

---

# Coresets for Decision Trees of Signals

---

**Ibrahim Jubran**  
Department of Computer Science  
University of Haifa, Israel  
ibrahim.jub@gmail.com

**Ernesto Evgeniy Sanches Shayda**  
Department of Computer Science  
University of Haifa, Israel  
ernestosanches@gmail.com

**Ilan Newman**  
Department of Computer Science  
University of Haifa, Israel  
ilan@cs.haifa.ac.il

**Dan Feldman**  
Department of Computer Science  
University of Haifa, Israel  
dannyf.post@gmail.com

## Abstract

A  $k$ -decision tree  $t$  (or  $k$ -tree) is a recursive partition of a matrix (2D-signal) into  $k \geq 1$  block matrices (axis-parallel rectangles, leaves) where each rectangle is assigned a real label. Its regression or classification loss to a given matrix  $D$  of  $N$  entries (labels) is the sum of squared differences over every label in  $D$  and its assigned label by  $t$ . Given an error parameter  $\varepsilon \in (0, 1)$ , a  $(k, \varepsilon)$ -coreset  $C$  of  $D$  is a small summarization that provably approximates this loss to *every* such tree, up to a multiplicative factor of  $1 \pm \varepsilon$ . In particular, the optimal  $k$ -tree of  $C$  is a  $(1 + \varepsilon)$ -approximation to the optimal  $k$ -tree of  $D$ .

We provide the first algorithm that outputs such a  $(k, \varepsilon)$ -coreset for *every* such matrix  $D$ . The size  $|C|$  of the coreset is polynomial in  $k \log(N)/\varepsilon$ , and its construction takes  $O(Nk)$  time. This is by forging a link between decision trees from machine learning – to partition trees in computational geometry.

Experimental results on `sklearn` and `lightGBM` show that applying our coresets on real-world data-sets boosts the computation time of random forests and their parameter tuning by up to  $\times 10$ , while keeping similar accuracy. Full open source code is provided.

## 1 Introduction

Decision trees are one of the most common algorithms used in machine learning today, both in the academy and industry, for classification and regression problems [52]. Informally, a decision tree is a recursive binary partition of the input feature space into hyper-rectangles, where each such hyper-rectangle is assigned a label. If the labels are given from a discrete set, the trees are usually called *classification trees*, and otherwise they are usually called *regression trees*. Variants include non-binary partitions and forests [29].

**Why decision trees?** Advantages of decision trees, especially compared to deep learning, include: **(i)** Interpretability. They are among the most popular algorithms for interpretable (transparent) machine learning [31]. **(ii)** Usually require small memory space, which also implies fast classification time. **(iii)** Accuracy. Decision trees are considered as one of the few competitors of deep networks. In competitions, such as the ones in Kaggle [36], they are one of the favorite classifiers [10], especially on small or traditional tabular data. **(iv)** May learn from small training data.

**The goal** is usually to compute the optimal  $k$ -tree  $t^*$  for a given dataset  $D$  and a given number  $k$  of leaves, according to some given loss function. In practice, researchers usually use ensemble of trees called forests, e.g., a Random Forest [12], which are usually learned from different subsets of the

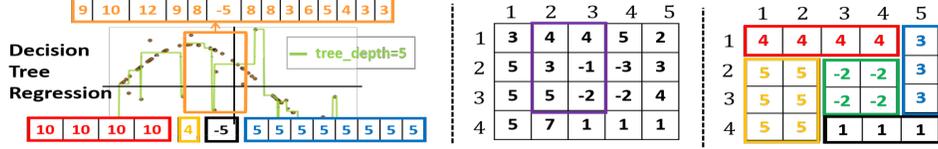


Figure 1: **(Left):** A one dimensional signal (orange points) and its segmentation into 25 “smooth” segments / leaves (green lines). Image taken from Section 1.10 (“Decision Trees”) of the sklearn’s User Guide [56]. The vector  $v$  on top represents a subset of the signal’s values. The bottom vectors represent a 4-segmentation of  $v$ , similar to the horizontal green line segments. Each segment contains the average value of its corresponding segment from  $v$ . **(Middle):** A matrix that represents the  $4 \times 5$  signal  $D = \{((1, 1), 3), ((1, 2), 4), ((1, 3), 5), \dots\}$  (in black) and a  $3 \times 2$  matrix that represents a  $3 \times 2$  sub-signal  $B = \{((1, 2), 4), ((1, 3), 4), ((2, 2), 3), \dots\}$  (in purple). **(Right):** A matrix that represents a 5-segmentation  $s$  of  $A = [4] \times [5]$ ; see Definition 1. Since  $s$  is a 5-segmentation, it partitions the  $[4] \times [5]$  matrix into 5 distinct block matrices  $B_1$  (red),  $B_2$  (blue),  $B_3$  (yellow),  $B_4$  (green), and  $B_5$  (black) such that  $s$  assigns the same number for all the entries in the same block. The SSE fitting loss  $\ell(D, s)$  is the sum of squared differences over every entry in the left matrix to its corresponding entry on the right matrix. Also, there is no  $k$ -tree that can obtain the same partition.

training set. The final classification is then based on a combination rule, such as majority or average vote. Since both the training and classification of each tree are computed independently and possibly in parallel, we focus on the construction of a single tree.

**A dataset**  $D$  in this paper is a set  $D = \{(x_1, y_1), \dots, (x_N, y_N)\} \subseteq A \times \mathbb{R}$  of pairs, where  $A$  is the *feature space*. Each pair  $(x, y) \in D$  consists of a database record (vector / sample)  $x \in A$  and its real label  $y \in \mathbb{R}$ . As common, we assume that non-real features, such as categorical features, are converted to real numbers; see e.g. [27]. For example, in common classification problems  $A = \mathbb{R}^d$  for some  $d \geq 1$  and  $y \in \{0, 1\}$  is a binary number. The resulting model may be used for prediction on another test dataset, completion of missing values, or efficient storage of the original dataset by replacing the label  $y$  of each pair  $(x, y) \in D$  with the label  $t(x)$  that was assigned to it by the tree  $t$ . The last technique is used e.g. in the MPEG4 encoder [46], where decision trees of a specific structure (quad-trees) are used to compress an image  $D$  that consists of pixel-grayscale pairs  $(x, y)$ .

**Challenges.** The motivation for this paper originated from the following challenges:

- (i) Sub-optimality.** Hardness of decision tree optimization is both a theoretical and practical obstacle [31]. It is NP-hard to compute the optimal  $k$ -tree, or its approximation, when the number  $k$  is not fixed [14, 39]. There were several attempts to improve the optimality of decision tree algorithms, from binary-split decision trees as in [6, 8], in a line of work of e.g. [9, 61]. Nevertheless, greedy implementations e.g., CART [40] and C4.5 [51] have remained the dominant methods in practice.
- (ii) Computation time.** Due to this lack of optimality, finding a decision tree that provides a good accuracy usually requires many runs, since each of them returns only a local minimum that might be arbitrarily far from the global optimum. The final model usually consists of a forest containing many trees. Popular forest implementations include the famous Sklearn, XGBoost, LightGBM, and CatBoost libraries [49, 15, 37, 19], which all utilize (as default) an ensemble of at least 100 trees. Moreover, there is a list of dozen parameters to calibrate including: number of trees, depth of each tree, pruning/splitting strategies on each tree and between them, and many others. To this end, the running time for obtaining reasonable results even on moderate size datasets might be impractical.
- (iii) Scalability.** Existing techniques tend not to scale to realistically-sized problems unless simplified to trees of a specific form as stated in [31].
- (iv) Streaming, parallel, and dynamic updates.** The common algorithms mentioned above do not support continuous learning or updating of the modeled tree when an input sample is either added or removed from the dataset, e.g., when the dataset does not fit into memory or arrives on-the-fly. Similarly, we do not know techniques to train a single tree in parallel on multiple machines.

## 1.1 Coresets

“Coresets are one of the central methods to facilitate the analysis of large data sets.” [47]. Informally, for an input dataset  $D$ , a set  $T$  of models, an approximation error  $\varepsilon \in (0, 1)$ , and a loss function  $\ell$ , a coreset  $C$  is a data structure that approximates the loss  $\ell(D, t)$  for every model  $t \in T$ , up to

a multiplicative factor of  $1 \pm \varepsilon$ , in time that depends only on  $|C|$ . Hence, ideally,  $C$  is also much smaller than their original input  $D$ .

**Why coresets?** The main motivation for constructing a coreset is to compute the optimal model or its approximation, much faster, while sacrificing little accuracy. Furthermore, a coreset for a family of classifiers is many times a “silver bullet” that provides a unified solution to all Challenges (i)-(iv) above. Combining the two main coreset properties: merge and reduce [32, 7, 26, 1], which are usually satisfied, with the fact that a coreset approximates every model, and not just the optimal model, enables it to support streaming and distributed data [11, 41], parallel computation [21], handle constrained versions of the problem [25], model compression [20], parameter tuning [43] and more.

**Coreset construction techniques.** There are many different techniques for coreset construction, ranging from loss-less to lossy, from deterministic to randomized, and from greedy to non-greedy constructions. Examples include accurate coresets [33] and non-accurate coresets [23] via computational geometry, random sampling-based coresets [22, 17, 44], and greedy deterministic coresets via the Frank-Wolfe algorithm [16]. Recently, many works focus on developing frameworks for general families of loss functions, e.g., [22, 57]. We refer the interested reader to the surveys [2, 1, 50, 11, 4, 21] with references therein.

**Practical usage.** Since a coreset is not just another solver that competes with existing solutions, but a data structure for approximating any given model in the family, we can apply *existing* approximation algorithms or heuristics on the coreset to obtain similar results compared to the original (full) data. Since the coreset is small, we may run these heuristics multiple times, or apply the hyperparameter tuning using the coreset [43], thus reducing the computational burden by orders of magnitude.

**Main challenges:** (i) No coreset. Unfortunately, it is not clear at all that a small coreset exists for a given family of models. In fact, we can conclude from [54] that a coreset for decision trees does not exist in general; see details below. In this case, we can either give up on the coreset paradigm and develop a new solver, or add assumption on the input dataset, instead of targeting every possible dataset that may be very artificial and unrealistic, as the counter example in [54]. In this paper, we choose the latter option. (ii) Unlike, say, uniform sampling, every problem formulation requires a different coreset construction, which may take years of research to design.

## 1.2 First coreset for decision trees and their generalization

In this paper, we tackle a generalized and more complex set of models than decision trees, where, rather than a recursive binary partition, we allow the input feature space  $\mathbb{R}^d$  to be partitioned into any  $k$  disjoint axis-parallel hyper-rectangles; see Fig. 1. This generalization is essential in order to support future non-recursive and not necessarily binary classification models, e.g., ID3 and C4.5 [30, 51]. To our knowledge, this is the first coreset with provable guarantees for constructing decision trees.

**Definition 1** ( $k$ -segmentation). *For an integer  $d \geq 1$  that denotes the dimension of the feature space  $A = \mathbb{R}^d$ , and an integer  $k \geq 1$  that denotes the size of the partition (number of leaves), a function  $s : A \rightarrow \mathbb{R}$  is a  $k$ -segmentation if there is a partition  $\mathcal{B} = \{B_1, \dots, B_k\}$  of  $A$  into  $k$  disjoint axis-parallel hyper-rectangles (blocks), such that  $|\{s(b) \mid b \in B\}| = 1$  for every block  $B \in \mathcal{B}$ , i.e.,  $s$  assigns a unique value for all the entries in each of its  $k$  rectangles; see Fig. 1. We define the union over all possible such  $k$ -segmentations by  $\text{SEG}_{(k,d)}$ .*

We now define our loss function, and an optimal  $k$ -segmentation model over some set  $A \subseteq \mathbb{R}^d$ .

**Definition 2** (Loss function). *For a dataset  $D = \{(x_1, y_1), \dots, (x_N, y_N)\} \subseteq A \times \mathbb{R}$ , an integer  $k \geq 1$ , and a  $k$ -segmentation  $s \in \text{SEG}_{(k,d)}$ , we define the sum of squared error (SSE) loss*

$$\ell(D, s) := \sum_{(x,y) \in D} (s(x) - y)^2$$

*as the loss of fitting  $s$  to  $D$ . A  $k$ -segmentation  $s^*$  is an optimal  $k$ -segmentation of  $D$  if it minimizes  $\ell(D, s)$  over every  $k$ -segmentation  $s \in \text{SEG}_{(k,d)}$  i.e.,  $s^* \in \arg \min_{s \in \text{SEG}_{(k,d)}} \ell(D, s)$ . The optimal SSE loss is denoted by  $\text{opt}_k(D) := \ell(D, s^*)$ .*

For example, the optimal 1-segmentation  $s^*$  of  $D$  is the constant function  $s^* \equiv \frac{1}{|D|} \sum_{(x,y) \in D} y$  since the mean of a set of numbers minimizes the sum of squared distances to the elements of the set. Also,  $\text{opt}_{|D|}(D) = 0$  for every dataset  $D$ .

We are now ready to formally define a coreset for the  $k$ -segmentation problem (and  $k$ -decision trees of at most  $k$  leaves, in particular).

**Definition 3 (Coreset).** Let  $D = \{(x_1, y_1), \dots, (x_n, y_n)\} \subseteq A \times \mathbb{R}$  be an input dataset. Let  $k \geq 1$  be an integer and  $\varepsilon \in (0, 1)$  be the desired approximation error. A  $(k, \varepsilon)$ -coreset for  $D$  is a data structure  $(C, u)$  where  $C \subseteq A \times \mathbb{R}$  is an ordered set, and  $u : C \rightarrow [0, \infty)$  is called a weight function, such that  $(C, u)$  suffices to approximate the loss  $\ell(D, s)$  of the original dataset  $D$ , up to a multiplicative factor of  $1 \pm \varepsilon$ , in time that depends only on  $|C|$  and  $k$ , for any  $k$ -segmentation  $s$ .

**Practical usage.** As defined above and discussed in Section 1.1, a coreset approximates every model in our set of models  $\text{SEG}_{(k,d)}$ . Hence, a coreset for decision trees is clearly also a coreset for forests with an appropriate tuning for  $k$ , since every tree in the forest is approximated independently by the coreset. We expect that applying existing heuristics (not necessarily with provable guarantees) such as sklearn [49] or LightGBM [37] on the coreset, would yield similar results compared to the original data. Indeed, our experimental results in Section 5 validate those claims.

**No coreset for general datasets.** Unfortunately, even for the case of  $k = 4$  and  $A \subseteq \mathbb{R}$ , i.e., when the input is simply a one dimensional dataset  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$  where  $x_1, \dots, x_n$  are real numbers, and the labels  $y_1, \dots, y_n \in \{0, 1\}$  have only binary values, it is easy to construct datasets which have provably no  $k$ -segmentation (or even  $k$ -tree) coreset of size smaller than  $n$ ; see e.g. [54]. Hence, there is no non-trivial decision tree coreset for general datasets of  $n$  vectors in any dimension. However, as we prove in the rest of the paper, a coreset does exist for datasets where the input is a matrix, i.e., a discrete signal where every coordinate in the domain is assigned a label (value), rather than a random set of  $n$  vectors.

**The first coreset for  $n \times m$ -signals.** To overcome the above problem, while still obtaining a small coreset, we assume a discretization of the dataset so that every coordinate has a label. We also assume, mainly for simplicity and lack of space, that the input feature space is  $A = [n] \times [m] \subseteq \mathbb{R}^2$ . That is, the input can be represented by an  $n \times m$  matrix. The output coreset may contain fraction of entries, as in Fig. 3, which is called an  $n \times m$  signal; see Section 1.5. Our assumption on the input data seems to be the weakest assumption that can enable us to have a provably small coreset for any input. Furthermore, it seems natural for e.g. images, matrices, or any input data from sensors (such as GPS) that has a value in every cell or continuous in some other sense.

**Previous work.** The prior works [54, 24, 62], which only handle the case of segmenting a 1-dimensional signal, use relaxations similar to our relaxation above to obtain a coreset of size  $O(k/\varepsilon^2)$ . However, our results apply easily for the case of vectors (1-dimensional signals) as in [54] and generalize for tensors if  $d \geq 3$ . We also give further applications, and provide extensive experiments with popular state of the art software.

A special case for  $d = 2$  includes image compression, where quadtrees are usually used in e.g. MPEG4 to replace the image by smooth blocks of different sizes [55], or for completion of missing values [58] Using dynamic programming, it is easy to compute the optimal tree of a 2D-signal  $D$  in  $O(k^2 n^5)$  time [5], which is impractical even for small datasets, *unless applied on a small coreset of  $D$* . However we do not know of any such coreset construction, for  $d \geq 2$ , with provable guarantees on its size. To this end, the following questions are the motivation for this paper: **(i)**: Is there a small coreset for any  $n \times m$  signal (e.g. of sub-linear size)? **(ii)**: If so, can it be computed efficiently? **(iii)**: Can it be used on real-world datasets to boost the performance of existing random forest implementations?

**Extensions.** For simplicity, we focus on the classic sum of squared distances (SSE) or the risk minimization model [59]. However, our suggested techniques mainly assume that a coreset for the case  $k = 1$  is known, which is trivial for SSE, but exists for many other loss functions e.g., non-squared distances; see Section 6.

### 1.3 Our Contribution

For any given error parameter  $\varepsilon \in (0, 1)$ , and an integer  $k \geq 1$ , this paper answers affirmatively the above three questions. More formally, in this paper we provide:

**(i)**: A proof that every  $n \times m$  signal  $D$  has a  $(k, \varepsilon)$ -coreset  $(C, u)$  of size  $|C|$  polynomial in  $k \log(nm)/\varepsilon$ . To our knowledge, this is the first coreset for decision trees whose size is smaller than the input; see Theorem 8. Due to lack of space, our full proofs are given in the appendix.

