Efficient Haar+ Synopsis Construction for the Maximum Absolute Error Measure

Jinhyun Kim
Seoul National University
Seoul, Korea
jhkim@kdd.snu.ac.kr

Jun-Ki Min
Korea Univ. of Technology and Education, Cheonan, Korea
jkmin@koreatech.ac.kr

Kyuseok Shim*
Seoul National University
Seoul, Korea
kshim@snu.ac.kr

ABSTRACT

Several wavelet synopsis construction algorithms were previously proposed based on dynamic programming for unrestricted Haar wavelet synopses as well as Haar$^+$ synopses. However, they find an optimal synopsis for every incoming value in each node of a coefficient tree, even if different incoming values share an identical optimal synopsis. To alleviate the limitation, we present novel algorithms, which keep only a minimal set of the distinct optimal synopses in each node of the tree, for the error-bounded synopsis problem. Furthermore, we propose the methods to restrict coefficient values to be considered to compute the optimal synopses in each node. In addition, by partitioning all optimal synopses in each node into a set of groups, such that every group can be represented by a compact representation, we significantly improve the performance of the proposed algorithms.

PVLDB Reference Format:
Jinhyun Kim, Jun-Ki Min, and Kyuseok Shim. Efficient Haar+ Synopsis Construction for the Maximum Absolute Error Measure. *PVLDB*, 11(1): 40 - 52, 2017.
DOI: https://doi.org/10.14778/3136610.3136614

1. INTRODUCTION

Synopsis structures are widely used for approximate query answering to handle big data. Among various synopsis structures, the variants of Haar wavelet synopses have been used widely in diverse applications such as image processing [20, 24], OLAP/DSS systems [21, 27], time-series mining [2], query optimization [21, 22], approximate query answering [7, 12] and stream data processing [3, 7].

Since Haar wavelet synopses minimizing L_2-error suffer from wide variance and severe bias in the quality of approximations [4], dynamic programming algorithms [5, 6, 8, 9] to minimize L_∞-error have been investigated. Such algorithms consider *restricted* Haar wavelet synopses consisting of the Haar wavelet coefficients only. To improve the quality of approximations, *unrestricted* Haar wavelet synopses [10, 11, 16, 26], whose coefficients can take any real value, have been studied. In addition, the Haar$^+$ tree [16] extends the Haar tree structure and achieves the highest quality among unrestricted Haar wavelet synopses. The existing algorithms focus on the space-bounded synopsis problem whose goal is to minimize error measures satisfying a space budget. Recently, the algorithms [17, 18, 23] for the *error-bounded synopsis problem* to find a space-optimal synopsis, which is a synopsis with the minimum size satisfying a maximum error threshold, are proposed and utilized to solve efficiently the space-bounded synopsis problem. Thus, we develop efficient algorithms to find a space-optimal Haar$^+$ synopsis for the error-bounded synopsis problem. For ease of presentation, we mainly present our algorithms to obtain a space-optimal unrestricted Haar wavelet synopsis and then describe how to extend them for the space of Haar$^+$ synopses.

The state-of-the-art algorithms [17, 18] consider only multiples of a resolution step δ as the coefficient values since it is impractical to enumerate every real value. A space-optimal synopsis by taking the multiples of δ rather than any real value is called a *space-δ-optimal* synopsis. In the rest of the paper, if a synopsis is space-δ-optimal, we simply say that it is optimal whenever the context is clear.

In the state-of-the-art algorithms [17, 18], a table E_i is kept in each node c_i of a coefficient tree. An entry $E_i[v]$ stores the minimum size of a synopsis satisfying the error bound ϵ for an incoming value v where v is computed by the coefficients of c_i's ancestor nodes. Since there are many incoming values v to c_i, we need to maintain a large number of entries $E_i[v]$, even if the space-optimal synopses for different incoming values could be identical. Thus, to compute space-optimal synopses for c_i, we examine a lot of entry pairs $E_{2i}[v]$ and $E_{2i+1}[v]$ which store the sizes of optimal synopses in c_i's left and right child nodes, respectively.

To alleviate the limitation of the state-of-the-art algorithms [17, 18], we propose the algorithm `OptExt-EB` for unrestricted Haar wavelet synopses to solve the error-bounded synopsis problem. Our algorithm `OptExt-EB` keeps only a minimal set of the synopses each of which is optimal for an incoming value in each node c_i. For each optimal synopsis, we annotate it with its *canonical error range* which allows us to compute the error of the synopsis for every incoming value to c_i. Then, we compute a set of the optimal synopses in c_i by examining only the distinct pairs of optimal synopses in c_i's left and right child nodes, respectively.

We observe that the optimal synopses in a node c_i, which are constructed from a pair of synopses in its child nodes by varying c_i's values, have the property that their canonical

*Primary contact author.

Proceedings of the VLDB Endowment, Vol. 11, No. 1

DOI: https://doi.org/10.14778/3136610.3136614

error ranges have the same length and are shifted by δ from each other. Thus, we can represent all optimal synopses in c_i by a set of such sets, called *extended synopses*, and denote each extended synopsis by its *compact representation*. Then, for a pair of extended synopses S_L and S_R in c_{2i} and c_{2i+1} respectively, we show that the optimal synopses, which are produced from every pair of synopses in S_L and S_R respectively, form an extended synopsis whose compact representation can be computed directly with $O(1)$ time.

Among various synopsis structures, since a Haar$^+$ tree is known as the most effective wavelet-inspired structure, `MinHaarSpace-HP` [18] was previously developed to find an optimal Haar$^+$ synopsis satisfying an error bound. Thus, we also develop `OptExtHP-EB` by extending `OptExt-EB` to search the space of Haar$^+$ synopses. As a resolution step δ decreases, although the quality of an optimal synopsis is enhanced, the execution times of both `MinHaarSpace` and `MinHaarSpace-HP` in [18] increase quadratically. On the contrary, our `OptExt-EB` and `OptExtHP-EB` have the desirable property that their execution times are less affected by δ. By utilizing the notion of extended synopses, our algorithms speed up the execution times up to orders of magnitude compared to the state-of-the-art algorithms.

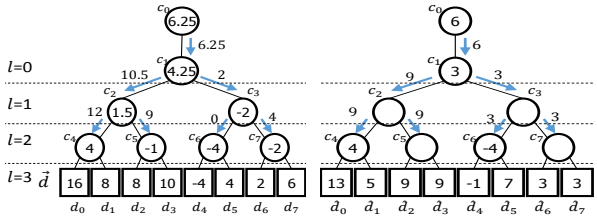
By conducting performance study with both synthetic and real-life data sets, we demonstrate that our `OptExtHP-EB` to find an optimal Haar$^+$ synopsis is the best performer compared to `MinHaarSpace-HP` [18] and `OptExt-EB` for the error-bounded synopsis problem. In addition, we show that the indirect algorithm [18] by invoking `MinHaarSpace-HP` with binary search to solve the space-bounded synopsis problem becomes much faster by utilizing our `OptExtHP-EB` instead.

2. RELATED WORK

The research on constructing wavelet synopses has a long and rich history. Since the quality of wavelet synopsis algorithms to minimize the overall root-mean squared error (i.e., L_2-error) usually varies widely [4], the algorithms minimizing the maximum-error (e.g., L_∞-error) are proposed in [5, 8]. To construct optimal wavelet synopses, two problems have been studied. The *space-bounded synopsis problem* finds an *error-optimal synopsis* minimizing some error measures for a given space budget. On the other hand, the *error-bounded synopsis problem* discovers a *space-optimal synopsis* with the smallest size satisfying a given error bound.

The space-bounded synopsis problem is studied in [5, 6, 8, 9, 13, 15]. Some algorithms [5, 13, 15] find an error-optimal *restricted* Haar wavelet synopsis consisting of Haar wavelet coefficients only. The other class of algorithms [10, 11, 16, 26] discovers an error-optimal *unrestricted* Haar wavelet synopsis whose coefficient values take any real value but take only multiples of a resolution step δ to delimit the domain of coefficient values. The unrestricted wavelet algorithms lead to higher quality than restricted wavelet algorithms [10, 11, 16, 26]. The wavelet synopses based on a refined wavelet-inspired data structure, called a *Haar$^+$ tree*, is shown to be more accurate than the other unrestricted wavelet synopses [16, 18]. A simplified variant of the Haar$^+$ tree, called the compact hierarchical histogram (CHH), is introduced in [17, 26]. The concept of winning intervals used in our algorithms was previously investigated in [14, 17].

Karras et al. [17, 18] propose the state-of-the-art Haar$^+$ synopsis algorithm to minimize the weighted maximum error for the error-bounded synopsis problem and show that the space-bounded synopsis problem can be more efficiently solved *indirectly* by utilizing the algorithm. The algorithms are proposed in [25] to find an unrestricted wavelet synopsis whose size is close to that of a space-optimal synopsis for the error-bounded synopsis problem. However, since these algorithms do not generate the synopsis with the smallest size satisfying the error bound, they cannot be utilized to solve indirectly the space-bounded synopsis problem.



(a) A coefficient tree T (b) An unrestricted synopsis

Figure 1: A coefficient tree and a synopsis

3. PRELIMINARY

3.1 The Error-bounded Synopsis Problem

A coefficient tree is a hierarchical structure [12] where each internal (i.e., coefficient) node and leaf node are associated with a wavelet coefficient and a data value, respectively. We refer to each leaf node and its data value, as d_j. When a coefficient with a value x is selected in a node c_i, we represent it by $c_i{:}x$. Every internal node c_i except the root node c_0 has two child nodes c_{2i} and c_{2i+1}. Let T be the coefficient tree obtained from a data vector $d{=}\langle d_0, \ldots, d_{N\text{-}1}\rangle$ and T_i be the subtree of T rooted at c_i. We denote the sets of all internal nodes (i.e., coefficients) and all leaf nodes (i.e., data values) in T_i as $\mathsf{coeff}(T_i)$ and $\mathsf{data}(T_i)$, respectively.

For a coefficient tree T, when the coefficient in the root node has the overall average value of data in $\mathsf{data}(T_0)$ and the coefficient in every other node c_i has the half of the difference between the averages of data values in $\mathsf{data}(T_{2i})$ and $\mathsf{data}(T_{2i+1})$ respectively, T becomes a Haar wavelet coefficient tree. Figure 1(a) shows an example of a coefficient tree for a data vector $d{=}\langle 16, 8, 8, 10, \text{-}4, 4, 2, 6\rangle$. Let A be a subset of the ancestor nodes of c_i in T. Then, the *incoming value* v to c_i is calculated as $v{=}\sum_{c_k:x\in A} sign_{ik}\cdot x$ where $sign_{ik}{=}$-1 if c_i appears in T_{2k+1}; otherwise, $sign_{ik}{=}1$.

Given a coefficient subtree T_i, the set of non-zero coefficients is called a *synopsis of* T_i (or T_i-synopsis). When a synopsis s has no coefficient in c_i, the incoming values v_ℓ to c_{2i} and v_r to c_{2i+1} are v. On the contrary, when a coefficient $c_i{:}x$ appears in a synopsis s, the incoming values v_ℓ to c_{2i} and v_r to c_{2i+1} become $v{+}x$ and $v{-}x$. Given a T_0-synopsis s, the incoming value to d_j computed from the ancestor nodes of d_j appearing in s becomes the reconstructed value of d_j. For example, for the synopsis $s{=}\{c_0{:}6, c_1{:}3, c_4{:}4, c_6{:}\text{-}4\}$ of the coefficient tree T in Figure 1(b), since the incoming value to d_4 is $6{-}3{+}(\text{-}4){=}\text{-}1$, the reconstructed value of d_4 is -1.

DEFINITION 3.1.: *For a coefficient subtree T_i, we define*

(1) $\hat{d}_j(T_i, s, v)$ *is the* ***reconstructed value*** *of $d_j{\in}\mathsf{data}(T_i)$ for the incoming value v to c_i with a synopsis s.*

(2) $F_{T_i,s}(v){=}\max_{d_j\in\mathsf{data}(T_i)}|d_j{-}\hat{d}_j(T_i,s,v)|$ *is the error function to compute the* ***maximum absolute error*** *of a synopsis s with $\mathsf{data}(T_i)$ for the incoming value v to c_i.*

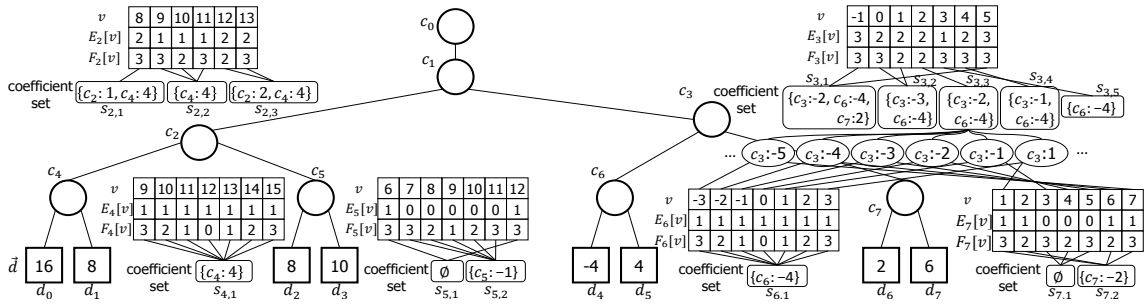


Figure 2: The tables stored in each node by MinHaarSpace when ϵ=3 and δ=1

EXAMPLE 3.2.: *Consider a synopsis* s={c_0:6,c_1:3,c_4:4, c_6:-4} *in* Figure 1(b). *Since the incoming value to* c_1 *is* 6 *and the coefficient value* 3 *is chosen in* c_1*, the incoming values to* c_2 *and* c_3 *are* 9 *and* 3*, respectively. Since* c_2 *and* c_5 *are not chosen in* s*, the incoming values to* d_2 *and* d_3 *are* 9*. Thus,* $\hat{d}_2(T,s,0)=\hat{d}_3(T,s,0)=9$ *and we get* $F_{T_5,s}(0)=1$. ∎

DEFINITION 3.3.: *Given a data vector* d *and an error bound* ϵ*, the error-bounded synopsis problem is to find a synopsis* s *of the coefficient tree for* d *such that* (1) *the maximum absolute error (i.e.,* $\max_{d_j \in \mathtt{data}(T_0)} |d_j - \hat{d}_j(T_0,s,0)|$*) is at most* ϵ *and* (2) *the number of coefficients in* s *is minimized. When there are several such synopses, the one with the minimum error is selected.*

We next review the MinHaarSpace algorithm [18] which is the state-of-the-art to find an optimal unrestricted Haar wavelet synopsis for the error-bounded synopsis problem.

