
Sliding Window Algorithms for k -Clustering Problems

Michele Borassi
Google Zürich
borassi@google.com

Alessandro Epasto
Google Research New York
aepasto@google.com

Silvio Lattanzi
Google Research Zürich
silviol@google.com

Sergei Vassilvskii
Google Research New York
sergeiv@google.com

Morteza Zadimoghaddam
Google Research Cambridge
zadim@google.com

Abstract

The sliding window model of computation captures scenarios in which data is arriving continuously, but only the latest w elements should be used for analysis. The goal is to design algorithms that update the solution efficiently with each arrival rather than recomputing it from scratch. In this work, we focus on k -clustering problems such as k -means and k -median. In this setting, we provide simple and practical algorithms that offer stronger performance guarantees than previous results. Empirically, we show that our methods store only a small fraction of the data, are orders of magnitude faster, and find solutions with costs only slightly higher than those returned by algorithms with access to the full dataset.

1 Introduction

Data clustering is a central tenet of unsupervised machine learning. One version of the problem can be phrased as grouping data into k clusters so that elements within the same cluster are similar to each other. Classic formulations of this question include the k -median and k -means problems for which good approximation algorithms are known [1, 45]. Unfortunately, these algorithms often do not scale to large modern datasets requiring researchers to turn to parallel [8], distributed [9], and streaming methods. In the latter model, points arrive one at a time and the goal is to find algorithms that quickly update a small *sketch* (or summary) of the input data that can then be used to compute an approximately optimal solution.

One significant limitation of the classic data stream model is that it ignores the time when a data point arrived; in fact, all of the points in the input are treated with equal significance. However, in practice, it is often important (and sometimes necessary) to restrict the computation to very recent data. This restriction may be due to data freshness—e.g., when training a model on recent events, data from many days ago may be less relevant compared to data from the previous hour. Another motivation arises from legal reasons, e.g., data privacy laws such as the General Data Protection Regulation (GDPR), encourage and mandate that companies not retain certain user data beyond a specified period. This has resulted in many products including a data retention policy [55]. Such recency requirements can be modeled by the *sliding window* model. Here the goal is to maintain a small sketch of the input data, just as with the streaming model, and then use only this sketch to approximate the solution on the last w elements of the stream.

Clustering in the sliding window model is the main question that we study in this work. A trivial solution simply maintains the w elements in the window and recomputes the clusters from scratch at each step. We intend to find solutions that use less space, and are more efficient at processing each

new element. In particular, we present an algorithm which uses space linear in k , and polylogarithmic in w , but still attains a constant factor approximation.

Related Work *Clustering.* Clustering is a fundamental problem in unsupervised machine learning and has application in a disparate variety of settings, including data summarization, exploratory data analysis, matrix approximations and outlier detection [40, 42, 47, 51]. One of the most studied formulations in clustering of metric spaces is that of finding k centers that minimize an objective consisting of the ℓ_p norm of the distances of all points to their closest center. For $p \in \{1, 2, \infty\}$ this problem corresponds to k -median, k -means, and k -center, respectively, which are NP-hard, but constant factor approximation algorithms are known [1, 35, 45]. Several techniques have been used to tackle these problems at scale, including dimensionality reduction [46], core-sets [6], distributed algorithms [5], and streaming methods reviewed later. To clarify between Euclidean or general metric spaces, we note that our results work on arbitrary general metric spaces. The hardness results in the literature hold even for special case of Euclidean metrics and the constant factor approximation algorithms hold for the general metric spaces.

Streaming model. Significant attention has been devoted to models for analyzing large-scale datasets that evolve over time. The streaming model of computation is of the most well-known (see [50] for a survey) and focuses on defining low-memory algorithms for processing data arriving one item at a time. A number of interesting results are known in this model ranging from the estimation of stream statistics [3, 10], to submodular optimization [7], to graph problems [2, 31, 43], and many others. Clustering is also well studied in this setting, including algorithms for k -median, k -means, and k -center in the insertion-only stream case [6, 20, 36].

Sliding window streaming model. The sliding window model significantly increases the difficulty of the problem, since deletions need to be handled as well. Several techniques are known, including the exponential histogram framework [28] that addresses weakly additive set functions, and the smooth histogram framework [18] that is suited for functions that are well-behaved and possesses a sufficiently small constant approximation. Since many problems, such as k -clustering, do not fit into these two categories, a number of algorithms have been developed for specific problems such as submodular optimization [14, 21, 30], graph sparsification [27], minimizing the enclosing ball [56], private heavy hitters [55], diversity maximization [14] and linear algebra operations [15]. Sliding window algorithms find also applications in data summarization [23].

Turning to sliding window algorithms for clustering, for the k -center problem Cohen et al. [25] show a $(6 + \epsilon)$ -approximation using $O(k \log \Delta)$ space and per point update time of $O(k^2 \log \Delta)$, where Δ is the spread of the metric, i.e. the ratio of the largest to the smallest pairwise distances. For k -median and k -means, [17] give constant factor approximation algorithms that use $O(k^3 \log^6 w)$ space and per point update time of $O(\text{poly}(k, \log w))$.¹ Their bound is polylogarithmic in w , but *cubic* in k , making it impractical unless $k \ll w$.² In this paper we improve their bounds and give a simpler algorithm with only linear dependency of k . Furthermore we show experimentally (Figure 1 and Table 1) that our algorithm is faster and uses significantly less memory than the one presented in [17] even with very small values k (i.e., $k \geq 4$). In a different approach, [57] study a variant where one receives points in batches and uses heuristics to reduce the space and time. Their approach does provide approximation guarantees but it applies only to the Euclidean k -means case. Recently, [33] studied clustering problems in the distributed sliding window model, but these results are not applicable to our setting.

The more challenging fully-dynamic stream case has also received attention [16, 39]. Contrary to our result for the sliding window case, in the fully-dynamic case, obtaining a $\tilde{O}(k)$ memory, low update time algorithm, for the *arbitrary* metric k -clustering case with general ℓ_p norms is an open problem. For the special case of d -dimensional Euclidean spaces for k -means, there are positive results—[39] give $\tilde{O}(kd^4)$ -space core-set with $1 + \epsilon$ approximation.

Dynamic algorithms have also been studied in a consistent model [24, 44], but there the objective is to minimize the number of changes to the solution as the input evolves, rather than minimizing the approximation ratio and space used. Finally, a relaxation of the fully dynamic model that allows only

¹We note that the authors assume that the cost of any solution is polynomial in w . We chose to state our bounds explicitly, which introduces a dependence on the ratio of the max and min costs of the solution.

²We note here that in some practical applications k can be large. For instance, in spam and abuse [54], near-duplicate detection [38] or reconciliation tasks [53].

a limited number of deletions has also been addressed [34, 49]. The only work related to clustering is that of submodular maximization [49] which includes exemplar-based clustering as a special case.

Our Contributions We simplify and improve the state-of-the-art of k -clustering sliding window algorithms, resulting in lower memory algorithms. Specifically, we:

- Introduce a simple new algorithm for k -clustering in the sliding window setting (Section 3.2). The algorithm is an example of a more general technique that we develop for minimization problems in this setting. (Section 3).
- Prove that the algorithm needs space linear in k to obtain a constant approximate solution (Theorem 3.4), thus improving over the best previously known result which required $\Omega(k^3)$ space.
- Show empirically that the algorithm is orders of magnitude faster, more space efficient, and more accurate than previous solutions, even for small values of k (Section 4).

2 Preliminaries

Let X be a set of arbitrary points, and $d : X \times X \rightarrow \mathbb{R}$ be a distance function. We assume that (X, d) is an arbitrary metric space, that is, d is non-negative, symmetric, and satisfies the triangle inequality. For simplicity of exposition we will make a series of additional assumptions, in Appendix F, we explain how we can remove all these assumptions. We assume that the distances are normalized to lie between 1 and Δ . We will also consider weighted instances of our problem where, in addition, we are given a function $\text{weight} : X \rightarrow \mathbb{Z}$ denoting the multiplicity of the point.

The k -clustering family of problems asks to find a set of k cluster centers that minimizes a particular objective function. For a point x and a set of points $Y = \{y_1, y_2, \dots, y_m\}$, we let $d(x, Y) = \min_{y \in Y} d(x, y)$, and let $\text{cl}(x, Y)$ be the point that realizes it, $\arg \min_{y \in Y} d(x, y)$. The cost of a set of centers \mathcal{C} is: $f_p(X, \mathcal{C}) = \sum_{x \in X} d^p(x, \mathcal{C})$. Similarly for weighted instances, we have $f_p(X, \text{weight}, \mathcal{C}) = \sum_{x \in X} \text{weight}(x) d^p(x, \mathcal{C})$.

Note that for $p = 2$, this is precisely the k -MEDOIDS problem.³ For $p = 1$, the above encodes the k -MEDIAN problem. When p is clear from the context, we will drop the subscript. We also refer to the optimum cost for a particular instance (X, d) as $\text{OPT}_p(X)$, and the optimal clustering as $\mathcal{C}_p^*(X) = \{c_1^*, c_2^*, \dots, c_k^*\}$, shortening to \mathcal{C}^* when clear from context. Throughout the paper, we assume that p is a constant with $p \geq 1$.

While mapping a point to its nearest cluster is optimal, any map $\mu : X \rightarrow X$ will produce a valid clustering. In a slight abuse of notation we extend the definition of f_p to say $f_p(X, \mu) = \sum_{x \in X} d(x, \mu(x))^p$.

In this work, we are interested in algorithms for sliding window problems, we refer to the window size as w and to the set of elements in the active window as W , and we use n for the size of the entire stream, typically $n \gg w$. We denote by X_t the t -th element of the stream and by $X_{[a,b]}$ the subset of the stream from time a to b (both included). For simplicity of exposition, we assume that we have access to a lower bound m and upper bound M of the cost of the optimal solution in any sliding window.⁴

We use two tools repeatedly in our analysis. The first is the relaxed triangle inequality. For $p \geq 1$ and any $x, y, z \in X$, we have: $d(x, y)^p \leq 2^{p-1}(d(x, z)^p + d(z, y)^p)$. The second is the fact that the value of the optimum solution of a clustering problem does not change drastically if the points are shifted around by a small amount. This is captured by Lemma 2.1 which was first proved in [36]. For completeness we present its proof in Appendix A.

Lemma 2.1. *Given a set of points $X = \{x_1, \dots, x_n\}$ consider a multiset $Y = \{y_1, \dots, y_n\}$ such that $\sum_i d^p(x_i, y_i) \leq \alpha \text{OPT}_p(X)$, for a constant α . Let \mathcal{B}^* be the optimal k -clustering solution for Y . Then $f_p(X, \mathcal{B}^*) \in O((1 + \alpha)\text{OPT}_p(X))$.*

³In the Euclidean space, if the centers do not need to be part of the input, then setting $p = 2$ recovers the k -MEANS problem.

⁴These assumptions are not necessary. In section E, we explain how we estimate them in our experiments and in appendix F we show how we can remove them.

Given a set of points X , a mapping $\mu : X \rightarrow Y$, and a weighted instance defined by (Y, weight) , we say that the weighted instance is *consistent* with μ , if for all $y \in Y$, we have that $\text{weight}(y) = |\{x \in X \mid \mu(x) = y\}|$. We say it is ϵ -consistent (for constant $\epsilon \geq 0$), if for all $y \in Y$, we have that $|\{x \in X \mid \mu(x) = y\}| \leq \text{weight}(y) \leq (1 + \epsilon)|\{x \in X \mid \mu(x) = y\}|$.

Finally, we remark that the k -clustering problem is NP-hard, so our focus will be on finding efficient approximation algorithms. We say that we obtain an α approximation for a clustering problem if $f_p(X, \mathcal{C}) \leq \alpha \cdot \text{OPT}_p(X)$. The best-known approximation factor for all the problems that we consider are constant [1, 19, 37]. Additionally, since the algorithms work in arbitrary metric spaces, we measure update time in terms of distance function evaluations and use the number of points as space cost (all other costs are negligible).

3 Algorithm and Analysis

The starting point of our clustering is the development of efficient sketching technique that, given a stream of points, X , a mapping μ , and a time, τ , returns a weighted instance that is ϵ -consistent with μ for the points inserted at or after τ . To see why having such a sketch is useful, suppose μ has a cost a constant factor larger than the cost of the optimal solution. Then we could get an approximation to the sliding window problem by computing an approximately optimal clustering on the weighted instance (see Lemma 2.1).

To develop such a sketch, we begin by relaxing our goal by allowing our sketch to return a weighted instance that is ϵ -consistent with μ for the *entire stream* X as opposed to the substream starting at X_τ . Although a single sketch with this property is not enough to obtain a good algorithm for the overall problem, we design a sliding window algorithm that builds multiple such sketches in parallel. We can show that it is enough to maintain a polylogarithmic number of carefully chosen sketches to guarantee that we can return a good approximation to the optimal solution in the active window.

In subsection 3.1 we describe how we construct a single efficient sketch. Then, in the subsection 3.2, we describe how we can combine different sketches to obtain a good approximation. All of the missing proofs of the lemmas and the pseudo-code for all the missing algorithms are presented in the appendix.

3.1 Augmented Meyerson Sketch

Our sketching technique builds upon previous clustering algorithms developed for the streaming model of computation. Among these, a powerful approach is the sketch introduced for facility location problems by Meyerson [48].

At its core, given an approximate lower bound to the value of the optimum solution, Meyerson's algorithm constructs a set \mathcal{C} of size $O(k \log \Delta)$, known as a *sketch*, and a consistent weighted instance, such that, with constant probability, $f_p(X, \mathcal{C}) \in O(\text{OPT}_p(X))$. Given such a sketch, it is easy to both: amplify the success probability to be arbitrarily close to 1 by running multiple copies in parallel, and reduce the number of centers to k by keeping track of the number of points assigned to each $c \in \mathcal{C}$ and then clustering this weighted instance into k groups.