**(ii)**: A novel coreset construction algorithm that outputs such a coreset  $(C, u)$  with the above guarantees, for every given input signal  $D$ . Its running time is  $O(nmk)$ , i.e., linear in the input size

- $|D|$ . Unlike common coresets constructions, our algorithm is deterministic; see Algorithm 3.
- (iii): AutoML for decision trees: since the suggested coreset approximates *every* tree of *at most*  $k$  leaves, we may use the same coreset for hyperparameter tuning. We demonstrate this in Section 5, by calibrating the parameter  $k$  using only the coreset, as compared to using the original (big) data.
  - (iv): Experimental results that apply modern solvers, such as the sklearn and LightGBM libraries, on this coreset for real-world public datasets. We show that our coreset can help boost the computation time of the above forest implementations and their parameter tuning by up to x10, while keeping similar accuracy; see Section 5.
  - (v): Open source code for our algorithms [35]. We expect that it will be used by both the academic and industry communities, not only to improve the running time of existing projects, but also to extend the algorithms and experimental results to other libraries and cost functions; see Section 6.

#### 1.4 Novel technique: partition trees meet decision trees

In a seminal paper [28] during the 80’s of the previous century, Haussler and Welz introduced the importance of VC-dimension by Vapnik–Chervonenkis [60]. Their main application was partition trees for answering range queries.

Informally, a *partition tree* of a given set of points on the plane is the result of computing recursively a *simplicial partition* which is defined as follows. For a set  $D$  of  $N$  points on the plane, a  $(1/\varepsilon)$ -*simplicial partition* is the partition of  $D$  into  $O(1/\varepsilon)$  subsets, such that: (i) each subset has at most  $2\varepsilon N$  points, and (ii) Every line in the plane intersects the convex hull of at most  $\sqrt{1/\varepsilon}$  sets. Answering range queries of the form “how many points in  $D$  are in a given rectangular” in sub-linear time, using partition trees, is straightforward: We can sum in  $O(1/\varepsilon)$  time the number of points in the subsets of the above simplicial partition that are not intersected by the query rectangular. We then continue recursively to count the points on each of the  $\sqrt{1/\varepsilon}$  intersected sets.

In other words, the main idea behind the above work is to partition the input into a (relatively small) number of subsets, each containing a fraction of the input, such that each query (in this case, a rectangular shape) might intersect only a small fraction of those subsets. Such a partition is termed a simplicial partition. The number of points contained in non-intersected subsets can be easily computed, while the sum of points in intersected subsets require a more involved solution. The novelty in that work is how to achieve such a partition of the input.

Our paper closes a loop in the sense that it forges links between decision trees in machine learning – to partition trees from computational geometry. We aim to generalize the above technique from covering problems to regression and classification problems. This is by devising an algorithm which achieves the above requirements, but where the query is a decision tree (and not a rectangular shape), and the cost function is the sum of squared distances to the query and not the number of points.

More precisely, We partition the input dataset  $D$  into a relatively small number of subsets, such that every possible decision tree (query) intersects at most few of these subsets. We then independently compress every subset via another novel algorithm such that the cost (sum of squared distances, in this case) of points contained in non-intersected subsets can be easily and accurately estimated, while the cost of points in intersected subsets can be provably approximated via a more involved calculation. This is very different from existing coreset techniques that are sampling-based [38], or utilize convex optimization greedy algorithms [16]. Our main challenge was to define and design such a “simplicial partition for sum of squared distances”, and the coreset to be computed for each subset in this partition.

#### 1.5 Preliminaries

In this section we define the notation that will be used in the next sections.

Let  $n \geq 1$  and denote  $[n] = \{1, \dots, n\}$ . An  $n$ -signal is a set  $\{(x, f(x)) \mid x \in [n]\}$  that is defined by a function  $f : [n] \rightarrow \mathbb{R}$  (known as the graph of  $f$ ). For an additional integer  $m \geq 1$ , an  $n \times m$  signal  $D = \{(x, g(x)) \mid x \in [n] \times [m]\}$  is the set that corresponds to a function  $g : [n] \times [m] \rightarrow \mathbb{R}$ . That is,  $D$  represents an  $n \times m$  real matrix whose size is  $|D| = N = nm$ . For integers  $i_1, i_2, j_1, j_2$  such that  $1 \leq i_1 \leq i_2 \leq n$  and  $1 \leq j_1 \leq j_2 \leq m$ , an  $n \times m$  sub-signal is the set  $B = \{(x, g(x)) \mid x \in \{i_1, \dots, i_2\} \times \{j_1, \dots, j_2\}\} \subseteq D$ ; see Fig. 1. A sub-signal  $B$  is called a *row* (respectively, *column*) if  $i_1 = i_2$  (respectively,  $j_1 = j_2$ ). For a sub-signal  $B$ , we

denote by  $B^T = \{((j, i), y) \mid ((i, j), y) \in B\}$  the *transposed sub-signal*. A  $k$ -segmentation  $s$  is said to *intersect* an  $n \times m$  sub-signal  $B$  if  $s$  assigns at least two distinct values to the entries of  $B$ , i.e.,  $|\{s(x) \mid (x, y) \in B\}| \geq 2$ . Furthermore, by definition, a  $k$ -segmentation  $s$  induces a partition of an  $n \times m$  sub-signal  $B$  into at most  $k$   $n \times m$  sub-signals. For two functions  $f, g : \mathbb{R} \rightarrow \mathbb{R}$  we use the big  $O$  notation  $f(x) \in O(g(x))$ , thinking of  $O(g(x))$  as the class of all functions  $h(x)$  such that  $|h(x)| \leq c|g(x)|$  for every  $x > x_0$ , for some constants  $c$  and  $x_0$ . Lastly, we denote  $\text{SEG}_k := \text{SEG}_{(k,2)}$  for brevity.

**Paper organization.** Section 2 provides a rough approximation to the  $k$ -segmentation problem. Section 3 provides an algorithm for computing a simplicial partition for the  $k$ -segmentation problem. Each region in this partition will be then compressed individually in Section 4 to obtain our desired coresnet. Experimental results and discussions are given in Section 5, and a conclusion in Section 6.

## 2 Bi-criteria Approximation

A coresnet construction usually requires some rough approximation to the optimal solution as its input. Unfortunately, we do not know how to *efficiently* compute even a constant factor approximation to the optimal  $k$ -segmentation problem in Definition 2, as explained in Section 1. Instead, we provide an  $(\alpha, \beta)_k$  or bi-criteria approximation [22], where the approximation is with respect to a pair of parameters: the number of segments in the partition may be up to  $\beta k$  instead of  $k$ , and the loss may be up to  $\alpha \cdot \text{opt}_k(D)$  instead of  $\text{opt}_k(D)$ .

**Definition 4** ( $(\alpha, \beta)_k$ -approximation.). *Let  $D$  be an  $n \times m$  sub-signal,  $k \geq 1$  be an integer and let  $\alpha, \beta > 1$ . A function  $s : [n] \times [m] \rightarrow \mathbb{R}$  is an  $(\alpha, \beta)_k$ -approximation of  $D$ , if  $s$  is a  $\beta k$ -segmentation whose fitting loss to  $D$  is at most  $\alpha$  times the loss of the optimal  $k$ -segmentation of  $D$ , i.e.,  $s \in \text{SEG}_{(\beta k)}$  and  $\ell(D, s) = \sum_{(x,y) \in D} (s(x) - y)^2 \leq \alpha \cdot \text{opt}_k(D)$ .*

We now describe an algorithm that computes such an approximation, in time only linear in the input's size  $|D| = nm$ . The following lemma gives the formal statement. A suggested implementation for the algorithm is given in the appendix, as well as the full proof of the lemma; see Section B.

**Lemma 5.** *Let  $D = \{(x_1, y_1), \dots, (x_{nm}, y_{nm})\}$  be an  $n \times m$  sub-signal and  $k \geq 1$  be an integer. Then, there is an algorithm that can compute, in  $O(knm)$  time, an  $(\alpha, \beta)_k$ -approximation for  $D$ , where  $\alpha \in O(k \log(nm))$  and  $\beta \in k^{O(1)} \log^2(nm)$ .*

**Overview of the bicriteria algorithm from Lemma 5:** The algorithm is iterative and works as follows. At the  $i$ th iteration, we find a collection  $\mathcal{B}_i$  of at most  $t$  disjoint sub-signals in  $D_i$  (where  $D_0 = D$  is the input), for which: (i)  $\sum_{B \in \mathcal{B}_i} \text{opt}_1(B) \leq \text{opt}_k(D_i) \leq \text{opt}_k(D)$ , and (ii)  $\cup_{B \in \mathcal{B}_i} B$  has size  $|\cup_{B \in \mathcal{B}_i} B| \geq |D_i|/c$  for some parameter  $c$  that depends on  $k$ , i.e., those sub-signals contain at least a  $1/c$  fraction of  $D_i$ . We then define  $D_{i+1} = D_i \setminus \cup_{B \in \mathcal{B}_i} B$ . After repeating this for at most  $\psi \in O(c \log(nm))$  iterations, we end up covering all entries of  $D$  with sub-signals where the overall loss of the sub-signals in each iteration is at most  $\text{opt}_k(D)$ . This defines a partition of  $D$  into a collection of at most  $t\psi$  disjoint sets  $\mathcal{B}'$ , which, in turn, define a set of at most  $(t\psi)^2$  distinct sub-signals. The output is now simply the function  $s$  that assigns, for every  $B \in \mathcal{B}'$  and  $b \in B$ , the mean value  $s(b) = \frac{1}{|B|} \sum_{((i,j),y) \in B} y$  of  $B$ . See Pseudo-code in Algorithm 4 at the appendix.

## 3 Balanced Partition

In this section we present Algorithm 2, which computes a partition similar to the simplicial partition described in Section 1.4; It computes, in  $O(|D|)$  time, a partition  $\mathcal{B}$  of the input  $D$  that satisfies the following properties: (i)  $|\mathcal{B}|$  depends on  $k/\varepsilon$  but independent of  $|D|$ , (ii) the loss  $\text{opt}_1(B)$  of every  $B \in \mathcal{B}$  is small, and (iii) every  $k$ -segmentation  $s$  intersects only few sub-signals  $B \in \mathcal{B}$ ; see Definition 6, Fig. 2, and Lemma 7. A full proof is given at the appendix; see Section C.

**Definition 6** (Balanced Partition). *Let  $D$  be an  $n \times m$  signal,  $k \geq 1$  be an integer, and  $c_1, c_2, c_3 > 0$ . A  $(c_1, c_2, c_3)_k$ -balanced partition of  $D$  is a partition  $\mathcal{B}$  of  $D$  such that: (i)  $\mathcal{B}$  contains  $|\mathcal{B}| \leq c_1 n \times m$  sub-signals, (ii)  $\text{opt}_1(B) \leq c_2$  for every  $B \in \mathcal{B}$ , and (iii) every  $k$ -segmentation  $\hat{s}$  intersects at most  $c_3$  sub-signals  $B \in \mathcal{B}$  (i.e., assigns more than one unique number to those sub-signals).*

**Lemma 7.** *Let  $D$  be an  $n \times m$  signal,  $k \geq 1$  be an integer,  $\varepsilon \in (0, 1/4)$  be an error parameter, and  $s : [n] \times [m] \rightarrow \mathbb{R}$  be an  $(\alpha, \beta)_k$ -approximation of  $D$ , where  $\alpha, \beta > 1$ . Define  $\sigma := \frac{\ell(D, s)}{\alpha}$*

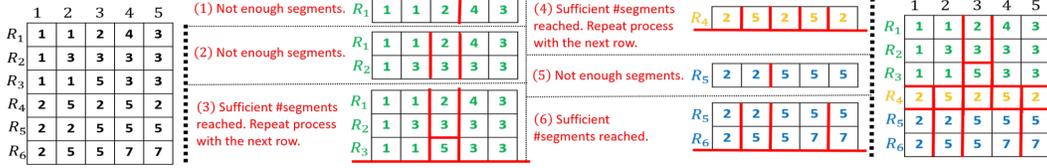


Figure 2: **(Left):** A  $6 \times 5$  signal  $D$  consisting of 6 rows  $R_1, \dots, R_6$ . **(Middle):** A step by step illustration of the call to  $\mathcal{B} := \text{PARTITION}(D, 1/4, 64)$  which partitions  $D$  into  $|\mathcal{B}| = 13$  sub-signals (sub-matrices) as follows. (1)  $\mathcal{B}' := \text{SLICEPARTITION}(\{R_1\}, 4)$  (top green row). (2)  $\mathcal{B}' := \text{SLICEPARTITION}(R_1 \cup R_2, 4)$  (middle green matrix), and so on as long as the output contains at most  $|\mathcal{B}'| \leq 1/\gamma = 4$  sub-signals (as in the bottom green matrix (3)). We then append  $\mathcal{B}'$  to the output  $\mathcal{B}$ , and repeat with the remaining  $\{R_4, R_5, R_6\}$ . (4)  $\mathcal{B}' := \text{SLICEPARTITION}(\{R_4\}, 4)$  which already returns  $|\mathcal{B}'| = 5 > 1/\gamma$  signals (yellow matrix). We append them to  $\mathcal{B}$  and repeat. **(Right):** The final partition  $\mathcal{B}$ , where  $\text{opt}_1(B) \leq \gamma^2 \sigma = 1/4^2 \cdot 64 = 4$  for every  $B \in \mathcal{B}$ .

and  $\gamma := \frac{\varepsilon^2}{\beta k}$ . Let  $\mathcal{B}$  be an output of a call to  $\text{PARTITION}(D, \gamma, \sigma)$ ; see Algorithm 2. Then  $\mathcal{B}$  an  $\left(O\left(\frac{\alpha}{\gamma^2}\right), \gamma^2 \sigma, O\left(\frac{k\alpha}{\gamma}\right)\right)_k$ -balanced partition of  $D$ . Moreover,  $\mathcal{B}$  can be computed in  $O(nm)$  time.

**Overview of Algorithms 1 and 2.** Algorithm 2 gets as input an  $n \times m$ -signal  $D$  and two parameters  $\sigma, \gamma$ . Algorithm 2 aims to compute a balanced partition of  $D$ ; see Fig. 2. In turn, it calls Algorithm 1, which takes as input an  $n \times m$  sub-signal  $R$  that is defined by several contiguous rows of the original dataset  $D$ , and a parameter  $\sigma > 0$ , and aims to compute a partition  $\mathcal{B}$  of  $R$ . To do so, Algorithm 1 partitions  $R$  along the vertical dimension (e.g., into vertical slices), in a greedy fashion, such that for every  $B \in \mathcal{B}$ ,  $\text{opt}_1(B)$  is as large as possible, while still upper bounded by  $\sigma$ . This will ensure that the partition is into a relatively small number of slices. In the case where one of the sub-signals  $B$  in this vertical partition of  $R$  contains only one column, and already exceeds the maximum tolerance  $\text{opt}_1(B) > \sigma$ , we recursively apply Algorithm 1 to  $B^T$  in order to partition  $B$  horizontally. As long as the total number of slices returned by Algorithm 1 is smaller than  $1/\gamma$ , Algorithm 2 adds yet another row to the previous set of rows, and repeats the above process. At this point, the partition of the current horizontal slice (collection of rows)  $R$  is final, and is added to the output partition of Algorithm 2. In turn, a new horizontal slice  $R$  of just one row, the first row of  $D$  that is not included in the previous  $R$ , is initiated on which we again call Algorithm 1.

Algorithm 1: SLICEPARTITION( $D, \sigma$ )	Algorithm 2: PARTITION( $D, \gamma, \sigma$ )
<b>Input</b> : A parameter $\sigma > 0$ and an $n \times m$ signal $D = \{(x_i, y_i)\}_{i=1}^N$ .	<b>Input</b> : An $n \times m$ signal $D$ , a parameter $\gamma \in (0, 1)$ , and a lower bound $\sigma \in [0, \text{opt}_k(D)]$ .
<b>Output</b> : A partition $\mathcal{B}$ of $D$ .	<b>Output</b> : A partition $\mathcal{B}$ of $D$ ; see Lemma 7
1 $\mathcal{B} := \emptyset$ and $c_{begin} := 1$	1 $\mathcal{B} := \emptyset$ and $r_{begin} := 1$
2 <b>while</b> $c_{begin} \leq m$ <b>do</b>	2 <b>while</b> $r_{begin} \leq n$ <b>do</b>
3 $B := \{(i, j), y \in D \mid j = c_{begin}\}$ // extract first column	3 $R := \{(i, j), y \in D \mid i = r_{begin}\}$ // extract first row
4 <b>if</b> $\text{opt}_1(B) > \sigma$ <b>then</b>	4 $\mathcal{B}' := \text{SLICEPARTITION}(R, \gamma^2 \sigma)$
5 $\mathcal{B}' := \text{SLICEPARTITION}(B^T, \sigma)$	5 $r_{end} := r_{begin}$
6 $\mathcal{B} := \mathcal{B} \cup \{B'^T \mid B' \in \mathcal{B}'\}$	6 $last\mathcal{B}' := \mathcal{B}'$
7 $c_{begin} := c_{begin} + 1$	7 <b>while</b> $ \mathcal{B}'  \leq 1/\gamma$ and $r_{end} < n$ <b>do</b>
8 <b>else</b>	8 $r_{end} := r_{end} + 1$
9 $c_{end} := c_{begin}$	9 $last\mathcal{B}' := \mathcal{B}'$
10 <b>while</b> $\text{opt}_1(B) \leq \sigma$ and $c_{end} < m$ <b>do</b>	10 $S :=$
11 $c_{end} := c_{end} + 1$ and $lastB := B$	$\{(i, j), y \in D \mid i \in [r_{begin}, r_{end}]\}$
12 $B :=$	// extract a slice
13 $\{(i, j), y \in D \mid i \in [c_{begin}, c_{end}]\}$	$\mathcal{B}' = \text{SLICEPARTITION}(S, \gamma^2 \sigma)$
14 <b>return</b> $\mathcal{B}$	12 $\mathcal{B} := \mathcal{B} \cup last\mathcal{B}'$
	13 $r_{begin} := r_{end}$
	14 <b>return</b> $\mathcal{B}$

## 4 Coreset Construction

In this section, we present our main algorithm (Algorithm 3), which outputs a  $(k, \varepsilon)$ -coreset for a given  $n \times m$  signal  $D$ , the number of leaves  $k \geq 1$ , and an approximation error  $\varepsilon \in (0, 1)$ .