3.2 The *MinHaarSpace* Algorithm

Let v_i^H be the incoming value to a node c_i when every ancestor node c_j of c_i has its Haar wavelet coefficient value x_j^H in a coefficient tree T. The incoming values and coefficient values in a node c_i can be restricted as follows.

LEMMA 3.4.: [18] *For a* T_i*-synopsis* s *satisfying an error bound* ϵ*, let* v *be the incoming value to* c_i *computed by only* s*' coefficients appearing in* c_i*'s ancestor nodes. Then, the incoming value* v *is always located in* $[v_i^H-\epsilon, v_i^H+\epsilon]$*. In addition, if the synopsis* s *has a coefficient value* x_i *in* c_i*, the value* x_i *always lies in* $[x_i^H-(\epsilon-|v-v_i^H|), x_i^H+(\epsilon-|v-v_i^H|)]$.

Since MinHaarSpace considers the δ multiples for coefficient values in c_i, by Lemma 3.4, the *candidate incoming value set* $\mathtt{IV}(i)$ consists of every multiple of δ in $[v_i^H-\epsilon, v_i^H+\epsilon]$ and *the candidate coefficient value set* $\mathtt{CV}(i,v)$ consists of every multiple of δ in $[x_i^H-(\epsilon-|v-v_i^H|), x_i^H+(\epsilon-|v-v_i^H|)]$.

For an incoming value v to c_i, MinHaarSpace finds a T_i-synopsis with the minimum error among all T_i-synopses with the minimum size satisfying the error bound ϵ. It utilizes $E_i[v]$ and $F_i[v]$ to store the minimum size and the minimum error of the synopsis discovered, respectively. Due to lack of space, we only provide the recursive equation of $E_i[v]$. Given a data vector $d=\langle d_0,\ldots,d_{N-1}\rangle$, we compute $E_i[v]$ with every incoming value v in $\mathtt{IV}(i)$ as follows.

When c_i is an internal node ($0<i<N$):

$$E_i[v] = \min \begin{cases} \min\limits_{x \in \mathtt{CV}(i,v)} \{E_{2i}[v+x]+E_{2i+1}[v-x]+1\}. & (1) \\ E_{2i}[v]+E_{2i+1}[v] & (2) \end{cases}$$

(a) If the coefficient value x is selected in c_i, since the incoming values to c_{2i} and c_{2i+1} become $v+x$ and $v-x$, we obtain Equation (1).

(b) If no coefficient is chosen, since the incoming values to c_{2i} and c_{2i+1} are v, we get Equation (2).

When c_i is the root node ($i=0$): Since the root node has a single child node c_1, $E_0[0]=\min(E_1[0], \min\limits_{x \in \mathtt{CV}(0,0)} E_1[x]+1)$.

When c_i is a leaf node $d_{(i-N)}$ ($i \geq N$): If $|v-d_{(i-N)}| \leq \epsilon$, $E_i[v]=0$. Otherwise, $E_i[v]=\infty$.

Note that MinHaarSpace takes $O(N(\epsilon/\delta)^2)$ time since it considers only the multiples of δ for coefficient values in c_i.

EXAMPLE 3.5: *For a data vector* $d=\langle 16,8,8,10,-4,4,2,6\rangle$ *with the coefficient tree* T *in* Figure 1*, we show how to compute* $E_3[v]$ *in the node* c_3 *with an error bound* ϵ=3 *and a resolution step* δ=1*. From* Lemma 3.4, $\mathtt{IV}(3)$={-1, 0, 1, 2, 3, 4, 5} *and* $\mathtt{CV}(3,2)$={-5, -4, -3, -2, -1, 0, 1} *for the incoming value* $2 \in \mathtt{IV}(3)$*. By* Equations (1)-(2)*,* $E_3[2]=\min(E_6[2]+E_7[2], \min_{x \in \mathtt{CV}(3,2)}\{E_6[2+x]+E_7[2-x]+1\})=2$*. We provide the values of* $E_i[v]$*s and* $F_i[v]$*s in* Figure 2*. We also show an optimal synopsis for each incoming value* v *in* Figure 2*, although* MinHaarSpace *does not store the optimal synopses.* ∎

The space-bounded synopsis problem: We can solve the space-bounded synopsis problem more efficiently by using the error-bounded synopsis algorithm [18] rather than *directly* computing an error-optimal synopsis. The indirect algorithm IndirectHaar [18] computes an error-optimal synopsis by invoking MinHaarSpace repeatedly with binary search on the error value. It takes $O(N(\epsilon/\delta)^2(\log\epsilon^* + \log N))$ time where ϵ^* is the minimal error in the required space B.

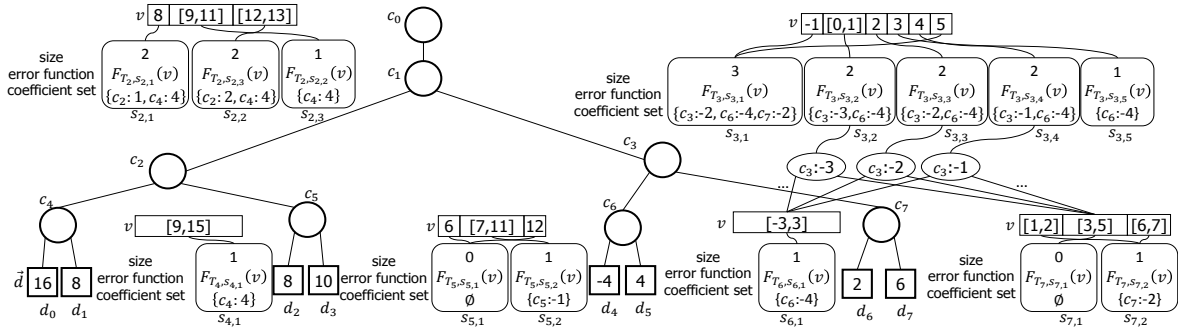
4. THE OUTLINE OF OUR ALGORITHM

In this section, we present an overview and pseudocode of our proposed algorithm OptExt-EB to compute a space-optimal unrestricted Haar wavelet synopsis.

4.1 An Overview of How OptExt-EB works

MinHaarSpace[18] stores $E_i[v]$ and $F_i[v]$ in each node c_i, even if different incoming values share an identical T_i-synopsis. For instance, in Figure 2, when ϵ=3 and δ=1, although the synopsis $s_{7,1}$ in c_7 is optimal for the incoming values 3, 4 and 5, it stores $E_7[v]$ and $F_7[v]$ individually for v=3, 4, 5.

Representing by distinct synopses: To alleviate the inefficiency of MinHaarSpace [18], we store the size and error function of each distinct synopsis *only once* for the incoming values where the synopsis is optimal. With the error function for a T_i-synopsis, we compute its maximum absolute error for any incoming value v. For the tables stored by MinHaarSpace in Figure 2, our new representations of the tables are shown in Figure 3(a). For example, the size 0 and error function $F_{T_7,s_{7,1}}(v)$ of the synopsis $s_{7,1}$ are kept for the incoming value interval $[3,5]$ only once in the node c_7.



(a) Representing by distinct synopses in each node

(b) Representing by extended synopses in each node

Figure 3: An example of how our proposed algorithm `OptExt-EB` works when ϵ=3 and δ=1

Representing by extended synopses: Consider the optimal synopses $s_{3,2}$, $s_{3,3}$ and $s_{3,4}$, whose error functions are $F_{T_3,s_{3,2}}(v)$, $F_{T_3,s_{3,3}}(v)$ and $F_{T_3,s_{3,4}}(v)$ respectively, for the node c_3 in Figure 3(a). The synopses have the same size 2 and are constructed from the same pair of $s_{6,1}$ and $s_{7,1}$ by varying the coefficient value in c_3 from -3 to -1 with resolution step δ=1. Although the constructed synopses are distinct synopses, we show later that $F_{T_3,s_{3,4}}(v)$=$F_{T_3,s_{3,3}}(v-\delta)$ =$F_{T_3,s_{3,2}}(v-2\delta)$ in Example 5.12. Thus, we represent the three synopses by an *extended synopsis* which can be denoted by a compact representation of the *synopsis size* (i.e., the number of selected coefficients), a *single error function* and the *group size* (i.e., the number of the represented synopses). For the stored tables in Figure 3(a), the extended synopses stored by `OptExt-EB` are shown in Figure 3(b). For instance, the extended synopsis $S_{3,2}$ representing the synopses $s_{3,2}$, $s_{3,3}$ and $s_{3,4}$ is represented by the synopsis size 2, a single error function $F_{T_3,s_{3,2}}(v)$, and the group size 3.

We show later that the synopses, which are obtained by pruning the useless synopses from the synopses constructed from a pair of extended synopses, form an extended synopsis as well as its compact representation is constructed in $O(1)$ time. Since a single synopsis is a special case of an extended synopsis, our `OptExt-EB` simply represents the optimal synopses stored in each node c_i by a *set of extended synopses*, called a *T_i-optimal extended synopsis set*. By using T_i-optimal extended synopsis sets, our proposed `OptExt-EB` becomes much faster than `MinHaarSpace`.

4.2 The Pseudocode of OptExt-EB

The pseudocode of `OptExt-EB` is presented in Figure 4. For each coefficient node c_i, it computes a T_i-optimal extended synopsis set from a pair of a T_{2i}-optimal extended synopsis set and a T_{2i+1}-optimal extended synopsis set in a bottom up fashion as follows.

```
Procedure OptExt-EB(i)
Input:    the coefficient node id i of a subtree T_i
1.  if i ≥ N then O_i = ∅    // for a leaf node
2.  else if 0 < i < N then  // for an internal node
3.    O_2i = OptExt-EB(2i); O_2i+1 =OptExt-EB(2i+1);
4.    ⟨m_2i, M_2i⟩=GetSizes(O_2i); ⟨m_2i+1, M_2i+1⟩=GetSizes(O_2i+1);
5.    for b=m_2i+m_2i+1 to M_2i+M_2i+1
6.      D^c_(i,b) = ExtSynCoef(O_2i, O_2i+1, b)      // c_i is selected
7.      D^∅_(i,b) = ExtSynNoCoef(O_2i, O_2i+1, b)    // c_i is not chosen
8.      D_(i,b) = RequiredSet(D^c_(i,b) ∪ D^∅_(i,b))
9.      ⟨O_(i,b), U_b⟩ = StrictOptSet(D_(i,b), U_b-1)
10.     if U_b=IV(i) then break
11. else  // for the root node
12.   O_1 = OptExt-EB(1); O_0 = OptExtRoot(O_1)
13. return O_i = ⋃_b O_(i,b)
```

Figure 4: The pseudocode of `OptExt-EB`

(1) When c_i is a leaf node (i.e., $i \geq N$): Since coeff(T_i) is empty, `OptExt-EB` returns the empty set (line 1).