What makes the sketch appealing in practice is its easy construction—each arriving point is added as a new center with some carefully chosen probability. If a new point does not make it as a center, it is assigned to the nearest existing center, and the latter's weight is incremented by 1.

Meyerson algorithm was initially designed for online problems, and then adapted to algorithms in the streaming computation model, where points arrive one at a time but are never deleted. To solve the sliding window problem naively, one can simply start a new sketch with every newly arriving point, but this is inefficient. To overcome these limitations we extend the Meyerson sketch. In particular, there are two challenges that we face in sliding window models:

1. The weight of each cluster is not monotonically increasing, as points that are assigned to the cluster time out and are dropped from the window.
2. The designated center of each cluster may itself expire and be removed from the window, requiring us to pick a new representative for the cluster.

Using some auxiliary bookkeeping we can augment the classic Meyerson sketch to return a weighted instance that is ϵ -consistent with a mapping μ whose cost is a constant factor larger than the cost of the optimal solution for the entire stream X . More precisely,

Lemma 3.1. *Let w be the size of the sliding window, $\epsilon \in (0, 1)$ be a constant and t the current time. Let (X, d) be a metric space and fix $\gamma \in (0, 1)$. The augmented Meyerson algorithm computes an implicit mapping $\mu : X \rightarrow \mathcal{C}$, and an ϵ -consistent weighted instance $(\mathcal{C}, \widehat{\text{weight}})$ for all substreams $X_{[\tau, t]}$ with $\tau \geq t - w$, such that, with probability $1 - \gamma$, we have: $|\mathcal{C}| \leq 2^{2p+8} k \log \gamma^{-1} \log \Delta$ and $f_p(X_{[\tau, t]}, \mathcal{C}) \leq 2^{2p+8} \text{OPT}_p(X)$.*

The algorithm uses space $O(k \log \gamma^{-1} \log \Delta \log(M/m)(\log M + \log w + \log \Delta))$ and stores the cost of the consistent mapping, $f(X, \mu)$, and allows a $1 + \epsilon$ approximation to the cost of the ϵ -consistent mapping, denoted by $\widehat{f}(X_{[\tau, t]}, \mu)$. This is the ϵ -consistent mapping that is computed by the augmented Meyerson algorithm. In section 2, M and m are defined as the upper and lower bounds on the cost of the optimal solution.

Note that when M/m and Δ are polynomial in w ,⁵ the above space bound is $O(k \log \gamma^{-1} \log^3(w))$.

3.2 Sliding Window Algorithm

In the previous section we have shown that we can use the Meyerson sketch to have enough information to output a solution using the points in the active window whose cost is comparable to the cost of the optimum computed on the whole stream. However, we need an algorithm that is competitive with the cost of the optimum solution computed solely on the elements in the sliding window.

We give some intuition behind our algorithm before delving into the details. Suppose we had a good guess on the value of the optimum solution, λ^* and imagine splitting the input x_1, x_2, \dots, x_t into blocks $A_1 = \{x_1, x_2, \dots, x_{b_1}\}$, $A_2 = \{x_{b_1+1}, \dots, x_{b_2}\}$, etc. with the constraints that (i) each block has optimum cost smaller than λ^* , and (ii) is also maximal, that is adding the next element to the block causes its cost to exceed λ^* . It is easy to see, that any sliding window of optimal solution of cost λ^* overlaps at most two blocks. The idea behind our algorithm is that, if we started an augmented Meyerson sketch in each block, and we obtain a good mapping for the suffix of the first of these two blocks, we can recover a good approximate solution for the sliding window.

We now show how to formalize this idea. During the execution of the algorithm, we first discretize the possible values of the optimum solution, and run a set of sketches for each value of λ . Specifically, for each guess λ , we run Algorithm 1 to compute the AugmentedMeyerson for two consecutive substreams, A_λ and B_λ , of the input stream X . (The full pseudocode of AugmentedMeyerson is available in Appendix C.) When a new point, x , arrives we check whether the k -clustering cost of the solution computed on the sketch after adding x to B_λ exceeds λ . If not, we add it to the sketch for B_λ , if so we reset the B_λ substream to x , and rename the old sketch of B_λ as A_λ . Thus the algorithm maintains two sketches, on consecutive subintervals. Notice that the cost of each sketch is at most λ , and each sketch is grown to be maximal before being reset.

We remark that to convert the Meyerson sketch to a k -clustering solution, we need to run a k -clustering algorithm on the weighted instance given by the sketch. Since the problem is NP-hard, let ALG denote any ρ -approximate algorithm, such as the one by [37]. Let $S(Z) = (Y(Z), \text{weight}(Z))$ denote the augmented Meyerson sketch built on a (sub)stream Z , with $Y(Z)$ as the centers, and $\text{weight}(Z)$ as the (approximate) weight function. We denote by $\text{ALG}(S(Z))$ the solution obtained by running ALG over the weighted instance $S(Z)$. Let $\widehat{f}_p(S(Z), \text{ALG}(S(Z)))$ be the estimated cost of the solution $\text{ALG}(S(Z))$ over the stream Z obtained by the sketch $S(Z)$.

We show that we can implement a function \widehat{f}_p that operates only on the information in the augmented Meyerson sketch $S(Z)$ and gives a $\beta \in O(\rho)$ approximation to the cost on the unabridged input.

Lemma 3.2 (Approximate solution and approximate cost from a sketch). *Using an approximation algorithm ALG, from the augmented Meyerson sketch $S(Z)$, with probability $\geq 1 - \gamma$, we can output a solution $\text{ALG}(S(Z))$ and an estimate $\widehat{f}_p(S(Z), \text{ALG}(S(Z)))$ of its cost s.t. $f_p(Z, \text{ALG}(S(Z))) \leq \widehat{f}_p(S(Z), \text{ALG}(S(Z))) \leq \beta(\rho) f_p(Z, \text{OPT}(Z))$ for a constant $\beta(\rho) \leq 2^{3p+6} \rho$ depending only the approximation factor ρ of ALG.*

⁵We note that prior work [17, 25] makes similar assumptions to get a bound depending on w .

Algorithm 1 Meyerson Sketches, $ComputeSketches(X, w, \lambda, m, M, \Delta)$

- 1: Input: A sequence of points $X = x_0, x_1, x_2, \dots, x_n$. The size of the window w . Cost threshold λ . A lower bound m and upper bound M of the cost of the optimal solution and upper bound on distances Δ .
 - 2: Output: Two sketches for the stream S_1 and S_2 .
 - 3: $S_1 \leftarrow \text{AugmentedMeyerson}(\emptyset, w, m, M, \Delta)$; $S_2 \leftarrow \text{AugmentedMeyerson}(\emptyset, w, m, M, \Delta)$
 - 4: $A_\lambda \leftarrow \emptyset$; $B_\lambda \leftarrow \emptyset$ (Recall that A_λ, B_λ are sets and S_1 and S_2 the corresponding sketches. Note that the content of the sets is not stored explicitly.)
 - 5: **for** $x \in X$ **do**
 - 6: Let S_{temp} be computed by $\text{AugmentedMeyerson}(B_\lambda \cup \{x\}, w, m, M, \Delta)$. (Note: it can be computed by adding x to a copy of the sketch maintained by S_2)
 - 7: **if** $\widehat{f}_p(S_{temp}, \text{ALG}(S_{temp})) \leq \lambda$ **then**
 - 8: Add x to the stream of the sketch S_2 . ($B_\lambda \leftarrow B_\lambda \cup \{x\}$, $S_2 \leftarrow \text{AugmentedMeyerson}(B_\lambda, w, m, M, \Delta)$)
 - 9: **else**
 - 10: $S_1 \leftarrow S_2$; $S_2 \leftarrow \text{AugmentedMeyerson}(\{x\}, w, m, M, \Delta)$. ($A_\lambda \leftarrow B_\lambda$; $B_\lambda \leftarrow \{x\}$)
 - 11: **end if**
 - 12: **end for**
 - 13: Return (S_1, S_2) , and start and end times of A_λ and B_λ
-

Composition of sketches from sub-streams Before presenting the global sliding window algorithm that uses these pairs of sketches, we introduce some additional notation. Let $S(Z)$ be the augmented Meyerson sketch computed over the stream Z . Let $\text{Suffix}_\tau(S(Z))$ denote the sketch obtained from a sketch S for the points that arrived after τ . This can be done using the operations defined in Appendix C.3.

We say that a time τ is contained in a substream A if A contains elements inserted on or after time τ . Finally we define A_τ as the suffix of A that contains elements starting at time τ . Given two sketches $S(A)$, and $S(B)$ computed over two disjoint substreams A, B , let $S(A) \cup S(B)$ be the sketch obtained by joining the centers of $S(A)$ and $S(B)$ (and summing their respective weights) in a single instance. We now prove a key property of the augmented Meyerson sketches we defined before.

Lemma 3.3 (Composition with a Suffix of stream). *Given two substreams A, B (with possibly $B = \emptyset$) and a time τ in A , let ALG be a constant approximation algorithm for the k -clustering problem. Then if $\text{OPT}_p(A) \leq O(\text{OPT}_p(A_\tau \cup B))$, then, with probability $\geq 1 - O(\gamma)$, we have $\widehat{f}_p(A_\tau \cup B, \text{ALG}(\text{Suffix}_\tau(S(A)) \cup S(B))) \leq O(\text{OPT}_p(A_\tau \cup B))$.*

The main idea of the proof is to show that $\text{Suffix}_\tau(S(A)) \cup S(B)$ is ϵ -consistent with a good mapping from $A_\tau \cup B$ and then by using a technique similar to Lemma 2.1 show that we can compute a constant approximation from an ϵ -consistent sketch.

Algorithm 2 Our main algorithm. Input: X, m, M, Δ , approx. factor of ALG (β) and δ .

- 1: $\Lambda \leftarrow \{m, (1 + \delta)m, \dots, 2^p \beta(1 + \delta)M\}$
 - 2: **for** $\lambda \in \Lambda$ **do**
 - 3: $S_{\lambda,1}, S_{\lambda,2} \leftarrow \text{ComputeSketches}(X, w, \lambda, m, M, \Delta)$
 - 4: **end for**
 - 5: **if** $B_{\lambda^*} = W$ for some λ^* **then return** $\text{ALG}(S_{\lambda^*,2})$
 - 6: $\lambda^* \leftarrow \min(\{\lambda : A_\lambda \not\subseteq W\})$
 - 7: $\tau \leftarrow \max(|X| - w, 1)$
 - 8: **if** $W \cap A_{\lambda^*} \neq \emptyset$ **then return** $\text{ALG}(\text{Suffix}_\tau(S_{\lambda^*,1}) \cup S_{\lambda^*,2})$
 - 9: **else return** $\text{ALG}(S_{\lambda^*,2})$
-

Final algorithm. We can now present the full algorithm in Algorithm 2. As mentioned before, we run multiple copies of $ComputeSketches$ in parallel, for geometrically increasing values of λ .

For each value of λ , we maintain the pair of sketches over the stream X . Finally, we compute the centers using such sketches. If we get lucky, and for the sliding window W there exists a subsequence where B_{λ^*} is precisely W , we use the appropriate sketch and return $\text{ALG}(S_{\lambda^*,2})$. Otherwise, we find the smallest λ^* for which A_λ is not a subset of W . We then use the pair of sketches associated with A_{λ^*} and B_{λ^*} , combining the sketch of the suffix of A_{λ^*} that intersects with W , and the sketch on B_{λ^*} .

The main result is that this algorithm provides a constant approximation of the k -clustering problem, for any $p \geq 1$, with probability at least $1 - \gamma$, using space linear in k and logarithmic in other parameters. The total running time of the algorithm depends on the complexity of ALG. Let $T(n, k)$ be the complexity of solving an instance of k -clustering with size n points using ALG.

Theorem 3.4. *With probability $1 - \gamma$, Algorithm 2, outputs an $O(1)$ -approximation for the sliding window k -clustering problem using space: $O(k \log(\Delta)(\log(\Delta) + \log(w) + \log(M)) \log^2(M/m) \log(\gamma^{-1} \log(M/m)))$ and total update time $O(T(k \log(\Delta), k) \log^2(M/m) \log(\gamma^{-1} \log(M/m)) (\log(\Delta) + \log(w) + \log(M)))$.*

We remark that if M and Δ are polynomial in w , then the total space is $O(k \log^4 w \log(\log w / \gamma))$ and the total update time is $O(T(k \log w, k) \log^3(w) \log(\log w / \gamma))$. The main component in the constant approximation factor of Theorem 3.4 statement comes from the $2^{3p+5} \rho$ approximation for the insertion-only case [44]. Here p is the norm, and ρ is the offline algorithm factor. Given the composition operation in our analysis in addition to applying triangle inequality and some other steps, we end up with an approximation factor $\approx 2^{8p+6} \rho$. We do not aim to optimize for this approximation factor, however it could be an interesting future direction.

4 Empirical Evaluation

We now describe the methodology of our empirical evaluation before providing our experiments results. We report only the main results in the section, more details on the experiments and results are in Appendix E. Our code is available open-source on github⁶. All datasets used are *publicly-available*.

Datasets. We used 3 real-world datasets from the UCI Repository [29] that have been used in previous experiments on k -clustering for data streams settings: SKINTYPE [12], $n = 245057, d = 4$, SHUTTLE, $n = 58000, d = 9$, and COVERTYPE [13], $n = 581012, d = 54$. Consistent with previous work, we stream all points in the natural order (as they are stored in the dataset). We also use 4 publicly-available synthetic dataset from [32] (the S-Set series) that have ground-truth clusters. We use 4 datasets (s1, s2, s3, s4) that are increasingly harder to cluster and have each $k = 15$ ground-truth clusters. Consistent with previous work, we stream the points in random order (as they are sorted by ground truth in the dataset). In all datasets, we pre-process each dataset to have zero mean and unit standard deviation in each dimension. All experiments use Euclidean distance, we focus on the the K-MEANS objective ($p = 2$) which we use as cost. We use k -means++ [4] as the solver ALG to extract the solution from our sketch.