**Overview of Algorithm 3:** The algorithm first utilizes the  $(\alpha, \beta)_k$ -approximation from Section 2 to obtain a lower bound  $\sigma \leq \text{opt}_k(D)$  for the optimal  $k$ -segmentation. It then computes, as described in Section 3, a balanced partition  $\mathcal{B}$  of  $D$ , where  $\text{opt}_1(B)$  is small and depends on  $\sigma$ , for every  $B \in \mathcal{B}$ . Finally, it computes a small representation  $(C_B, u_B)$  for every  $B \in \mathcal{B}$ , and returns the union of those representations. Each such pair  $(C_B, u_B)$  satisfies: (i)  $|C_B| = 4$ , and (ii) has the same weighted sum of values, weighted sum of squared values, and sum of weights, as  $B$ , i.e.,  $\sum_{(a,b) \in C_B} u_B((a,b)) \cdot (b \mid b^2 \mid 1) = \sum_{(x,y) \in B} (y \mid y^2 \mid 1)$ ; see Fig. 3. Such a representation can be computed using Caratheodory’s theorem, as explained in Section E of the supplementary material.

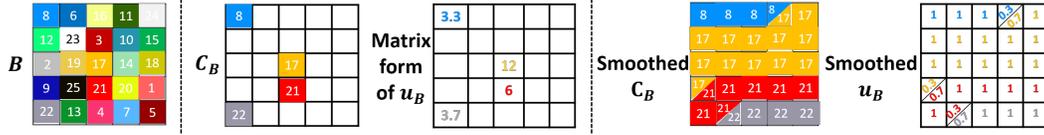


Figure 3: **(Left):** A matrix representing of a  $5 \times 5$  sub-signal  $B$  where  $y$  is mapped into unique colors for every  $(x, y) \in B$ . **(Middle):** A representative (coreset) pair  $(C_B, u_B)$  for  $B$  where  $C_B \subseteq B$  is a (small) subset and  $u_B : C_B \rightarrow [0, \infty)$  is a weight function. That is, the pair  $(C_B, u_B)$  satisfies  $\sum_{(a,b) \in C_B} u_B((a,b)) \cdot (b \mid b^2 \mid 1) = \sum_{(x,y) \in B} (y \mid y^2 \mid 1)$ . **(Right):** A duplication of the coreset points according to their weight. We call the resulting pair a “smoothed” version of  $(C_B, u_B)$ ; see more details and formal definition in Section D of the supplementary material.

**Some intuition behind Algorithm 3:** Consider some  $k$ -segmentation  $s$ . By the properties of the balanced partition  $\mathcal{B}$  of  $D$ , only a small number of sub-signals  $B \in \mathcal{B}$  are intersected by  $s$ , i.e., assigned at least 2 distinct values. For every non-intersected sub-signal  $B \in \mathcal{B}$ , the loss  $\ell(B, s)$  is accurately estimated by the (coreset) pair  $(C_B, u_B)$ . On the other hand, for every sub-signal  $B \in \mathcal{B}$  which is intersected by  $s$ , by the guarantees of the representation  $(C_B, u_B)$ , the loss  $\ell(B, s)$  will be approximated, using only  $(C_B, u_B)$ , up to some small error that depends on  $\text{opt}_1(B)$ . However, again by the properties of  $\mathcal{B}$ , we have that  $\text{opt}_1(B)$  is small. Hence, using the union  $(C, u)$  of the representations we can approximate  $\ell(D, s)$  as required. Furthermore, combining that  $|C_B| \in O(1)$  for every  $B \in \mathcal{B}$  with the fact that  $|\mathcal{B}|$  is small yields that  $|C|$  is indeed small; see Theorem 8.

---

**Algorithm 3:** SIGNAL-CORESET( $D, k, \varepsilon$ ); see Theorem 8

---

**Input :** An  $n \times m$  signal  $D$ , an integer  $k \geq 1$ , and an error parameter  $\varepsilon \in (0, 1/4)$ .

**Output :** A  $(k, \varepsilon)$ -coreset  $(C, u)$  for  $D$ .

- 1  $s :=$  an  $(\alpha, \beta)_k$  approximation of  $D$  for  $\alpha \in O(k \log(nm))$  and  $\beta \in k^{O(1)} \log^2(nm)$ ; see Lemma 5 for suggested implementation.
  - 2  $\gamma := \varepsilon^2 / (\beta k)$ ,  $\sigma := \frac{\ell(D, s)}{\alpha}$  and  $C := \emptyset$
  - 3  $\mathcal{B} :=$  PARTITION( $D, \gamma, \sigma$ ) // see Algorithm 2.
  - 4 **for every set**  $B \in \mathcal{B}$  **do**
  - 5      $(C_B, u_B) :=$  a  $(1, 0)$ -coreset for  $B$ , (a zero error coreset for  $k = 1$ ), such that  $C_B \subseteq B$ ,  $|C_B| = 4$ , and  $\sum_{(a,b) \in C_B} u_B((a,b)) \cdot (b \mid b^2 \mid 1) = \sum_{(x,y) \in B} (y \mid y^2 \mid 1)$  this is done using Caratheodory’s theorem; see Corollary 17 in the appendix.
  - 6     Replace each of the coordinates  $a$  of the 4 pairs  $(a, b) \in C$  with one of the 4 corner coordinates of the pairs in  $B$ ; see detailed explanation in the proof of Theorem 8.
  - 7      $C := C \cup C_B$  and  $u((a, b)) := u_B((a, b))$  for every  $(a, b) \in C_B$ .
  - 8 **return**  $(C, u)$
- 

**Theorem 8 (Coreset).** Let  $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$  be an  $n \times m$  signal i.e.,  $N := nm$ . Let  $k \geq 1$  be an integer (that corresponds to the number of leaves/rectangles), and  $\varepsilon \in (0, 1/4)$  be an error parameter. Let  $(C, u)$  be the output of a call to SIGNAL-CORESET( $D, k, \varepsilon/\Delta$ ) for a

sufficiently large constant  $\Delta \geq 1$ ; see Algorithm 3. Then,  $(C, u)$  is a  $(k, \varepsilon)$ -coreset for  $D$  of size  $|C| \in \frac{(k \log(N))^{O(1)}}{\varepsilon^4}$ ; see Definition 3. Moreover,  $(C, u)$  can be computed in  $O(kN)$  time.

**Coreset size.** While Theorem 8 gives a worst-case theoretical upper bound, this bound is too pessimistic in practice, as common in coreset papers [42, 34]. This phenomenon is well known for coresets; see discussion e.g., in [21, 53]. The reasons might include: worst-case artificial examples vs. average behaviour on structured real-world data, noise removing/smoothing by coresets, the fact that in practice we run heuristics that output a local minima (and not optimal solutions with global minimum), non-tight analysis (especially when it comes to constants), etc.

Our experiments in Section 5 show that, empirically, the constructed coresets are significantly smaller: for  $N := nm \sim 140,000$ ,  $k = 1000$ , and  $\varepsilon = 0.2$ , Theorem 8 predicts, in the worst case, a coreset of size larger than the full dataset size  $N$ . However, such an  $\varepsilon$  error is obtained with a coreset of size at most 1% of the input; see Fig. 4.

## 5 Experimental Results

We implemented our coreset construction from Algorithm 3 in Python 3.7, and in this section we evaluate its empirical results, both on synthetic and real-world datasets. More results are placed in the supplementary material; see Section A. Open-source code can be found in [35]. The hardware used was a standard MSI Prestige 14 laptop with an Intel Core i7-10710U and 16GB of RAM. Since our coreset construction algorithm does not compete with existing solvers, but improves them by reducing their input as a pre-processing step, we apply existing solvers as a black box on the small coreset returned by Algorithm 3. The results show that our coreset can boost, by up to x10 times, the running time and storage cost of common random forest implementations.

**Implementations for forests.** We used the following common implementations: (i) the function `RandomForestRegressor` from the `sklearn.ensemble` package, and (ii) the function `LGBMRegressor` from the `lightGBM` package that implements a forest of gradient boosted trees. Both functions were used with their default hyperparameters, unless states otherwise.

**Data summarizations.** We consider the following compression schemes:

(i): `DT-coreset` $(D, k, \varepsilon)$  - The implementation based on Algorithm 3. In all experiments we used a constant  $k = 2000$  for computing the coreset, regardless of the (larger) actual  $k$  value in each test, since  $k = 2000$  was sufficient to obtain a sufficiently small empirical approximation error. Hence, the parameter  $\varepsilon$  controls the trade-off between size and accuracy.

(ii): `RandomSample` $(D, \tau)$  - returns a uniform random sample of size  $\tau$  from  $D$ . In all tests  $\tau$  was set to the size of the coreset `DT-coreset` $(D, k, \varepsilon)$  for fair comparison.

**Datasets.** We used the following pair of datasets from the public UCI Machine Learning Repository [3], each of which was normalized to have zero mean and unit variance for every feature:

(i): **Air Quality Dataset** [18] - contains  $n = 9358$  instances and  $m = 15$  features.

(ii) **Gesture Phase Segmentation Dataset** [45] - contains  $n = 9900$  instances and  $m = 18$  features.

**The experiment.** The goal was to predict missing entries in every given dataset, by training random forests on the available data. The test set (missing values) consists of 30% of the dataset, and was extracted from the input dataset matrix by randomly and uniformly choosing a sufficient number of  $5 \times 5$  patches in the input dataset, and defining them as missing values. The final loss of a trained forest is the sum of squared distances between the forest predictions for the missing values, and the ground truth values. To tune the hyperparameter  $k$ , we randomly generate a set  $\mathcal{K}$  of possible values for  $k$  on a logarithmic scale. Then, we either: (i) apply the standard tuning (train the forest on the full data, for each value in  $\mathcal{K}$ , and pick the one with the smallest test set error), or (ii) compress the input (only once) into a small representative set, and then apply the standard tuning on the small, rather than the full, data. The experiment was repeated 10 times. All the results are averaged over all 10 tests; see Fig. 4.

**Discussion.** While the size and accuracy of our coreset are independent of our exact implementation of Algorithm 3, the running time is heavily based on our naive implementation, as compared to the very efficient professional Python libraries. This explains why most of the running time is still devoted to the coreset construction rather than the forest training. Nevertheless, even our simple implementation yielded improvements of up to x10 in both computational time and storage, for a relatively small accuracy drop of 0.03 in the SSE. Tuning more than one hyperparameter will result in a bigger improvement. Furthermore, Fig. 4 empirically shows that tuning a hyperparameter on the

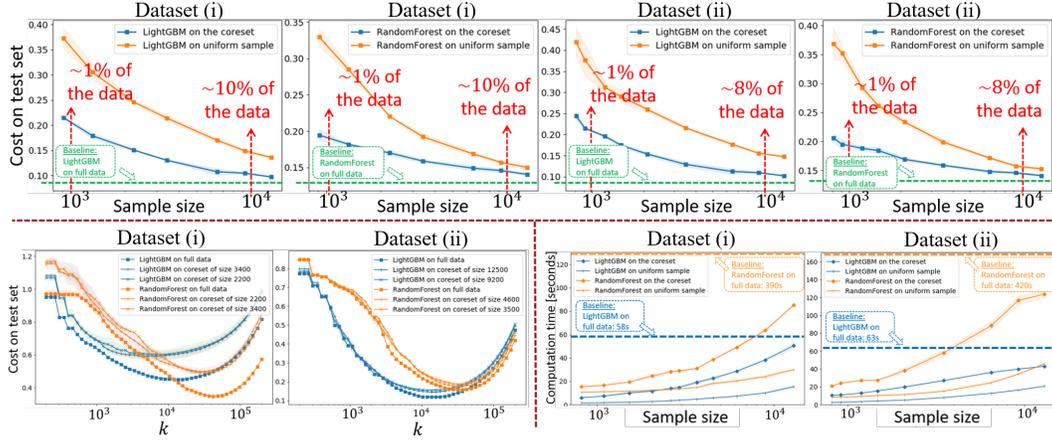


Figure 4: **Experimental results. (Top):** The  $X$ -axis is the compression size. For every compression size  $\gamma$ , hyperparameter tuning is applied on both the coreset and the uniform sample (which are both of size  $\gamma$ ). A random forest is then trained, on the full data, using those tuned parameters. The  $Y$ -axis presents the test set SSE loss of the trained forests. **(Bottom left):** Hyperparameter tuning. For every different value of  $k$  ( $X$ -axis), a forest is trained using this parameter value either on the compression (of two different sizes) or on the full data. The  $Y$ -axis presents  $\ell + k/10^5$ , where  $\ell$  is the normal SSE loss of the trained forest on the test set. **(Bottom right):** Time comparison. The  $Y$ -axis presents the total running time of both to compute the compression and to tune the parameter  $k$  on the compression (out of 50 different values). Note that the bottom right figures measure the total time to tune the parameter  $k$  in the bottom left figures, but using many more compression sizes. The optimal obtained parameter was then used to train the random forest in the top figures.

coreset yields a loss curve very similar to the loss curve of tuning on the full data. Lastly, we observe that, in practice, our coresets have size much smaller than predicted in the pessimistic theory.

## 6 Conclusions and Future Work

While coresets for  $k$ -trees do not exist in general, we provided an algorithm that computes such a coreset for every input  $n \times m$  signal. The coreset size depends polynomially on  $k \log(nm)/\epsilon$  and can be computed in  $O(nmk)$  time. Our experimental results on real and synthetic datasets demonstrates how to apply existing forest implementations and tune their hyperparameters on our coreset to boost their running time and storage cost by up to  $\times 10$ . In practice our coreset works very well also on non-signal datasets, probably since they have “real-world” properties that do not exist in the artificial worst-case example from Section 1.2. An open problem is to define these properties. Moreover, while this paper focuses on the sum of squares distances loss, we expect that the results can be generalized to support other loss functions; see Section 1.2. Lastly, supporting high-dimensional data (tensors), instead of matrices, is a straightforward generalization that can be achieved via minor modifications to our algorithms. We also leave this to future work.

## Acknowledgments and Disclosure of Funding

This research was supported by The ISRAEL SCIENCE FOUNDATION, grant number 379/21.

## References

- [1] Pankaj K Agarwal, Graham Cormode, Zengfeng Huang, Jeff M Phillips, Zhewei Wei, and Ke Yi. Mergeable summaries. *ACM Transactions on Database Systems (TODS)*, 38(4):1–28, 2013.
- [2] Pankaj K Agarwal, Sariel Har-Peled, Kasturi R Varadarajan, et al. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52:1–30, 2005.
- [3] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

- [4] Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coresets constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.
- [5] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.
- [6] Kristin P Bennett. Decision tree construction via linear programming. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 1992.
- [7] Jon Louis Bentley and James B Saxe. Decomposable searching problems i. static-to-dynamic transformation. *Journal of Algorithms*, 1(4):301–358, 1980.
- [8] Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.
- [9] Rafael Blanquero, Emilio Carrizosa, Cristina Molero-Río, and Dolores Romero Morales. Optimal randomized classification trees. In *Technical Report*. 2018.
- [10] Casper Solheim Bojer and Jens Peder Meldgaard. Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, 37(2):587–603, 2021.
- [11] Vladimir Braverman, Dan Feldman, and Harry Lang. New frameworks for offline and streaming coresets constructions. *arXiv preprint arXiv:1612.00889*, 2016.
- [12] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [13] Constantin Carathéodory. Über den variabilitätsbereich der koeffizienten von potenzreihen, die gegebene werte nicht annehmen. *Mathematische Annalen*, 64(1):95–115, 1907.
- [14] Venkatesan T Chakaravarthy, Vinayaka Pandit, Sambuddha Roy, Pranjal Awasthi, and Mukesh Mohania. Decision trees for entity identification: Approximation algorithms and hardness results. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 53–62, 2007.
- [15] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [16] Kenneth L Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):1–30, 2010.
- [17] Michael B Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 181–190, 2015.
- [18] Saverio De Vito, Ettore Massera, Marco Piga, Luca Martinotto, and Girolamo Di Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2):750–757, 2008.
- [19] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.
- [20] Abhimanyu Dubey, Moitrey Chatterjee, and Narendra Ahuja. Coreset-based neural network compression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 454–470, 2018.
- [21] Dan Feldman. Core-sets: Updated survey. *Sampling Techniques for Supervised or Unsupervised Tasks*, pages 23–44, 2020.
- [22] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 569–578, 2011.
- [23] Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering. *SIAM Journal on Computing*, 49(3):601–657, 2020.
- [24] Dan Feldman, Cynthia Sung, and Daniela Rus. The single pixel gps: learning big data signals from tiny coresets. In *Proceedings of the 20th international conference on advances in geographic information systems*, pages 23–32, 2012.
- [25] Dan Feldman and Tamir Tassa. More constraints, smaller coresets: Constrained matrix approximation of sparse big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 249–258, 2015.