(2) When c_i is an internal node (i.e., $1 \leq i < N$): To compute a T_i-optimal extended synopsis set O_i, `OptExt-EB` first calculates a T_{2i}-optimal extended synopsis set O_{2i} and a T_{2i+1}-optimal extended synopsis set O_{2i+1} (line 3). It next finds the smallest and largest synopsis sizes m_{2i} and M_{2i} (respectively, m_{2i+1} and M_{2i+1}) of the extended synopses in O_{2i} (respectively, O_{2i+1}) by invoking `GetSizes` (line 4). When a T_i-extended synopsis s is constructed from a pair of extended synopses in O_{2i} and O_{2i+1} respectively, the smallest (respectively, largest) synopsis size of s is $m_{2i}+m_{2i+1}$ (respectively, $M_{2i}+M_{2i+1}+1$). Let $\mathsf{O}_{(i,b)}$ be a *$T_{(i,b)}$-strictly optimal extended synopsis set*, which is the set of all extended synopses with synopsis size b in O_i. Then, the **for**

Table 1: Notations

Notations	Description
$\texttt{CER}(T_i, s)$	the T_i-CER of a T_i-synopsis s
$S_i^C(s_L, s_R)$ (resp., $S_i^{NC}(s_L, s_R)$)	the T_i-constructed coef (resp., nocoef) synopsis set constructed from a T_{2i}-synopsis s_L and a T_{2i+1}-synopsis s_R
$S_i^C(S_L, S_R)$ (resp., $S_i^{NC}(S_L, S_R)$)	the T_i-constructed coef (resp., nocoef) synopsis set constructed from a T_{2i}-extended synopsis S_L and a T_{2i+1}-extended synopsis S_R
$I_W(s, S)$ (resp., $I_W(S, \mathsf{S}_E)$)	the winning interval of a T_i-synopsis s (resp., a T_i-extended synopsis S) wrt a T_i-synopsis set S (resp., a T_i-extended synopsis set S_E)

loop (lines 5-9) generates each $T_{(i,b)}$-strictly optimal extended synopsis set $\mathsf{O}_{(i,b)}$ one by one starting the index b from $(m_{2i}+m_{2i+1})$ to $(M_{2i}+M_{2i+1}+1)$.

To compute $\mathsf{O}_{(i,b)}$, the procedure `ExtSynCoef` (respectively, `ExtSynNoCoef`) calculates the set $D^c_{(i,b)}$ (respectively, $D^{\emptyset}_{(i,b)}$), which is the set of the extended synopses with synopsis size b generated from every pair of extended synopses in O_{2i} and O_{2i+1} respectively, when a coefficient (respectively, no coefficient) is selected in c_i (lines 6-7). Based on the properties of extended synopses to be explored later, `OptExt-EB` removes the extended synopses, which cannot be optimal in c_i, from $D^c_{(i,b)} \cup D^{\emptyset}_{(i,b)}$ by invoking the procedures `RequiredSet` and `StrictOptSet` (lines 8-9).

The procedure `RequiredSet` generates $D_{(i,b)}$ by pruning the useless extended synopses among the b-sized extended synopses from $D^c_{(i,b)} \cup D^{\emptyset}_{(i,b)}$. Let U_b be the set of incoming values whose optimal synopses exist in a $T_{(i,b')}$-strictly optimal extended synopsis set $\mathsf{O}_{(i,b')}$ with $b' \leq b$. For an incoming value to c_i, if there exists a b'-sized synopsis satisfying an error bound ϵ, since a b-sized synopsis with $b > b'$ is not optimal by Definition 3.3. Thus, the procedure `StrictOptSet` computes $\mathsf{O}_{(i,b)}$ by removing such useless synopses from $D_{(i,b)}$ by examining U_{b-1}, and returns $\mathsf{O}_{(i,b)}$ and U_b.

After computing $\mathsf{O}_{(i,b)}$, if $U_b = \texttt{IV}(i)$, since there is an optimal synopsis for every incoming value to c_i, we do not need to compute $\mathsf{O}_{(i,b'')}$ with $b'' > b$ (line 10). Thus, we stop the **for** loop and union every $\mathsf{O}_{(i,b)}$ to compute O_i (line 13).

(3) When c_i is the root node (i.e., $i=0$): We first call `OptExt-EB`(1) to obtain a T_1-optimal extended synopsis set O_1. Then, `OptExt-EB` computes an optimal synopsis by invoking `OptExtRoot` with O_1 (line 12).

More detailed descriptions of the procedures invoked by `OptExt-EB` are presented in Section 6.2.

5. EXTENDED SYNOPSES

We introduce an extended synopsis to denote a set of the same sized synopses whose error functions are δ-shifted from each other. The notations to be used are in Table 1. Due to the lack of space, we omit the proofs in the rest of the paper. The proofs can be found in our technical report [19].

5.1 Properties of Synopses

We first provide the error function $F_{T_i,s}(v)$ of a T_i-synopsis s with its properties. We next show how to obtain the error function of a T_i-synopsis s and how to limit the coefficient values in each node c_i when the T_i-synopsis s is constructed from a pair of a T_{2i}-synopsis s_L and a T_{2i+1}-synopsis s_R.

Error functions of synopses: The canonical error range of a T_i-synopsis s is used to define the error function $F_{T_i,s}(v)$.

DEFINITION 5.1.: *For a T_i-synopsis s and an incoming value v to c_i, $e_{\mathsf{min}}(T_i, s, v)$ and $e_{\mathsf{max}}(T_i, s, v)$ are the minimum and maximum* ***signed*** *errors of all data values d_j in $\mathsf{data}(T_i)$, respectively. That is,*

$$e_{\mathsf{min}}(T_i, s, v) = \min_{d_j \in \mathsf{data}(T_i)}(d_j - \hat{d}_j(T_i, s, v)),$$

$$e_{\mathsf{max}}(T_i, s, v) = \max_{d_j \in \mathsf{data}(T_i)}(d_j - \hat{d}_j(T_i, s, v)).$$

*Moreover, $[e_{\mathsf{min}}(T_i, s, 0), e_{\mathsf{max}}(T_i, s, 0)]$ is the T_i-**canonical error range** (abbreviated by T_i-CER) of s, denoted by $\texttt{CER}(T_i, s)$.*

Since $e_{\mathsf{min}}(T_i, s, v) \leq d_j - \hat{d}_j(T_i, s, v) \leq e_{\mathsf{max}}(T_i, s, v)$, the maximum absolute error defined in Definition 3.1 is the larger of $|e_{\mathsf{min}}(T_i, s, v)|$ and $|e_{\mathsf{max}}(T_i, s, v)|$. Because an incoming value v to a node c_i contributes positively to the reconstructed value of every d_j in $\mathsf{data}(T_i)$, $\hat{d}_j(T_i, s, v) = \hat{d}_j(T_i, s, 0) + v$ resulting in that $e_{\mathsf{min}}(T_i, s, v) = e_{\mathsf{min}}(T_i, s, 0) - v$ and $e_{\mathsf{max}}(T_i, s, v) = e_{\mathsf{max}}(T_i, s, 0) - v$. Thus, we can compute $F_{T_i,s}(v)$ as follows.

PROPOSITION 5.2.: *For a T_i-synopsis s with $\texttt{CER}(T_i, s) = [e_{\mathsf{min}}, e_{\mathsf{max}}]$, $F_{T_i,s}(v) = \max(|e_{\mathsf{min}} - v|, |e_{\mathsf{max}} - v|)$.*

By Proposition 5.2, for a T_i-synopsis s, we obtain $F_{T_i,s}(v)$ for every incoming value v to c_i from $\texttt{CER}(T_i, s)$ only, even if we do not compute the reconstructed value of every data in $\mathsf{data}(T_i)$ with the selected coefficient values in s. Thus, when a T_i-synopsis s is optimal for several incoming values to c_i, we store its T_i-CER only in each coefficient node c_i.

EXAMPLE 5.3: *For a coefficient node c_7, consider the synopsis $s_{7,1}$ in* Figure 3(a). *The subtree T_7 has two leaf nodes d_6 and d_7. Since $d_6 - \hat{d}_6(T_7, s_{7,1}, 0) = 2$ and $d_7 - \hat{d}_7(T_7, s_{7,1}, 0) = 6$, the T_7-CER of $s_{7,1}$ is $[2, 6]$. Then, $F_{T_7,s_{7,1}}(3)$ becomes 3 $(= \max(|2-3|, |6-3|))$ by* Proposition 5.2. ∎

The error function of a constructed synopsis: When a T_i-synopsis s is constructed from a pair of a T_{2i}-synopsis s_L and a T_{2i+1}-synopsis s_R, $\texttt{CER}(T_i, s)$ can be computed from $\texttt{CER}(T_{2i}, s_L)$ and $\texttt{CER}(T_{2i+1}, s_R)$ in $O(1)$ time as follows.

LEMMA 5.4.: *When a T_i-synopsis s is constructed from a pair of a T_{2i}-synopsis s_L and a T_{2i+1}-synopsis s_R with $\texttt{CER}(T_{2i}, s_L) = [e_{L,\mathsf{min}}, e_{L,\mathsf{max}}]$ and $\texttt{CER}(T_{2i+1}, s_R) = [e_{R,\mathsf{min}}, e_{R,\mathsf{max}}]$, we have $\texttt{CER}(T_i, s) = [e_{\mathsf{min}}, e_{\mathsf{max}}]$ where*

- *When a coefficient value x is selected in c_i,*
  $e_{\mathsf{min}} = \min(e_{L,\mathsf{min}} - x, e_{R,\mathsf{min}} + x)$, $e_{\mathsf{max}} = \max(e_{L,\mathsf{max}} - x, e_{R,\mathsf{max}} + x)$.
- *When no coefficient is selected in c_i,*
  $e_{\mathsf{min}} = \min(e_{L,\mathsf{min}}, e_{R,\mathsf{min}})$, $e_{\mathsf{max}} = \max(e_{L,\mathsf{max}}, e_{R,\mathsf{max}})$.

EXAMPLE 5.5: *Consider the T_3-synopsis $s_{3,2}$ in the coefficient node c_3* in Figure 3(a). *The T_3-synopsis $s_{3,2}$ is constructed from the pair of a T_6-synopsis $s_{6,1}$ and a T_7-synopsis $s_{7,1}$, whose $\texttt{CER}(T_6, s_{6,1}) = [0, 0]$ and $\texttt{CER}(T_7, s_{7,1}) = [2, 6]$ respectively, with the coefficient value -3 in c_3. By* Lemma 5.4, *the T_3-CER of the synopsis $s_{3,2}$ becomes* [-1, 3]. ∎

Properties of error functions: To discard useless synopses which cannot contribute to generate optimal synopses, we devise the following proposition, definition and lemma.

PROPOSITION 5.6.: *For a T_i-synopsis s whose T_i-CER is $[e_{\mathsf{min}}, e_{\mathsf{max}}]$, we have $F_{T_i,s}(v) \leq \epsilon$ for every v in $[e_{\mathsf{max}} - \epsilon, e_{\mathsf{min}} + \epsilon]$. In addition, if $e_{\mathsf{max}} - e_{\mathsf{min}} > 2\epsilon$, we get $F_{T_i,s}(v) > \epsilon$ for every v.*

By Proposition 5.6, if the T_i-CER's length of a T_i-synopsis s is larger than 2ϵ, we can safely prune the T_i-synopsis s.

DEFINITION 5.7.: *For T_i-synopses s and s' whose T_i-CERs are $[e_{\mathsf{min}}, e_{\mathsf{max}}]$ and $[e'_{\mathsf{min}}, e'_{\mathsf{max}}]$ respectively, if $e'_{\mathsf{min}} \leq e_{\mathsf{min}}$ and $e_{\mathsf{max}} \leq e'_{\mathsf{max}}$, we say the T_i-CER of s'* ***contains*** *that of s.*

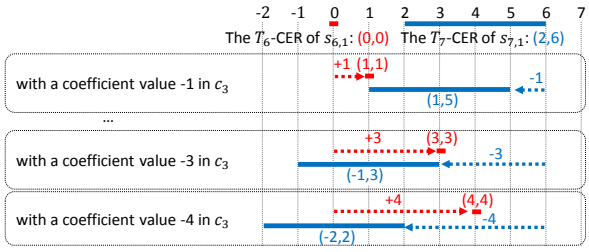


Figure 5: CERs of $s_{6,1}$ and $s_{7,1}$ shifted by coefficients

LEMMA 5.8.: *If the T_i-CER of a T_i-synopsis s is contained by that of another T_i-synopsis s', $F_{T_i,s}(v) \leq F_{T_i,s'}(v)$ for all incoming values v to c_i.*

Based on Lemma 5.8, for a pair of the same sized T_i-synopses s and s', if $\texttt{CER}(T_i, s')$ contains $\texttt{CER}(T_i, s)$, we can further prune s' even if the T_i-CER length of s' is at most 2ϵ.

Restricting the coefficient values: Consider the T_3-CERs of the T_3-synopses constructed from a T_6-synopsis $s_{6,1}$ and a T_7-synopsis $s_{7,1}$ with a coefficient value x in $[-\infty, \infty]$. As shown in Figure 5, when x is smaller than -3 (e.g., x=-4), the T_3-CER of the constructed T_3-synopsis always contains that of the constructed T_3-synopsis with x=-3. Similarly, the T_3-CER of the constructed T_3-synopsis with x>-1 (e.g., x=0) also always contains that of the T_3-synopsis with x=-1. Thus, by Lemma 5.8, we only consider the coefficient values in [-3, -1]. Based on the observation, we obtain the following.

LEMMA 5.9.: *Consider a T_{2i}-synopsis s_L and a T_{2i+1}-synopsis s_R with $\texttt{CER}(T_{2i}, s_L) = [e_{L,\min}, e_{L,\max}]$ and $\texttt{CER}(T_{2i+1}, s_R) = [e_{R,\min}, e_{R,\max}]$. When a T_i-synopsis s is constructed from s_L and s_R with coefficient value x in c_i, it is sufficient to consider x only in $[\min(a_{\min}, a_{\max}), \max(a_{\min}, a_{\max})]$ where $a_{\min} = (e_{L,\min} - e_{R,\min})/2$ and $a_{\max} = (e_{L,\max} - e_{R,\max})/2$.*

Similar to `MinHaarSpace` [18], we consider only the multiples of δ for coefficient values. By Lemma 5.9, we define the *candidate coefficient value set* to be used in Lemma 5.11 and to define a *T_i-constructed coef synopsis set* later.