Parameters. We vary the number of centers, k , from 4 to 40 and window size, w , from 10,000 to 40,000. We experiment with $\delta = [0.1, 0.2]$ and set $\epsilon = 0.05$ (empirically the results are robust to wide settings of ϵ).

Metrics. We focus on three key metrics: cost of the clustering, maximum space requirement of our sketch, and average running time of the update function. To give an implementation independent view into space and update time, we report as space usage the number of points stored, and as update time the number of distance evaluations. All of the other costs are negligible by comparison.

Baselines. We consider the following baselines.

Batch K-Means++: We use k -means++ over the entire window as a proxy for the optimum, since the latter is NP-hard to compute. At every insertion, we report the best solution over 10 runs of k -means++ on the window. Observe that this is inefficient as it requires $\Omega(w)$ space and $\Omega(kw)$ run time per update. **Sampling:** We maintain a random sample of points from the active window, and then run k -means++ on the sample. This allows us to evaluate the performance of a baseline, at the same space cost of our algorithm. **SODA16:** We also evaluated the only previously published algorithm for this setting in [17].

We note that we made some practical modifications to further improve the performance of our algorithm which we report in the Appendix.

Comparison with previous work. We begin by comparing our algorithm to the previously published algorithm of [17]. The baseline in this paragraph is **SODA16** algorithm in [17]. We confirm

⁶https://github.com/google-research/google-research/tree/master/sliding_window_clustering/

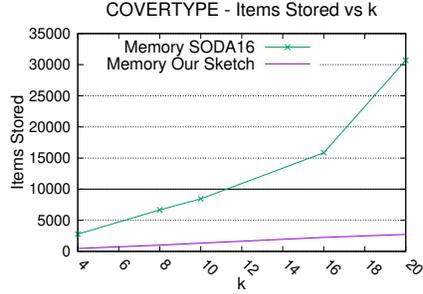


Figure 1: Comparison of the max memory use of SODA16 and our algorithm for $W = 10,000$.

Dataset	k	Space Decr. Factor	Speed-Up Factor	Cost (ratio)
COVER	4	5.23	10.88	99.5%
	16	6.04	13.10	95.1%
SHUTTLE	4	5.07	9.09	106.8%
	16	15.32	15.14	118.9%

Table 1: Decrease of space use, decrease in (speed-up) and ratio of mean cost of the solutions of our algorithm vs the SODA16 baseline (100% means same cost, < 100% means a reduction in cost).

empirically that the memory use of this baseline already exceeds the size of the sliding window for very small k , and that it is significantly slower than our algorithm. Figure 1 shows the space used by our algorithm and by the baseline over the COVERTYPE dataset for a $|W| = 10,000$ and different k . We confirm that our algorithm’s memory grows linearly in k while the baseline grows super-linearly in k and that for $k > 10$ the baseline costs more than storing the entire window. In Table 1 we show that our algorithm is significantly faster and uses less memory than the **SODA16** already for small values of k . In the appendix we show that the difference is even larger for bigger values of k . Given the inefficiency of the SODA16 baseline, for the rest of the section we do not run experiments with it.

Cost of the solution. We now take a look at how the cost of the solution evolves over time during the execution of our algorithm. In Figure 2 we plot the cost of the solution obtained by our algorithm (Sketch), our proxy for the optimum (KM++) and the sampling baseline (Sampling Baseline) on the COVERTYPE dataset. The sampling baseline is allowed to store the same number of points stored by our algorithm (at the same point in time). We use $k = 20$, $|W| = 40,000$, and $\delta = 0.2$. The plot is obtained by computing the cost of the algorithms every 100 timesteps. Observe that our algorithm closely tracks that of the offline algorithm result, even as the cost fluctuates up and down. Our algorithm’s cost is always close to that of the off-line algorithm and significantly better than the random sampling baseline

Update time and space tradeoff. We now investigate the time and space tradeoff of our algorithm. As a baseline we look at the cost required simply to recompute the solution using k -means++ at every time step. In Table 2 ($\delta = 0.2$) we focus on the COVERTYPE dataset, the other results are similar. Table 2 shows the percent of the sliding window data points stored (Space) and the percent of update time (Time) of our algorithm vs a single run of k -means++ over the window. The Table 5 in Appendix E shows that the savings become larger (at parity of k) as $|W|$ grows and that we always

W	k	Space	Time
40000	10	3.5% (0.39%)	0.45% (0.29%)
	20	6.5% (0.87%)	0.93% (0.63%)
	40	11.3% (1.74%)	1.58% (1.23%)

Table 2: Max percentage of sliding window (length W) stored (Space) and median percentage of time (Time) vs. one run of k -means++ (stddev in parantesis).

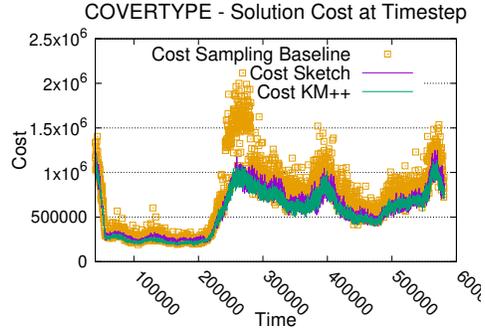


Figure 2: Cost of the solution obtained by our algorithm (Sketch) and the two baselines for $k = 20$, $|W| = 40,000$ and $\delta = 0.2$ on COVERTYPE.

store a small fraction of the window, providing order-of-magnitude speed ups (e.g., we use $< 0.5\%$ of the time of the baseline for $k = 10$, $|W| = 40,000$). Here the baseline is the k -means++ algorithm.

Recovering ground-truth clusters. We evaluated the accuracy of the clusters produced by our algorithm on a dataset with ground-truth clusters using the well known V-Measure accuracy definition for clustering [52]. We observe that on all datasets our algorithm performs better than the sampling baseline and in line with the offline k -means++. For example, on the s1 our algorithm gets V-Measure of 0.969, while k -means++ gets 0.969 and sampling gets 0.933. The full results are available in the appendix E.

5 Conclusion

We present the first algorithms for the k -clustering problem on sliding windows with space linear in k . Empirically we observe that the algorithm performs much better than the analytic bounds, and it allows to store only a small fraction of the input. A natural avenue for future work is to give a tighter analysis, and reduce this gap between theory and practice.

Broader Impact

Clustering is a fundamental unsupervised machine learning problem that lies at the core of multiple real-world applications. In this paper, we address the problem of clustering in a sliding window setting. As we argued in the introduction, the sliding window model allows us to discard old data which is a core principle in data retention policies.

Whenever a clustering algorithm is used on user data it is important to consider the impact it may have on the users. In this work we focus on the algorithmic aspects of the problem and we do not address other considerations of using clustering that may be needed in practical settings. For instance, there is a burgeoning literature on fairness considerations in unsupervised methods, including clustering, which further delves into these issues. We refer to this literature [22, 41, 11] for addressing such issues.

References

- [1] Sara Ahmadian, Ashkan Norouzi-Fard, Ola Svensson, and Justin Ward. Better guarantees for k -means and euclidean k -median by primal-dual algorithms. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 61–72, 2017.
- [2] Kook Jin Ahn, Sudipto Guha, and Andrew McGregor. Analyzing graph structure via linear measurements. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '12*, pages 459–467, Philadelphia, PA, USA, 2012. Society for Industrial and Applied Mathematics.

- [3] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 20–29, New York, NY, USA, 1996. ACM.
- [4] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.
- [5] Olivier Bachem, Mario Lucic, and Andreas Krause. Distributed and provably good seedings for k-means in constant rounds. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 292–300. JMLR. org, 2017.
- [6] Olivier Bachem, Mario Lucic, and Silvio Lattanzi. One-shot coresets: The case of k-clustering. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 784–792, 2018.
- [7] Ashwinkumar Badanidiyuru, Baharan Mirzasoleiman, Amin Karbasi, and Andreas Krause. Streaming submodular maximization: Massive data summarization on the fly. In *ACM SIGKDD*, pages 671–680. ACM, 2014.
- [8] Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. Scalable k-means++. *Proc. VLDB Endow.*, 5(7):622–633, March 2012.
- [9] Maria-Florina F Balcan, Steven Ehrlich, and Yingyu Liang. Distributed k-means and k-median clustering on general topologies. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1995–2003. Curran Associates, Inc., 2013.
- [10] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, D. Sivakumar, and Luca Trevisan. Counting distinct elements in a data stream. In *Proceedings of the 6th International Workshop on Randomization and Approximation Techniques*, RANDOM '02, pages 1–10, Berlin, Heidelberg, 2002. Springer-Verlag.
- [11] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairml-book.org, 2019. <http://www.fairmlbook.org>.
- [12] Rajen Bhatt and Abhinav Dhall. Skin segmentation dataset. *UCI Machine Learning Repository*, 2010.
- [13] Jock A Blackard and Denis J Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, 24(3):131–151, 1999.
- [14] Borassi, Epasto, Lattanzi, Vassilvitskii, and Zadimoghaddam. Better sliding window algorithms to maximize subadditive and diversity objectives. In *The Principles of Database Systems (PODS)*, 2019.
- [15] Vladimir Braverman, Petros Drineas, Cameron Musco, Christopher Musco, Jalaj Upadhyay, David P Woodruff, and Samson Zhou. Near optimal linear algebra in the online and sliding window models. *arXiv preprint arXiv:1805.03765*, 2018.
- [16] Vladimir Braverman, Gereon Frahling, Harry Lang, Christian Sohler, and Lin F Yang. Clustering high dimensional dynamic data streams. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 576–585. JMLR. org, 2017.
- [17] Vladimir Braverman, Harry Lang, Keith Levin, and Morteza Monemizadeh. Clustering problems on sliding windows. In *SODA*, pages 1374–1390, 2016.
- [18] Vladimir Braverman and Rafail Ostrovsky. Smooth histograms for sliding windows. In *FOCS*, pages 283–293, 2007.
- [19] Jaroslav Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for k-median, and positive correlation in budgeted optimization. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pages 737–756, 2015.
- [20] Moses Charikar, Liadan O’Callaghan, and Rina Panigrahy. Better streaming algorithms for clustering problems. In *Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing*, STOC '03, pages 30–39, New York, NY, USA, 2003. ACM.

- [21] Jiecao Chen, Huy L. Nguyen, and Qin Zhang. Submodular maximization over sliding windows. *CoRR*, abs/1611.00129, 2016.
- [22] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pages 5029–5037, 2017.
- [23] G. Chrysos, O. Papapetrou, D. Pnevmatikatos, A. Dollas, and M. Garofalakis. Data stream statistics over sliding windows: How to summarize 150 million updates per second on a single node. In *2019 29th International Conference on Field Programmable Logic and Applications (FPL)*, pages 278–285, 2019.
- [24] Vincent Cohen-Addad, Niklas Oskar D Hjuler, Nikos Parotsidis, David Saulpic, and Chris Schwiegelshohn. Fully dynamic consistent facility location. In *Advances in Neural Information Processing Systems*, pages 3250–3260, 2019.
- [25] Vincent Cohen-Addad, Chris Schwiegelshohn, and Christian Sohler. Diameter and k-Center in Sliding Windows. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, pages 19:1–19:12, 2016.
- [26] Vincent Cohen-Addad, Chris Schwiegelshohn, and Christian Sohler. Diameter and k-center in sliding windows. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.
- [27] Michael S Crouch, Andrew McGregor, and Daniel Stubbs. Dynamic graphs in the sliding-window model. In *European Symposium on Algorithms*, pages 337–348. Springer, 2013.
- [28] Mayur Datar, Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Maintaining stream statistics over sliding windows. *SIAM journal on computing*, 31(6):1794–1813, 2002.
- [29] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017.
- [30] Alessandro Epasto, Silvio Lattanzi, Sergei Vassilvitskii, and Morteza Zadimoghaddam. Submodular optimization over sliding windows. In *Proceedings of the 26th International Conference on World Wide Web, WWW '17*, pages 421–430, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.
- [31] Joan Feigenbaum, Sampath Kannan, Andrew McGregor, Siddharth Suri, and Jian Zhang. On graph problems in a semi-streaming model. *Theor. Comput. Sci.*, 348(2):207–216, December 2005.
- [32] Pasi Fränti and Sami Sieranoja. K-means properties on six clustering benchmark datasets, 2018.
- [33] Sutanu Gayen and NV Vinodchandran. New algorithms for distributed sliding windows. In *LIPICs-Leibniz International Proceedings in Informatics*, volume 101. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
- [34] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems*, pages 3513–3526, 2019.
- [35] Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *TCS*, 38:293–306, 1985.
- [36] S. Guha, N. Mishra, R. Motwani, and L. O’Callaghan. Clustering data streams. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science, FOCS '00*, pages 359–, Washington, DC, USA, 2000. IEEE Computer Society.
- [37] Anupam Gupta and Kanat Tangwongsan. Simpler analyses of local search algorithms for facility location. *CoRR*, abs/0809.2554, 2008.
- [38] Oktie Hassanzadeh, Fei Chiang, Hyun Chul Lee, and Renée J Miller. Framework for evaluating clustering algorithms in duplicate detection. *Proceedings of the VLDB Endowment*, 2(1):1282–1293, 2009.
- [39] Wei Hu, Zhao Song, Lin F Yang, and Peilin Zhong. Nearly optimal dynamic k -means clustering for high-dimensional data. *arXiv preprint arXiv:1802.00459*, 2018.
- [40] Anil K Jain. Data clustering: 50 years beyond k -means. *Pattern Recognition Letters*, 31(8):651–666, 2010.
- [41] Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019.