- [26] Sarel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 291–300, 2004.
- [27] Melissa A Hardy. *Regression with dummy variables*, volume 93. Sage, 1993.
- [28] D Haussler and E Welzl. Epsilon-nets and simplex range queries. In *Proceedings of the Second Annual Symposium on Computational Geometry*, SCG '86, page 61–71, New York, NY, USA, 1986. Association for Computing Machinery.
- [29] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.
- [30] Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri, and Mohammed Erritali. A comparative study of decision tree id3 and c4. 5. *International Journal of Advanced Computer Science and Applications*, 4(2):13–19, 2014.
- [31] Xiyang Hu, Cynthia Rudin, and Margo Seltzer. Optimal sparse decision trees. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [32] Piotr Indyk, Sepideh Mahabadi, Mohammad Mahdian, and Vahab S Mirrokni. Composable core-sets for diversity and coverage maximization. In *Proceedings of the 33rd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 100–108, 2014.
- [33] Ibrahim Jubran, Alaa Maalouf, and Dan Feldman. Overview of accurate coresets. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1429.
- [34] Ibrahim Jubran, Murad Tukan, Alaa Maalouf, and Dan Feldman. Sets clustering. In *International Conference on Machine Learning*, pages 4994–5005. PMLR, 2020.
- [35] Jubran, Ibrahim and Sanches, Ernesto and Newman, Ilan and Feldman, Dan. Open source code for the algorithms presented in this paper, 2021. Link for open-source code.
- [36] Kaggle. Kaggle website. <https://www.kaggle.com/>.
- [37] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017.
- [38] Michael Langberg and Leonard J Schulman. Universal  $\varepsilon$ -approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 598–607. SIAM, 2010.
- [39] Hyafil Laurent and Ronald L Rivest. Constructing optimal binary decision trees is np-complete. *Information processing letters*, 5(1):15–17, 1976.
- [40] Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.
- [41] Hanlin Lu, Ming-Ju Li, Ting He, Shiqiang Wang, Vijaykrishnan Narayanan, and Kevin S Chan. Robust coreset construction for distributed machine learning. *IEEE Journal on Selected Areas in Communications*, 38(10):2400–2417, 2020.
- [42] Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. Training gaussian mixture models at scale via coresets. *The Journal of Machine Learning Research*, 18(1):5885–5909, 2017.
- [43] Alaa Maalouf, Ibrahim Jubran, and Dan Feldman. Fast and accurate least-mean-squares solvers. In *Advances in Neural Information Processing Systems*, pages 8305–8316, 2019.
- [44] Alaa Maalouf, Ibrahim Jubran, Murad Tukan, and Dan Feldman. Faster pac learning and smaller coresets via smoothed analysis. *arXiv preprint arXiv:2006.05441*, 2020.
- [45] Renata CB Madeo, Clodoaldo AM Lima, and Sarajane M Peres. Gesture unit segmentation using support vector machines: segmenting gestures from rest positions. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 46–52, 2013.
- [46] Detlev Marpe, Thomas Wiegand, and Gary J Sullivan. The h. 264/mpeg4 advanced video coding standard and its applications. *IEEE communications magazine*, 44(8):134–143, 2006.
- [47] Alexander Munteanu, Chris Schwiigelshohn, Christian Sohler, and David Woodruff. On coresets for logistic regression. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

- [48] Soliman Nasser, Ibrahim Jubran, and Dan Feldman. Autonomous toy drone via coresets for pose estimation. *Sensors*, 20(11):3042, 2020.
- [49] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [50] Jeff M Phillips. Coresets and sketches. *arXiv preprint arXiv:1601.00617*, 2016.
- [51] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.
- [52] Lior Rokach and Oded Maimon. Decision trees. In *Data mining and knowledge discovery handbook*, pages 165–192. Springer, 2005.
- [53] Frédéric Ros and Serge Guillaume. *Sampling techniques for supervised or unsupervised tasks*. Springer, 2020.
- [54] Guy Rosman, Mikhail Volkov, Danny Feldman, John W Fisher III, and Daniela Rus. Coresets for k-segmentation of streaming data. In *Neural Information Processing Systems Foundation*, 2014.
- [55] Eli Shusterman and Meir Feder. Image compression via improved quadtree decomposition algorithms. *IEEE Transactions on Image Processing*, 3(2):207–215, 1994.
- [56] sklearn. sklearn user guide - decision trees. <https://scikit-learn.org/stable/modules/tree.html>.
- [57] Morad Tukan, Alaa Maalouf, and Dan Feldman. Coresets for near-convex functions. *Advances in Neural Information Processing Systems*, 33, 2020.
- [58] Bhekisipho Twala. An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, 23(5):373–405, 2009.
- [59] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.
- [60] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.
- [61] Sicco Verwer and Yingqian Zhang. Learning optimal classification trees using a binary linear program formulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1625–1632, 2019.
- [62] Mikhail Volkov, Guy Rosman, Dan Feldman, John W Fisher, and Daniela Rus. Coresets for visual summarization with applications to loop closure. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3638–3645. IEEE, 2015.

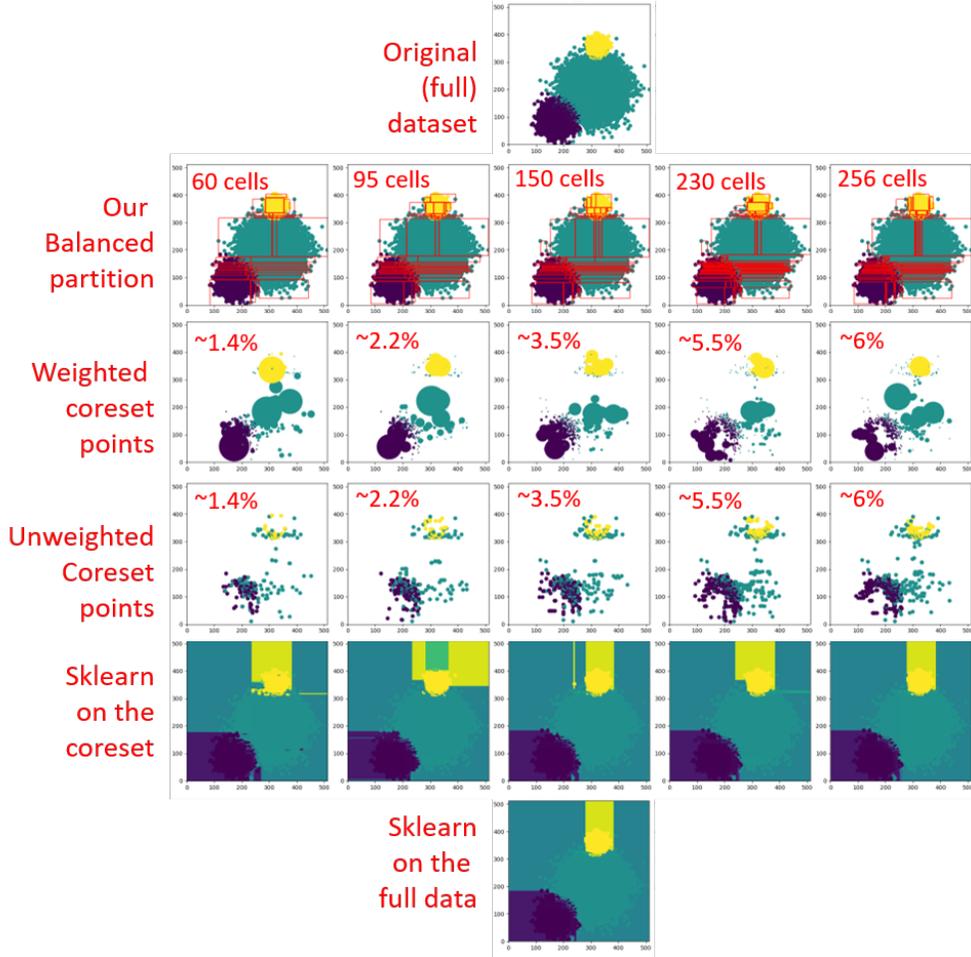


Figure 5: The blobs dataset. The dataset  $D$  was generated using the function `sklearn.datasets.make_blobs`, and contains  $n = 17,000$  points clustered into 3 clusters (containing 8500, 5800, and 2700 points), each with a different label. A coreset  $(C, u)$  was constructed using Algorithm 3. From top down, the rows illustrate: (i) The input dataset  $D$ , (ii) The balanced partition of  $D$ , including the number of sets in the partition (iii) The weighted coreset points. Each point  $(x, y) \in C$  is plotted at location  $x$ , colored according to its label  $y$ , and its radius is proportional to its weight  $u(x, y)$ . The percentage of the coreset size relative to the full data is presented. (iv) The unweighted coreset points. Each point  $(x, y) \in C$  is plotted at location  $x$ , colored according to  $y$ , and has a fixed radius. The percentage of the coreset size relative to the full data is presented. (v) The partition of the space via a decision tree computed using a call to `DecisionTreeRegressor` from the `sklearn.tree` package, where the input was the weighted coreset points only. Each region is assigned a color according to the label assigned to it by the computed tree. (vi) Similar to Row (v), but where the decision tree is trained on the full data. Algorithm 5 was used during training of the decision tree to evaluate the loss of each model.

## A Additional Experiments

In this section, we present some additional experiments conducted using our algorithm from Sections 3-4. We give visual illustration both of our coreset itself and of the result of applying a very common decision tree implementation on the coreset, as compared to running the same function on the original (full) data; see Fig. 5-6.

**Discussion. Visual representation.** As seen in Fig. 5-6, the balanced partition in the second row partitions the input data into multiple subsets, where, as expected, flat and relatively smooth regions are partitioned into a smaller number of large cells, while more complex regions are partitioned into a

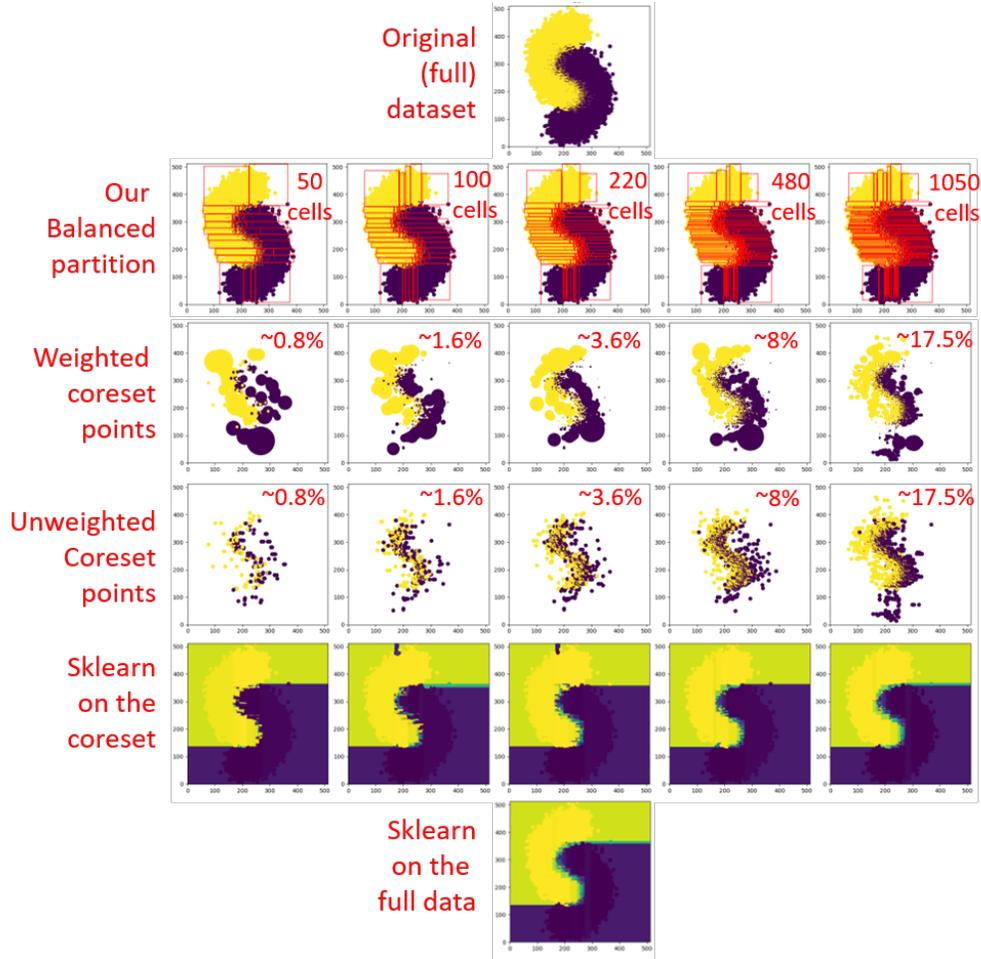


Figure 6: The moons dataset. The dataset was generated using the function `sklearn.datasets.make_moons`. The dataset contains  $n = 24,000$  points spread across two interleaving half circles (12,000 points for each half circle), each with a different label. See caption of Fig. 5 for a detailed explanation about the rows.

larger number of finer cells. This is expected since the balanced partition insures a small variance inside each cell.

Furthermore, as seen in the third row of the above figures, the weighted coreset contains a small number of “large” circles (points with large weight) in the flat and relatively smooth regions, while it contains large number of “small” circles (points with small weight) in the more complex regions of the input.

**Accuracy.** As seen in the last two rows of Fig. 5-6, the decision tree trained only on the coreset points resembles the decision tree trained on the full data, even for coresets of size only 6%, 8%, and 14% of the input data, as seen in Fig. 5,6, and 7 respectively. This implies a x10 faster training time of a decision tree (or, similarly, a forest) on a given coreset, compared to training it on the full data, with almost no compromises to the accuracy.

The difference in the coreset size required in order to accurately represent the full data depends on the complexity of the input dataset. Indeed, the dataset in Fig. 5 is a much simpler dataset for a decision tree to classify, compared to the dataset in Fig. 7. Hence, the coreset sizes required in Fig. 5 are smaller than the ones in Fig. 7.

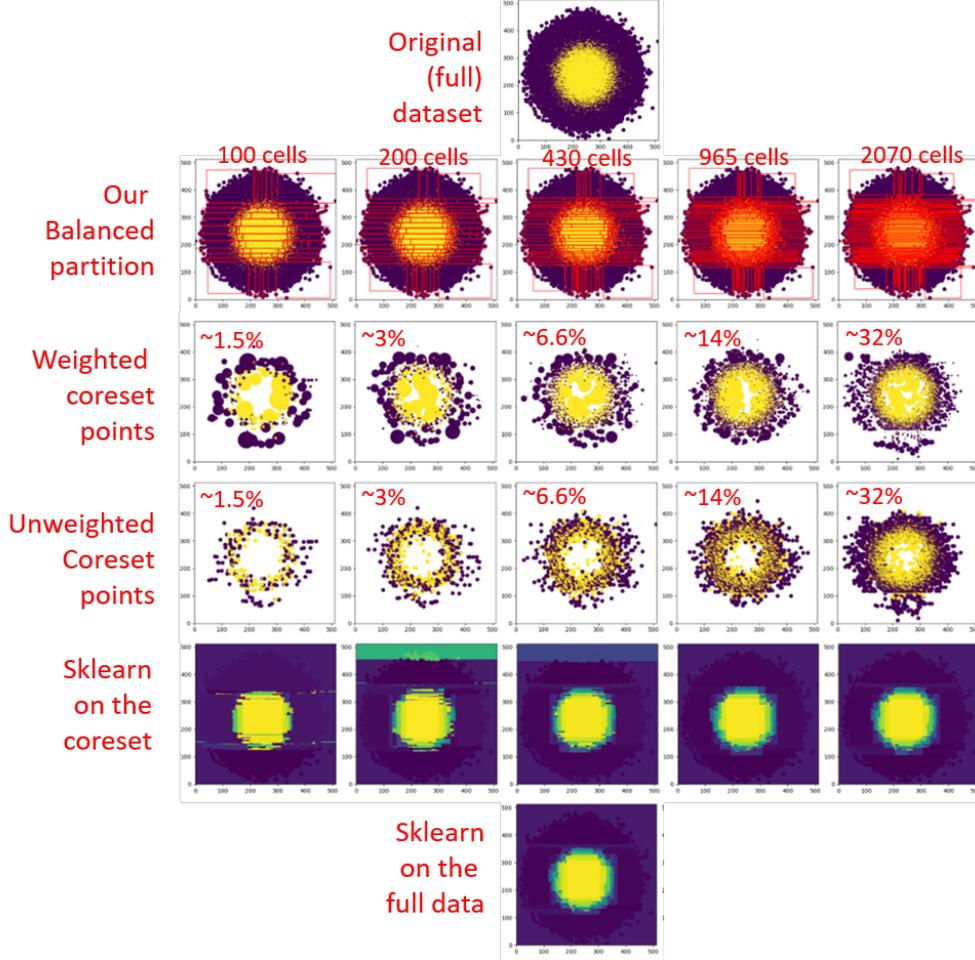


Figure 7: The circles dataset. The dataset was generated using the function `sklearn.datasets.make_circles`. The dataset contains  $n = 26,000$  points spread across a big circle (14,000 points) and a small circle (12,000 points), each with a different label. See caption of Fig. 5 for a detailed explanation about the rows.

## B Bi-criteria Approximation

**Notations.** A sub-signal  $B$  is said to be *horizontally intersected* by a  $k$ -segmentation function  $s$  if there are  $((i_1, j_1), y_1), ((i_2, j_2), y_2) \in B$  where  $i_1 \neq i_2$  such that  $s(i_1, j_1) \neq s(i_2, j_2)$ . Similarly, a block  $B$  of  $D$  is said to be *vertically intersected* by  $s$  if there are  $((i_1, j_1), y_1), ((i_2, j_2), y_2) \in B$  where  $j_1 \neq j_2$  such that  $s(i_1, j_1) \neq s(i_2, j_2)$ .  $B$  is said to be *intersected* by  $s$  if  $B$  is either horizontally or vertically intersected, i.e.,  $|\{s(x) \mid (x, y) \in B\}| > 1$ . A set of sub-signals  $\mathcal{B}$  is said to be horizontally (vertically) intersected by  $s$  if it contains a sub-signal  $B \in \mathcal{B}$  that is horizontally (vertically) intersected by  $s$ .

Also, we might abuse notation and denote by signal (sub-signal) an  $n \times m$  signal (sub-signal) and by  $k$ -segmentation an  $n \times m$   $k$ -segmentation.

In this section we give a constructive proof for Lemma 5. A suggested implementation for this constructive proof is given in Algorithm 4.

We first prove a small technical observation (see Observation 9), and then we prove Lemma 10, which will be used throughout the proof of Lemma 5.

**Observation 9.** Let  $A$  and  $B$  be two  $n \times m$  sub-signals. Then it holds that

$$\text{opt}_1(A \cup B) \geq \text{opt}_1(A) + \text{opt}_1(B).$$

---

**Algorithm 4:** BICRITERIA( $D, k$ ); Lemma 5

---

**Input** : An  $n \times m$  sub-signal  $D = \{(x_i, y_i)\}_{i=1}^N$  and an integer  $k \geq 1$ .  
**Output** : An  $(\alpha, \beta)_k$ -approximation for  $D$ .