DEFINITION 5.10.: *Consider a T_{2i}-synopsis s_L and a T_{2i+1}-synopsis s_R with $\texttt{CER}(T_{2i}, s_L) = [e_{L,\min}, e_{L,\max}]$ and $\texttt{CER}(T_{2i+1}, s_R) = [e_{R,\min}, e_{R,\max}]$. Let $a_{\min} = (e_{L,\min} - e_{R,\min})/2$ and $a_{\max} = (e_{L,\max} - e_{R,\max})/2$. The **candidate coefficient value set** $\texttt{C}_i(s_L, s_R)$ is $\{x_s + j \cdot \delta \mid 0 \leq j < m\}$ where $m = (x_e - x_s)/\delta + 1$, $x_s = \lceil \min(a_{\min}, a_{\max})/\delta \rceil \cdot \delta$ and $x_e = \lfloor \max(a_{\min}, a_{\max})/\delta \rfloor \cdot \delta$.*

For two ranges $[e_{\min}, e_{\max}]$ and $[e'_{\min}, e'_{\max}]$, if $e'_{\min} = e_{\min} + j \cdot \delta$ and $e'_{\max} = e_{\max} + j \cdot \delta$, we say $[e'_{\min}, e'_{\max}]$ is ***shifted*** by $j \cdot \delta$ from $[e_{\min}, e_{\max}]$. By restricting the coefficient values in $\texttt{C}_i(s_L, s_R)$, we obtain the following property of constructed synopses.

LEMMA 5.11.: *Consider a pair of T_i-synopses s and s' constructed from a T_{2i}-synopsis s_L and a T_{2i+1}-synopsis s_R with the coefficient values x and $x + j \cdot \delta$ in $\texttt{C}_i(s_L, s_R)$, respectively. Let $[e_{\min}, e_{\max}]$ and $[e'_{\min}, e'_{\max}]$ be the T_i-CERs of s and s', respectively. Then, if the length of $\texttt{CER}(T_{2i}, s_L)$ is at most (respectively, larger than) that of $\texttt{CER}(T_{2i+1}, s_R)$, $[e'_{\min}, e'_{\max}]$ is shifted by $j \cdot \delta$ (respectively, by -$j \cdot \delta$) from $[e_{\min}, e_{\max}]$.*

Since $F_{T_i,s}(v) = \max(|e_{\min} - v|, |e_{\max} - v|)$ by Proposition 5.2 where $[e_{\min}, e_{\max}] = \texttt{CER}(T_i, s)$, the error function of a synopsis s' is shifted from that of s by $j \cdot \delta$ (i.e., $F_{T_i,s'}(v) = F_{T_i,s}(v - j \cdot \delta)$), if the T_i-CER of s' is shifted from that of s by $j \cdot \delta$.

EXAMPLE 5.12.: *For a T_6-synopsis $s_{6,1}$ and a T_7-synopsis $s_{7,1}$ whose T_i-CERs are $[0, 0]$ and $[2, 6]$ respectively in* Figure 3(a), $\texttt{C}_3(s_{6,1}, s_{7,1}) = \{-3, -2, -1\}$ *with* $\delta = 1$. *By* Lemma 5.9, *the synopses $s_{3,2}$, $s_{3,3}$ and $s_{3,4}$ are constructed from the synopses $s_{6,1}$ and $s_{7,1}$ with values -3, -2 and -1, respectively. As* Lemma 5.11 *states, the T_3-CERs of $s_{3,3}$ and $s_{3,4}$ are shifted from that of $s_{3,2}$ by δ and 2δ respectively in* Figure 3(a). ∎

5.2 Properties of Extended Synopses

For a pair of a T_{2i}-synopsis s_L and a T_{2i+1}-synopsis s_R, the ***T_i-constructed coef synopsis set***, denoted by $S_i^C(s_L, s_R)$, is $\{s_1, \ldots, s_{|\texttt{C}_i(s_L, s_R)|}\}$ where every T_i-synopsis s_j is constructed from s_L and s_R with the j-th coefficient (respectively, $(|\texttt{C}_i(s_L, s_R)| + 1 - j)$-th coefficient) in $\texttt{C}_i(s_L, s_R)$, if the length of $\texttt{CER}(T_{2i}, s_L)$ is at most (respectively, larger than) that of $\texttt{CER}(T_{2i+1}, s_R)$. In addition, the ***$T_i$-constructed nocoef synopsis set***, denoted by $S_i^{NC}(s_L, s_R)$, is $\{s\}$ where s is constructed from s_L and s_R without selecting any coefficient in c_i. Note that the synopsis set $\{s_{3,2}, s_{3,3}, s_{3,4}\}$, which constructed from a pair of $s_{6,1}$ and $s_{7,1}$ in Example 5.12, is the T_i-constructed coef synopsis set of $s_{6,1}$ and $s_{7,1}$.

To denote a set of T_i-synopses whose T_i-CERs are shifted by δ from each other, we introduce an *extended synopsis*.

DEFINITION 5.13.: *A T_i-CER set $P = \{p_1, \ldots, p_m\}$ is called an **extended T_i-CER set** if every p_j is shifted from p_1 by $(j-1) \cdot \delta$. For a T_i-synopsis set S, its T_i-**CER set** is the set of the T_i-CERs of all synopses in S. A T_i-synopsis set is called a **T_i-extended synopsis** (or simply **extended synopsis**) if (1) its T_i-CER set forms an extended T_i-CER set and (2) the sizes of all its synopses are the same. For an extended synopsis $S = \{s_1, \ldots, s_m\}$, its **head** and **tail synopses** are s_1 and s_m, respectively. In addition, its **head** (respectively, **tail**) T_i-CER is the T_i-CER of s_1 (respectively, s_m).*

Since the T_i-CERs of the synopses in a T_i-extended synopsis S are shifted from each other by δ, we can obtain the T_i-CERs of all synopses in S from that of its head synopsis. Thus, instead of keeping the T_i-CER of every synopsis in S, we denote compactly an T_i-extended synopsis S by *its head T_i-CER*, *its synopsis size* and *its group size* only.

EXAMPLE 5.14.: *Consider $S_3^C(s_{6,1}, s_{7,1})$ of $s_{6,1}$ and $s_{7,1}$ in* Figure 3(a). *The group size of $S_3^C(s_{6,1}, s_{7,1})$ is $|\texttt{C}_3(s_{6,1}, s_{7,1})| = 3$ by* Definition 5.10 *and the head T_3-CER of $S_3^C(s_{6,1}, s_{7,1})$, is* [-1, 3] *as shown in* Example 5.5. ∎

We generalize the notions of T_i-constructed coef and nocoef synopsis sets for extended synopses as follows.

DEFINITION 5.15.: *For a pair of a T_{2i}-extended synopsis S_L and a T_{2i+1}-extended synopsis S_R, the **T_i-constructed coef synopsis set**, denoted by $S_i^C(S_L, S_R)$, is $S_i^C(S_L, S_R) = \bigcup_{s_L \in S_L, s_R \in S_R} S_i^C(s_L, s_R)$. In addition, the **$T_i$-constructed nocoef synopsis set**, denoted by $S_i^{NC}(S_L, S_R)$, is $S_i^{NC}(S_L, S_R) = \bigcup_{s_L \in S_L, s_R \in S_R} S_i^{NC}(s_L, s_R)$.*

5.3 Computing Required Synopsis Sets

Given a set S of T_i-synopses with the same size, to find every synopsis in S which is optimal for an incoming value to c_i, we define a *required synopsis set* of S below.

DEFINITION 5.16.: *For a T_i-synopsis set S whose synopsis sizes are the same, let P be the set of all **distinct** T_i-CERs of the synopses in S. The **required** T_i-CER **set** of S in T_i,*

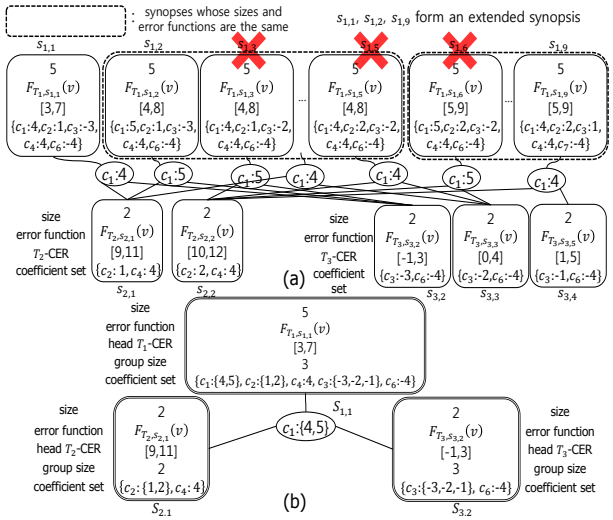


Figure 6: Synopses built from extended synopses

denoted by $P_{req}(T_i, S) = \{[e_{1,\min}, e_{1,\max}], \ldots, [e_{m,\min}, e_{m,\max}]\}$, *is the maximal subset of* P *such that* (1) $e_{j,\min} < e_{j+1,\min}$ *(with* $1 \le j < m$*),* (2) $e_{j,\max} - e_{j,\min} \le 2\epsilon$ *and* (3) *every* $[e_{j,\min}, e_{j,\max}]$ *does not contain any other* T_i-CER *in* P*. Let* $S_P(p_j, S)$ *be the set of every synopsis* $s \in S$ *such that its* T_i-CER *is a* p_j*. Then, a* ***required synopsis set*** *of* S *is* $\{s_1, \ldots, s_{|P_{req}(S)|}\}$ *where* $s_j \in S_P(p_j, S)$ *with the* j*-th* p_j *in* $P_{req}(T_i, S)$*.*

EXAMPLE 5.17**:** *For* T_1*-constructed coef synopsis set* $S_1^C(S_{2,1}, S_{3,2}) = \{s_{1,1}, \ldots, s_{1,9}\}$ *in* Figure 6(a)*, since* CER$(T_1, s_{1,1}) = [3, 7]$*,* CER$(T_1, s_{1,j}) = [4, 8]$ *with* $2 \le j \le 5$ *and* CER $(T_1, s_{1,j}) = [5, 9]$ *with* $6 \le j \le 9$*, the required* T_i-CER *set of* $S_1^C(S_{2,1}, S_{3,2})$ *is* $\{[3, 7], [4, 8], [5, 9]\}$ *with* $\epsilon = 3$*. Thus, a required synopsis set of* $S_1^C(S_{2,1}, S_{3,2})$ *has a single synopsis from* $\{s_{1,1}\}$*,* $\{s_{1,2}, \ldots, s_{1,5}\}$ *and* $\{s_{1,6}, \ldots, s_{1,9}\}$ *respectively.* ∎

We next show that a required synopsis set of a T_i-synopsis set S contains every optimal synopsis appearing in S.

LEMMA 5.18**:** *For a* T_i*-synopsis set* S *with the same synopsis size, if* S *has an optimal synopsis* s *for an incoming value* v *to* c_i*, every required synopsis set of* S *contains a* T_i*-synopsis* s' *such that* $F_{T_i,s'}(v) = F_{T_i,s}(v)$*.*

To compute the optimal T_i-synopses, we need only required synopsis sets of $S_i^C(S_L, S_R)$ and $S_i^{NC}(S_L, S_R)$ by Lemma 5.18. We will show that the required synopsis set also forms an extended synopsis, and present how to compute its compact representation in $O(1)$ time.

(a) A required synopsis set of $S_i^C(S_L, S_R)$**:** If we blindly union all T_i-constructed coef synopsis sets $S_i^C(s_L, s_R)$ for every pair of synopses $s_L \in S_L$ and $s_R \in S_R$, there exist many T_i-synopses whose T_i-CERs are the same (i.e., their error functions are the same too). However, we need to keep only a single synopsis among them, as illustrated in Example 5.17. Based on the observation, we obtain the following lemma.