- [42] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. Fair k-center clustering for data summarization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3448–3457, Long Beach, California, USA, 09–15 Jun 2019. PMLR.
- [43] Robert Krauthgamer and David Reitblat. Almost-smooth histograms and sliding-window graph algorithms. *arXiv preprint arXiv:1904.07957*, 2019.
- [44] Silvio Lattanzi and Sergei Vassilvitskii. Consistent k-clustering. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1975–1984. JMLR. org, 2017.
- [45] Shi Li and Ola Svensson. Approximating k-median via pseudo-approximation. *SIAM J. Comput.*, 45(2):530–547, 2016.
- [46] Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. Performance of johnson-lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1027–1038, 2019.
- [47] Gustavo Malkomes, Matt J Kusner, Wenlin Chen, Kilian Q Weinberger, and Benjamin Moseley. Fast distributed k-center clustering with outliers on massive data. In *Advances in Neural Information Processing Systems*, pages 1063–1071, 2015.
- [48] Adam Meyerson. Online facility location. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 426–431. IEEE, 2001.
- [49] Baharan Mirzasoleiman, Amin Karbasi, and Andreas Krause. Deletion-robust submodular maximization: Data summarization with the right to be forgotten. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2449–2458. JMLR. org, 2017.
- [50] S. Muthukrishnan. Data streams: Algorithms and applications. *Found. Trends Theor. Comput. Sci.*, 1(2):117–236, August 2005.
- [51] Dino Oglic and Thomas Gärtner. Nyström method with kernel k-means++ samples as landmarks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2652–2660. JMLR. org, 2017.
- [52] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420, 2007.
- [53] M Ali Rostami, Alieh Saeedi, Eric Peukert, and Erhard Rahm. Interactive visualization of large similarity graphs and entity resolution clusters. In *EDBT*, pages 690–693, 2018.
- [54] Mina Sheikhalishahi, Andrea Saracino, Mohamed Mejri, Nadia Tawbi, and Fabio Martinelli. Fast and effective clustering of spam emails based on structural similarity. In *International Symposium on Foundations and Practice of Security*, pages 195–211. Springer, 2015.
- [55] Jalaj Upadhyay. Sublinear space private algorithms under the sliding window model. In *International Conference on Machine Learning*, pages 6363–6372, 2019.
- [56] Yanhao Wang, Yuchen Li, and Kian-Lee Tan. Coresets for minimum enclosing balls over sliding windows. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 314–323, 2019.
- [57] Jonghem Youn, Junho Shim, and Sang-Goo Lee. Efficient data stream clustering with sliding windows based on locality-sensitive hashing. *IEEE Access*, 6:63757–63776, 2018.

A Omitted Proofs from Preliminaries

Lemma A.1 (Restatement of Lemma 2.1). *Given a set of points $X = \{x_1, \dots, x_n\}$ consider a multiset $Y = \{y_1, \dots, y_n\}$ such that $\sum_i d^p(x_i, y_i) \leq \alpha \text{OPT}_p(X)$, for a constant α . Let \mathcal{B}^* be the optimal solution for Y . Then $f_p(X, \mathcal{B}^*) \in O(\text{OPT}_p(X))$.*

Proof. This proof follows through several applications of triangle inequality. First, observe that:

$$f_p(X, \mathcal{B}^*) = \sum_i d^p(x_i, \mathcal{B}^*) \leq 2^{p-1} \sum_i \left(d^p(x_i, y_i) + d^p(y_i, \mathcal{B}^*) \right) \quad (1)$$

instance (X, d) as $\text{OPT}_p(X)$, and the optimal clustering as $\mathcal{C}_p^*(X) = \{c_1^*, c_2^*, \dots, c_k^*\}$, shortening to \mathcal{C}^* when clear from context. Notice that everywhere in the paper, we assume that p is a constant with $p \geq 1$.

Let $\mathcal{C}^* = \{c_1^*, \dots, c_k^*\}$ be the optimal solution for X . In a slight abuse of notation, let $\mathcal{C}^*(x_i)$ represent the optimal center closest to x_i . In addition, for every c_j^* , let t_j be its closest neighbor in Y , $t_j = \text{cl}(c_j^*, Y)$. Observe that the set $T = \{t_1, \dots, t_k\}$ represents a feasible solution.

Therefore:

$$\sum_i d^p(y_i, \mathcal{B}^*) \leq d^p(y_i, T). \quad (2)$$

Finally, let $t(i) = \text{cl}(\mathcal{C}^*(x_i), Y)$ represent the point in Y which is closest to the optimal center assigned to x_i .

$$\begin{aligned} \sum_i d^p(y_i, \mathcal{B}^*) &\leq \sum_i d^p(y_i, t(i)) \\ &\leq 2^{p-1} \sum_i \left(d^p(y_i, \mathcal{C}^*(x_i)) + d^p(\mathcal{C}^*(x_i), t(i)) \right) \\ &\leq 2^{p-1} \sum_i 2 \cdot d^p(y_i, \mathcal{C}^*(x_i)) \\ &\leq 2^{2p-1} \left(\sum_i d^p(y_i, x_i) + d^p(x_i, \mathcal{C}^*(x_i)) \right) \\ &\leq 2^{2p-1} (1 + \alpha) \text{OPT}_p(X), \end{aligned}$$

Where the first inequality follows because T is a feasible solution, the second and fourth by the generalized triangle inequality, and the third by the definition of $t(i)$. Putting this together with Equation 1 completes the proof. \square

For completeness we present the next lemma that would be useful to show our main theorem

Lemma A.2. *Given a set of points $X = \{x_1, \dots, x_n\}$ by adding points to the set X the optimal cost of the k -clustering can decrease at most by a 2^p .*

Proof. Let $Y = \{x_1, \dots, x_n, y_1, \dots, y_m\}$ be the set obtained by adding $\{y_1, \dots, y_m\}$ arbitrary points to X . Let $\mathcal{C}^* = \{c_1^*, c_2^*, \dots, c_k^*\}$ be the optimal set of centers for Y and let $\mathcal{C}' = \{x_{c_1^*}, \dots, x_{c_k^*}\}$ the (multi-)set where $x_{c_i^*}$ is the closest point to c_i^* in X . Furthermore let $c^*(x)$ be the closest center in \mathcal{C}^* to a point x . Then we have $\text{OPT}_p(X) \leq \sum_{x \in X} d(x, \mathcal{C}') \leq \sum_{x \in X} 2^{p-1} (d(x, c^*(x)) + d(c^*(x), x_{c^*(x)})) \leq \sum_{x \in X} 2^p d(x, c^*(x)) \leq 2^p \sum_{x \in X} d(x, \mathcal{C}^*) \leq 2^p \sum_{y \in Y} d(y, \mathcal{C}^*) \leq 2^p \text{OPT}_p(Y)$. \square

B Meyerson Sketches

In this section we define the Meyerson sketch algorithm and show some basic facts about the property of any mapping μ and properties of the Meyerson sketch.

The results presented here follow from well-known results [20, 44], but we present them here for completeness.

B.1 Meyerson algorithm

We begin by formally stating the guarantees of the Meyerson sketch, giving a constant factor approximation for any constant p .⁷ The following Lemma, captures the main properties of the sketch.

Lemma B.1. *Let (X, d) be a metric space and fix $\gamma \in (0, 1)$. Then the Meyerson sketch algorithm computes a mapping $\mu : X \rightarrow \mathcal{C}$, and a consistent weighted instance $(\mathcal{C}, \text{weight})$, such that $\mathcal{C} \subseteq X$ and, with probability $1 - \gamma$: $|\mathcal{C}| \leq 2^{2p+8} k \log^{1/\gamma} \log \Delta$; $f_p(X, \mathcal{C}) \leq 2^{p+7} \text{OPT}_p(X)$. The algorithm uses at most $2^{2p+8} k \log^{1/\gamma} \log \Delta \lceil \log M/m \rceil$ space and in the consistent mapping μ , each point is mapped to a center inserted before it into the stream.*

We now present the pseudocode of the Meyerson sketch algorithm. Fix the parameter p of the problem. Consider for now to have access to a lower bound to the cost L^p of the optimal solution for the problem, such that $L^p \geq \alpha \text{OPT}_p$, for some constant $0 < \alpha < 1$. We will remove this assumption at the end of the section.

We start by presenting the code for a single Meyerson sketch as described in Algorithm 3. It picks each point as a new center with the probability proportional to the distance of the point from the selected centers so far. We note that a point that is far from the current centers should be added to avoid a large cost for that point. To obtain the results in Lemma B.1 we will run multiple instances of Algorithm 3.

Algorithm 3 Single Meyerson sketch

- 1: Input: A sequence of points $x_0, x_1, x_2, \dots, x_n$. A finite p , a guess L^p , an upper bound Δ of the max distance, and the parameter k .
 - 2: Output: A mapping $\mu : X \rightarrow S$, and a weighted instance (S, weight) that is consistent with X and cost of the moving cost_μ .
 - 3: $S \leftarrow \emptyset$
 - 4: Let X be a set of points and assume L^p is such that $L^p \geq \alpha \text{OPT}_p(X)$, for some constant $0 < \alpha < 1$
 - 5: $\text{cost}_\mu = 0$
 - 6: **for** $x \in X$ **do**
 - 7: **if** $S = \emptyset$ **then**
 - 8: $S \leftarrow \{x\}$
 - 9: **else**
 - 10: With probability $\min\left(\frac{k(1+\log \Delta)d(x,S)^p}{L^p}, 1\right)$ add x to S and set $\text{weight}(x) = 1$
 - 11: Otherwise, let $z \leftarrow \text{argmin}_{y \in S} d(x, y)$, set $\text{weight}(z) \leftarrow \text{weight}(z) + 1$, $\text{cost}_\mu \leftarrow \text{cost}_\mu + d(x, y)^p$
 - 12: **end if**
 - 13: **end for**
 - 14: Return $S, \text{weight}, \text{cost}_\mu$
-

Using a single sketch we can show the following property.

Lemma B.2. *For a constant $\alpha \in (0, 1)$ and input X , Algorithm 3 computes a mapping μ and a consistent weighted instance (S, weight) with probability at least $\frac{1}{2}$:*

$$|S| \leq 4k(1 + \log \Delta) \left(\frac{2^{p+3}}{\alpha^p} + 1 \right) \quad \text{and}$$

$$f_p(X, S) \leq \text{cost}_\mu \leq 2^{p+5} \text{OPT}_p(X).$$

Furthermore, in the consistent mapping μ between X and S , every point in X is mapped to a point in S inserted before it into the stream. The algorithm computes the cost of the mapping cost_μ .

Proof. Our mapping μ is defined by $z = \text{argmin}_{y \in S} d(x, y)$ in Algorithm 3, note that every point is mapped to a point inserted earlier than it. We will show that in expectation S has size $k(1 + \log \Delta) \left(\frac{2^{2p+1}}{\alpha^p} + 1 \right)$ and $f_p(X, S) \leq 2^{p+3} \text{OPT}_p(X)$. The lemma then follows by an application of Markov inequality. Let $c_1^*, c_2^*, \dots, c_k^*$ be the centers in the optimal solution, $C_1^*, C_2^*, \dots, C_k^*$ be the respective clusters and \mathcal{C}^* be the optimal clustering. Let $f_p(C_i^*, \mathcal{C}^*)$ be the cost for cluster i and a_i^* be: $a_i^* = \left(\frac{f_p(C_i^*, \mathcal{C}^*)}{|C_i^*|} \right)^{1/p}$.

⁷We do not optimize constants throughout the paper to keep proofs simpler. We empirically evaluate the algorithm's performance and show the efficacy of our approach.

Consider the set of points in C_i^* with distance at most a_i^* to the center c_i^* . Let $D = \{v_1, v_2, \dots, v_q\}$ be the set of points with distance at most a_i^* to the center c_i^* . Let p_i be the probability that element v_i is added to S , i.e., $p_i = \min(1, \frac{d(v_i, S)^p k(1+\log \Delta)}{L^p})$. With some abuse of notation we say that $p_{q+1} = 1$. We want to estimate the cost of the subset of D in the stream before any point in D is added to S . For the points $v_1, \dots, v_{q'}$, $q' \leq q$, before the first point added to the set, the probability of them being added to the set is $\frac{d(v_i, S)^p k(1+\log \Delta)}{L^p} < 1$ and its cost is $d(v_i, S)^p$. So the expected cost for these points is:

$$\begin{aligned}
& \sum_{i=1}^{q'} \left(\sum_{j=1}^i d(v_j, S)^p \right) \left(p_{i+1} \prod_{j=1}^i (1 - p_j) \right) \\
&= \sum_{i=1}^{q'} d(v_i, S)^p \left(\prod_{j=1}^i (1 - p_j) \right) \left(\sum_{j=i+1}^{q'+1} p_j \prod_{l=i+1}^{j-1} (1 - p_l) \right) \\
&\leq \sum_{i=1}^{q'} d(v_i, S)^p \left(\prod_{j=1}^i (1 - p_j) \right) \\
&\leq \frac{L^p}{k(1 + \log \Delta)} \sum_{i=1}^{q'} \prod_{j=1}^i (1 - p_j) \leq \frac{L^p}{k(1 + \log \Delta)}.
\end{aligned}$$

The first inequality follows from the fact that we are dividing by a number smaller than 1 ($\frac{d(v, S)^p k(1+\log \Delta)}{L^p} < 1$).

After we add a point to the set, we can use the generalized triangle inequality to bound the cost of any point v in the set with $(2^{p-1}((a_i^*)^p + d(v, c_i^*)^p))$. Therefore, the probability of adding v as a center is bounded by $(2^{p-1}((a_i^*)^p + d(v, c_i^*)^p)) k(1 + \log \Delta)/L^p$.