- 1  $\mathcal{B} := \emptyset$
- 2  $\nu, \gamma :=$  sufficiently large constants // see proof of Lemma 5
- 3 **while**  $|D| > k \log N$  **do**
- 4     **if**  $D$  contains a row  $R$  with  $|R| \geq \frac{|D|}{\nu k}$  **then**
- 5         Partition  $[m]$  into  $t' = \gamma k$  intervals  $[m] = \cup_{j=1}^{t'} I_j$  such that every  $j \in [t']$ , the size of each corresponding sub-signal  $R_j = \{(x_1, x_2), y) \in R \mid x_2 \in I_j\}$  is  $|R_j| \in \left\{ \frac{|R|}{t'} - 1, \frac{|R|}{t'} + 1 \right\}$ . // e.g., by a greedy pass over  $[m]$ .
- 6          $\mathcal{B} :=$  the set of  $t' - 2k$  signals  $R_j$  with the smallest  $\text{opt}_1(R_j)$ .
- 7     **else**
- 8         Partition  $[n]$  into  $\psi$  intervals  $[n] = \cup_{j=1}^{\psi} I_j$  such that for every  $j \in [\psi]$ , the size of each corresponding sub-signal  $D_j = \{(x_1, x_2), y) \in D \mid x_1 \in I_j\}$  is  $\frac{|D|}{\nu k} \leq |D_j| \leq \frac{2|D|}{\nu k}$ . // e.g., by a greedy pass over  $[n]$ .
- 9         **if** at least  $\psi/2$  of the sub-signals  $D_j$  do not contain a column  $col$  of size  $|col| \geq \frac{|D_j|}{2(\nu k)^2}$  **then**
- 10             Vertically partition each of the (at least  $\psi/2$ ) sub-signals  $D_j$  into  $\psi_j$  sub-signals, each such sub-signal  $B$  of size  $\frac{|D_j|}{2(\nu k)^2} \leq |B| \leq \frac{|D_j|}{(\nu k)^2}$ , and let  $\mathcal{B}'$  contain the union of all those sub-signals. e.g., via a greedy algorithm.
- 11              $\mathcal{B} :=$  the set of  $|\mathcal{B}'| - 4\nu^2 k^3 - 2k\psi$  signals  $B \in \mathcal{B}'$  with the smallest  $\text{opt}_1(B)$ .
- 12         **else**
- 13              $\mathcal{B} := \left\{ C \mid C \text{ is a column of } D_j, |C| \geq \frac{|D_j|}{2(\nu k)^2} \text{ and } j \in [\psi] \right\}$
- 14          $D := D \setminus \cup_{B \in \mathcal{B}'} B$  and  $\mathcal{B} := \mathcal{B} \cup \mathcal{B}$
- 15      $s(b) := 1/|B| \sum_{(x,y) \in B} y$  for every  $b \in B$  and  $B \in \mathcal{B}$ .
- 16 **return**  $s$

---

*Proof.* Let  $C = A \cup B$ . Let  $\mu = \frac{1}{|C|} \sum_{(x,y) \in C} y$  be the weighted mean of  $A \cup B$ . By Definition of  $\text{opt}$  we have that

$$\begin{aligned}
 \text{opt}_1(A \cup B) &= \sum_{(x,y) \in C} (y - \mu)^2 \\
 &= \sum_{(x,y) \in A} (y - \mu)^2 + \sum_{(x,y) \in B} (y - \mu)^2 \\
 &\geq \text{opt}_1(A) + \text{opt}_1(B),
 \end{aligned}$$

where the first derivation holds since the mean of a points minimizes the sum of squared distances to those points, and the last derivation is by the definition of  $\text{opt}_1$ .  $\square$

**Lemma 10.** Let  $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$  be an  $n \times m$  sub-signal and let  $k \geq 1$  be an integer. Then, in  $O(N)$  time we can find a set  $\mathcal{B}$  of  $|\mathcal{B}| = t \in O(k^3)$  mutually disjoint blocks with respect to  $D$ , for which

- (i)  $\sum_{B \in \mathcal{B}} \text{opt}_1(B) \leq \text{opt}_k(D)$ .
- (ii)  $|\cup_{B \in \mathcal{B}} B| \in \Omega\left(\frac{N}{k}\right)$ .

*Proof.* Let  $\nu > 50$  be an arbitrary parameter and let  $\gamma \geq 8$  be a parameter that will be defined later. We will prove Lemma 10 for  $t \leq 2\nu^3 k^3$  and for  $|\cup_{B \in \mathcal{B}} B| \geq \frac{N}{8\nu k}$ .

We start with the simple 1-dimensional case, namely – we assume that  $m = 1$ . In this case, we just partition  $[n]$  into  $t' = \gamma k$  consecutive intervals  $[n] = \cup_{j=1}^{t'} E_j$ , such that each corresponding sub-signal  $D_j = \{(a, b), y) \in D \mid a \in E_j\}$ ,  $j \in [t']$  of  $D$  has equal share of elements in  $D$  (up to  $\pm 1$ ), i.e.,

$|D_j| \in \{|D|/t' - 1, |D|, |D|/t' + 1\}$ . This can be done by moving point by point along the elements of  $D$ , which are assumed to be sorted in ascending order according to  $a$  for every  $((a, b), y) \in D$ , and defining a new interval at the first moment the current interval contains more than  $|D|/t'$  elements of  $D$  or just at the prior point. We then compute for each  $D_j$ ,  $j \in [t']$  the optimal  $\text{opt}_1(D_j)$ , and return as output the set  $\mathcal{B}$  containing the  $t = t' - 2k$  sub-signals  $D_j$  with the smallest loss  $\text{opt}_1(D_j)$ , among all the  $t'$  sub-signals  $\{D_1, \dots, D_{t'}\}$ .

To see what guarantees we get, note that the computation takes  $O(nm) = O(N)$  time to sort the elements  $((a, b), y)$  of  $D$  according to  $a$ , since  $a \in [nm]$  is a bounded integer. Afterwards, the above greedy partition also takes  $O(nm) = O(N)$  time. Furthermore,

$$|\cup_{B \in \mathcal{B}} B| \geq (\gamma - 2)k \cdot \frac{|D|}{\gamma k} \geq \frac{\gamma - 2}{\gamma} |D| \geq \frac{N}{8\nu k}.$$

For  $\gamma \geq 8$  this proves Property (ii) above.

Finally, any  $n \times 1$   $k$ -segmentation function intersects at most  $2k$  sub-signals from  $\{D_j\}_{j=1}^{t'}$  (i.e., at most  $2k$  sub-signals are assigned more than 1 distinct value via the  $k$ -segmentation function). Hence, the optimal  $k$ -segmentation  $s^*$  of  $D$  intersects at most  $2k$  of those sub-signals as well. This implies that at least  $t' - 2k \geq (\gamma - 2)k$  of the intervals  $\{D_j\}_{j=1}^{t'}$  which are assigned 1 distinct value  $|\{s(x) \mid (x, y) \in D_j\}| = 1$ . Hence, by Observation 9, we conclude that  $\text{opt}_k(D) \geq \sum_{B \in \mathcal{B}} \text{opt}_1(D)$  which verifies the 1st item above.

We remark that for the 1-dimensional case we can do much better (there is an overall  $(1 + \varepsilon)$ -approximation of the  $k$ -segmentation using logarithmic number of blocks), but this will not be used here. Another remark is that we have ignored the  $\pm 1$  slack in the sizes above, making the actual part of  $D$  that is removed at least  $\frac{\gamma-2}{\gamma}|D| - t'$ . This is insignificant in the 1-dimensional case above, as for  $t = O(1)$  and for  $|D| \geq \log n$  this would be an insignificant fraction, while for smaller  $D$ , we can just use single point sub-signals.

**The 2-dim case :** Consider a row  $R$  of  $D$ , say the  $i$ 'th row  $R = \{(a, b), y) \in D \mid a = i\}$ . We call a row  $R$  of  $D$   $r$ -heavy if  $|R| \geq |D|/r$ , namely  $R$  contains at least  $|D|/r$  elements from  $D$ . Analogously, we define a column to be  $r$ -heavy.

Assume first that our  $D$  contains a  $\nu k$ -heavy row  $R$ . We choose  $R$  and use it as in the 1-dimensional case. As explained above we can find in  $R$  a set of disjoint sub-signals  $\mathcal{B}$  containing  $\gamma k$  blocks and for which the first item above holds for  $\text{opt}_k(R)$  and in particular for  $\text{opt}_k(D)$ , i.e.,  $\sum_{B \in \mathcal{B}} \text{opt}_1(B) \leq \text{opt}_k(R) \leq \text{opt}_k(D)$ . Further, using the above guarantees for the 1-dim case,  $|\cup_{B \in \mathcal{B}} B| \geq \frac{\gamma-2}{\gamma} \cdot |R| \geq \frac{\gamma-2}{\gamma} \cdot \frac{|D|}{\nu k}$ . This proves 2nd item for this case (with  $\gamma \geq 3$ ).

Otherwise, let  $e_i = |R_i|$  where  $R_i = \{(a, b), y) \in D \mid a = i\}$ . By our assumption  $e_i \leq |D|/\nu k$  for every  $i \in [n]$ . Our algorithm is essentially identical to the 1-dimensional case, on the 1-dim array  $L = (e_1, \dots, e_n)$  where we weight the  $i$ th element by  $e_i$ . Namely, we find a partition of  $L$  into  $\psi$  contiguous subintervals  $\mathcal{E} = \{E_1, \dots, E_\psi\}$  such that the corresponding sub-signals  $D_j = \{(a, b), y) \in D \mid a \in E_j\}$  of  $D$  are as equal as possible. By our assumption this could be done so that for any  $j \in \psi$ , the number of elements in  $D_j$  is between  $\frac{|D|}{\nu k}$  and at most  $\frac{2|D|}{\nu k}$ . This is since adding a new 'point' from the list to an existing interval may increase the sum by at most  $|D|/\nu k$ . this implies that  $\nu k/2 \leq \psi \leq \nu k$ .

Next we perform the above algorithm again on  $D_1, \dots, D_\psi$  with the intention to "vertically partition" each such  $D_j$ , i.e., split each  $D_j$  into sets according to the value  $b$  for every  $((a, b), y) \in D_j$  (rather than considering the value  $a$  above). Let  $r = 2\nu^2 k^2$ , we continue with the following case analysis: (i) at least a  $\frac{1}{2}$ -fraction of  $\{D_1, \dots, D_\psi\}$  contain no a  $r$ -heavy column, and (ii) at most a  $\psi/2$  of  $\{D_1, \dots, D_\psi\}$  contain no a  $r$ -heavy column.

**Case (i):** At least a  $\frac{1}{2}$ -fraction of  $\{D_1, \dots, D_\psi\}$  contain no a  $r$ -heavy column. Then we partition each set  $D_j$  with no  $r$ -heavy column into  $\psi_i$  sub-signals  $\{D_j^{(1)}, \dots, D_j^{(\psi_i)}\}$  of nearly equal number of points, where the partition is applied onto the values  $b$  of every  $((a, b), y) \in D_j$ . By a reasoning similar to that above, each  $r/2 \leq \psi_i \leq r$ , and each such block  $B \in \{D_j^{(1)}, \dots, D_j^{(\psi_i)}\}$  contains  $|D_i|/r \leq |B| \leq 2|D_i|/r$ . Using the bounds on  $|D_i|$  and  $r$  we get  $\frac{|D|}{2\nu^3 k^3} \leq |B| \leq \frac{2|D|}{\nu^3 k^3}$ . In particular

we conclude that the total number of such sub-signals is at most  $t'$ , and  $t' \leq 2\nu^3 k^3$ . On the other hand,  $t' \geq \frac{\psi}{2} \frac{|D|}{\nu k} \cdot \frac{\nu^3 k^3}{2|D|} \geq \nu^3 k^3 / 8$ . Let  $\mathcal{B}_2$  be the collection of these sub-signals.

We choose for our output collection,  $\mathcal{B}$ , the set of  $t' - z$  sub-signals  $B \in \mathcal{B}_2$  with the smallest  $\text{opt}_1(B)$ , for  $z = 2k(r + \psi) \leq 6k^3 \nu^2$ .

We note that  $\mathcal{B}$  contains  $t$  sub-signals, where  $t' - z \leq t \leq t'$ . Further, by the lower bound on  $t$  it follows that  $|\cup_{B \in \mathcal{B}} B| \geq (t' - z) \cdot \frac{|D|}{2\nu^3 k^3} \geq (\nu^3 k^3 / 8 - 6k^3 \nu^2) \frac{|D|}{2\nu^3 k^3} \geq \frac{|D|}{8} (1 - 48/\nu)$ .

This verifies the 2nd item for this case. Finally, we note that any row of  $D$  is shared by at most  $r$  sub-signals of  $\mathcal{B}$ , and each column of  $D$  is shared by at most  $\psi$  sub-signals of  $\mathcal{B}$ . Hence, any  $k$ -segmentation function may intersect at most  $z = 2k(r + \psi)$  sub-signals from  $\mathcal{B}$ . Therefore, for any  $k$ -segmentation  $s$  there are at least  $t' - z$  sub-signals in  $\mathcal{B}_2$  which are not-intersected by  $s$ . By our definition of  $\mathcal{B}$  to be the set of  $t' - z$  sub-signals in  $\mathcal{B}_2$  with the smallest loss, we obtain that the loss  $\ell(D, s)$  is at least  $\sum_{B \in \mathcal{B}} \ell(B)$  proving the 1st item of the lemma for this case.

**Case (ii):** At most  $\psi/2$  of  $\{D_1, \dots, D_\psi\}$  contain no a  $r$ -heavy column, namely – at least  $\psi/2$  of the  $D_i$  have a  $r$ -heavy column. In this case we take the heavy column from each  $D_i$  as its own sub-signal. We get a collection  $\mathcal{B}_1$  of  $\psi_1 \geq \psi/2$  blocks. We now return as output the set  $\mathcal{B}$  of the  $\psi_1 - 2k$  sub-signals  $B \in \mathcal{B}_1$  with the smallest  $\text{opt}_1(B)$ . We note that number of blocks we output in this case is at most  $\psi \leq \nu k$ .

Note also that  $\mathcal{B}$  contains at least  $\psi/2$  blocks, it follows that  $|\cup_{B \in \mathcal{B}} B| \geq \frac{\psi}{2} \cdot \frac{|D|}{2\nu^2 k^2} \geq \frac{|D|}{8\nu k}$  which proves the 2nd item in the lemma.

Finally, note that any  $k$ -segmentation  $s$  can intersect at most  $2k$  intervals from  $\mathcal{B}_1$  (similarly to the 1-dim case). Hence, there are at least  $|\mathcal{B}_1| - 2k = \psi_1 - 2k$  sub-signals in  $\mathcal{B}_1$  that are not-intersected by  $s$ , which implies that its loss  $\ell(D, s)$  is at least the sum  $\sum_{B \in \mathcal{B}} \text{opt}_1(B)$ , which proves the 1st item of this lemma.

**Remark:** we did not optimise the parameter. A slightly better partition can be obtains (less blocks), but this is good enough for our purposes.

**Computational time:** Note that the elements of the input  $D$  can be sorted in lexicographic order in  $O(nm) = O(N)$  time since the coordinates  $a$  and  $b$  for every  $((a, b), y) \in D$  are bounded integers. Then, a linear-time preprocessing can be applied to the input  $D$  to store some statistics, e.g., the number of elements in each non-empty row or column, and the index of the next non-empty row or column for every element in  $D$ . Afterwards, the above greedy partition also takes  $O(nm) = O(N)$  time.  $\square$

We now restate and prove Lemma 5 from Section 2.

**Lemma 11** (Lemma 5). *Let  $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$  be an  $n \times m$  signal and  $k \geq 1$  be an integer. Then, in  $O(kN)$  time we can compute an  $(\alpha, \beta)_k$ -approximation for  $D$ , where  $\alpha \in O(k \log N)$  and  $\beta \in O(k^{O(1)} \log^2 N)$ .*

The proof of the above claim is a constructive proof. A suggested implementation is provided in Algorithm 4.

*Proof.* The top level idea of the algorithm is as follows. We suggest an iterative algorithm. We start the first iteration with  $D_1 = D$  and, using Lemma 10, find a collection of disjoint sub-signals  $\mathcal{B}_1 = \{B_1, \dots, B_t\}$  in  $D_1$  (which will not necessarily cover the entire signal  $D_1$ ), such that: (i) the sum of their 1-segmentation loss satisfies,  $\sum_{i=1}^t \text{opt}_1(B_i) \leq \text{opt}_k(D_1) = \text{opt}_k(D)$ , and (ii)  $\cup_{B \in \mathcal{B}_1} B$  has size  $|\cup_{B \in \mathcal{B}_1} B| \geq |D_1|/c$  for some  $c$  (e.g.,  $c \in O(k)$  in the lemma). Let  $\mathcal{B}_1$  be such a collection. We then ‘delete’ from  $D_1$  the elements of  $\cup_{B \in \mathcal{B}_1} B$ , and set  $D_2 = D_1 \setminus \cup_{B \in \mathcal{B}_1} B$ .

In the  $i$ th iteration, we repeat the same process with respect to the current set  $D_i$ . Namely, we find a collection  $\mathcal{B}_i$  of at most  $t$  disjoint sub-signals in  $D$ , for which: (i)  $\sum_{B \in \mathcal{B}_i} \text{opt}_1(B) \leq \text{opt}_k(D_i) \leq \text{opt}_k(D)$ , and (ii)  $\cup_{B \in \mathcal{B}_i} B$  has size  $|\cup_{B \in \mathcal{B}_i} B| \geq |D_i|/c$ , i.e., those blocks cover at least a constant fraction of  $D_i$ .

Repeating these iterative procedure for at most  $\psi = O(c \log(nm))$  times, we end up covering all entries of  $D$  with sub-signals where the overall loss of the sub-signals in each iteration is at most

$\text{opt}_k(D)$ . This defines a collection of at most  $t\psi$  sub-signals  $\mathcal{B}$  that cover the entire original set  $D$ . Hence, the total overall loss over those sub-signals is  $\sum_{B \in \mathcal{B}'} \text{OPT}_1(B) \leq \psi \text{opt}_k(D)$ . By defining the output function  $s$  such that for every  $B \in \mathcal{B}$  and  $b \in B$ ,  $s$  assigns to  $b$  the mean value of  $B$ , i.e.,  $s(b) = 1/|B| \sum_{((i,j),y) \in B} y$ , we obtain that  $\ell(B, s) = \text{opt}_1(B)$  for every  $B \in \mathcal{B}$ , and that the  $\ell(D, s) \leq \sum_{B \in \mathcal{B}} \ell(B, s) \leq \psi \text{opt}_k(D)$ .