LEMMA 5.19**:** *For a pair of a* T_{2i}*-extended synopsis* S_L *and a* T_{2i+1}*-extended synopsis* S_R*, every required synopsis set* S_{req} *of* $S_i^C(S_L, S_R)$ *is a* T_i*-extended synopsis such that the head (respectively, tail)* T_i-CER *of* S_{req} *is the same as that of* $S_i^C(s_L, s_R)$ *where* s_L *and* s_R *are* S_L *and* S_R*'s head (respectively, tail) synopses, respectively.*

EXAMPLE 5.20**:** *Consider* Example 5.17 *again. Every required synopsis set of* $S_1^C(S_{2,1}, S_{3,2})$ *is an extended synopsis whose head* T_1-CER *is* $[3, 7]$ *which is computed by* Lemma 5.4*. By* Lemma 5.19*, the head* T_1-CER *is computed from the* T_2-CER *of* $S_{2,1}$*'s head synopsis* $s_{2,1}$ *and the* T_3-CER *of* $S_{3,2}$*'s head synopsis* $s_{3,2}$ *with the smallest value* 4 *in* $\mathtt{C}_1(s_{2,1}, s_{3,2})$*.* Figure 6(a) *shows the* T_1-CERs *of the synopses in* $S_i^C(S_{2,1}, S_{3,2})$*. The set* $\{s_{1,1}, s_{1,2}, s_{1,9}\}$ *is a required synopsis set of* $S_1^C(S_{2,1}, S_{3,2})$ *computed by* Lemma 5.19 *and it forms an extended synopsis* $S_{1,1}$ *as illustrated in* Figure 6(b)*.* ∎

The smallest (and largest) coefficient value in $\mathtt{C}_i(s_L, s_R)$ can be calculated in $O(1)$ time by Definition 5.10. Moreover, the T_i-CER of the T_i-synopsis constructed from a pair of a T_{2i}-synopsis s_L and a T_{2i+1}-synopsis s_R with a coefficient value can be computed in $O(1)$ time by Lemma 5.4. Thus, the head and tail T_i-CERs of a required synopsis set of $S_i^C(S_L, S_R)$ in Lemma 5.19 can be obtained in $O(1)$ time.

We next consider the case of constructing a required synopsis set of $S_i^{NC}(S_L, S_R)$.

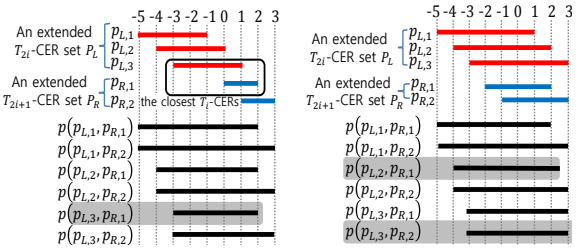
(b) A required synopsis set of $S_i^{NC}(S_L, S_R)$**:** For a pair of a T_{2i}-extended synopsis $S_L = \{s_{L,1}, \ldots, s_{L,n_1}\}$ and a T_{2i+1}-extended synopsis $S_R = \{s_{R,1}, \ldots, s_{R,n_2}\}$, let $P_L = \{p_{L,1}, \ldots, p_{L,n_1}\}$ be the T_{2i}-CER set of S_L and $P_R = \{p_{R,1}, \ldots, p_{R,n_2}\}$ be the T_{2i+1}-CER set of S_R as well as $p(p_{L,j}, p_{R,k})$ be the T_i-CER constructed from $p_{L,j}$ and $p_{R,k}$. We generate a required synopsis set of $S_i^{NC}(S_L, S_R)$ differently depending on whether there exists a containment relationship between every pair of $p_{L,j} \in P_L$ and $p_{R,k} \in P_R$ or not as follows.

Case (b-1): When there is no containment relationship between any pair of $p_{L,j} \in P_L$ and $p_{R,k} \in P_R$, the T_i-CER $p_{closest}$ constructed from the closest pair (i.e., the pair whose min values are the closest) of a T_{2i}-CER in P_L and a T_{2i+1}-CER in P_R is contained by those constructed from every other pair of $p_{L,j} \in P_L$ and $p_{R,k} \in P_R$. Thus, if the length of $p_{closest}$ is at most 2ϵ, the required T_i-CER set $P_{req}(T_i, S_i^{NC}(S_L, S_R))$ defined in Definition 5.16 contains a single T_i-CER $p_{closest}$ and is trivially an extended synopsis. For instance, in Figure 7(a), the closest pair in $P_L = \{p_{L,1}, p_{L,2}, p_{L,3}\}$ and $P_R = \{p_{R,1}, p_{R,2}\}$ is ($p_{L,3} = [-3, 1]$, $p_{R,1} = [0, 2]$). By Lemma 5.4, the T_i-CER $p(p_{L,3}, p_{R,1})$ is $[-3, 2]$ and is contained by the T_i-CER constructed from every other pair of $p_{L,j}$ and $p_{R,k}$. Thus, $P_{req}(T_i, S_i^{NC}(S_L, S_R))$ contains $p(p_{L,3}, p_{R,1})$ only.

Case (b-2): We next examine the case of when there is a containment relationship between pairs of $p_{L,j} \in P_L$ and $p_{R,k} \in P_R$. Without loss of generality, assume that $p_{L,j}$ contains $p_{R,k}$. It implies that the length of $p_{L,j}$ is at least that of $p_{R,k}$. By Lemma 5.4, $p(p_{L,j}, p_{R,k})$ is the same as $p_{L,j}$ and (1) $p_{L,j} = p(p_{L,j}, p_{R,k})$ is contained by $p(p_{L,j}, p_{R,k'})$ with $1 \le k' \le n_2$. Furthermore, (2) $p(p_{L,j}, p_{R,k})$ is contained by $p(p_{L,j'}, p_{R,k})$ for $p_{L,j'}$ which does not contain $p_{R,k}$. Based on (1) and (2), to compute the required T_i-CER set $P_{req}(T_i, S_i^{NC}(S_L, S_R))$, we only need to consider the set P'_L which is the set of every $p(p_{L,j}, p_{R,k})$ such that $p_{L,j} \in P_L$ contains $p_{R,k} \in P_R$.

Since $p(p_{L,j}, p_{R,k}) = p_{L,j}$, we have $P'_L \subseteq P_L$. Thus, there is no containment relationship between any pair of elements in P'_L. In other words, P'_L becomes the required T_i-CER set $P_{req}(T_i, S_i^{NC}(S_L, S_R))$. In addition, since P_L (respectively, P_R) is an extended T_{2i}-CER set (respectively, extended T_{2i+1}-CER set), when $p_{L,j}$ contains $p_{R,k}$, $p_{L,j+x}$ also contains $p_{R,k+x}$ with $1 \le j+x \le n_1$ and $1 \le k+x \le n_2$. Therefore, P'_L is an extended T_i-CER set.

From the cases of (b-1) and (b-2), we obtain the following.



(a) Without any containment relationship (b) With a containment relationship

Figure 7: A required synopsis set of $S_i^{NC}(S_L, S_R)$

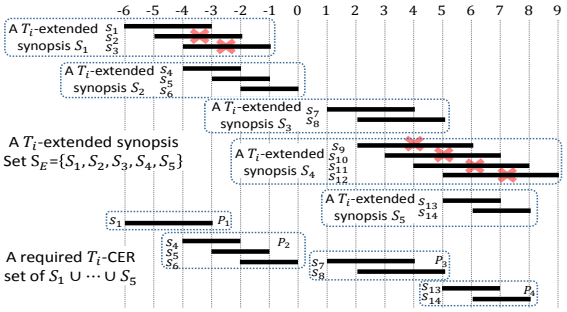


Figure 8: A required extended synopsis set

LEMMA 5.21.: *For a pair of a T_{2i}-extended synopsis S_L and a T_{2i+1}-extended synopsis S_R, every required synopsis set of the T_i-constructed nocoef synopsis set $S_i^{NC}(S_L, S_R)$ is a T_i-extended synopsis satisfying the following properties.*

(a) ***If there is no containment relationship between any pair of elements in S_L's T_{2i}-CER set and S_R's T_{2i+1}-CER set:*** *Let s_L and s_R be the pair of a T_{2i}-synopsis and a T_{2i+1}-synopsis, whose T_{2i}-CER and T_{2i+1}-CER are the closest, and s be the T_i-synopsis constructed from s_L and s_R with no coefficient in c_i. If the T_i-CER length of s is at most 2ϵ, every required synopsis set S_{req} of $S_i^{NC}(S_L, S_R)$ has a single synopsis s. Otherwise, $S_{req}=\emptyset$.*

(b) ***Otherwise:*** *Without loss of generality, assume that the T_{2i}-CER's length of a synopsis in S_L is at least that of S_R. Let m_1 (respectively, m_2) be the smallest (respectively, largest) integer such that the T_{2i}-CER of s_{m_1} (respectively, s_{m_2}) in $S_L=\{s_1, \ldots, s_m\}$ contains that of a synopsis in S_R. The head (respectively, tail) T_i-CER of every required synopsis set of $S_i^{NC}(S_L, S_R)$ is the same as the T_{2i}-CER of s_{m_1} (respectively, s_{m_2}).*

EXAMPLE 5.22.: *Consider a T_{2i}-extended synopsis $S_L=\{s_{L,1}, s_{L,2}, s_{L,3}\}$, whose head and tail T_{2i}-CERs are* [-5, 1] *and* [-3, 3]*, and a T_{2i+1}-extended synopsis $S_R=\{s_{R,1}, s_{R,2}\}$, whose head and tail T_{2i+1}-CERs are* [-2, 2] *and* [-1, 3]*, in* Figure 7(b)*. Since $p_{L,2}$=*[-4, 2] *has $p_{R,1}$=*[-2, 2] *and $p_{L,3}$=*[-3, 3] *has $p_{R,2}$=*[-1, 3]*, $P_{req}(T_i, S_i^{NC}(S_L, S_R))$ is* {[-4, 2], [-3, 3]} *which forms an extended T_i-CER set.* ∎

We can check in constant time if there is a containment relationship between a pair of $p_{L,j} \in P_L$ and $p_{R,k} \in P_R$. In addition, if there is a pair of $p_{L,j}$ and $p_{R,k}$ with a containment relationship, the head and tail T_i-CERs of a required synopsis set f $S_i^{NC}(S_L, S_R)$ can be obtained in $O(1)$ time. For instance, when the T_i-CER's length of S_L is larger than that of S_R, the head T_{2i}-CER $s_{L,H}$ of S_L is the head T_i-CER of S_{req} if $s_{L,H}$ contains the T_{2i+1}-CER of a synopsis in S_R. Otherwise, we can obtain the first T_{2i}-CER of S_L containing the T_{2i+1}-CER of S_R in $O(1)$ time.

6. COMPUTING AN OPTIMAL EXTENDED SYNOPSIS SET

In this section, given a pair of optimal extended synopsis sets in c_{2i} and c_{2i+1} respectively, we show how to generate an optimal extended synopsis set in a node c_i.

6.1 A Required Extended Synopsis Set

We define a *required extended synopsis set* in each node c_i which is the result of eliminating the synopses in every generated extended synopsis whose T_i-CERs contain that of a synopsis in another generated extended synopsis.

DEFINITION 6.1.: *A* ***$T_{(i,b)}$-extended synopsis set*** *is a set of T_i-extended synopses with the same synopsis size b. Let $S_{ALL}(\mathsf{S}_E)$ be the set of all synopses represented by each extended synopsis in a $T_{(i,b)}$-extended synopsis set S_E. Also, let $P_1, \ldots, P_m$ be a disjoint partition of the required T_i-CER set $P_{req}(T_i, S_{ALL}(\mathsf{S}_E))$ such that (1) all elements in P_j form an extended T_i-CER set contained by the T_i-CER set of an extended synopsis in S_E and (2) the minimum error of the first element in P_j is smaller than that in P_{j+1} with $1 \le j < m$. Then, let S_{P_j} be an extended synopsis in S_E whose T_i-CER set contains P_j and let S'_{P_j} be an extended synopsis consisting of every synopsis in S_{P_j} whose T_i-CER appears in P_j. Then, a* ***required extended synopsis set*** *of the $T_{(i,b)}$-extended synopsis set S_E is the set of every S'_{P_j} with $1 \le j \le m$.*

EXAMPLE 6.2.: *For an extended synopsis set $\mathsf{S}_E=\{S_1, \ldots, S_5\}$ with their extended T_i-CER sets in* Figure 8*, $S_{ALL}(\mathsf{S}_E)$ is $\{s_1, \ldots, s_{14}\}$.* Figure 8 *shows the required T_i-CER set $P_{req}(T_i, S_{ALL}(\mathsf{S}_E))$ by pruning the X-marked T_i-CERs, based on containment relationships. Then, $P_{req}(T_i, S_{ALL}(\mathsf{S}_E))$ is partitioned into four subsets $P_1, \ldots, P_4$ each of which forms an extended T_i-CER set, as in* Figure 8*. $\{\{s_1\}, \{s_4, s_5, s_6\}, \{s_7, s_8\}, \{s_{13}, s_{14}\}\}$ is a required extended synopsis set of S_E.* ∎

For a set S of synopses with the same size, we define the ***winning interval*** of a synopsis s in S as the incoming value interval such that s has the smallest error among all synopses in S. Note that the concept of winning intervals was previously investigated for a hierarchical structure, such as CHH or LH, in [14, 17]. Meanwhile, we compute the winning interval of each synopsis in the synopses constructed in each node of a coefficient tree. The following lemma allows us to calculate the winning interval of every synopsis in S. Note that v_i^H is defined in Section 3.2 and used in Lemma 3.4.