Similarly for any $j > 0$ when we consider the points in C_i^* that have distance between $2^j a_i^*$ and $2^{j+1} a_i^*$ to the center c_i^* , the expected cost for all points added before inserting any element to S is at most $\left(\frac{L^p}{k(1+\log \Delta)} \right)$. After we add a point from this annulus to set S , we can use the generalized triangle inequality to bound the cost of any other point v by $(2^{p-1}((2^{j+1} a_i^* + d(v, c_i^*)^p)) \leq (2^{p-1}((2d(v, c_i^*)^p + d(v, c_i^*)^p))$. Again, the probability of adding this point as a center is $\frac{(2^{p-1}(2d(v, c_i^*)^p + d(v, c_i^*)^p) k(1+\log \Delta))}{L^p}$.

Therefore, we can bound the expected number of centers added by:

$$\begin{aligned}
& \sum_{C_i^*} \left(1 + \log \Delta + \frac{k(1 + \log \Delta)}{L^p} \sum_{v \in C_i^*} 2^{p-1} \left((a_i^*)^p + (2d(v, c_i^*))^p + d(v, c_i^*)^p \right) \right) \\
&\leq \sum_{C_i^*} \left(1 + \log \Delta + \frac{k(1 + \log \Delta)}{L^p} 2^{p+1} \sum_{v \in C_i^*} \left((a_i^*)^p + d(v, c_i^*)^p \right) \right) \\
&\leq k(1 + \log \Delta) + \sum_{C_i^*} \left(k(1 + \log \Delta) 2^{p+2} \frac{f_p(C_i^*, \mathcal{C}^p)}{L^p} \right) \\
&= k(1 + \log \Delta) + \left(k(1 + \log \Delta) 2^{p+2} \frac{\text{OPT}_p(X)}{L^p} \right) \\
&\leq \left(1 + \frac{2^{p+3}}{\alpha^p} \right) k(1 + \log \Delta).
\end{aligned}$$

In a similar manner, we can bound the cost by

$$\begin{aligned}
& \sum_{C_i^*} \left(\frac{L^p}{k(1 + \log \Delta)} (1 + \log \Delta) + \sum_{v \in C_i^*} 2^{p-1} \left((a_i^*)^p + (2d(v, c_i^*))^p + d(v, c_i^*)^p \right) \right) \\
& \leq \sum_{C_i^*} \left(\frac{L^p}{k} + 2^{p+1} \sum_{v \in C_i} \left((a_i^*)^p + d(v, c_i^*)^p \right) \right) \\
& \leq (L^p + 2^{p+2} \text{OPT}_p(X)) \leq 2^{p+3} \text{OPT}_p(X)
\end{aligned}$$

Thus the claim follows. \square

First step in proving Lemma B.1 using Lemma B.2 is to show how to get our guarantees with probability $1 - \gamma$. This is done in Algorithm 4. Lemma B.3 states the properties of this algorithm.

Algorithm 4 *ComputeMeyerson*($X, L^p, \alpha, \gamma, \Delta, p, k$)

- 1: **Input:** A sequence of points X , a lower bound to the optimum L^p , a constant α such a that $L_p \geq \alpha \text{OPT}_p$, a constant γ , an upper bound Δ on the max distance, the parameters p and k of the problem.
 - 2: **Output:** A mapping $\mu : X \rightarrow M$, and a weighted instance (M, weight) that is consistent with X and cost of the moving cost_μ
 - 3: **for** $i \in [2 \log \gamma^{-1}]$ **do**
 - 4: $M_i \leftarrow x_0$
 - 5: $\text{cost}_{\mu_i} = 0$
 - 6: **end for**
 - 7: **for** $x \in X_i$ **do**
 - 8: **for** $i \in [2 \log \gamma^{-1}]$ **do**
 - 9: **if** $|M_i| \leq 4k(1 + \log \Delta) \left(\frac{2^{p+3}}{\alpha^p} + 1 \right)$ **then**
 - 10: **if** $M_i == \emptyset$ **then**
 - 11: $M_i \leftarrow \{x\}$
 - 12: **else**
 - 13: With probability $\min \left(\frac{k(1 + \log \Delta)d(x, S)^p}{L^p}, 1 \right)$ add x to M_i and set $\text{weight}_i(x) = 1$
 - 14: Otherwise, let $z \leftarrow \text{argmin}_{y \in S} d(x, y)$, set $\text{weight}_i(z) \leftarrow \text{weight}_i(z) + 1$, $\text{cost}_{\mu_i} \leftarrow \text{cost}_{\mu_i} + d(x, y)^p$
 - 15: **end if**
 - 16: **end if**
 - 17: **end for**
 - 18: **end for**
 - 19: Let j be the index of the Meyerson sketch of minimum cost cost_{μ_j} such that $|M_j| \leq 4k(1 + \log \Delta) \left(\frac{2^{p+3}}{\alpha^p} + 1 \right)$, if such j does not exist return $M = \cup_{i=1}^{2 \log \gamma^{-1}} M_i$, weight_1, ∞
 - 20: Extend weight_j to give weight 0 to all the points in $M = \cup_{i=1}^{2 \log \gamma^{-1}} M_i$ not contained in M_j
 - 21: Return $M = \cup_{i=1}^{2 \log \gamma^{-1}} M_i$, weight_j , cost_{μ_j}
-

Lemma B.3. *For a constant $\alpha \in (0, 1)$ and input X , Algorithm 4 computes a mapping μ and a consistent weighted instance $M = \cup_{i=1}^{2 \log \gamma^{-1}} (M_i, \text{weight}_i)$ such that with probability at least $1 - \gamma$, we have:*

$$|M| \leq 8k \log \gamma^{-1} (1 + \log \Delta) \left(\frac{2^{p+3}}{\alpha^p} + 1 \right) \in O(k \log \Delta \log \gamma^{-1})$$

$$\text{and } f_p(X, M) \leq 2^{p+5} \text{OPT}_p(X)$$

Furthermore, in the consistent mapping μ between X and M every point in X is mapped to a point in M inserted before it into the stream. The algorithm computes the cost of the mapping cost_μ .

Proof. As mentioned above, Lemma B.2 implies that if we construct $2 \log \gamma^{-1}$ single Meyerson sketches in parallel, with probability in $1 - \gamma$, at least one of them gives a constant approximation to the optimum at every point in time and furthermore the single Meyerson contains only $4k(1 + \log \Delta) \left(\frac{2^{p+3}}{\alpha^p} + 1 \right)$ points.

Now in Algorithm 4 we are almost building $2 \log \gamma^{-1}$ single Meyerson sketches, the only difference is that we stop adding points to a single sketch if that becomes too large. This modification does not change the probability that there exist at least one single sketch that gives a constant approximation to the optimum at every point in time and that contains only $4k(1 + \log \Delta) \left(\frac{2^{p+3}}{\alpha^p} + 1 \right)$ points.

Thus with probability $1 - \gamma$ at least one of the sketches constructed in B.2 gives a constant approximation to the optimum at every point in time. Merging other sketches to this sketch does not affect this property. Furthermore the number of points in each sketch is explicitly bounded by $4k(1 + \log \Delta) \left(\frac{2^{p+3}}{\alpha^p} + 1 \right)$ so the total number of points in M is bounded by $8k \log \gamma^{-1} (1 + \log \Delta) \left(\frac{2^{p+3}}{\alpha^p} + 1 \right)$.

Now we have to prove the existence of a consistent mapping, note that for this we can just use the weighting of a Meyerson sketch with less $4k(1 + \log \Delta) \left(\frac{2^{p+3}}{\alpha^p} + 1 \right)$ and cost less $2^{p+5} \text{OPT}_p(X)$ and assign the weight to all the other points equal to 0. \square

Finally in order to complete the proof of Lemma B.1, we need to remove the assumption of knowing a good lower bound for the optimal solution. We do so by trying different guess of the OPT in particular we try for different L^p in $\{m, 2m, 4m, 8m, \dots, 2^{\lceil \log M/m \rceil} m\}$. This guarantees that we also run it for a guess L^p such that $L^p \geq \frac{1}{2} \text{OPT}_p(X)$. Now we can detect the correct guess by checking the cost of the Meyerson sketches. The pseudo-code of the final algorithm is presented in Algorithm 5.

Algorithm 5 *SimpleMeyerson*($X, m, M, \gamma, \Delta, p, k$)

- 1: **Input:** A sequence of points X , lower bound m and upper bound M to the optimum, and γ , an upper bound Δ on the max distance, the parameters p and k of the problem.
 - 2: **Output:** A mapping $\mu : X \rightarrow X'$, and a weighted instance (X', weight) that is consistent with X with cost of the moving cost_μ
 - 3: **for** $L^p \in \{m, 2m, 4m, 8m, \dots, 2^{\lceil \log M/m \rceil} m\}$ **do**
 - 4: In parallel *ComputeMeyerson*($X, L^p, \alpha = 1/2, \gamma, \Delta, p, k$)
 - 5: **end for**
 - 6: Let ℓ be the smallest index (if it exists, otherwise it is an arbitrary index) in $\{m, 2m, 4m, 8m, \dots, 2^{\lceil \log M/m \rceil} m\}$ for which the output *ComputeMeyerson* called with $L_p = \ell$ has size smaller than $8k \log \gamma^{-1} (1 + \log \Delta) (2^{2p+3} + 1)$ and cost smaller than $2^{p+6} \ell$
 - 7: Return the result of the index ℓ call for *ComputeMeyerson*.
-

We are now ready to prove the main Lemma of this section.

Proof of Lemma B.1. The bound on the size of the set and cost follow from the check in Algorithm 5. The probability of success is larger than or equal to the probability of success for the $L^p \in \{m, 2m, 4m, 8m, \dots, 2^{\lceil \log M/m \rceil} m\}$ such that $L^p \geq \frac{1}{2} \text{OPT}_p(X)$. The total space is a result of running $O(\log M/m)$ times *ComputeMeyerson* calls in parallel and the result in Lemma B.3. \square

C Augmented Meyerson Sketches

We now show how we can augment the sketch data structure with additional information for clustering in the sliding window model.

C.1 Maintaining Weights.

As we mentioned above, one of the challenges in adapting the Meyerson sketch to the sliding window setting lies in tracking the weight of a center as points expire. In fact, some of the points mapped initially to a center by μ may no longer be part of the sliding window. Formally, given a stream $X = x_1, x_2, \dots, x_t$, we would like to maintain an estimate of the weight of each center in \mathcal{C} restricted to the sliding window, i.e., for $c \in \mathcal{C}$, we want to estimate $|\{x \in W : \mu(x) = c\}|$. We denote this

quantity by $\text{weight}_\mu(c, X_{[t-w, t]})$. Maintaining an estimate of such a weight function falls directly into the Smooth Histograms framework introduced by [18]. We leverage their approach to maintain a $(1 + \epsilon)$ estimate using only $\log_{1+\epsilon} w$ additional overhead per cluster center.

Lemma C.1. *Fix constant $\epsilon > 0$. Using additional space $O(|\mathcal{C}| \log_{1+\epsilon}(w))$, we can extend the Meyerson sketch \mathcal{C} , weight with a function $\widehat{\text{weight}} : \mathcal{C} \times [n] \rightarrow \mathbb{Z}$ such that, at every time t , for every $c \in \mathcal{C}$, and every time τ in the active window: $\text{weight}_\mu(c, X_{[\tau, t]}) \leq \widehat{\text{weight}}(c, \tau) \leq (1 + \epsilon) \text{weight}_\mu(c, X_{[\tau, t]})$.*

Proof. In the sketch, for each point $v \in \mathcal{C}$, we maintain a sequence of weights $R_v = (r_{v,1}, r_{v,2}, \dots)$ corresponding to the number of points assigned to v by μ and a sequence of times $T_v = (t_{v,1}, t_{v,2}, \dots)$. The two sequences are initialized as empty. We preserve the invariant that for each time $t_{v,j} \in T_v$, the number of points assigned to center v by μ from time $t_{v,j}$ (inclusive) to the end of the stream is equal to $r_{v,j}$. To do so, when at time i , a point is assigned to v by μ (recall that the assignment is fixed), we increase by one all of the weights stored in R_v and then we add a new weight initialized to 1 to R_v and a new time equal to i to T_v .

To reduce the size of the structure, we maintain only the significant changes in weights in R_v . More precisely, at any time, we delete $r_{v,l}$ and $t_{v,l}$ for any $l \in [2, \dots, |R_v| - 1]$, if $r_{v,l-1} \leq (1 + \epsilon)r_{v,l+1}$. We also renumber the indices to be consecutive. Finally, we remove the $r_{v,l}$ (and the corresponding $t_{v,l}$) for which the weights are larger than $(1 + \epsilon)|W|$.⁸ Notice that at any time, for each v and l , either $r_{v,l} = r_{v,l+1} + 1$ or $r_{v,l} \leq (1 + \epsilon)r_{v,l+1}$. In fact, if $r_{v,l}$ and $r_{v,l+1}$ refer to consecutive assignments of points to the center, the first case is true. If $r_{v,l}$ and $r_{v,l+1}$ became consecutive after the removal of a point between them, the latter condition is true at the time of the removal, and is preserved by adding 1 to both elements.

Now, in order to answer $\widehat{\text{weight}}(c, \tau)$ we return the value r_{i_v} in the corresponding R_v array, where i_v is the index of the largest value smaller or equal to τ in T_v . Then $\text{weight}_\mu(c, X_{[\tau, t]}) \leq \widehat{\text{weight}}(c, \tau) \leq (1 + \epsilon) \text{weight}_\mu(c, X_{[\tau, t]})$. Finally, note that for each $l \in [|R_v| - 2]$, $r_{v,l} > (1 + \epsilon)r_{v,l+2}$, hence, the sequence is decreasing by a factor of $(1 + \epsilon)$ every 2 steps, so that the total length is at most $O(\log_{1+\epsilon}(w))$. \square

C.2 Maintaining Centers.