While for every  $B \in \mathcal{B}$   $s$  assigns the same value for every element  $b \in B$ , there is not necessarily a partition of  $n \times m$  into  $|\mathcal{B}| \in O(t\psi)$  distinct axis-parallel blocks that correspond to the sub-signals of  $\mathcal{B}$ . Therefore,  $s$  is not necessarily a  $|\mathcal{B}|$ -segmentation function. However, looking at all possible intersections of the sub-signals in  $\mathcal{B}$ , it is easy to realize that the  $t\psi$  sub-signals in  $\mathcal{B}$  define a partition of  $D$  into at most  $O(t^2\psi^2)$  sub-signals  $\mathcal{B}'$  that indeed correspond to a distinct partition of  $n \times m$  into  $|\mathcal{B}'|$  distinct axis-parallel blocks. Hence,  $s$  is guaranteed to be a  $|\mathcal{B}'|$ -segmentation function.

The parameters that are guaranteed by the lemma are  $c \in O(k)$ ,  $t \in O(k^3)$ . This implies that  $\beta = |\mathcal{B}'| \in O(t^2\psi^2) = O(k^8 \log^2 nm)$  and  $\alpha = \psi \in O(k \log nm)$ .

**Computational time:** By Lemma 10, each of the  $\psi$  iterations above takes time linear in the input size. The input size in the  $i$ 'th iteration is  $O(N((k-1)/k)^i)$  since at each iteration we remove at least a  $1/k$  fraction of the input. Hence, the total running time is the sum of the geometric series  $N \cdot \sum_{i \in \psi} ((k-1)/k)^i \in O(kN)$ .  $\square$

## C Balanced Partition

In this section we give our full proof for Lemma 7. We first prove the following lemma regarding the output of Algorithm 1.

**Lemma 12.** *Let  $D$  be an  $n \times m$  sub-signal, and  $\sigma > 0$  be a parameter. Let  $\mathcal{B} = \{B_1, \dots, B_{|\mathcal{B}|}\}$  be the output of a call to  $\text{SLICEPARTITION}(D, \sigma)$ , where the sub-signals in  $\mathcal{B}$  are numbered according to the order in which each of them was added to  $\mathcal{B}$ ; see Algorithm 1. Then the following properties hold:*

- (i)  $\mathcal{B}$  is a partition of  $D$ .
- (ii)  $\text{opt}_1(B) \leq \sigma$  for every sub-signal  $B \in \mathcal{B}$ .
- (iii) If  $|\mathcal{B}| > 8k$  then for any  $k$ -segmentation  $s$  that does not horizontally intersect  $D$  we have that  $\ell(D, s) \geq \left(\frac{|\mathcal{B}|}{4} - 2k\right) \sigma$ .
- (iv)  $\mathcal{B}$  can be computed in  $O(|D|)$  time.

*Proof.* We consider the variables defined in Algorithm 1.

**Proof of (i):** By construction it immediately follows that  $\mathcal{B}$  is a partition of  $D$ .

**Proof of (ii):** Consider a sub-signal  $B \in \mathcal{B}$ . We prove (ii) for each of the following cases: **Case (a):**  $B$  was added to  $\mathcal{B}$  at Line 12, and **Case (b):**  $B$  was either added to  $\mathcal{B}$  at Line 6

**Case (a):** In this case, by the condition at Line 9,  $B$  must satisfy that  $\text{opt}_1(B) \leq \sigma$ .

**Case (b):** In this case,  $B$  was returned via a recursive call. Hence, this case holds trivially by Case (a) above.

Therefore, (ii) above holds by combining Cases (a)–(b).

**Proof of (iii):** Let  $t = |\mathcal{B}|$  and assume for simplicity that  $t$  is an even number. Recall that the index of each sub-signal in  $\mathcal{B}$  indicates its order of insertion to  $\mathcal{B}$ , i.e.,  $B_1$  is the first sub-signal that was inserted to  $\mathcal{B}$  and  $B_t$  was the last such sub-signal to be inserted to  $\mathcal{B}$ .

Observe that each recursive call  $\mathcal{B}' := \text{SLICEPARTITION}(B^T, \sigma)$  at Line 5 returns at least  $|\mathcal{B}'| \geq 2$  sub-signals. This is because the recursive call happens only when  $\text{opt}_1(B^T) = \text{opt}_1(B) > \sigma$ , which can only happen if  $|\{(i, j) \mid ((i, j), y) \in B\}| > 1$ , i.e.,  $B^T$  exceeds the maximum tolerance, and can indeed be partitioned into sub-signals. Hence, there are at least  $t/4$  distinct pairs of consecutive sub-signals  $B_i$  and  $B_{i+1}$  that were both either computed via the recursive call or both were not computed via the recursive call. We now show that each such pair satisfies  $\text{opt}_1(B_i \cup B_{i+1}) > \sigma$ .

Consider a pair of consecutive sub-signals  $B_i$  and  $B_{i+1}$  that were both not computed via the recursive call at Line 5. Let  $B' \subseteq B_{i+1}$  contain the elements  $((i, j), y) \in B_{i+1}$  with the smallest value of  $i$  over all elements of  $B_{i+1}$ . By the greedy partition loop at Line 9 we obtain that  $\text{opt}_1(B_i \cup B') > \sigma$ . We now have that

$$\text{opt}_1(B_i \cup B_{i+1}) \geq \text{opt}_1(B_i \cup B') > \sigma,$$

where the first inequality is by Claim 9.

Consider a pair of consecutive sub-signals  $B_i$  and  $B_{i+1}$  that were both computed via the recursive call at Line 5. Then, similarly to the previous argument, we obtain that  $\text{opt}_1(B_i \cup B_{i+1}) > \sigma$ .

Now, let  $s$  be a  $k$ -segmentation that does not horizontally intersect  $\mathcal{B}$ , i.e., it does not horizontally intersect any  $B \in \mathcal{B}$ . By the definition of  $s$ , there might be at most  $2k$  sub-signals in  $\mathcal{B}$  which are vertically intersected by  $s$ . Hence, among the  $t/4$  distinct consecutive pairs of sub-signals discussed above there are at least  $t/4 - 2k$  such pairs that are not intersected by  $s$ .

Since  $\mathcal{B}$  is a partition of  $D$ , we have that  $\ell(D, s)$  is at least the sum of  $\text{opt}_1(B_i \cup B_{i+1}) \geq \sigma$ , over the above  $t/4 - 2k$  non-intersected pairs of sub-signals. Hence,

$$\ell(D, s) \geq \left(\frac{t}{4} - 2k\right) \sigma = \left(\frac{|\mathcal{B}|}{4} - 2k\right) \sigma.$$

**Proof of (iv):** The greedy Algorithm 1 can be implemented so that it computes only  $O(|D|)$  operations. The most costly operation is the computation of  $\text{opt}_1(B)$  for some sub-signal  $B$ . We now argue that this can be computed in  $O(1)$  time. Let  $B$  be a sub-signal and let  $\mu_B = 1/|B| \sum_{(x,y) \in B} y$  be its mean value. Observe that

$$\begin{aligned} \text{opt}_1(B) &= \sum_{(x,y) \in B} (y - \mu_B)^2 \\ &= \sum_{(x,y) \in B} y^2 + |B| \cdot \mu_B - 2\mu_B \sum_{(x,y) \in B} y. \end{aligned} \tag{1}$$

By precomputing and storing some statistics at each of the signal's elements, then the three terms on the right hand side of (1) can all be evaluated in  $O(1)$  time for any sub-signal  $B$ . Hence, the total running time of Algorithm 1 is linear in the input size. □

We now restate and prove Lemma 7 from Section 3.

**Lemma 13.** *Let  $D$  be an  $n \times m$  signal,  $k \geq 1$  be an integer,  $\varepsilon \in (0, 1/4)$  be an error parameter, and  $s : [n] \times [m] \rightarrow \mathbb{R}$  be an  $(\alpha, \beta)_k$ -approximation of  $D$ . Define  $\sigma := \frac{\ell(D, s)}{\alpha}$  and  $\gamma := \frac{\varepsilon^2}{\beta k}$ . Then algorithm  $\text{PARTITION}(D, \gamma, \sigma)$  outputs a partition  $\mathcal{B}$  of  $D$  that is an  $\left(O\left(\frac{\alpha}{\gamma^2}\right), \gamma^2 \sigma, O\left(\frac{k\alpha}{\gamma}\right)\right)_k$ -balanced partition in  $O(|D|)$  time.*

*Proof.* To prove that  $\mathcal{B}$  is a  $\left(O\left(\frac{\alpha}{\gamma^2}\right), \gamma^2 \sigma, O\left(\frac{k\alpha}{\gamma}\right)\right)_k$ -balanced partition as in Definition 6, we need to prove the following properties:

- (i)  $\text{opt}_1(B) \leq \gamma^2 \sigma$  for every  $B \in \mathcal{B}$ .
- (ii)  $\mathcal{B}$  is a partition of  $D$  whose size is  $|\mathcal{B}| \in O\left(\frac{\alpha}{\gamma^2}\right)$ .
- (iii) For every  $k$ -segmentation  $\hat{s}$  there are  $O\left(\frac{k\alpha}{\gamma}\right)$  sub-signals  $B \in \mathcal{B}$  for which  $\hat{s}$  assigns at least 2 distinct values, i.e.,  $|\{\hat{s}(x) \mid (x, y) \in B\}| \geq 2$ .

**Proof of (i):** Observe that the output set  $\mathcal{B}$  contains the the union of multiple output sets  $\mathcal{B}' := \text{SLICEPARTITION}(\cdot, \gamma^2 \sigma)$  computed via calls to Algorithm 1. By Property (ii) of Lemma 12, every sub-signal  $B \in \mathcal{B}'$  in such output set  $\mathcal{B}'$  satisfies that  $\text{opt}_1(B) \leq \gamma^2 \sigma$ . Hence, Property (i) of Lemma 7 immediately holds.

**Proof of (ii):** By the greedy construction it holds that  $\mathcal{B}$  is a partition of  $D$ . We now prove that the number of times that Line 12 was executed is  $t \in O(\alpha/\gamma)$ , i.e., the number of times we append a set of signals  $last\mathcal{B}'$  to  $\mathcal{B}$  is at most  $O(\alpha/\gamma)$ . Let  $\mathcal{B}_1, \dots, \mathcal{B}_t$  denote the set of sub-signals  $last\mathcal{B}'$  in each of the  $t$  executions of Line 12, i.e.,  $\mathcal{B}_1 := last\mathcal{B}'$  at the first time Line 12 was executed. Each such set is called a *horizontal set*.

Recall that  $s \in \text{SEG}_{\beta k}$  is a  $\beta k$ -segmentation. First of all, by the definition of a  $\beta k$ -segmentation function there are at most  $2\beta k$  sets among the horizontal sets  $\mathcal{B}_1, \dots, \mathcal{B}_t$  which can be horizontally intersected by  $s$ .

Consider two consecutive horizontal sets  $\mathcal{B}_i, \mathcal{B}_{i+1}$  that are not horizontally intersected by  $s$ , and let  $H = \bigcup_{B \in \mathcal{B}_i \cup \mathcal{B}_{i+1}} B$  be the union of all the sub-signals in  $\mathcal{B}_i \cup \mathcal{B}_{i+1}$ . We now argue that the loss  $\ell(H, s)$  is at least  $O(\gamma\sigma)$ . Since  $\mathcal{B}_i$  and  $\mathcal{B}_{i+1}$  are two different horizontal sets, by the greedy construction we know that their union  $H$  could have been partitioned via a call to  $\mathcal{E} := \text{SLICEPARTITION}(H, \gamma^2\sigma)$  into a set  $\mathcal{E} = \{E_1, \dots, E_{|\mathcal{E}|}\}$  of at least  $|\mathcal{E}| \geq \frac{1}{\gamma}$  blocks. By substituting  $D = H, k = \beta k, \mathcal{B} = \mathcal{E}$  and  $\sigma = \gamma^2\sigma$  in Property (iii) of Lemma 12, for a  $\beta k$ -segmentation  $s$ , we have that

$$\begin{aligned} \ell(H, s) &\geq \left( \frac{|\mathcal{E}|}{4} - 2\beta k \right) \gamma^2\sigma \geq \left( \frac{1}{4\gamma} - 2\beta k \right) \gamma^2\sigma \\ &\geq \left( \frac{1}{4\gamma} - \frac{\beta k}{9\varepsilon^2} \right) \gamma^2\sigma = \left( \frac{1}{4\gamma} - \frac{1}{9\gamma} \right) \gamma^2\sigma \\ &\geq \gamma\sigma/2, \end{aligned}$$

where the second derivation holds for  $\varepsilon \in (0, 1/3)$ , and the third derivation is by the definition of  $\gamma$ .

Assume by contradiction that there are more than  $\frac{2\alpha}{\gamma}$  such pairs of consecutive horizontal sets  $\mathcal{B}_i, \mathcal{B}_{i+1}$ , which are not horizontally intersected by  $s$ . The loss of those slices to  $s$  would be bigger than  $\frac{2\alpha}{\gamma} \cdot \frac{\gamma\sigma}{2} = \alpha\sigma = \ell(D, s)$ , which is a contradiction. Therefore, the number of pairs of consecutive horizontal sets, which are not horizontally intersected by  $s$ , cannot exceed  $O\left(\frac{\alpha}{\gamma}\right)$ . Observe that the total number of horizontal sets that can be intersected by  $s$  is at most  $2\beta k$ . Hence, the total number of horizontal sets is at most

$$m \in O\left(\frac{\alpha}{\gamma} + 2\beta k\right) \in O\left(\frac{\alpha}{\gamma}\right). \quad (2)$$

We now prove that the number of output cells is at most  $|\mathcal{B}| \in O\left(\frac{\alpha}{\gamma^2}\right)$  in two steps. In step (i) we consider the horizontal sets that contain more than one row of  $D$  and show that they contain a total of  $O\left(\frac{\alpha}{\gamma^2}\right)$  sub-signals. In step (ii) we consider the horizontal sets that contain exactly one row of  $D$  and prove that they also contain a total of  $O\left(\frac{\alpha}{\gamma^2}\right)$  sub-signals.

**Step (i):** By (2), the total number of horizontal sets is at most  $m \in O\left(\frac{\alpha}{\gamma}\right)$ . Therefore, the total number of horizontal slices that contain more than one row of  $A$  is also at most  $O\left(\frac{\alpha}{\gamma}\right)$ . By the construction in Algorithm 2, each such horizontal set  $\mathcal{B}_i$  with more than 1 row of  $A$  is partitioned into at most  $\frac{1}{\gamma}$  sub-signals. Hence, the total number of blocks in horizontal sets that contain more than one row of  $D$  is at most  $O\left(\frac{\alpha}{\gamma} \cdot \frac{2}{\gamma}\right) = O\left(\frac{\alpha}{\gamma^2}\right)$ .

**Step (ii):** Consider all the horizontal sets  $\mathcal{B}_i$  which contain one row of  $D$ , and which have been partitioned into  $|\mathcal{B}_i| \leq 2\beta k \leq 1/\gamma$  blocks. The total number of blocks in such horizontal slices is thus bounded by the maximum number of horizontal slices  $m \in O(\alpha/\gamma)$  times  $1/\gamma$  for a total of at most  $O(\alpha/\gamma^2)$  blocks.

For the rest of this step, we assume that all horizontal sets  $\mathcal{B}_i$  have  $|\mathcal{B}_i| \geq 2\beta k$ . Let  $G \subseteq [t]$  contain the indices of the horizontal sets which contain exactly one row of  $D$ , and let  $i \in G$ . Observe that  $\mathcal{B}_i$  was computed, at some point, via a call  $\mathcal{B}_i := \text{SLICEPARTITION}(\bigcup_{B \in \mathcal{B}_i} B, \gamma^2\sigma)$ . Also, since the points  $\bigcup_{B \in \mathcal{B}_i} B$  of  $\mathcal{B}_i$  all have the same row index, observe that  $\mathcal{B}_i$  cannot be horizontally intersected by  $s$ . Therefore, by substituting  $D = \bigcup_{B \in \mathcal{B}_i} B, k = \beta k$  and  $\sigma = \gamma^2\sigma$  in Property (iii) of Lemma 12

we obtain that

$$\ell(\mathcal{B}_i, s) \geq \left( \frac{|\mathcal{B}_i|}{4} - 2\beta k \right) \cdot \gamma^2 \sigma. \quad (3)$$

Furthermore, we have that

$$\alpha \cdot \sigma = \ell(D, s) \geq \sum_{i \in G} \ell(\mathcal{B}_i, s) \geq \sum_{i \in G} \left( \frac{|\mathcal{B}_i|}{4} - 2\beta k \right) \cdot \gamma^2 \sigma, \quad (4)$$

where the first derivation is by the definition of  $\sigma$ , the second derivation holds since  $\{(x, y) \in B \mid B \in \mathcal{B}_i, i \in G\} \subseteq D$ , and the third derivation is by (3). Rearranging terms in (4) concludes Step (ii) as

$$\begin{aligned} \sum_{i \in G} |\mathcal{B}_i| &\leq \frac{4\alpha}{\gamma^2} + 8 \sum_{i \in G} \beta k \leq \frac{4\alpha}{\gamma^2} + 8t\beta k \\ &\leq \frac{4\alpha}{\gamma^2} + \frac{8\alpha\beta k}{\gamma} \in O\left(\frac{\alpha}{\gamma^2}\right). \end{aligned}$$

Therefore, the total number of blocks in horizontal sets that contain exactly one row of  $D$  is at most  $O\left(\frac{\alpha}{\gamma^2}\right)$ .

**Proof of (iii):** By the properties above we have that: (i) there are at most  $O(\alpha/\gamma)$  horizontal sets, and (ii) each horizontal set  $\mathcal{B}_i$  either contains at most  $O(1/\gamma)$  sub-signals, or all the points  $((i, j), y) \in B$  of all the blocks  $B \in \mathcal{B}_i$  have the same row index  $i$ . Let  $\hat{s}$  be a  $k$ -segmentation.  $\hat{s}$  can horizontally intersect all the (at most)  $1/\gamma$  sub-signals of at most  $k$  horizontal sets, and can vertically intersect at most 1 block from each of the  $O(\alpha/\gamma)$  horizontal sets. Hence, the total number of intersected sub-signals is  $O(k\alpha/\gamma)$ .