LEMMA 6.3.: *For a synopsis set S with the same-sized synopses, let $S_{req}=\{s_1, \ldots, s_m\}$ be a required synopsis set of S where $\mathsf{CER}(T_i, s_j)=[e_{j,\mathsf{min}}, e_{j,\mathsf{max}}]$. Also, let* (1) $\mathsf{mid}(s_0, s_1)=v_i^H-\epsilon$, (2) $\mathsf{mid}(s_j, s_{j+1})=(e_{j,\mathsf{min}}+e_{j+1,\mathsf{max}})/2$ *with $1 \le j < m$, and* (3) $\mathsf{mid}(s_m, s_{m+1}) = v_i^H+\epsilon$ *where $[v_i^H-\epsilon, v_i^H+\epsilon]$ is the candidate incoming value interval in c_i by* Lemma 3.4*. Then, the synopsis s_j has the minimum error for every value v in $[\mathsf{mid}(s_{j-1}, s_j), \mathsf{mid}(s_j, s_{j+1})]$ among all synopses in S.*

To generalize Lemma 6.3 to extended synopsis sets, partition a required synopsis set S_{req} into disjoint subsets $S_1, \ldots, S_m$, where S_j is an extended synopsis. Let $\mathsf{mid}(S_j, S_{j+1})=(e^T_{j,\mathsf{min}}+e^H_{j+1,\mathsf{max}})/2$ where $e^H_{j,\mathsf{max}}$ (respectively, $e^T_{j,\mathsf{min}}$) is the

maximum (respectively, minimum) error of S_j's head (respectively, tail) T_i-CER. In addition, let $\mathsf{mid}(S_0, S_1) = v_i^H - \epsilon$ and $\mathsf{mid}(S_m, S_{m+1}) = v_i^H + \epsilon$. Then, by Lemma 6.3, each extended synopsis S_j has a synopsis with the smallest error for every value v in $[\mathsf{mid}(S_{j\text{-}1}, S_j), \mathsf{mid}(S_j, S_{j+1})]$. Thus, we obtain the following corollary.

COROLLARY 6.4.: *For an extended synopsis set S_E with the same synopsis size, consider a required extended synopsis set $\mathsf{S}_{req} = \{S_1, \ldots, S_m\}$ of S_E. Then, each extended synopsis S_j in S_{req} has a synopsis with the smallest error for every value v in $[\mathsf{mid}(S_{j\text{-}1}, S_j), \mathsf{mid}(S_j, S_{j+1})]$.*

By Corollary 6.4, for a required extended synopsis set $\mathsf{S}_{req} = \{S_1, \ldots, S_m\}$, the ***winning interval*** of the extended synopsis S_j with respect to S_{req}, denoted by $I_W(S_j, \mathsf{S}_{req})$, is $[\mathsf{mid}(S_{j\text{-}1}, S_j), \mathsf{mid}(S_j, S_{j+1})]$.

EXAMPLE 6.5.: *Consider S_E's required extended synopsis set $\mathsf{S}_{req} = \{S'_1, S'_2, S'_3, S'_4\}$ where $S'_1 = \{s_1\}$, $S'_2 = \{s_4, s_5, s_6\}$, $S'_3 = \{s_7, s_8\}$ and $S'_4 = \{s_{13}, s_{14}\}$ in* Example 6.2. *As shown in* Figure 8, $\mathsf{mid}(S'_1, S'_2) = (e^T_{1,min} + e^H_{2,max})/2 = -4$ *and* $\mathsf{mid}(S'_2, S'_3) = (e^T_{2,min} + e^H_{3,max})/2 = 1$. *Thus,* $I_W(S'_2, \mathsf{S}_{req}) = [-4, 1]$. ∎

6.2 An Optimal Extended Synopsis Set

If there exists an extended synopsis with the synopsis size smaller than b such that its winning interval contains an incoming value v to c_i, an extended synopsis with the synopsis size b cannot have an optimal synopsis for v to c_i by Definition 3.3. Thus, we eliminate such non-optimal extended synopses for computing a T_i-extended synopsis set such that there exists an optimal T_i-synopsis for every incoming value to c_i. We define a *T_i-optimal extended synopsis set* constructed from a pair of a T_{2i}-optimal extended synopsis set and a T_{2i+1}-optimal extended synopsis set as follows.

DEFINITION 6.6.: *Consider a pair of a T_{2i}-extended synopsis set S_{EL} and a T_{2i+1}-extended synopsis set S_{ER}.*

- *Let U_i^C (respectively, U_i^{NC}) be the union of the T_i-constructed coef (respectively, nocoef) synopsis set $S_i^C(S_L, S_R)$ (respectively, $S_i^{NC}(S_L, S_R)$) for all pairs of S_L in S_{EL} and S_R in S_{ER}. The **$T_{(i,b)}$-extended synopsis set** constructed from S_{EL} and S_{ER}, denoted by $\mathsf{U}_{(i,b)}(S_{EL}, S_{ER})$, is the set of every extended synopsis with the synopsis size b in $\mathsf{U}_i^C \cup \mathsf{U}_i^{NC}$.*
- *Let $S_{(i,b)} = \{S_{(i,b),1}, \ldots, S_{(i,b),m}\}$ be a required extended synopsis set of $\mathsf{U}_{(i,b)}(S_{EL}, S_{ER})$. A **$T_{(i,b)}$-strictly optimal extended synopsis set** $\mathsf{O}_{(i,b)}$ constructed from S_{EL} and S_{ER} is a set of every $S_{(i,b),j}$ whose winning interval $I_W(S_{(i,b),j}, S_{(i,b)})$ contains a value v that is not contained by the winning interval of any extended synopsis in $S_{(i,0)} \cup S_{(i,1)} \cup \ldots \cup S_{(i,b\text{-}1)}$.*
- *A **T_i-optimal extended synopsis set** constructed from S_{EL} and S_{ER} is the union of the $T_{(i,b)}$-strictly optimal extended synopsis sets constructed from S_{EL} and S_{ER} with $0 \le b \le |\mathsf{coeff}(T_i)|$.*

To find an optimal synopsis efficiently, we compute a T_i-optimal extended synopsis set for every coefficient node c_i by invoking OptExt-EB (in Section 4.2) which calls the procedures presented below. Let N_b be the number of pairs of extended synopses $\langle S_L, S_R \rangle$ such that $S_L \in \mathsf{O}_{(2i,p)}$, $S_R \in \mathsf{O}_{(2i+1,q)}$ and $p+q=b$ (i.e., $N_b = \sum_{p=0}^{b} |\mathsf{O}_{(2i,p)}| \cdot |\mathsf{O}_{(2i+1,b-p)}|$).

The procedure ExtSynNoCoef: When no coefficient is chosen in c_i, MinHaarSpace considers the pairs of $E_{2i}[v]$ and $E_{2i+1}[v]$ to fill $E_i[v]$. Similarly, we return $D^{\emptyset}_{(i,b)}$, which is the set of the extended synopses with synopsis size b generated from every pair of extended synopses in O_{2i} and O_{2i+1} respectively without a coefficient in c_i. It is produced by unioning $S_i^{NC}(S_L, S_R)$ for every pair of $S_L \in \mathsf{O}_{(2i,p)}$ and $S_R \in \mathsf{O}_{(2i+1,q)}$ such that $I_W(S_L, \mathsf{O}_{(2i,p)})$ and $I_W(S_R, \mathsf{O}_{(2i+1,q)})$ overlap and $p+q=b$. Thus, ExtSynNoCoef takes $O(N_b)$ time.

The procedure ExtSynCoef: For every pair of extended synopses $S_L \in \mathsf{O}_{(2i,p)}$ and $S_R \in \mathsf{O}_{(2i+1,q)}$ with $p+q=b-1$, there always exist an incoming value v and a coefficient value x in c_i such that $v+x \in I_W(S_{EL}, \mathsf{O}_{(2i,p)})$ and $v-x \in I_W(S_{ER}, \mathsf{O}_{(2i+1,q)})$. Thus, we return $D^c_{(i,b)}$, which is the set of the extended synopses with synopsis size b generated from every pair of extended synopses in O_{2i} and O_{2i+1} respectively with a coefficient in c_i (i.e., $\bigcup_{S_L \in \mathsf{O}_{(2i,p)}, S_R \in \mathsf{O}_{(2i+1,b-p-1)}} S_i^C(S_L, S_R)$). This procedure requires $O(N_{b-1})$ time.

The procedure RequiredSet: It takes a T_i-extended synopsis set $D^c_{(i,b)} \cup D^{\emptyset}_{(i,b)}$ and returns its required extended synopsis set. We first obtain $\mathsf{S}_E = \{S_1, \ldots, S_m\}$ from $D^c_{(i,b)} \cup D^{\emptyset}_{(i,b)}$ by sorting the elements in $D^c_{(i,b)} \cup D^{\emptyset}_{(i,b)}$ in increasing order of the minimum errors of their head T_i-CERs. For each extended synopsis $S_j \in \mathsf{S}_E$, we next examine every extended synopsis $S_k \in \mathsf{S}_E$ with $j < k$ until the maximum error of S_j's tail T_i-CER is at least the minimum error of S_k's head T_i-CER. For a pair $\langle S_j, S_k \rangle$, if there is any synopsis $s \in S_j$ (respectively, $s' \in S_k$) whose T_i-CER contains that of $s' \in S_k$ (respectively, $s \in S_j$), we remove every such synopsis s (respectively, s') from S_j (respectively, S_k) in $O(1)$ time. Since $|D^c_{(i,b)} \cup D^{\emptyset}_{(i,b)}| = N_{b-1} + N_b$, for each S_j, there may exist $(N_{b-1} N_b)$ number of S_ks. Thus, the time complexity of RequiredSet is $O((N_{b-1} + N_b)^2)$ in the worst case. However, for each S_j, because S_E is sorted by the minimum errors of S_k's head T_i-CERs, the number of S_ks to be checked is much smaller than $(N_{b-1} + N_b)$ in practice. Thus, RequiredSet takes $O((N_{b-1} + N_b) \cdot \log(N_{b-1} + N_b))$ time due to sorting.

The procedure StrictOptSet: From the result $D_{(i,b)}$ of RequiredSet as well as $U_{b-1} = \{I_W(S, \mathsf{O}_{(i,b')}) | S \in \mathsf{O}_{(i,b')}$ and $0 \le b' \le b-1\}$ which is the union of the winning intervals of all T_i-extended synopses contained in $\mathsf{O}_{(i,0)} \cup \ldots \cup \mathsf{O}_{(i,b\text{-}1)}$, StrictOptSet computes a $T_{(i,b)}$-strictly optimal extended synopsis set $\mathsf{O}_{(i,b)}$. Every T_i-extended synopsis S in $D_{(i,b)}$, whose $I_W(S, D_{(i,b)})$ contains an incoming value v that is not contained by U_{b-1}, is inserted into $\mathsf{O}_{(i,b)}$. Finally, StrictOptSet returns $\mathsf{O}_{(i,b)}$ and $U_b = U_{b-1} \cup \{I_W(S, D_{(i,b)}) | S \in \mathsf{O}_{(i,b)}\}$. Since $D_{(i,b)}$ is a required synopsis set of $D^c_{(i,b)} \cup D^{\emptyset}_{(i,b)}$, we have $|D_{(i,b)}| \le |D^c_{(i,b)} \cup D^{\emptyset}_{(i,b)}| = (N_{b-1} + N_b)$. Thus, the time complexity of StrictOptSet is $O(N_{b-1} + N_b)$.

The procedure OptExtRoot: It computes a T_0-optimal extended synopsis set O_0 from a T_1-optimal extended synopsis set O_1. Let B_i be the smallest synopsis size of all extended synopses in a T_i-optimal extended synopsis set O_i. If there is an extended synopsis S with $zero \in I_W(S, \mathsf{O}_{(1,B_1)})$ in a $T_{(1,B_1)}$-strictly optimal extended synopsis set $\mathsf{O}_{(1,B_1)}$, there is an optimal synopsis $s_\emptyset$ with $zero \in I_W(s_\emptyset, \mathsf{O}_{(1,B_1)})$ in S. Thus, a $T_{(0,B_0)}$-strictly optimal extended synopsis set $\mathsf{O}_{(0,B_0)}$ becomes $\{s_\emptyset\}$. Otherwise (i.e., there is no such an extended synopsis in $\mathsf{O}_{(1,B_1)}$), there is no optimal synopsis

with the size B_1 and we need to find an optimal synopsis with the size B_1+1 by selecting a coefficient in the root node. Thus, we first select an extended synopsis S in $\mathsf{O}_{(1,B_1)}$ with the shortest T_1-CER length. We next set $\mathsf{O}_{(0,B_0)}=\{s_{c_0}\}$ where s_{c_0} is the synopsis constructed from S's head synopsis with a root coefficient. `OptExtRoot` takes $O(|\mathsf{O}_1|)$ time.