Our second task is making sure each cluster has a good center. Once again, this issue is unique to the sliding window setting, as in traditional data streams, points (and thus centers) never expire. Specifically, whenever a center c expires, we aim to replace it with a center c' such that, $d(c, c')$ is small. In our context, small means comparable to the distance between the center c and any point x in the window that was mapped to it.

Definition C.2. *Fix a mapping μ and a center c . We say that $y \in X_{[\tau, t]}$ is an ϵ -replacement for c at time τ if: $d(c, y)^p \leq (1 + \epsilon) d(c, x)^p$ for any $x \in X_{[\tau, t]}$ with $\mu(x) = c$, where t is the last time in the current stream.*

To obtain an ϵ -replacement, we consider a sequence of concentric shells with geometrically increasing radii around each center c , and keep the last occurring point in every shell. Since distances are between 1 and Δ , we can bound the total number of shells by $O(\log_{1+\epsilon} \Delta)$ per center. When c needs to be replaced, we simply look for the smallest non-empty shell and return its representative. By construction, we will ensure that the ϵ -replacement property is satisfied.

Lemma C.3. *Let $\epsilon \in (0, 1)$, then with space $O(|\mathcal{C}| \log_{1+\epsilon}(\Delta))$, it is possible to construct a function $\text{last} : \mathcal{C} \times [n] \rightarrow X$ that will always return an ϵ -replacement for $c \in \mathcal{C}$.*

Proof. Recall that the minimum distance between two distinct points is 1, and that Δ is the maximum distance between two points in the dataset. Let Λ to be the set $\{1, (1 + \epsilon), (1 + \epsilon)^2, \dots, \Delta^p\}$ with both extremes included. For each point $y \in \mathcal{C}$, we maintain a sequence of elements $L_y = \{l_{y,\lambda} | \lambda \in \Lambda\}$ and a sequence of times $H_y = \{h_{y,\lambda} | \lambda \in \Lambda\}$ indexed by $\lambda \in \Lambda$. When a point y is added to \mathcal{C} , we

⁸Note that we can do this because no point has weight larger than W inside the sliding window at any point in the algorithm.

initialize the sequence L_y with y and H_y with its insertion time. Whenever a point x is mapped by the Meyerson sketch to y at time i , let $d = d(y, x)^p$, we set $l_{y,\lambda} = x$ and $h_{y,\lambda} = i$ for all $\lambda \geq d$. First note that we never add a point x with distance larger than Δ by definition, so each point is added to at least one set.

Now, to answer $\text{last}(y, \tau)$, we return the point $l_{y,\lambda}$ for the smallest λ such that $h_{y,\lambda} \geq \tau$. If no such λ exists, we return \emptyset .

Note that by construction, we return a point in $\{x_\tau, \dots, x_t\}$ that is mapped to y if and only if the set is non-empty. Now, suppose that we return y' and $y' \in l_{y,\lambda}$. We know that there exists no point $x \in \{x_\tau, \dots, x_t\}$ mapped to y by the Meyerson sketch such that $d(x, y)^p < \frac{\lambda}{1+\epsilon}$; otherwise, the point would be stored in $l_{y, \frac{\lambda}{1+\epsilon}}$, which is a contradiction as we returned $l_{y,\lambda}$. Hence, we have $d(y, y')^p \leq (1 + \epsilon) d(x, y)^p$ for any $x \in \{x_\tau, \dots, x_t\}$ mapped to y by the Meyerson sketch.

Finally, notice that we store $O(\log_{1+\epsilon} \Delta)$ elements for each element in \mathcal{C} (notice that p is a constant). \square

C.3 Suffix Sketch

The two augmentations before allow us to modify the Meyerson sketch in such a way that it can return an approximate solution for any suffix of the stream with length at most w . More precisely, we will show that we maintain a valid solution for the suffix with a cost comparable to the optimum of the entire stream.

Before proving the main result of this section, we need to add some additional bookkeeping to estimate the cost of the sketch inside the sliding window. The idea is similar to that used to maintain the approximate weight of a point.

Lemma C.4. *Let $\epsilon \in (0, 1)$ be a constant. Then using additional space $O(\log_{1+\epsilon}(M/m))$ we can extend the Meyerson sketch $(\mathcal{C}, \text{weight})$ with a function $\widehat{\text{cost}}_\mu : [n] \rightarrow \mathbb{R}$ such that, for every time τ in the active window: $\text{cost}_\mu(\tau) \leq \widehat{\text{cost}}_\mu(\tau) \leq (1 + \epsilon)\text{cost}_\mu(\tau)$, where $\text{cost}_\mu(\tau)$ is the mapping cost for points inserted after time τ*

Proof. The proof is very similar to the proof of Lemma C.1 and we include it here for completeness.

In the sketch, we maintain a sequence of cost $B = (b_1, b_2, \dots)$ corresponding to different moving costs of μ and a sequence of times $G = (g_1, g_2, \dots)$. The two sequences are initialized as empty. We preserve the invariant that for each time $g_j \in G$, the moving cost of μ for points inserted after time g_j (inclusive) to the end of the stream is equal to b_j . To do so, when at time i , a point, x_i is assigned to some center y by μ (recall that the assignment is fixed), we increase by $d(x_i, y)^p$ all of the weights stored in B and then we add a new weight initialized to $d(x_i, y)^p$ to B and a new time equal to i to G .

To reduce the size of the structure, we maintain only the significant changes in weights in B . More precisely, at any time, we delete b_l and g_l for any $l \in [2, \dots, |B| - 1]$, if $b_{l-1} \leq (1 + \epsilon)b_{l+1}$. We also renumber the indices to be consecutive. Finally, we remove the b_l (and the corresponding g_l) for which the weights are larger than $(1 + \epsilon)2^{p+7}M$.⁹ Notice that at any time, for each l , either $b_l = b_{l+1} + 1$ or $b_l \leq (1 + \epsilon)b_{l+1}$. In fact, if b_l and b_{l+1} refer to consecutive assignments of points to the center, the first case is true. If b_l and b_{l+1} became consecutive after the removal of a point between them, the latter condition is true at the time of the removal, and is preserved by adding 1 to both elements.

Now, to compute the cost of mapping for every suffix, $\widehat{\text{cost}}_\mu(\tau)$, we return the value b_i in the B array, where i is the index of largest value smaller or equal to τ in G . Then $\text{cost}_\mu(\tau) \leq \widehat{\text{cost}}_\mu(\tau) \leq (1 + \epsilon)\text{cost}_\mu(\tau)$. Finally, note that for each $l \in [|B| - 2]$, $b_l > (1 + \epsilon)b_{l+2}$, hence, the sequence is decreasing by a factor of $(1 + \epsilon)$ every 2 steps, so that the total length is at most $O(\log_{1+\epsilon}(M/m))$. \square

We note that the the algorithm defined previously can be easily modified to add the bookkeeping described in the previous lemmas, in particular we need only to modify the *ComputeMeyerson* function.

⁹Note that we can do this because mapping of a good Meyerson sketch cost more than $(1 + \epsilon)2^{p+7}M$ by Lemma B.1.

Algorithm 6 AugmentedMeyerson(X, w, m, M, Δ)

- 1: **Input:** A sequence of points X , lower bound m and upper bound M to the optimum, and an upper bound Δ on the max distance.
 - 2: **Output:** A data structure allowing to extract a ϵ -consistent weighted instance of the substream $X_{[\tau, t]}$ with centers belonging to $X_{[\tau, t]}$.
 - 3: Let parameters p and k be the parameters of the k -clustering problem.
 - 4: Let γ be the desired probability of success of the individual Meyerson algorithm.
 - 5: Let ϵ be the parameter of the ϵ -consistent mapping and ϵ -replacement of centers as described in sections C.1 and C.2.
 - 6: **for** $L^p \in \{m, 2m, 4m, 8m, \dots, 2^{\lceil \log M/m \rceil} m\}$ **do**
 - 7: In parallel run $ComputeMeyerson(X, L^p, \alpha = 1/2, \gamma, \Delta, p, k)$ with bookkeeping for maintaining ϵ -consistent mappings and ϵ -replacement of centers as described in sections C.1 and C.2.
 - 8: **end for**
 - 9: Let ℓ be the smallest index (if it exists, otherwise it is an arbitrary index) in $\{m, 2m, 4m, 8m, \dots, 2^{\lceil \log M/m \rceil} m\}$ for which the output $ComputeMeyerson$ called with $L_p = \ell$ has size smaller than $8k \log \gamma^{-1} (1 + \log \Delta) (2^{2p+3} + 1)$ and cost smaller than $2^{p+6} \ell$
 - 10: Return the result of the index ℓ call for $ComputeMeyerson$ with the needed booking for allowing the operation $Suffix_\tau$.
-

Now we can formally prove the main properties of our augmented Meyerson sketch.

Lemma C.5 (Lemma 3.1 restated). *Let w be the size of the sliding window, $\epsilon \in (0, 1)$ be a constant and t the current time. Let (X, d) be a metric space and fix $\gamma \in (0, 1)$. The augmented Meyerson algorithm computes an implicit mapping $\mu : X \rightarrow \mathcal{C}$, and an ϵ -consistent weighted instance $(\mathcal{C}, \widehat{\text{weight}})$ for all substreams $X_{[\tau, t]}$ with $\tau \geq t - w$ such that, with probability $1 - \gamma$, we have:*

$$|\mathcal{C}| \leq 2^{2p+8} k \log \gamma^{-1} \log \Delta \quad \text{and} \\ f_p(X_{[\tau, t]}, \mathcal{C}) \leq 2^{2p+8} \text{OPT}_p(X).$$

The algorithm uses space $O(k \log \gamma^{-1} \log \Delta \log M/m (\log M + \log w + \log \Delta))$ and stores the cost of the consistent mapping, $f(X, \mu)$, and also a $1 + \epsilon$ approximation to the cost of the ϵ -consistent mapping, $\widehat{f}(X_{[\tau, t]}, \mu)$.

Proof. The main idea of the proof is to use our three bookkeeping tricks to provide a good sketch also for an arbitrary suffix of a stream. In particular our algorithm will run the Meyerson sketch with the additional bookkeeping as presented in Algorithm 5 and output a weighted instance and its cost using the additional bookkeeping.

Note that by Lemma B.1 we know that with probability $1 - \gamma$ the sets of center that we use in a Meyerson sketch are at most $2^{2p+8} k \log \gamma^{-1} \log \Delta$, our augmented Meyerson sketch change the center as describe in Lemma C.3 to guarantee that $\mathcal{C} \subseteq X_{[\tau, t]}$ but it never increases their number so the final number of center is still bounded by $2^{2p+8} k \log \gamma^{-1} \log \Delta$.

To bound the cost of the moving, we note that $f_p(X_{[\tau, t]}, \mathcal{C})$ is initially bounded by $2^{p+7} \text{OPT}_p(X)$. Although in the augmented sketch we change the set of centers so the cost could increase, nevertheless for every point $y \in X_{[\tau, t]}$, we are guaranteed that there is a point in the final set of centers at distance at most $2^{p+1} d(y, \mu(y))^p$. This follows from Lemma C.3 and triangle inequality and because $2 + \epsilon < 4$. Finally the the factor 2^{2p+8} in the statements comes from multiplying 2^{p+1} with the approximation factor of Meyerson 2^{p+7} .

Finally we notice that we can compute $\widehat{\text{weight}}$ using the bookkeeping and the algorithm presented in Lemma C.1 and the $\widehat{\text{cost}}_\mu(\tau)$ using bookkeeping and the algorithm presented in Lemma C.4.

The space bound follows from the space of the simple Meyerson sketch times the required space for bookkeeping. \square

For completeness, we provide a pseudocode for the AugmentedMeyerson (Algorithm 6) which is used in Algorithm 1. The AugmentedMeyerson corresponds to SimpleMeyerson (Algorithm 5) with the additional bookkeeping described in Section C. More precisely, AugmentedMeyerson stores for each sketch the data structures need for maintaining the estimate of the weights and centers described

in sections C.1 and C.2. Using these data structure the algorithm we can extract the weighted set $\text{Suffix}_\tau(S) = (\mathcal{C}, \widehat{\text{weight}})$ for a given τ which represents an ϵ -consistent weighted instance of the substream $X_{[\tau,t]}$ with centers C belonging to $X_{[\tau,t]}$.

D Omitted Proofs from Sliding Windows subsection

Lemma D.1 (Lemma 3.2 restated). *Using an approximation algorithm ALG, from the augmented Meyerson sketch $S(Z)$, with probability $\geq 1 - \gamma$, we can output a solution $\text{ALG}(S(Z))$ and an estimate $\widehat{f}_p(S(Z), \text{ALG}(S(Z)))$ of its cost s.t.*

$$\begin{aligned} f_p(Z, \text{ALG}(S(Z))) &\leq \widehat{f}_p(S(Z), \text{ALG}(S(Z))) \\ &\leq \beta(\rho) f_p(Z, \text{OPT}(Z)) \end{aligned}$$

for a constant $\beta(\rho) \leq 2^{3p+6}\rho$ depending only the approximation factor ρ of ALG.

Proof. Given the sketch $S(Z)$ computed with Algorithm 5, we obtain a set of centers $Y(Z)$ and their weights $\text{weight}(Z)$, and a cost of the mapping $\text{cost}_\mu(Z)$ (as well as an implicitly defined consistent mapping μ).

So we can construct a weighted instance $(Y(Z), \text{weight}(Z))$ and solve it approximately with ALG to obtain centers $\mathcal{C} = \text{ALG}(S(Z))$ which are given in output. We now bound the cost of \mathcal{C} over Z .