**Computational time:** We now prove that  $\mathcal{B}$  can be computed in  $O(|D|)$  time. The computational time of Algorithm 2 is dominated by the computational time of Line 11 where we partition a slice  $S$ . Using Algorithm 1 we can partition each such slice  $S$  in linear  $O(|S|)$  time; see Lemma 12. Therefore, the naive implementation, i.e. by calling Algorithm 1 for every slice  $S$ , will result in  $O(|D|^2)$  overall time, since many rows of  $D$  participate many times in such a call to Algorithm 1.

However, we can implement Line 11 in  $O(m)$  time, rather than  $O(|S|)$  time, by preprocessing the input signal  $D$ , in linear time  $O(|D|)$ , and storing some statistics for every element  $((i, j), y) \in D$ . For example, one can store the sum of values and squared values over all elements  $((i', j'), y')$  where  $i' < i$  or  $j' < j$ . Using those values we can compute  $\text{opt}(B)$  in  $O(1)$  time for every sub-signal  $B$  of  $D$ . Now, using such statistics (and possibly more statistics), Line 11 can be implemented in  $O(m)$  time via a greedy algorithm that iterates over the points of the last row  $R = \{((i, j), y) \in D \mid i = r_{\text{end}}\}$  added to  $S$  (i.e. with no need to iterate over other elements of  $S$ ). We leave the small details to the reader. □

## D Coreset Construction

In this section, we provide the proof of correctness for our main coreset construction algorithm presented in Algorithm 3; see Theorem 15. Furthermore, we provide an algorithm that gets as input a  $k$ -segmentation  $s$ , as well as a  $(k, \varepsilon)$ -coreset for some input dataset  $D$ , which was computed using Algorithm 3. The algorithm returns a  $(1 + \varepsilon)$ -approximation to the loss  $\ell(D, s)$ , in  $O(k|C|)$  time; see Algorithm 5 and full details in Lemma 14.

In what follows, for an  $n \times m$  sub-signal  $B$  and a weight function  $u : B \rightarrow [0, \infty)$ , we abuse notation and denote  $u((a, b))$  by simply  $u(a, b)$  for  $(a, b) \in B$ .

**Some intuition behind Algorithm 5.** Given a  $(k, \varepsilon)$ -coreset  $(C, u)$  for an input dataset  $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$ , and a  $k$ -segmentation  $s$ , the algorithm outputs a  $(1 + \varepsilon)$ -approximation to  $\ell(D, s)$  in time that depends only on  $k$  and  $|C|$ .

During the computation of  $(C, u)$  in Algorithm 3, a partition  $\mathcal{B}$  of  $D$  was computed. Then, for every set  $B$  in the partition  $\mathcal{B}$ , a representative pair  $(C_B, u_B)$  for  $B$  was computed and added to  $C$ .

To approximate the loss  $\ell(D, s)$ , we will approximate individually  $\ell(B, s)$  for every  $B \in \mathcal{B}$ , and return the sum of those losses. Therefore, we now consider a single set  $B \in \mathcal{B}$ , and consider the following two cases.

**Case (i) :**  $s$  assigns the same value for all the elements of  $B$ . Then, by construction, it is guaranteed that  $\ell(B, s) = \sum_{(x,y) \in C_B} u_B(x, y)(s(y) - y)^2$ . Therefore, in this case,  $\ell(B, s)$  will be accurately estimated using  $(C_B, u_B)$ .

**Case (ii) :**  $s$  assigns more than one unique value to the elements of  $B$ . In this case, if we ignore the computational time, we would ideally want to compute a ‘‘smoothed version’’  $(S, w)$  of  $(C_B, u_B)$ , as shown in Fig. 3 (see (9)- (11) below for formal details). Then, we would return the loss  $\sum_{(x,y) \in S} w(x, y)(s(y) - y)^2$ . However, computing  $(S, w)$  is not necessary, since there are many subsets of  $B$  in which all the elements  $x \in B$  have simultaneously the same label in  $S$  and are assigned the same value by  $s$ . Combining this with the fact that those subsets are of rectangular (simple) shape, we obtain that the loss over those subsets can be evaluated efficiently, as computed in Algorithm 5.

---

**Algorithm 5:** FITTING-LOSS $((C, u), s)$ ; see Lemma 14

---

**Input :** A  $(k, \varepsilon)$ -coreset  $(C, u)$  which was returned from a call to SIGNAL-CORESET $(D, k, \varepsilon/\Delta)$  in Algorithm 3, for some  $n \times m$ -signal  $D, k \geq 1$ ,  $\varepsilon \in (0, 1)$  and a sufficiently large  $\Delta \geq 1$ .  
A  $k$ -segmentation (or  $k$ -tree)  $s$ .

**Output :** A  $(1 + \varepsilon)$ -approximation to the loss  $\ell(D, s)$ .

```

1 loss := 0
2 for every 4 consecutive elements  $\hat{C} = \{(a_i, b_i)\}_{i=1}^4$  in  $C$  do
3   Denote by  $B$  the sub-signal that corresponds to  $C'$ . // By construction in
   Algorithm 3, the coordinates  $a$  of the 4 elements  $(a, b) \in \hat{C}$  are the
   corners of  $B$ .
4    $z := |\{s(x) \mid (x, y) \in B\}|$ 
5   if  $z = 1$  // i.e.,  $s$  does not intersect  $B$ 
6     then
7        $loss_{\hat{C}} := \sum_{(x,y) \in C_B} u_B(x, y)(s(x) - y)^2$ . // note that  $s(x_1) = s(x_2)$  for
       every  $x_1, x_2 \in B$ 
8     else
9       // In this case,  $s$  intersects  $B$ 
10      Denote by  $S$  the partition that  $s$  induces onto  $[n] \times [m]$ . //  $S$  contains  $|S| \leq k$ 
       subsets of  $[n] \times [m]$ .
11       $i := 1$ 
12      for every  $S' \in S$  do
13        Denote by  $\ell$  the label that  $s$  assigns to the elements of  $S'$  i.e.,  $s(x, y) = \ell$  for every
         $(x, y) \in S'$ 
14         $z := |B \cap S'|$  // The number of element in the intersection of the
         $S'$  and the subset of  $[n] \times [m]$  that is represented by  $C'$ .
15         $loss_{\hat{C}} := 0$ 
16        while  $z \geq 1$  do
17          if  $u(a_i, b_i) \leq z$  then
18             $loss_{\hat{C}} := loss_{\hat{C}} + u(a_i, b_i) \cdot (\ell - b_i)^2$ 
19             $u(a_i, b_i) := 0$ 
20             $z := z - u(a_i, b_i)$ 
21             $i := i + 1$ 
22          else
23             $loss_{\hat{C}} := loss_{\hat{C}} + z \cdot (\ell - b_i)^2$ 
24             $u(a_i, b_i) := u(a_i, b_i) - z$ 
25             $z := 0$ 
26         $loss := loss + loss_{\hat{C}}$ 
27 return loss

```

---

**Lemma 14.** Let  $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$  be an  $n \times m$  signal i.e.,  $N := nm$ . Let  $k \geq 1$  be an integer,  $\varepsilon \in (0, 1/4)$  be an error parameter, and  $(C, u)$  be the output of a call to  $\text{SIGNAL-CORESET}(D, k, \varepsilon)$  (see Algorithm 3). Let  $s$  be a  $k$ -segmentation (in particular, a  $k$ -tree). Finally, let  $\text{loss}$  be the output of a call to  $\text{FITTING-LOSS}((C, u), s)$ ; see Algorithm 5. Then there is a sufficiently large constant  $\Delta \geq 1$  such that

$$|\ell(D, s) - \text{loss}| \leq \Delta \varepsilon \cdot \ell(D, s).$$

Moreover,  $\text{loss}$  can be computed in  $O(k|C|)$  time.

*Proof.* We consider the variables defined in Algorithm 5.

First, consider a subset  $\hat{C}$  of  $C$  from some iteration of the For loop at Line 2 of Algorithm 5, and let  $B$  be the sub-signal that corresponds to  $\hat{C}$ , as in Line 3. We now prove that the loss  $\text{loss}_{\hat{C}}$  computed in the same iteration of the For loop (i.e., at Lines 3-25) satisfies the following claim.

**Claim 14.1.** Let  $z = |\{s(x) \mid (x, y) \in B\}|$  be the number of distinct values  $s$  assigns to the coordinates of  $B$  (as computed in Line 4). Then,  $\text{loss}_{\hat{C}}$  satisfies that

$$\begin{cases} \ell(B, s) = \text{loss}_{\hat{C}} & \text{if } z = 1 \\ |\ell(B, s) - \text{loss}_{\hat{C}}| \leq \varepsilon \cdot \ell(B, s) + O\left(\frac{\text{opt}_1(B)}{\varepsilon}\right) & \text{otherwise} \end{cases}$$

*Proof.* We prove Claim 14.1 using the following case analysis: (i)  $z = 1$  and (ii)  $z \geq 2$ .

**Case (i):**  $z = 1$ . We prove that  $\ell(B, s) = \text{loss}_{\hat{C}}$ .

Since the input coresets  $(C, u)$  was computed using Algorithm 3, we know that the set  $\hat{C}$  was computed at Line 5 of Algorithm 3, along with a weight function  $\hat{u}$ . Hence, the pair  $(\hat{C}, \hat{u})$  satisfy, by construction, the following property:

$$\sum_{(a,b) \in \hat{C}} \hat{u}((a, b)) \cdot (b \mid b^2 \mid 1) = \sum_{(x,y) \in B} (y \mid y^2 \mid 1). \quad (5)$$

Now, for any constant  $\hat{s} \in \mathbb{R}$ , we have that

$$\begin{aligned} & \sum_{(a,b) \in \hat{C}} \hat{u}(a, b)(b - \hat{s})^2 \\ &= \sum_{(a,b) \in \hat{C}} \hat{u}(a, b) \cdot b^2 + \hat{s}^2 \sum_{(a,b) \in \hat{C}} \hat{u}(a, b) - 2\hat{s} \sum_{(a,b) \in \hat{C}} \hat{u}(a, b)b \\ &= \sum_{(x,y) \in B} y^2 + \hat{s}^2 \sum_{(x,y) \in B} 1 - 2\hat{s} \sum_{(x,y) \in B} y \\ &= \sum_{(x,y) \in B} (y - \hat{s})^2, \end{aligned} \quad (6)$$

where the second equality is by (5).

Since  $z = |\{s(x) \mid (x, y) \in B\}| = 1$ , there is a constant  $\hat{s} \in \mathbb{R}$  such that

$$\ell(B, s) = \sum_{(x,y) \in B} (y - \hat{s})^2. \quad (7)$$

Hence, we have that

$$\ell(B, s) = \sum_{(x,y) \in B} (y - \hat{s})^2 = \sum_{(a,b) \in \hat{C}} \hat{u}(a, b)(b - \hat{s})^2 = \text{loss}_{\hat{C}},$$

where the first derivation is by (7), the second derivation is by (6), and the last derivation is by the definition of  $\text{loss}_{\hat{C}}$  at Line 7 of Algorithm 5.

**Case (ii):**  $z \geq 2$ . We prove that

$$|\ell(B, s) - \text{loss}_{\hat{C}}| \leq \varepsilon \cdot \ell(B, s) + O\left(\frac{\text{opt}_1(B)}{\varepsilon}\right).$$

We first observe that, by the triangle inequality, for any  $a, b, c \in \mathbb{R}$  we have that

$$\begin{aligned} ||a - c|^2 - |b - c|^2| &= ||a - c| - |b - c|| \cdot (|a - c| + |b - c|) \\ &\leq |a - b| \cdot (2|a - c| + |a - b|) \\ &= |a - b|^2 + 2|a - c| \cdot |a - b| \\ &= |a - b|^2 + 2\sqrt{\varepsilon}|a - c| \cdot \frac{|a - b|}{\sqrt{\varepsilon}} \\ &\leq |a - b|^2 + \varepsilon \cdot |a - c|^2 + \frac{|a - b|^2}{\varepsilon} \\ &= \varepsilon \cdot |a - c|^2 + \left(1 + \frac{1}{\varepsilon}\right) \cdot (a - b)^2, \end{aligned} \tag{8}$$

where the second inequality holds since  $2xy \leq x^2 + y^2$  for every  $x, y \in \mathbb{R}$ .

**Smoothed coreset.** In Algorithm 3 we computed some small compression  $C_B$ , along with a weights function  $u_B$ , for every subset  $B$  in the partition of the input. The size  $|C_B|$  of this compression is a small constant, independent of the (potentially large) size of  $B$ . The pair  $(C_B, u_B)$  satisfy a set of properties, which we visually demonstrate via this “smoothed coreset” notion; see Fig. 3. Informally, the “smoothed version” of  $(C_B, u_B)$  is another pair  $(\hat{C}_B, \hat{u}_B)$ , such that  $\hat{C}_B$  contains a duplication of the elements of  $C_B$ . The number of duplications of every element  $c$  from  $C_B$  is according to its weight  $u_B(c)$ .

We now formally define a “smoothed version” of a pair  $(\hat{C}, \hat{u})$ . A pair  $(S, w)$  is said to be a *smoothed version* of the pair  $(\hat{C}, \hat{u})$  if it satisfies the following properties: (i)  $(S, w)$  has the same sum of weights, sum of labels, and sum of squared labels as  $(\hat{C}, \hat{u})$ , (ii) The set of coordinates  $\{a|(a, b) \in S\}$  in  $S$  covers the entire set of coordinates  $\{x|(x, y) \in B\}$  of the original set  $B$ , with possible duplicates, and (iii) The sum of weights over all elements in  $S$  with the same coordinate is 1. Formally,

$$\sum_{(a,b) \in S} w((a, b)) \cdot (b | b^2 | 1) = \sum_{(a,b) \in \hat{C}} \hat{u}((a, b))(b | b^2 | 1), \tag{9}$$

$$\{x|(x, y) \in B\} = \{a|(a, b) \in S\}, \tag{10}$$

and

$$\sum_{(a,b) \in S: a=x} w((a, b)) = 1 \text{ for every } (x, y) \in B. \tag{11}$$

In what follows, for every pair  $(S, w)$  which is a smoothed version of  $(\hat{C}, \hat{u})$ , we prove the following two properties: We now prove the following two properties: first, that

$$|\ell(B, s) - \ell((S, w), s)| \leq \varepsilon \cdot \ell(B, s) + O\left(\frac{\text{opt}_1(B)}{\varepsilon}\right), \tag{12}$$

for every every pair  $(S, w)$  which is a smoothed version of  $(\hat{C}, \hat{u})$ . Second, we need to prove there is some pair  $(\hat{S}, \hat{w})$  which is a smoothed version of  $(\hat{C}, \hat{u})$  that satisfies

$$\text{loss}_{\hat{C}} = \ell((\hat{S}, \hat{w}), s), \tag{13}$$

where  $\text{loss}_{\hat{C}}$  is the loss computed at Lines 3- 25, using only the pair  $(\hat{C}, \hat{u})$  (i.e., at the current iteration of the outer-most For loop of Algorithm 5), without actually computing  $(\hat{S}, \hat{w})$ . Case (ii) then immediately holds by combining (12) and (13) above.

**A proof of (12).** Let  $(S, w)$  be a pair which is a smoothed version of  $(\hat{C}, \hat{u})$ . By definition of  $(S, w)$ , we have that  $w$  sums to 1 over all  $(a, b) \in S$  with the same  $a$ , as in (11). Therefore, for every

$(x, y) \in B$  we can rewrite the term  $(y - s(x))^2$  as  $\sum_{(a,b) \in S: a=x} w(a, b)(y - s(x))^2$ . Now, define  $y_B(x) = y$  for every  $(x, y) \in B$ . We therefore have that

$$\begin{aligned} \ell(B, s) &= \sum_{(x,y) \in B} (y - s(x))^2 \\ &= \sum_{(x,y) \in B} \left( \sum_{(a,b) \in S: a=x} w(a, b) \right) \cdot (y - s(x))^2 \\ &= \sum_{(x,y) \in S} w(x, y)(y_B(x) - s(x))^2, \end{aligned} \quad (14)$$

where the last equality holds by (10) and by simply combining the two sums.

We now have that

$$\begin{aligned} &|\ell(B, s) - \ell((S, u), s)| \\ &= \left| \sum_{(x,y) \in S} w(x, y)(y_B(x) - s(x))^2 - \sum_{(x,y) \in S} w(x, y)(y - s(x))^2 \right| \end{aligned} \quad (15)$$

$$\begin{aligned} &= \left| \sum_{(x,y) \in S} w(x, y) \cdot ((y_B(x) - s(x))^2 - (y - s(x))^2) \right| \\ &\leq \sum_{(x,y) \in S} w(x, y) |(y_B(x) - s(x))^2 - (y - s(x))^2| \end{aligned} \quad (16)$$

$$\leq \sum_{(x,y) \in S} w(x, y) \left( \varepsilon \cdot (y_B(x) - s(x))^2 + \left(1 + \frac{1}{\varepsilon}\right) (y_B(x) - y)^2 \right) \quad (17)$$

$$\begin{aligned} &= \varepsilon \cdot \sum_{(x,y) \in S} u(x, y) \cdot (y_B(x) - s(x))^2 + \left(1 + \frac{1}{\varepsilon}\right) \sum_{(x,y) \in S} w(x, y) \cdot (y_B(x) - y)^2 \\ &= \varepsilon \cdot \ell(B, s) + \left(1 + \frac{1}{\varepsilon}\right) \sum_{(x,y) \in S} u(x, y) \cdot (y_B(x) - y)^2, \end{aligned} \quad (18)$$

where (15) is by combining (14) and the definition of  $\ell$ , (16) holds since the sum of absolute values is greater or equal than the absolute value of a sum, (17) holds by substituting in (8) every term in the sum, and (18) is by (14).

