\begin{aligned}
&\tilde{r}^+_{\mathcal{D}}(y) - \tilde{r}^-_{\mathcal{D}}(y) - \tilde{\omega}_{\mathcal{D}}(y) \\
=&[\tilde{r}^+_{\mathcal{D}_1}(y) + \tilde{r}^+_{\mathcal{D}_2}(y)] - [\tilde{r}^-_{\mathcal{D}_1}(y) + \tilde{r}^-_{\mathcal{D}_2}(y)] - [\tilde{\omega}_{\mathcal{D}_1}(y) + \tilde{\omega}_{\mathcal{D}_2}(y)] \\
\le&\epsilon_1\omega(\mathcal{D}_1) + \epsilon_2\omega(\mathcal{D}_2) \le \max(\epsilon_1, \epsilon_2)\omega(\mathcal{D}_1 \cup \mathcal{D}_2)
\end{aligned}$$

Here the first inequality is due to Lemma 1.3. □

1.4 Prune Operation

Before we start discussing the prune operation, we first introduce a query function $g(Q, d)$. The definition of function is shown in Algorithm 1. For a given rank d, the function returns a x whose rank is close to d. This property is formally described in the following Lemma.

LEMMA 1.4. *For a given ϵ-approximate summary $Q(\mathcal{D}) = (S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$, $x^* = g(Q, d)$ satisfies the following property*

$$\begin{aligned}
d &\ge \tilde{r}^+_{\mathcal{D}}(x^*) - \tilde{\omega}_{\mathcal{D}}(x^*) - \frac{\epsilon}{2}\omega(\mathcal{D}) \\
d &\le \tilde{r}^-_{\mathcal{D}}(x^*) + \tilde{\omega}_{\mathcal{D}}(x^*) + \frac{\epsilon}{2}\omega(\mathcal{D})
\end{aligned} \tag{24}$$

PROOF. We need to discuss four possible cases

- $d < \frac{1}{2}[\tilde{r}^-_{\mathcal{D}}(x_1) + \tilde{r}^+_{\mathcal{D}}(x_1)]$ and $x^* = x_1$. Note that the rank information for x_1 is accurate ($\tilde{\omega}_{\mathcal{D}}(x_1) = \tilde{r}^+_{\mathcal{D}}(x_1) = \omega(x_1)$, $\tilde{r}^-_{\mathcal{D}}(x_1) = 0$), we have

$$\begin{aligned}
d &\ge 0 - \frac{\epsilon}{2}\omega(\mathcal{D}) = \tilde{r}^+_{\mathcal{D}}(x_1) - \tilde{\omega}_{\mathcal{D}}(x_1) - \frac{\epsilon}{2}\omega(\mathcal{D}) \\
d &< \frac{1}{2}[\tilde{r}^-_{\mathcal{D}}(x_1) + \tilde{r}^+_{\mathcal{D}}(x_1)] \\
&\le \tilde{r}^-_{\mathcal{D}}(x_1) + \tilde{r}^+_{\mathcal{D}}(x_1) \\
&= \tilde{r}^-_{\mathcal{D}}(x_1) + \tilde{\omega}^+_{\mathcal{D}}(x_1)
\end{aligned}$$

- $d \ge \frac{1}{2}[\tilde{r}^-_{\mathcal{D}}(x_k) + \tilde{r}^+_{\mathcal{D}}(x_k)]$ and $x^* = x_k$, then

$$\begin{aligned}
d &\ge \frac{1}{2}[\tilde{r}^-_{\mathcal{D}}(x_k) + \tilde{r}^+_{\mathcal{D}}(x_k)] \\
&= \tilde{r}^+_{\mathcal{D}}(x_k) - \frac{1}{2}[\tilde{r}^+_{\mathcal{D}}(x_k) - \tilde{r}^-_{\mathcal{D}}(x_k)] \\
&= \tilde{r}^+_{\mathcal{D}}(x_k) - \frac{1}{2}\tilde{\omega}_{\mathcal{D}}(x_k) \\
d &< \omega(\mathcal{D}) + \frac{\epsilon}{2}\omega(\mathcal{D}) = \tilde{r}^-_{\mathcal{D}}(x_k) + \tilde{\omega}_{\mathcal{D}}(x_k) + \frac{\epsilon}{2}\omega(\mathcal{D})
\end{aligned}$$

- $x^* = x_i$ in the general case, then

$$\begin{aligned}
2d &< \tilde{r}^-_{\mathcal{D}}(x_i) + \tilde{\omega}_{\mathcal{D}}(x_i) + \tilde{r}^+_{\mathcal{D}}(x_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x_{i+1}) \\
&= 2[\tilde{r}^-_{\mathcal{D}}(x_i) + \tilde{\omega}_{\mathcal{D}}(x_i)] + [\tilde{r}^+_{\mathcal{D}}(x_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x_{i+1}) - \tilde{r}^-_{\mathcal{D}}(x_i) - \tilde{\omega}_{\mathcal{D}}(x_i)] \\
&\le 2[\tilde{r}^-_{\mathcal{D}}(x_i) + \tilde{\omega}_{\mathcal{D}}(x_i)] + \epsilon\omega(\mathcal{D}) \\
2d &\ge \tilde{r}^-_{\mathcal{D}}(x_i) + \tilde{r}^+_{\mathcal{D}}(x_i) \\
&= 2[\tilde{r}^+_{\mathcal{D}}(x_i) - \tilde{\omega}_{\mathcal{D}}(x_i)] - [\tilde{r}^+_{\mathcal{D}}(x_i) - \tilde{\omega}_{\mathcal{D}}(x_i) - \tilde{r}^-_{\mathcal{D}}(x_i)] + \tilde{\omega}_{\mathcal{D}}(x_i) \\
&\ge 2[\tilde{r}^+_{\mathcal{D}}(x_i) - \tilde{\omega}_{\mathcal{D}}(x_i)] - \epsilon\omega(\mathcal{D}) + 0
\end{aligned}$$