As we have shown in the proof of Lemma 2.1, we have that $f_p(Z, \mathcal{C}) \leq 2^{p-1}(\text{cost}_\mu(Z) + f_p(Y(Z), \text{weight}(Z), \mathcal{C}))$, where $f_p(Y(Z), \text{weight}(Z), \mathcal{C})$ is the cost of the solution \mathcal{C} over the weighted instance. We also know that the weighted instance has an optimum cost of $\leq 2^{2p-1}(\text{cost}_\mu(Z) + \text{OPT}_p(Z))$. So applying an ρ -approximate algorithm we get

$$f_p(Y(Z), \text{weight}(Z), \mathcal{C}) \leq \rho 2^{2p-1}(\text{cost}_\mu(Z) + \text{OPT}_p(Z))$$

Finally we get that $f_p(Z, \mathcal{C}) \leq \rho 2^{2p}(\text{cost}_\mu(Z) + \text{OPT}_p(Z))$ which by Lemma B.3 gives $f_p(Z, \mathcal{C}) \leq 2^{3p+6}\rho \text{OPT}_p(Z)$, with probability $\geq 1 - \gamma$, so the theorem holds with $\beta(\rho) := 2^{3p+6}\rho$.

Finally, notice that the value $\widehat{f}_p(S(Z), \text{ALG}(S(Z))) := 2^{p-1}(\text{cost}_\mu(Z) + f_p(Y(Z), \text{weight}(Z), \mathcal{C}))$, can be computed from $S(Z)$ (using ALG) so we can output it as an estimate of the cost. Notice, that this is valid a lowerbound of the actual cost on Z of the solution, and it is $\leq \beta(\rho) \text{OPT}_p(Z)$. \square

Now we prove some basic properties of Algorithm 1] that we will use in other prove.

Lemma D.2 (Invariants of Algorithm 1). *For a set A_λ , let A_λ^+ be the set A_λ together with the first element of the corresponding B_λ . The following are invariants maintained by the algorithm: (i) A_λ and B_λ are two disjoint consecutive substreams of X , (ii) A_λ precedes B_λ , and (iii) B_λ ends with the current time t and always contains the last element of the stream. (iv) When $|A_\lambda| \geq 1$: $\widehat{f}_p(S(A_\lambda^+), \text{ALG}(S(A_\lambda^+))) > \lambda$. (v) When $|A_\lambda| \geq 2$: $\widehat{f}_p(S(A_\lambda), \text{ALG}(S(A_\lambda))) \leq \lambda$. (vi) When $|B_\lambda| \geq 2$: $\widehat{f}_p(S(B_\lambda), \text{ALG}(S(B_\lambda))) \leq \lambda$*

Proof. The first three points follow from construction. The fourth follows from the fact that A_λ is set to a non-empty sketch only when the value associated with $B_\lambda \cup \{x\}$ is $> \lambda$ and this property is maintained thereafter. Equation (v) follows from the fact that $|A_\lambda| \geq 2$ implies that A_λ has been set as a copy of a prior B_λ which was not a singleton, and this happens only if the value of the cost of the sketch associated with B_λ is $< \lambda$. Equation (vi) follows from the same property. \square

Lemma D.3 (Composition with a Suffix of stream, restatement of Lemma 3.3). *Given two substreams A, B (with possibly $B = \emptyset$) and a time τ in A , let ALG be a constant approximation algorithm for the k -clustering problem. Then if $\text{OPT}_p(A) \leq O(\text{OPT}_p(A_\tau \cup B))$, then, with probability $\geq 1 - O(\gamma)$, we have $f_p(A_\tau \cup B, \text{ALG}(\text{Suffix}_\tau(S(A)) \cup S(B))) \leq O(\text{OPT}_p(A_\tau \cup B))$.*

Proof. Let μ_A be the mapping consistent with the sketch $S(A)$ and μ_B be the mapping consistent with the sketch $S(B)$. Note that from this two mappings we can construct a third mapping μ such that $\mu(x) = \mu_A(x)$ if $x \in A$ and $\mu(x) = \mu_B(x)$ if $x \in B$ (note that input points are in

general position so they are all in distinct position). The mapping μ is now a valid mapping for $A_\tau \cup B$, furthermore we can note that the number of points mapped to a point y by the mapping is equal to $|\{x \in A : \mu(x) = y\}| + |\{x \in B : \mu(x) = y\}|$, so by definition of the sketch $\text{Suffix}_\tau(S(A)) \cup S(B)$ and by the fact that $\text{Suffix}_\tau(S(A))$ is ϵ -consistent and $S(B)$ is consistent we obtain that $\text{Suffix}_\tau(S(A)) \cup S(B)$ is ϵ -consistent with μ .

Note from Lemma B.1 we know that the moving cost of μ_A on A is in $O(\text{OPT}_p(A))$ with probability $\geq 1 - \gamma$. Similarly the moving cost of μ_B on B is in $O(\text{OPT}_p(B))$ with probability $\geq 1 - \gamma$, so the moving cost of the function μ on $A \cup B$ is $O(\text{OPT}_p(A) + \text{OPT}_p(B)) \in O(\text{OPT}_p(A_\tau \cup B))$ by the hypothesis in our Lemma with probability $\geq 1 - 2\gamma$. Furthermore also the moving cost of μ on $A_\tau \cup B$ is in $O(\text{OPT}_p(A_\tau \cup B))$.

Now, let \mathcal{C}_{ALG} be the centers selected by a constant approximation algorithm on $\text{Suffix}_\tau(S(A)) \cup S(B)$. From generalized triangle inequality we get $f_p(A_\tau \cup B, \mathcal{C}_{\text{ALG}}) = \sum_{x \in A_\tau \cup B} d(x, \mathcal{C}_{\text{ALG}})^p \leq \sum_{x \in A_\tau \cup B} 2^{p-1} (d(x, \mu(x))^p + d(\mu(x), \mathcal{C}_{\text{ALG}})^p) \in O(\text{OPT}_p(A_\tau \cup B) + \sum_{x \in A_\tau \cup B} d(\mu(x), \mathcal{C}_{\text{ALG}})^p)$. Now from Lemma 2.1 we know that the instance obtained by applying the mapping μ to $A_\tau \cup B$ has optimal solution \mathcal{C}_μ with cost in $O(\text{OPT}_p(A_\tau \cup B))$. Unfortunately we do not apply the algorithm directly on an instance consistent with the mapping, nevertheless by ϵ -consistency we know that the cost of the set of centers \mathcal{C}_μ for the instance $\text{Suffix}_\tau(S(A)) \cup S(B)$ is most $(1 + \epsilon)$ times the cost for the instance obtained by the mapping. So also the optimal solution of $\text{Suffix}_\tau(S(A)) \cup S(B)$ has cost in $O(\text{OPT}_p(A_\tau \cup B))$. So also \mathcal{C}_{ALG} have cost in $O(\text{OPT}_p(A_\tau \cup B))$ because ALG is a constant approximation algorithm. The Lemma follows by noticing that by definition ϵ -consistent $\sum_{x \in A_\tau \cup B} d(\mu(x), \mathcal{C}_{\text{ALG}})^p \leq f_p(\text{Suffix}_\tau(S(A)) \cup S(B), \mathcal{C}_{\text{ALG}})$. \square

Theorem D.4 (Theorem 3.4 restated). *With probability $1 - \gamma$, Algorithm 2, outputs an $O(1)$ -approximation for the sliding window k -clustering problem using space: $O(k \log(\Delta)(\log(\Delta) + \log(w) + \log(M)) \log^2(M/m) \log(\gamma^{-1} \log(M/m)))$ and total update time $O(T(k \log(\Delta), k) \log^2(M/m) \log(\gamma^{-1} \log(M/m)) (\log(\Delta) + \log(w) + \log(M)))$.*

Proof. Notice that at any point in time we run at most $2 \log_{1+\delta}(M/m)$ Augmented Meyerson sketches. We can set each of them to use as probability of error bound $\frac{\gamma}{2 \log_{1+\delta}(M/m)}$ to have total probability of any of them failing $\leq 1 - \gamma$. We continue the analysis on assuming we are in the case all current sketches did not fail.

Consider the sketches maintained by the algorithm. If for some λ , the active window W is identical to the interval B_λ , we return $\text{ALG}(S_2)$ as the solution. Since S_2 is a Meyerson sketch we have that the weighted instance computed by the sketch is consistent with a mapping of cost at most a constant factor larger than the cost of the optimal solution in the active window and so by Lemma 2.1 we have that $\text{ALG}(S_2)$ is a constant approximation. (Note that this is independent of the value of λ , in fact λ is not used by the Meyerson sketch that independently tries several possible lower bounds for the cost of the solution)

Otherwise, we find the maximum $\lambda' \in \Lambda$ for which $A_{\lambda'}$ is not empty and fully contained in the current active window. As long as $|W| > 1$, for the sketches associated with the smallest λ , we have $A_m \neq \emptyset$, which guarantees that such a λ' exists. Here m is the lower bound on the optimum value as defined in Section 2. Furthermore, let $\lambda^* = \lambda'(1 + \delta)$. Note that the fact that for the largest λ , all elements are contained in B_λ , this guarantees that such a λ^* exists in the set of thresholds Λ as well.

We now show that $\lambda' < O(\text{OPT}(W))$, which implies that $\lambda^* < (1 + \delta)O(\text{OPT}(W))$. In fact, suppose λ' is such that $A_{\lambda'}$ is non-empty. We know that $A_{\lambda'} \subseteq W$. Let $A_{\lambda'}^+$ be $A_{\lambda'}$ plus the first element of B_λ , by definition of the algorithm, we have that $\lambda' < \widehat{f}_p(S_{A_{\lambda'}^+}, \text{ALG}(S(A_{\lambda'}^+)))$ where the inequality follows from the invariant (iv) of Lemma D.2, but now $\widehat{f}_p(S_{A_{\lambda'}^+}, \text{ALG}(S(A_{\lambda'}^+)))$ is the cost of a solution computed on the Meyerson sketch for $A_{\lambda'}^+$, for $A_{\lambda'}^+$, so using the fact that the Meyerson sketch provides a constant approximation we get $\lambda' < \widehat{f}_p(S_{A_{\lambda'}^+}, \text{ALG}(S(A_{\lambda'}^+))) \leq O(\text{OPT}(A_{\lambda'}^+))$ but now we can notice that by adding points to any set X we can decrease the cost of the solution by at most a factor 2^p (for a formal proof of this fact look at Lemma A.2) so we get $\lambda' < \widehat{f}_p(S_{A_{\lambda'}^+}, \text{ALG}(S(A_{\lambda'}^+))) \leq O(\text{OPT}(A_{\lambda'}^+)) \leq 2^p O(\text{OPT}(W)) \in O(\text{OPT}(W))$.

So, we can then assume that $\lambda^* < (1 + \delta)O(\text{OPT}(W))$. By definition of λ^* , $W \subseteq A_{\lambda^*} \cup B_{\lambda^*}$. There are three cases to consider.

- W is a strict suffix of B_{λ^*} . In this case, the W is equal to a suffix of B_{λ^*} starting at position τ for some τ . Notice that $|B_{\lambda^*}| \geq 2$ so we know that the sketch over B_{λ^*} has cost $\leq \lambda^*$. so we can apply the Lemma 3.3 to compose the suffix of B_{λ^*} with the empty sketch.
- $W = B_{\lambda^*}$, this ensures to a constant approximation.
- W intersects with A_{λ^*} . We start by computing $\text{Suffix}_t(S_1)$, and then we compute $\text{ALG}(\text{Suffix}_t(S_1) \cup S_2)$. Again we know that the A_{λ^*} has a sketch of small cost so we can apply the composition Lemma 3.3.

Notice that the space bound comes from the total number of points in all sketches used. Notice also that the number of distance function evaluations (and thus the update time) to update the Meyerson sketches per each point in the stream is bounded by the number of points in all sketches (we do one evaluation for each center). The total update time depends also on the complexity of ALG , but we only run it on instances of $k \log(\Delta)$ points, so the total update time is $O(T(k \log(\Delta), k) \log^2(M/m) \log(\gamma^{-1} \log(M/m))(\log(\Delta) + \log(w) + \log(M)))$. Notice also that at the last step, to return a solution, after the updates are done, we only require solving a constant number of k -clustering instances with ALG , again each of them of size $O(k \log(\Delta))$. \square

E Empirical analysis

E.1 Additional information on the experimental setup

Baselines. We consider the following baselines. **Off-Line K-Means++:** We use k -means++ over the entire window as a proxy for the optimum, since the latter is NP-hard to compute. At every insertion, we report the best solution over 10 runs of k -means++ on the window. Observe that this is inefficient as it requires $\Omega(w)$ space and $\Omega(kw)$ cost per update. Specifically, we run k -means++ 10 times and report the best solution obtained. To compare the cost of our algorithm to the optimum, we run k -means++ every time a new element is inserted. While this gives the best cost, this is obviously inefficient: such an approach requires $\Omega(w)$ space and $\Omega(kw)$ cost per update. **Sampling:** We maintain a random sample of points from the active window, and then run k -means++ on the sample. This allows us to evaluate the performance of a baseline, at the same space cost of our algorithm. **SODA16:** We also evaluated the only previously published algorithm for this setting in [17]. We note that we made some practical modifications to further improve the performance of our algorithm which we now describe.

For our algorithm we used a standard optimization from the literature [44]: running one copy of the Meyerson sketch instead of the $O(\log(1/\gamma))$ copies that are needed for high probability statements. We also developed a lazy evaluation of the cost of the Meyerson sketch that saves update time. For the constant α in the Meyerson algorithm we use .5, as shown in the pseudo-code.

Second, as we focus on the cost of the solution, instead of the centers output, we ignore the presence of the center in the window (i.e., we do not use the method to output centers in the window). This is consistent with the problem addressed by [17] in which they do not restrict the centers to be in the sliding window. This way we allow a fair comparison of our results with this baseline.