We now bound the rightmost term of (18). Let  $\hat{s} \equiv 1/|B| \sum_{(x,y) \in B} y$  be a 1-segmentation function that returns the label mean of  $B$ . We have that

$$\begin{aligned} &\sum_{(x,y) \in S} w(x, y) \cdot (y_B(x) - y)^2 \\ &\leq 2 \cdot \sum_{(x,y) \in S} w(x, y) \cdot ((y_B(x) - \hat{s}(x))^2 + (y - \hat{s}(x))^2) \end{aligned} \quad (19)$$

$$\begin{aligned} &= 2 \sum_{(x,y) \in S} w(x, y) \cdot (y_B(x) - \hat{s}(x))^2 \\ &\quad + 2 \sum_{(x,y) \in S} w(x, y) \cdot (y - \hat{s}(x))^2 \\ &= 2 \cdot (\ell(B, \hat{s}) + \ell((S, u), \hat{s})) \end{aligned} \quad (20)$$

$$= 2 \cdot (\ell(B, \hat{s}) + \ell(B, \hat{s})) \quad (21)$$

$$\begin{aligned} &= 4 \cdot \ell(B, \hat{s}) \\ &= 4 \cdot \text{opt}_1(B) \end{aligned} \quad (22)$$

where (19) is by the weak triangle inequality, (20) is by combining the definition of  $\ell$  with (14), (21) holds by Case (i) above, and (22) holds since the label means minimizes its sum of squared differences to the labels.

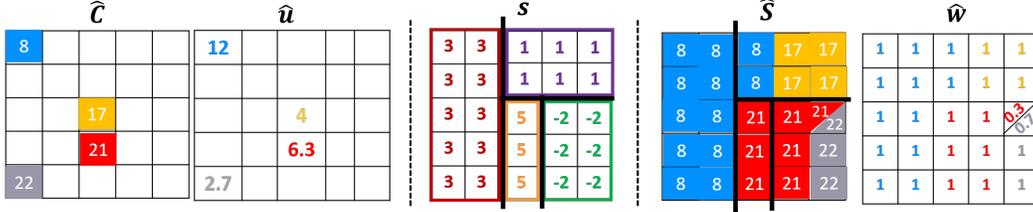


Figure 8: **(Left):** The pair  $(\hat{C}, \hat{u})$ . **(Middle):** A 4-segmentation  $s$ , which induces a partition of  $[5] \times [5]$  into 4 sets  $\mathcal{B} = B_1 \cup B_2 \cup B_3 \cup B_4$  where  $B_1 = \{(1, 1), (2, 1), (3, 1), (4, 2), \dots\}$ ,  $B_2 = \{(1, 3), (2, 3), (1, 4), (2, 4), \dots\}$ ,  $B_3 = \{(3, 3), (4, 3), (5, 3)\}$ ,  $B_4 = \{(3, 4), (4, 4), (4, 5), (3, 5), \dots\}$ . **(Right):** A smoothed version  $(\hat{S}, \hat{w})$  of  $(\hat{C}, \hat{u})$ . There can be more than one unique smoothed version for the same pair  $(\hat{C}, \hat{u})$ ; see Properties in (9)-(11). The pair  $(\hat{S}, \hat{w})$  is constructed by iterating over every set  $B \in \mathcal{B}$ . Every element in  $B$  is assigned to a label from the labels of the 4-coreset points (8, 17, 21, 22) as follows: If  $|B| > \hat{u}(l_i)$ , then  $\hat{u}(l_i)$  elements of  $B$  are assigned to  $l_1$  and  $|B| - \hat{u}(l_i)$  are assigned to  $l_{i+1}$ . If  $|B| \leq \hat{u}(l_i)$  then all the elements of  $B$  are assigned to  $\hat{u}(l_i)$ , and  $|B|$  is subtracted from  $\hat{u}(l_i)$ , and so on. If  $\hat{u}$  assigns fractional weights, then some elements of  $B$  might be assigned to more than one label, as long as the sum of weights over every element in  $B$  is 1. By construction,  $(\hat{S}, \hat{w})$  satisfies Properties (10)-(11). Hence,  $(\hat{S}, \hat{w})$  is a smoothed version of  $\hat{C}, \hat{u}$ . Computing  $\ell((\hat{S}, \hat{w}), s)$  can be trivially computed in time only  $O(k|\hat{C}|)$  (rather than  $O(n)$  where  $n$  is the size of the original data), since the sets in the partition  $\mathcal{B}$  contain a duplication of a constant number of labels.

Equation (12) now holds by combining (18) and (22).

**A proof of (13).** To prove (13), in Fig. 8 we construct a smoothed version  $(\hat{S}, \hat{w})$  of  $(\hat{C}, \hat{u})$  which satisfies (13).

Furthermore, by combining the construction of  $(\hat{S}, \hat{w})$  with the computation of  $loss_{\hat{C}}$  in Lines 9-25 of Algorithm 5, we obtain, as desired, that  $loss_{\hat{C}} = \ell((\hat{S}, \hat{w}), s)$ .

Claim 14.1 now holds by combining cases (i) and (ii) above. □

**We now prove Lemma 14.** Consider the construction of  $(C, u)$  in Algorithm 3. By definition and by Lemma 5, the function  $s'$  computed at Line 5 of Algorithm 3 is an  $O(k^8 \log^2 nm)$ -segmentation and satisfies that

$$\ell(D, s') \in O(k \log n \cdot \text{opt}_k(D)).$$

By the last derivation, let  $c_\alpha$  be the smallest constant such that

$$\ell(D, s') \leq c_\alpha \cdot k \log n \cdot \text{opt}_k(D).$$

By the last inequality and definitions of  $\sigma$  and  $\alpha$  we obtain that

$$\begin{aligned} \sigma &:= \frac{\ell(D, s')}{\alpha} \leq \frac{c_\alpha k \log nm \cdot \text{opt}_k(D)}{\alpha} \\ &= \frac{c_\alpha k \log nm \cdot \text{opt}_k(D)}{c_\alpha k \log nm} \\ &= \text{opt}_k(D). \end{aligned} \tag{23}$$

Now consider the partition  $\mathcal{B}$  computed at Line 3 of Algorithm 3 via a call to  $\text{PARTITION}(D, \gamma, \sigma)$ . By Lemma 7,  $\mathcal{B}$  satisfies that

(i)  $\text{opt}_1(B) \leq \gamma^2 \sigma$  for every  $B \in \mathcal{B}$ .

(ii)  $\mathcal{B}$  is a partition of  $D$  whose size is  $|\mathcal{B}| \in O\left(\frac{\alpha}{\gamma^2}\right)$ .

(iii) There are  $O\left(\frac{k\alpha}{\gamma}\right)$  sub-signals  $B \in \mathcal{B}$  where  $|\{s(x) \mid (x, y) \in B\}| > 1$ .

(iv)  $\mathcal{B}$  can be computed in  $O(|D|)$  time.

Consider the pair  $(C_B, u_B)$  computed at Line 5 of Algorithm 3 for some  $B \in \mathcal{B}$ . Now, consider the For loop at Line 2 of Algorithm 5. For every pair  $(C_B, u_B)$ , there is an iteration of this For loop for which  $C' = C_B$ . In this iteration, Algorithm 5 computes a loss  $loss_{C_B}$  that corresponds to  $(C_B, u_B)$ . We can plug  $B$ ,  $(C_B, u_B)$ , and  $loss_{C_B}$  in Claim 14.1 to obtain that:

$$\begin{cases} \ell(B, s) = loss_{C_B} & \text{if } z = 1 \\ |\ell(B, s) - loss_{C_B}| \leq \varepsilon \cdot \ell(B, s) + O\left(\frac{\text{opt}_1(B)}{\varepsilon}\right) & \text{otherwise} \end{cases} \quad (24)$$

Let  $\mathcal{B}_1 \subseteq \mathcal{B}$  contain the set of sub-signals in  $\mathcal{B}$  that are not intersected by  $s$ , i.e.,  $\mathcal{B}_1 = \{B \in \mathcal{B} \mid |s(B)| = 1\}$ , and let  $\mathcal{B}_2 = \mathcal{B} \setminus \mathcal{B}_1$  be the set of sub-signals which are partially intersected by  $s$ .

By Property (iii) above,

$$|\mathcal{B}_2| \in O\left(\frac{k\alpha}{\gamma}\right). \quad (25)$$

Furthermore, by combining (24) with Property (i) above, for every  $B \in \mathcal{B}_2$  we have that

$$|\ell(B, s) - loss_{C_B}| \leq \varepsilon \cdot \ell(B, s) + O\left(\frac{\text{opt}_1(B)}{\varepsilon}\right) \leq \varepsilon \cdot \ell(B, s) + O\left(\frac{\gamma^2 \sigma}{\varepsilon}\right). \quad (26)$$

In other words, the loss  $\ell(B, s)$  of every sub-signal  $B \in \mathcal{B}_2$  is approximated by  $loss_{C_B}$  up to some small error. Hence, by summing over all  $B \in \mathcal{B}_2$  we obtain that

$$\begin{aligned} & \sum_{B \in \mathcal{B}_2} |\ell(B, s) - loss_{C_B}| \\ & \in \sum_{B \in \mathcal{B}_2} \left( \varepsilon \cdot \ell(B, s) + O\left(\frac{\gamma^2 \sigma}{\varepsilon}\right) \right) \end{aligned} \quad (27)$$

$$\leq \varepsilon \cdot \ell(D, s) + O\left(|\mathcal{B}_2| \cdot \frac{\gamma^2 \sigma}{\varepsilon}\right) \quad (28)$$

$$\leq \varepsilon \cdot \ell(D, s) + O\left(\frac{k\alpha}{\gamma} \cdot \frac{\gamma^2 \sigma}{\varepsilon}\right) \quad (29)$$

$$\leq \varepsilon \cdot \ell(D, s) + O\left(\frac{k\alpha\gamma}{\varepsilon} \cdot \text{opt}_k(D)\right) \quad (29)$$

$$\leq \varepsilon \cdot \ell(D, s) + O(\varepsilon \cdot \text{opt}_k(D)) \quad (30)$$

$$\in O(\varepsilon \cdot \ell(D, s)), \quad (31)$$

where (27) follows from (26), (28) is by (25), (29) is by (23), (30) holds since  $k\alpha\gamma \leq \varepsilon^2$ , and (31) holds since  $\text{opt}_k(D) \leq \ell(D, s)$  for every  $k$ -segmentation  $s$ .

Furthermore, for every  $B \in \mathcal{B}_1$ , by (24) we have that  $\ell(B, s) = loss_{C_B}$ . Hence, by summing over every  $B \in \mathcal{B}_1$  we obtain that

$$\sum_{B \in \mathcal{B}_1} \ell(B, s) = \sum_{B \in \mathcal{B}_1} loss_{C_B}. \quad (32)$$

In other words, the loss  $\ell(B, s)$  of every sub-signal  $B \in \mathcal{B}_1$  is accurately estimated by  $loss_{C_B}$ .

Algorithm 5 then outputs the sum of losses

$$loss := \sum_{B \in \mathcal{B}} loss_{C_B}. \quad (33)$$

We hence obtain that

$$\begin{aligned} |\ell(D, s) - loss| &= \left| \sum_{B \in \mathcal{B}_1} (\ell(B, s) - loss_{C_B}) + \sum_{B \in \mathcal{B}_2} (\ell(B, s) - loss_{C_B}) \right| \\ &\leq \left| \sum_{B \in \mathcal{B}_1} (\ell(B, s) - loss_{C_B}) \right| + \left| \sum_{B \in \mathcal{B}_2} (\ell(B, s) - loss_{C_B}) \right| \in O(\varepsilon \cdot \ell(D, s)), \end{aligned} \quad (34)$$

where the first derivation is by 33, second derivation is by the triangle inequality, and the last is by combining (32) and (31).

By (34), there is a constant  $\Delta \geq 1$  such that

$$|\ell(D, s) - \text{loss}| \leq \Delta \varepsilon \ell(D, s).$$

This concludes the proof of the claim in Lemma 14.

**Computational time:** Line 2 of the Algorithm 5 is a loop with  $\frac{|C|}{|\hat{C}|}$  iterations. Inside this loop: if  $z > 1$  line 7 is computed in  $O(|\hat{C}|)$ , else line 11 is another loop with  $O(k)$  iterations, inside which the line 20 is executed at most  $|\hat{C}|$  times and line 24 can be executed only once because it results in  $z = 0$  and in exiting from the while loop.

In total the complexity of line 15 is  $O(|\hat{C}|)$ , of line 11 is  $O(k|\hat{C}|)$  and of line 2 and the whole algorithm:  $O(k|C|)$

**Space complexity:** Algorithm 5 uses only constant amount of additional storage space because in each line of the algorithm only numeric variables are created and variables are reused inside the loops. □

**Theorem 15 (Coreset).** *Let  $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$  be an  $n \times m$  signal i.e.,  $N := nm$ . Let  $k \geq 1$  be an integer (that corresponds to the number of leaves/rectangles), and  $\varepsilon \in (0, 1/4)$  be an error parameter. Let  $(C, u)$  be the output of a call to SIGNAL-CORESET( $D, k, \varepsilon/\Delta$ ) for a sufficiently large constant  $\Delta \geq 1$ ; see Algorithm 3. Then,  $(C, u)$  is a  $(k, \varepsilon)$ -coreset for  $D$  of size  $|C| \in \frac{(k \log(N))^{O(1)}}{\varepsilon^4}$ ; see Definition 3. Moreover,  $(C, u)$  can be computed in  $O(kN)$  time.*

*Proof.* To prove that  $(C, u)$  is a  $(k, \varepsilon)$ -coreset for  $D$ , we need to prove that for every  $k$ -segmentation  $s$ ,  $(C, u)$  suffices to approximate the loss  $\ell(D, s)$ , up to a multiplicative factor of  $1 + \varepsilon$ , in time that depends only on  $|C|$  and  $k$ .

Let  $s$  be a  $k$ -segmentation and let  $\text{loss} \geq 0$  be an output of a call to FITTING-LOSS( $(C, u), s$ ); see Algorithm 5. Then, by Lemma 14,  $\text{loss}$  can be computed in  $O(k|C|)$  time (i.e., in time that depends only on  $|C|$  and  $k$ ), and provides, as required, a  $(1 + \varepsilon)$ -approximation to  $\ell(D, s)$  as

$$|\ell(D, s) - \text{loss}| \leq \Delta \cdot \frac{\varepsilon}{\Delta} \cdot \ell(D, s) = \varepsilon \cdot \ell(D, s).$$

Hence,  $(C, u)$  is a  $(k, \varepsilon)$ -coreset for  $D$ .

Line 1 of Algorithm 3 can be computed in  $O(k \cdot |D|)$  time by Lemma 5. Line 3 of Algorithm 3 can be computed in  $O(|D|)$  time by Lemma 7. The loop at Line 4 can be computed in  $\sum_{B \in \mathcal{B}} O(|B|) = O(|D|)$  time by Section E. Hence, the call SIGNAL-CORESET( $D, k, \varepsilon$ ) can be implemented in  $O(k|D|) = O(kmn)$  time.

**Proof behind Line 6.** We now prove that the replacements of the coordinates applied at Line 6 does not violate the correctness of the algorithm. Observe that replacing the coordinates of entries inside each cell, while keeping the same labels, does not affect the variance of this subset. Therefore, the cost of this cell, which is computed in Algorithm 5) and depends only on the labels, remains exactly the same.

**Space complexity:** By construction, each pair  $(C_B, u_B)$  computed at Line 5 can be stored using only  $O(1)$  space. Hence, the concatenation  $(C, u)$  of the  $|B|$  pairs  $\{(C_B, u_B) \mid B \in \mathcal{B}\}$  can be stored using  $O(|\mathcal{B}|) \in O(\alpha/\gamma^2) = O(\alpha(\beta k)^2/\varepsilon^4) = O\left(\frac{k^{O(1)} \log^{O(1)} nm}{\varepsilon^4}\right)$  space. □

## E The Caratheodory Theorem

Given a point  $p \in \mathbb{R}^d$  inside the convex hull of a set of points  $P \subseteq \mathbb{R}^d$ , Caratheodory's Theorem proves that there is a subset of at most  $d + 1$  points in  $P$  whose convex hull also contains  $p$ .

**Theorem 16 (Caratheodory's Theorem [13, 48]).** *Let  $P \subseteq \mathbb{R}^d$  be a (multi)set of  $n$  points. Then in  $O(nd^3)$  time we can compute a subset  $Q \subseteq P$  and a weights functions  $u : Q \rightarrow [0, \infty)$  such that: (i)  $Q \subseteq P$ , (ii)  $|Q| = d + 1$ , (iii)  $\sum_{q \in Q} u(q) \cdot q = \frac{1}{n} \sum_{p \in P} p$ , and (iv)  $\sum_{q \in Q} u(q) = n$ .*

**Corollary 17.** *Let  $D$  be an  $n \times m$  sub-signal. Then, in  $O(|D|)$  time we can compute a weighted  $n \times m$  sub-signal  $(A, w)$  such that: (i)  $A \subseteq D$ , (ii)  $|A| = 4$ , (iii)  $\sum_{(a,b) \in A} w(a, b) \cdot (b \mid b^2 \mid 1) =$*

$$\sum_{(x,y) \in D} (y \mid y^2 \mid 1), \text{ and } \sum_{(a,b) \in A} w(a, b) = |D|.$$

*Proof.* Define the multi-set  $P = \{(y \mid y^2 \mid 1) \in \mathbb{R}^3 \mid (x, y) \in D\}$ . Now, substituting  $P$  in Theorem 16 yields that in  $O(n)$  time we can compute a subset  $Q \subseteq P$  and a weights functions  $u : Q \rightarrow [0, \infty)$  such that: (i)  $Q \subseteq P$ , (ii)  $|Q| = 4$ , (iii)  $\sum_{q \in Q} u(q) \cdot q = \frac{1}{|D|} \sum_{p \in P} p$ , and (iv)  $\sum_{q \in Q} u(q) = |D|$ .

Now, add to  $A$  a single element  $(x, y) \in D$  for every  $(y \mid y^2 \mid 1) \in Q$ . In other words, for every element chosen for the set  $Q$  by the Caratheodory theorem, add its corresponding element from  $D$  to  $A$ . Furthermore, define  $w(x, y) = |D| \cdot u((y \mid y^2 \mid 1))$  for every  $(x, y) \in A$ . Corollary 17 trivially holds for  $(A, w)$ .  $\square$