Time complexity of `OptExt-EB`: To obtain a T_i-optimal extended synopsis set, we need to calculate every $T_{(i,b)}$-strictly optimal extended synopsis set. Furthermore, since `RequiredSet` is the slowest among all procedures used in the **for** loop of lines 5-9, it takes $O((N_{b-1}+N_b)\cdot\log(N_{b-1}+N_b))$ time for an iteration of the **for** loop. Thus, the **for** loop takes $O(\sum_{b=(m_{2i}+m_{2i+1})}^{(M_{2i}+M_{2i+1})+1}(N_{b-1}+N_b)\cdot\log(N_{b-1}+N_b))$ time. Note that $\sum_b N_b=|\mathsf{O}_{2i}|\cdot|\mathsf{O}_{2i+1}|$ because every pair of extended synopses in O_{2i} and O_{2i+1} respectively is examined exactly once by each `ExtSynCoef` and `ExtSynNoCoef` during the course of `OptExt-EB`. Thus, `OptExt-EB` takes $O(|\mathsf{O}_{2i}|\cdot|\mathsf{O}_{2i+1}|\cdot\log(|\mathsf{O}_{2i}|\cdot|\mathsf{O}_{2i+1}|)$ time to compute a T_i-optimal extended synopsis set, and the overall time complexity of our `OptExt-EB` is $O(N\cdot|\mathsf{O}_{2i}|\cdot|\mathsf{O}_{2i+1}|\cdot\log(|\mathsf{O}_{2i}|\cdot|\mathsf{O}_{2i+1}|))$ where N is the number of nodes in the coefficient tree T.

Let m be the maximum number of extended synopses in every O_i. The running time of `OptExt-EB` is $O(N\cdot m^2\cdot\log m)$. Because $m\ll\lceil 2\epsilon/\delta\rceil$ in practice (See the experiments in Section 8.1), where $\lceil 2\epsilon/\delta\rceil$ is the size of the candidate incoming value set $\mathtt{IV}(i)$, the running time of `OptExt-EB` is much smaller than that of `MinHaarSpace` (i.e., $O(N(\epsilon/\delta)^2)$).

7. EXTENDING TO HAAR$^+$ SYNOPSES

In this section, we describe how to apply our techniques to a Haar$^+$ tree. The root node of a Haar$^+$ tree has a single coefficient whose value contributes positively to all data values. The internal nodes of a Haar$^+$ tree correspond to *triads* which consist of *head*, *left supplementary* and *right supplementary* coefficients. A head coefficient in a triad C_i acts in exactly the same way as an unrestricted Haar wavelet coefficient. Thus, our techniques can be applied directly for head coefficients in C_i. A left (respectively, right) supplementary coefficient contributes positively to the data values in its left (respectively, right) subtree only. We only provide the details for left supplementary coefficients since the details for right supplementary coefficients are symmetric.

To apply our techniques to the case of when a left supplementary coefficient is selected, the following lemma is used to restrict the left coefficient values in c_i and to calculate the T_i-CER of a T_i-synopsis from a T_{2i}-CER and a T_{2i+1}-CER.

LEMMA 7.1.: *Consider a T_{2i}-synopsis s_L and a T_{2i+1}-synopsis s_R with $\mathtt{CER}(T_{2i},s_L)=[e_{L,\mathsf{min}},e_{L,\mathsf{max}}]$ and $\mathtt{CER}(T_{2i+1},s_R)=[e_{R,\mathsf{min}},e_{R,\mathsf{max}}]$. To construct a T_i-synopsis s from s_L and s_R with a left supplementary coefficient value x in c_i, it is sufficient to consider x in $[\min(a^{\mathcal{L}}_{\mathsf{min}},a^{\mathcal{L}}_{\mathsf{max}}),\max(a^{\mathcal{L}}_{\mathsf{min}},a^{\mathcal{L}}_{\mathsf{max}})]$ where $a^{\mathcal{L}}_{\mathsf{min}}=e_{L,\mathsf{min}}-e_{R,\mathsf{min}}$ and $a^{\mathcal{L}}_{\mathsf{max}}=e_{L,\mathsf{max}}-e_{R,\mathsf{max}}$. Furthermore, for a left supplementary coefficient value x in c_i, the T_i-CER of s is $[\min(e_{L,\mathsf{min}}-x,e_{R,\mathsf{min}}),\max(e_{L,\mathsf{max}}-x,e_{R,\mathsf{max}})]$.*

Since we consider only the multiples of δ for coefficient values, we define the *candidate left supplementary coefficient value set* and *T_i-constructed left coef synopsis set*.

DEFINITION 7.2.: *Consider a T_{2i}-synopsis s_L and a T_{2i+1}-synopsis s_R with $\mathtt{CER}(T_{2i},s_L)=[e_{L,\mathsf{min}},e_{L,\mathsf{max}}]$ and $\mathtt{CER}(T_{2i+1},s_R)=[e_{R,\mathsf{min}},e_{R,\mathsf{max}}]$. Let $a_{\mathsf{min}}=e_{L,\mathsf{min}}-e_{R,\mathsf{min}}$ and $a_{\mathsf{max}}=e_{L,\mathsf{max}}-e_{R,\mathsf{max}}$. The **candidate left supplementary coefficient value set** $\mathtt{C}^{\mathcal{L}}_i(s_L,s_R)$ is $\{x_s+j\cdot\delta \mid 0\le j<m\}$ with $m=(x_e-x_s)/\delta+1$ where $x_s=\lceil\min(a_{\mathsf{min}},a_{\mathsf{max}})/\delta\rceil\cdot\delta$ and $x_e=\lfloor\max(a_{\mathsf{min}},a_{\mathsf{max}})/\delta\rfloor\cdot\delta$. Furthermore, the **$T_i$-constructed left coef synopsis set**, denoted by $S_i^{\mathcal{LC}}(s_L,s_R)$, is $\{s_1,\ldots,s_{|\mathtt{C}^{\mathcal{L}}_i(s_L,s_R)|}\}$ where each T_i-synopsis s_j is constructed from s_L and s_R with the j-th coefficient in $\mathtt{C}^{\mathcal{L}}_i(s_L,s_R)$. In addition, for a pair of a T_{2i}-extended synopsis S_L and a T_{2i+1}-extended synopsis S_R, the **T_i-constructed left coef synopsis set** is $\bigcup_{s_L\in S_L,s_R\in S_R}S_i^{\mathcal{LC}}(s_L,s_R)$ which is denoted by $S_i^{\mathcal{LC}}(S_L,S_R)$.*

LEMMA 7.3.: *For a T_{2i}-extended synopsis S_L and a T_{2i+1}-extended synopsis S_R, every required synopsis set of a T_i-constructed left coef synopsis set $S_i^{\mathcal{LC}}(S_L,S_R)$ is a T_i-extended synopsis such that the head (respectively, tail) T_i-CERs of S is the same as that of $S_i^{\mathcal{LC}}(s_L,s_R)$ where s_L and s_R are S_L and S_R's head (respectively, tail) synopses, respectively.*

Similarly, we can define the **T_i-constructed right coef synopsis set**, denoted by $S_i^{\mathcal{RC}}(S_L,S_R)$. By Lemma 7.3, we can obtain the required synopsis sets of $S_i^{\mathcal{LC}}(S_L,S_R)$ and $S_i^{\mathcal{RC}}(S_L,S_R)$ in $O(1)$ time. By utilizing the techniques presented in Sections 5 and 6, it is straightforward to obtain `OptExtHP-EB` extended from `OptExt-EB` for Haar$^+$ synopses. Since `OptExtHP-EB` also exploits extended synopses, the procedures `RequiredSet` and `StrictOptSet` of `OptExt-EB` are also utilized in `OptExtHP-EB`. However, to utilize the procedure `ExtSynCoef` in `OptExt-EB` for `OptExtHP-EB`, we need to compute $T_{(i,b)}$-extended synopsis set constructed from an optimal T_{2i}-extended synopsis set and an optimal T_{2i+1}-extended synopsis set based on $S_i^{\mathcal{LC}}(S_L,S_R)$ and $S_i^{\mathcal{RC}}(S_L,S_R)$. The details on how to extend our techniques to Haar$^+$ synopses can be found in our technical report [19].

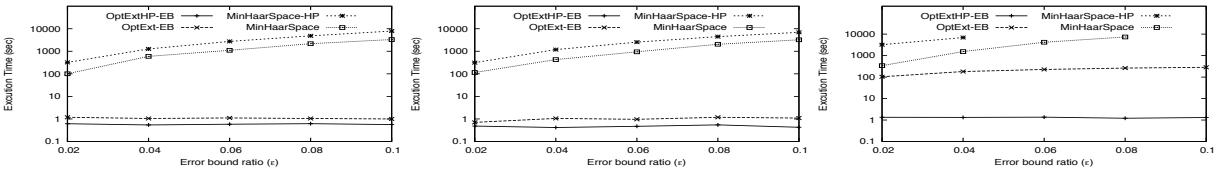
Table 2: Implemented algorithms

Algorithms	Description
	Error-bounded synopsis problem
`OptExt-EB`	Our unrestricted wavelet synopsis algorithm
`OptExtHP-EB`	Our algorithm for Haar$^+$ synopses
`MinHaarSpace`	The state-of-the-art [18] for unrestricted wavelets
`MinHaarSpace-HP`	The state-of-the-art [18] for Haar$^+$ synopses
	Space-bounded synopsis problem
`IndirectExt`	The indirect algorithm based on `OptExt-EB`
`IndirectExt-HP`	The indirect algorithm utilizing `OptExtHP-EB`
`IndirectHaar`	The indirect algorithm [18] using `MinHaarSpace`
`IndirectHaar-HP`	The indirect algorithm [18] with `MinHaarSpace-HP`

8. EXPERIMENTS

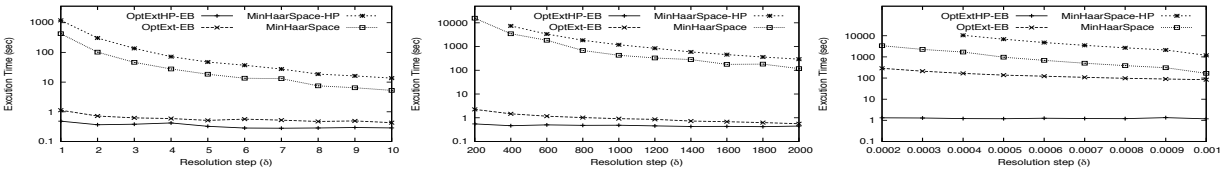
The tested algorithms are shown in Table 2. We use Javac 1.7 compiler for our implementations. Although we got the C++ source code of `MinHaarSpace-HP` used in [18], we reimplemented `MinHaarSpace-HP`. We conducted the experiments on the machine with Intel i3 3.3 GHz CPU and 8GB RAM running Linux. We compare the performance of our `OptExt-EB` (respectively, `OptExtHP-EB`) with `MinHaarSpace` (respectively, `MinHaarSpace-HP`) for unrestricted wavelet (respectively, Haar$^+$) synopses. In addition, we demonstrate the superiority of our `IndirectExt` and `IndirectExt-HP` for the space-bounded synopsis problem compared with the existing algorithms `IndirectHaar` and `IndirectHaar-HP` [18]. We ran our proposed algorithms ten times and report the average execution times. Some algorithms which do not terminate within 3 hours are not plotted. The parameters used in our experiments are summarized in Table 3.

Data sets: We used both synthetic and real-life datasets. The synthetic dataset was generated with a Zipf distribution



(a) Synthetic dataset (b) WC dataset (c) TM dataset

Figure 9: Varying the error bound ratio ε for the error-bounded synopsis problem



(a) Synthetic dataset (b) WC dataset (c) TM dataset

Figure 10: Varying the resolution step δ for the error-bounded synopsis problem

in the domain $[0, M_0]$ varying the data size n and the skewness α. The real-life datasets are WC [1] and TM datasets (http://kdd.ics.uci.edu). WC dataset contains the request frequency of 98 World Cup Web sitefrom Apr. 30 to Jul. 26 in 1998. TM dataset has the sea surface temperatures measured from the equatorial Pacific. WC and TM datasets include 89,997 and 161,072 values whose averages are 9,714 and 26.75 as well as standard deviations are 237,733 and 1.91 with the domains $[0, 16777216]$ and $[17.35, 31.26]$, respectively. WC dataset is far more skewed than TM dataset.

Recall that all algorithms consider the multiples of δ for the coefficient values. Thus, we varied δ from 1 to 10 for synthetic datasets, 200 to 2000 for WC dataset and 0.0002 to 0.001 for TM dataset. The default value of δ for the synthetic, WC and TM datasets are 1, 1000 and 0.0005, respectively. Since datasets have different data value domains, for the error-bounded synopsis algorithms, we used an error bound ratio $\varepsilon = \epsilon/(M_0 - m_0)$ instead of an error bound ϵ. The error bound ratio ε is varied from 0.02 to 0.1. For the space-bounded synopsis problem, we used the space bound ratio $\beta = B/n$ by a similar reason where B is a space bound and varied the space bound ratio β from 0.01 to 0.05.

8.1 The Error-bounded Synopsis Problem

Varying the error bound ratio ε: Figure 9 shows the experiment results varying ε on the synthetic, WC and TM datasets. For each node c_i, `MinHaarSpace` computes $E_i[v]$ and $F_i[v]$ for every incoming value v in $[v_i^H - \epsilon, v_i^H + \epsilon]$ as mentioned in Section 3.2 and the range grows with increasing an error bound ϵ. Thus, the execution times of `MinHaarSpace` increase over all datasets with incrementing an error bound ratio ε. In contrast, the performance of `OptExt-EB` is less affected by changing an error bound ratio ε since `OptExt-EB` keeps a T_i-optimal extended synopsis set in each node c_i and represents it as a set of T_i-extended synopses by their compact representations (i.e., the head T_i-`CER` and the group size). Furthermore, the compact representation of each T_i-extended synopsis in a T_i-optimal extended synopsis set is computed from those in a T_{2i}-optimal extended synopsis set and T_{2i+1}-optimal extended synopsis set in $O(1)$ time. Thus, `OptExt-EB` is superior to `MinHaarSpace` over all datasets with varying ε. Similarly, in the Haar$^+$ case, `OptExtHP-EB` is much faster than `MinHaarSpace-HP`.