- $x^* = x_{i+1}$ in the general case

$$\begin{aligned}
2d &\ge \tilde{r}^-_{\mathcal{D}}(x_i) + \tilde{\omega}_{\mathcal{D}}(x_i) + \tilde{r}^+_{\mathcal{D}}(x_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x_{i+1}) \\
&= 2[\tilde{r}^+_{\mathcal{D}}(x_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x_{i+1})] \\
&\quad - [\tilde{r}^+_{\mathcal{D}}(x_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x_{i+1}) - \tilde{r}^-_{\mathcal{D}}(x_i) - \tilde{\omega}_{\mathcal{D}}(x_i)] \\
&\ge 2[\tilde{r}^+_{\mathcal{D}}(x_{i+1}) + \tilde{\omega}_{\mathcal{D}}(x_{i+1})] - \epsilon\omega(\mathcal{D}) \\
2d &\le \tilde{r}^-_{\mathcal{D}}(x_{i+1}) + \tilde{r}^+_{\mathcal{D}}(x_{i+1}) \\
&= 2[\tilde{r}^-_{\mathcal{D}}(x_{i+1}) + \tilde{\omega}_{\mathcal{D}}(x_{i+1})] \\
&\quad + [\tilde{r}^+_{\mathcal{D}}(x_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x_{i+1}) - \tilde{r}^-_{\mathcal{D}}(x_{i+1})] - \tilde{\omega}_{\mathcal{D}}(x_{i+1}) \\
&\le 2[\tilde{r}^-_{\mathcal{D}}(x_{i+1}) + \tilde{\omega}_{\mathcal{D}}(x_{i+1})] + \epsilon\omega(\mathcal{D}) - 0
\end{aligned}$$

□

Now we are ready to introduce the prune operation. Given a quantile summary $Q(\mathcal{D}) = (S, \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$ with $S = \{x_1, x_2, \cdots, x_k\}$ elements, and a memory budget b. The prune operation creates another summary $Q'(\mathcal{D}) = (S', \tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}})$ with $S' = \{x'_1, x'_2, \cdots, x'_{b+1}\}$, where x'_i are selected by query the original summary such that

$$x'_i = g\left(Q, \frac{i-1}{b}\omega(\mathcal{D})\right).$$

The definition of $\tilde{r}^+_{\mathcal{D}}, \tilde{r}^-_{\mathcal{D}}, \tilde{\omega}_{\mathcal{D}}$ in Q' is copied from original summary Q, by restricting input domain from S to S'. There could be duplicated entries in the S'. These duplicated entries can be safely removed to further reduce the memory cost. Since all the elements in Q' comes from Q, we can verify that Q' satisfies all the constraints in Definition 1.1 and is a valid quantile summary.

THEOREM 1.2. *Let $Q'(\mathcal{D})$ be the summary pruned from an ϵ-approximate quantile summary $Q(\mathcal{D})$ with b memory budget. Then $Q'(\mathcal{D})$ is a $(\epsilon + \frac{1}{b})$-approximate summary.*

PROOF. We only need to prove the property in Eq. (14) for Q'. Using Lemma 1.4, we have

$$\frac{i-1}{b}\omega(\mathcal{D}) + \frac{\epsilon}{2}\omega(\mathcal{D}) \geq \tilde{r}^+_{\mathcal{D}}(x'_i) - \tilde{\omega}_{\mathcal{D}}(x'_i)$$
$$\frac{i-1}{b}\omega(\mathcal{D}) - \frac{\epsilon}{2}\omega(\mathcal{D}) \leq \tilde{r}^-_{\mathcal{D}}(x'_i) + \tilde{\omega}_{\mathcal{D}}(x'_i)$$

Combining these inequalities gives

$$\begin{aligned}&\tilde{r}^+_{\mathcal{D}}(x'_{i+1}) - \tilde{\omega}_{\mathcal{D}}(x'_{i+1}) - \tilde{r}^-_{\mathcal{D}}(x'_i) - \tilde{\omega}_{\mathcal{D}}(x'_i)\\ \leq&[\frac{i}{b}\omega(\mathcal{D}) + \frac{\epsilon}{2}\omega(\mathcal{D})] - [\frac{i-1}{b}\omega(\mathcal{D}) - \frac{\epsilon}{2}\omega(\mathcal{D})] = (\frac{1}{b} + \epsilon)\omega(\mathcal{D})\end{aligned}$$

□

2. REFERENCES

[1] M. Greenwald and S. Khanna. Space-efficient online computation of quantile summaries. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pages 58–66, 2001.

[2] Q. Zhang and W. Wang. A fast algorithm for approximate quantiles in high speed data streams. In *Proceedings of the 19th International Conference on Scientific and Statistical Database Management*, 2007.