As shown in Figure 9, `MinHaarSpace-HP` is slower than `MinHaarSpace`. The reason is that the time complexity of `MinHaarSpace` (i.e., $O(N(\epsilon/\delta)^2)$) is smaller than that of `MinHaarSpace-HP` (i.e., $O(N((\Delta+\epsilon)/\delta)^2)$) where $\Delta = M_0 - m_0$. To see how many entries are stored in each level of the tree T_0, we show the average numbers of entries $E_i[v]$s kept in a node of the top-3 levels ℓ of the tree by `MinHaarSpace-HP` and `MinHaarSpace` for WC dataset in Table 4.

Meanwhile, `OptExtHP-EB` is faster than `OptExt-EB`. Since each node of a Haar$^+$ tree has more opportunity to select a left or right supplementary coefficient, `OptExtHP-EB` is likely to generate an extended synopsis covering a wider winning interval. Thus, the number of extended synopses stored by `OptExtHP-EB` tends to be smaller than that by `OptExt-EB`. We report the average numbers of extended synopses stored in a node of the top-3 levels ℓ of the trees by both algorithms for WC dataset in Table 4. Note that the numbers of entries stored by the traditional algorithms are much larger than those by our algorithms. In addition, for the default value of ϵ on WC dataset, we found that the sizes of the optimal synopses generated by `OptExtHP-EB` and `MinHaarSpace-HP` are 126, while those of `OptExt-EB` and `MinHaarSpace` are 229. Thus, `OptExtHP-EB` is the best performer since it is the fastest and finds the smallest synopsis.

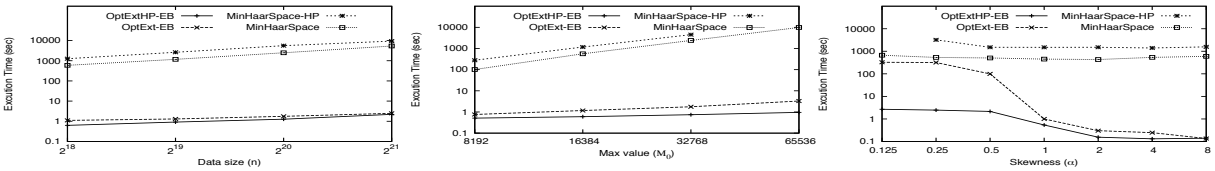
We also compare the memory usages of the implemented algorithms on WC dataset. The memory usages of `OptExtHp-EB`, `OptExt-EB`, `MinHaarSpace-HP` and `MinHaarSpace` are 9, 52, 552 and 76 KBs, respectively. Note that the memory usages of our algorithms are smaller than those of the traditional algorithms. The reason is that the numbers of en-

Table 3: Parameters for synthetic datasets

Parameter	Range	Default
Resolution step (δ)	$1, 2, 3, \ldots, 9, 10$	1
Data size (n)	$2^{18} \sim 2^{21}$	2^{18}
Domain of data values ($[m_0, M_0]$)	$[0, 8192 \sim 65536]$	[0,16384]
Skewness (α)	$0.125 \sim 8$	1

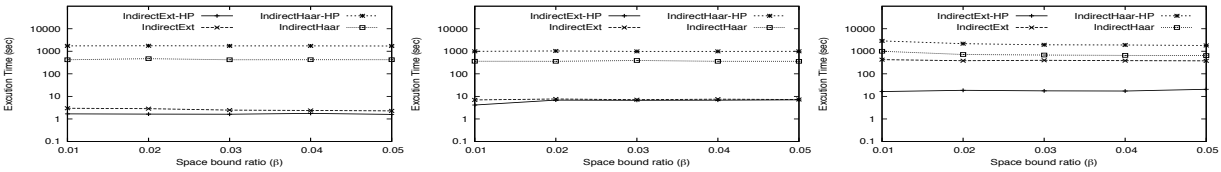
Table 4: Statistics for WC dataset

ℓ	# for δ=1000 (# for δ=500)			
	# of extended synopses		# of computed entries $E_i[v]$s	
	`OptExtHP-EB`	`OptExt-EB`	`MinHaarSpace-HP`	`MinHaarSpace`
0	18 (18)	160 (225)	18120 (36239)	1341 (2683)
1	10.5 (10.5)	111 (156.5)	9731.5 (19462)	1341 (2683)
2	7.8 (8.3)	76 (107)	5707.8 (11414.5)	1341.5 (2683.3)



(a) Varying n (b) Varying M_0 (c) Varying α

Figure 11: Varying n, M_0 and α on synthetic datasets for the error-bounded synopsis problem



(a) Synthetic dataset (b) WC dataset (c) TM dataset

Figure 12: Varying the space bound ratio β for the space-bounded synopsis problem

tries stored by our algorithms are much smaller than those by the traditional algorithms as shown in Table 4, even though a single T_i-extended synopsis used in `OptExt-EB` or `OptExtHP-EB` requires more memory than an entry (i.e., $E_i[v]$ or $F_i[v]$) used in `MinHaarSpace` or `MinHaarSpace-HP`.

Varying the resolution step δ: We report the experiment result with varying δ in Figure 10. `MinHaarSpace` and `MinHaarSpace-HP` become slower with decreasing δ since the numbers of incoming values and coefficient values to be considered increase in each node. When δ becomes half, the numbers of stored entries $E_i[v]$ in each node by both algorithms [18] are doubled. (See Table 4.) Meanwhile, in `OptExt-EB` and `OptExtHP-EB`, although the number of synopses in a extended synopsis increases with decreasing δ, we represent it by a single representation regardless of δ. Thus, the numbers of extended synopses stored in each level by our proposed algorithms change much less than the traditional algorithms, as shown in Table 4, and `OptExt-EB` as well as `OptExtHP-EB` tend to be slightly slower with decreasing δ.

Varying n, M_0 and α: We experimented with the synthetic datasets generated by varying the size n, maximum data value M_0 and skewness α of the Zipfian distribution. The execution times of the algorithms varying n, M_0 and α are plotted in Figure 11(a), (b) and (c), respectively. Similar to the experiments with varying ε or δ, our `OptExt-EB` (respectively, `OptExtHP-EB`) performs better than `MinHaarSpace` (respectively, `MinHaarSpace-HP`) with varying n, as shown in Figure 11(a), due to the use of the compact representations of extended synopses. Furthermore, as shown in Figure 11(b), increasing M_0 has a similar effect of decreasing δ since the number of incoming values to be considered for each node also becomes larger with growing M_0.

We reported the result with varying the skewness α of the Zipfian distribution in Figure 11(c). The higher α, the more skewed the data and the number of values belonging to the long tail grows. As values in the long tail are less frequent, the more such values there are and the less burdensome it is to represent them. Since we can represent the values in the long tail effectively with less coefficients, all algorithms become faster with higher α. Moreover, as the skewness α decreases, the gap of execution times between `OptExt-EB` and `OptExtHP-EB` widens. In short, `OptExtHP-EB` outperforms all the other algorithms regardless of the skewness.

8.2 The Space-bounded Synopsis Problem

To see the performance of the indirect algorithms, we varied the space bound ratio β and report the execution times in Figure 12. Since `IndirectHaar` (respectively, `IndirectHaar-HP`) invokes `MinHaarSpace` (respectively, `MinHaarSpace-HP`) repeatedly, they did not finish within three hours with the default δ values. Thus, we use 10, 5000 and 0.005 as the values of δ for the synthetic, WC and TM datasets, respectively. Note that the direct algorithm computing an error-optimal Haar$^+$ synopsis is faster than that calculating an error-optimal unrestricted Haar wavelet synopsis [16]. But, for the indirect algorithms, `IndirectHaar` is faster than `IndirectHaar-HP`, as shown in Figure 12. As we expected, `IndirectExt-HP` is the best performer for the space-bounded synopsis problem among all indirect algorithms.

9. CONCLUSION

We proposed the dynamic programming algorithms for the error-bounded synopsis problem based on unrestricted wavelet synopses and Haar$^+$ synopses. Our algorithms store a distinct set of optimal synopses and restrict the coefficient values to consider in each node. By partitioning all optimal synopses in each node into a set of extended synopses and representing each synopsis by its compact representation, we improve the performance of our algorithms significantly. By experiments on synthetic and real-life datasets, we demonstrate that our `OptExtHP-EB` (respectively, `IndirectExt-HP`) is the best performer for the error-bounded synopsis (respectively, the space-bounded synopsis) problem. We plan to study how to generalize our algorithms to handle the weighted maximum error as future work.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2016R1D1A1A02937186). It was also supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-2013-0-00881) supervised by the IITP(Institute for Information & communications Technology Promotion) and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2015R1D1A1A01058909).

10. REFERENCES

[1] M. Arlitt and T. Jin. A workload characterization study of the 1998 world cup web site. *IEEE Network*, 14(3):30–37, 2000.

[2] K. Chakrabarti, E. Keogh, S. Mehrotra, and M. Pazzani. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM TODS*, 27(2):188–228, 2002.

[3] G. Cormode, M. Garofalakis, and D. Sacharidis. Fast approximate wavelet tracking on streams. In *International Conference on Extending Database Technology*, pages 4–22. Springer, 2006.

[4] M. Garofalakis and P. B. Gibbons. Probabilistic wavelet synopses. *ACM TODS*, 29(1):43–90, 2004.

[5] M. Garofalakis and A. Kumar. Deterministic wavelet thresholding for maximum-error metrics. In *PODS*, pages 166–176, 2004.

[6] M. Garofalakis and A. Kumar. Wavelet synopses for general error metrics. *TODS*, 30(4):888–928, 2005.

[7] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. J. Strauss. One-pass wavelet decompositions of data streams. *TKDE*, 15(3):541–554, 2003.

[8] S. Guha. Space efficiency in synopsis construction algorithms. In *VLDB*, pages 409–420, 2005.

[9] S. Guha. On the space–time of optimal, approximate and streaming algorithms for synopsis construction problems. *VLDB journal*, 17(6):1509–1535, 2008.

[10] S. Guha and B. Harb. Wavelet synopsis for data streams: minimizing non-euclidean error. In *SIGKDD*, pages 88–97, 2005.

[11] S. Guha and B. Harb. Approximation algorithms for wavelet transform coding of data streams. *Information Theory*, 54(2):811–830, 2008.

[12] S. Guha, H. Park, and K. Shim. Wavelet synopsis for hierarchical range queries with workloads. *VLDB journal*, 17(5):1079–1099, 2008.

[13] J. Jestes, K. Yi, and F. Li. Building wavelet histograms on large data in mapreduce. *PVLDB*, 5(2):109–120, 2011.

[14] P. Karras. Optimality and scalability in lattice histogram construction. *PVLDB*, 2(1):670–681, 2009.

[15] P. Karras and N. Mamoulis. One-pass wavelet synopses for maximum-error metrics. In *VLDB*, pages 421–432, 2005.

[16] P. Karras and N. Mamoulis. The Haar+ tree: A refined synopsis data structure. In *ICDE*, pages 436–445, 2007.

[17] P. Karras and N. Mamoulis. Hierarchical synopses with optimal error guarantees. *TODS*, 33(3):18, 2008.

[18] P. Karras, D. Sacharidis, and N. Mamoulis. Exploiting duality in summarization with deterministic guarantees. In *SIGKDD*, pages 380–389. ACM, 2007.

[19] J. Kim, J. Min, and K. Shim. Efficient Haar$^+$ synopsis construction for the maximum absolute error measure. Technical report, Seoul National University, 2017. http://kdd.snu.ac.kr/~shim/TR/TRHaarP.pdf.

[20] S. Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

[21] Y. Matias, J. S. Vitter, and M. Wang. Wavelet-based histograms for selectivity estimation. In *SIGMOD*, volume 27, pages 448–459. ACM, 1998.

[22] Y. Matias, J. S. Vitter, and M. Wang. Dynamic maintenance of wavelet-based histograms. In *VLDB*, pages 101–110, 2000.

[23] S. Muthukrishnan. Subquadratic algorithms for workload-aware haar wavelet synopses. In *FSTTCS*, pages 285–296, 2005.

[24] A. Natsev, R. Rastogi, and K. Shim. Walrus: A similarity retrieval algorithm for image databases. In *SIGMOD*, volume 28, pages 395–406, 1999.

[25] C. Pang, Q. Zhang, X. Zhou, D. Hansen, S. Wang, and A. Maeder. Computing unrestricted synopses under maximum error bound. *Algorithmica*, 65(1):1–42, 2013.

[26] F. Reiss, M. Garofalakis, and J. M. Hellerstein. Compact histograms for hierarchical identifiers. In *VLDB*, pages 870–881, 2006.

[27] J. S. Vitter and M. Wang. Approximate computation of multidimensional aggregates of sparse data using wavelets. In *SIGMOD*, volume 28, pages 193–204. ACM, 1999.