Third, similarly to [44], in the implementation of the update function, the cost of the sketch is estimated through a lazy evaluation in which the cost is not recomputed from the last time unless at least one new center is added or the number of points or total cost associated with a center is increased by a factor $(1 + \eta)$ from the last full evaluation (we use $\eta = 0.05$); we stop a Meyerson sketch only when the size bound is exceeded, not the cost bound.

Finally, to output a solution, instead of the specific pair of summaries described in our algorithm, we make best-effort use of all available pairs of sketches associated with all guesses of the optimum and return the best one with lowest estimated cost—this also makes the algorithm more resistant to error in estimating the lower and upper bounds of the cost. Our empirical analysis shows that these techniques speed up the computation significantly while retaining very strong approximation factors in real dataset, as we show in this section.

Dataset	k	Space Decr. Factor	Speed-Up Factor	Cost (ratio)
COVER	4	5.23	10.88	99.5%
	8	4.86	10.30	99.9%
	10	4.86	10.36	97.6%
	16	6.04	13.10	95.1%
	20	7.09	16.45	94.3%
SHUTTLE	4	5.07	9.09	106.8%
	8	6.09	7.89	102.4%
	10	6.28	7.79	103.2%
	16	15.32	15.14	118.9%
	20	13.30	12.32	143.7%

Table 3: Decrease of space use, decrease in (speed-up) and ratio of mean cost of the solutions of our algorithm vs the SODA16 baseline. Space decrease and the speed-up factors are multiplicative, the cost column reports the ratio of the cost of our algorithm over that of the SODA16 algorithm.

Upper and lower bounds on cost. Our algorithms described before require an estimate of the lower and upper bounds for the cost to initialize the thresholds used in the summaries. In the Appendix F we show that such assumptions can be removed at the expenses of a more complicated algorithm. In our experiments instead of removing the assumptions we have implemented the simpler algorithms described in the paper and we used an heuristic to estimate the bounds m, M in input to the algorithms. Here we describe the heuristic, which empirically results in strong approximation results and efficiency: sample a number of window-sized substreams from (a prefix) of the stream and compute an approximate solution for each window with any algorithm (e.g., kmeans++ or an insertion-only stream algorithm); then use the min and max cost found (with a slack) for the bounds.

More precisely, let m', M' be the minimum and maximum cost observed in the samples and let μ, σ be the empirical mean and standard deviation of the samples' costs. To estimate m, M we use $m := \max(\frac{m'}{3}, \mu - 3\sigma)$ and $M := \max(3M', \mu + 3\sigma)$. We use 10 samples in our experiments for the estimation. Note that the max with $\frac{m'}{3}$ is used to ensure that the lower bound is set to be > 0 . We use the same technique in the evaluation of the SODA16 baseline [17]. For the Δ input of our algorithms we use the M' obtained from the heuristic.

Implementation details for [17]. To the best of our knowledge, this algorithm has not been evaluated empirically. In implementing this algorithm we used the same speed up techniques used for our algorithm: we used one instead of logarithmic many Meyerson sketches, we used the same implementation of the Meyerson sketch used for our algorithm for fairness of comparison. We also provide the same upper and lower bounds on the optimum used for our algorithm. Finally, we set the grid width parameter δ of the algorithm, as the equivalent δ parameter of our algorithm. We set similarly the β parameter [17].

E.2 More detailed comparison with previous work

Here we report a more detailed analysis in Table 3 where we show that our algorithm uses up to $15x$ less space, does up to $16x$ fewer distance calculations and achieves a solution up to 7% cheaper. Notice how the improvements over the baseline grow with k as predicted by theory. We observe that for $k = 16$ on the SODA16 baseline always has roughly the same space or more that that required to store the entire sliding window making the sliding window algorithm trivial.

Notice how our algorithm result in substantial improvements even for small values of k .

E.3 Additional experiments on cost of the solution

Observe that our algorithm closely tracks that of the optimum, even as the cost fluctuates up and down. Notice how the random sampling baseline presents a particularly high approximation factor in COVERTYPE (Fig. 3(a)) during timesteps when there is a shift in the cost magnitude; this confirms the importance of sliding window algorithms to detect correctly (and re-cluster appropriately) shifts in data promptly. Similar results sre observed in all datasets Figure 3. To capture this analytically, we show the median relative error in Table 4. While our algorithm has a median error within 10% of KM++, the median error of the sampling baseline is 30-40% higher than KM++. Notice how the size of the sliding window does not affect substantially the experimental results.

Dataset	W	Sketch	Sampling	Reduction
COVER	10000	1.09 (0.14)	1.29 (0.31)	-15.9%
	20000	1.05 (0.12)	1.31 (0.30)	-19.4%
	40000	1.06 (0.13)	1.41 (0.44)	-25.0%
SHUTTLE	10000	1.11 (0.16)	1.48 (0.25)	-24.8%
	20000	1.09 (0.14)	1.32 (0.14)	-17.4%
	40000	1.07 (0.11)	1.26 (0.13)	-15.0%
SKIN	10000	1.13 (0.19)	1.40 (0.40)	-19.7%
	20000	1.10 (0.17)	1.34 (0.32)	-17.8%
	40000	1.09 (0.14)	1.31 (0.28)	-16.4%

Table 4: Mean cost ratio of our algorithm (std-dev in parenthesis) and of the sampling baseline over the k -means++ gold standard for $k = 10$, $\delta = 0.2$

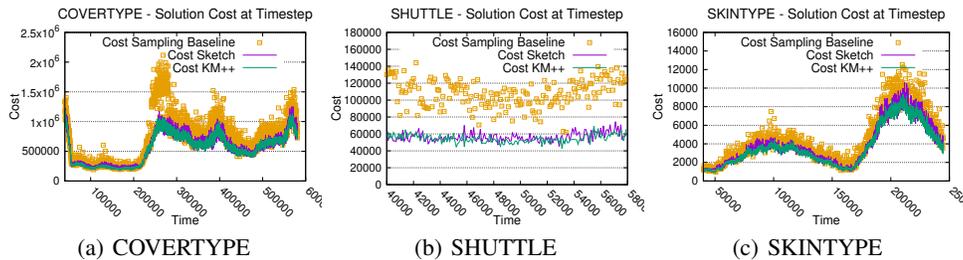


Figure 3: Cost of the solution obtained by our algorithm (Sketch) and the two baselines for $k = 20$, $|W| = 40,000$ and $\delta = 0.2$. Notice that our algorithm’s cost is close to that of the off-line algorithm and significantly better than the random sampling baseline

E.4 Additional experiments on update time and space tradeoff

Additional experiments are presented in Table 5 where we observe substantial savings w.r.t. the naive method. Notice how the savings increase when the size w grows.

E.5 Recovering ground-truth clusters.

In this section we evaluate the accuracy of the clusters produced by our algorithm on a dataset with ground-truth clusters. Note that the ground truth is for all the all set so in our setting we use the cluster assignments of the original ground truth restricted to the particular sliding window. We compared the offline KM++ baseline (KM++), our algorithm (Sketch), and the same-space (Sampling) baseline. We do not evaluate the SODA16 baseline because it is dominated by KM++ in space use and solution quality. We use the well known V-Measure accuracy definition for clustering [52].

We report in Table 6 the results for our empirical evaluation of the accuracy of our algorithm on identifying the ground truth clusters in synthetic datasets.

We observe that our algorithm always dominates the Sampling baseline for all parameters and datasets and it has results comparable to the gold standard offline algorithm while using significantly less space and running time.

W	k	Space	Time
20000	10	7.2% (0.75%)	1.2% (0.68%)
	20	13.72% (2.06%)	2.6% (1.5%)
	40	21.3% (3.4%)	3.5% (2.8%)
40000	10	3.5% (0.39%)	0.45% (0.29%)
	20	6.5% (0.87%)	0.93% (0.63%)
	40	11.3% (1.74%)	1.58% (1.23%)

Table 5: Max percentage of sliding window stored (Space) and median percentage of time (Time) vs. one run of k -means++ (stddev in parenthesis).

Dataset	W	Off-line	Sketch	Sampling
s1	2000	0.958 (0.016)	0.964 (0.019)	0.927 (0.026)
	3000	0.965 (0.015)	0.969 (0.015)	0.926 (0.025)
	4000	0.969 (0.012)	0.969 (0.017)	0.933 (0.030)
s2	2000	0.901 (0.018)	0.903 (0.015)	0.856 (0.026)
	3000	0.889 (0.014)	0.903 (0.014)	0.860 (0.028)
	4000	0.901 (0.014)	0.903 (0.021)	0.851 (0.025)
s3	2000	0.741 (0.018)	0.748 (0.013)	0.717 (0.019)
	3000	0.740 (0.017)	0.745 (0.011)	0.717 (0.015)
	4000	0.735 (0.018)	0.746 (0.017)	0.712 (0.013)
s4	2000	0.678 (0.015)	0.677 (0.017)	0.665 (0.019)
	3000	0.678 (0.017)	0.674 (0.015)	0.659 (0.019)
	4000	0.676 (0.009)	0.675 (0.013)	0.666 (0.013)

Table 6: Average V-Measure of the cluster using the k -means++, sampling baselines and, our algorithm as compared to the ground truth clusters in synthetic datasets (stddev in paranthesis). We use $\delta = 0.2$.

Dataset	W	Sketch	Sampling	Reduction
COVER	10000	1.64	4.09	-60.0%
	20000	1.49	4.05	-63.1%
	40000	1.62	3.79	-57.2%
SHUTTLE	10000	2.03	2.53	-20.0%
	20000	1.69	1.80	-6.3%
	40000	1.45	1.75	-17.00%
SKIN	10000	2.28	4.16	-45.3%
	20000	1.19	4.13	-54.2%
	40000	1.76	3.27	-46.3%

Table 7: Maximum cost ratio of our algorithm and the sampling baseline over the k -means++ gold standard for $k = 10$, $\delta = 0.2$

E.6 Additional empirical results on comparisons with ths baselines

Max cost ratios vs gold standard Our improvement over this baseline is even larger if we look at max error (Table 7) over the entire stream. Notice that, in these experiments, our algorithm always has less than 2.2 cost factor, while the sampling baseline can return costs that are $> 4x$ higher than the gold standard, thus resulting in a reduction of up to -60% in cost. Note also that the empirical approximation ratios are much better than what the worst case theoretical analysis pessimistically predicts.

We report in Table 8 additional results.

Comparison of costs with baseline We now analyze quantitatively the cost of the solutions obtained by our algorithm and the other baselines. Table 4 reports the median ratio (over all timesteps evaluated) between the cost obtained by our algorithm and the k -means++ gold standard (column Sketch) as well as the median ratio of the sampling baseline over the gold standard (Sampling) for $k = 10$, $\delta = 0.2$, and various $|W|$'s. Notice how our algorithm has a median approximation factor (w.r.t. the gold standard) of $5 - 13\%$ higher cost and a reduction of cost w.r.t. the sampling baseline of up to -25% . An approximation factor of < 1.2 is significantly better than the pessimistic worst-case theoretical analysis and shows that our algorithm is highly precise in practice.

Fraction of window stored Figure 5 we show the percent of the sliding window data points stored as a function of the desired number of clusters, k , for the COVERTYPE dataset and $W = 40,000$ (other results are similar). Observe that we always store a small fraction of the dataset, and that the bounds grow linearly with k , as suggested by the theory. In Finally, In Figure 4 we show the average number of points stored per individual sketch. Notice how the number of points stored is mostly unaffected by w , thus confirming that our algorithm can scale to large window sizes.

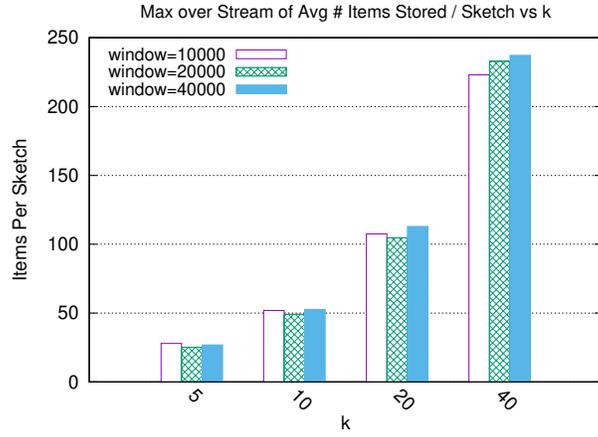


Figure 4: Number of items per sketch stored by our algorithm.

Dataset	W	k	Space	Time
COVER	10000	5	7.24	1.57
		10	14.43	3.12
		20	28.33	6.56
		40	37.22	6.33
	20000	5	3.66	0.62
		10	7.28	1.22
		20	13.72	2.62
		40	21.08	3.51
	40000	5	1.83	0.24
		10	3.51	0.46
		20	6.53	0.93
		40	11.34	1.59
SHUTTLE	10000	5	3.86	0.57
		10	5.77	0.88
		20	8.23	1.60
		40	4.81	0.58
	20000	5	1.60	0.18
		10	1.72	0.17
		20	2.22	0.20
		40	2.22	0.18
	40000	5	0.72	0.09
		10	0.69	0.05
		20	0.90	0.07
		40	0.94	0.05
SKIN	10000	5	7.08	1.61
		10	9.81	1.18
		20	7.09	0.43
		40	9.14	0.29
	20000	5	3.26	0.62
		10	4.15	0.51
		20	3.38	0.22
		40	4.64	0.16
	40000	5	1.67	0.26
		10	2.46	0.28
		20	1.85	0.12
		40	2.10	0.08

Table 8: Max percentage of the items stored w.r.t. the sliding window (Space) and median percentage of distance evaluation vs. one run of k -means++ for $\delta = 0.2$

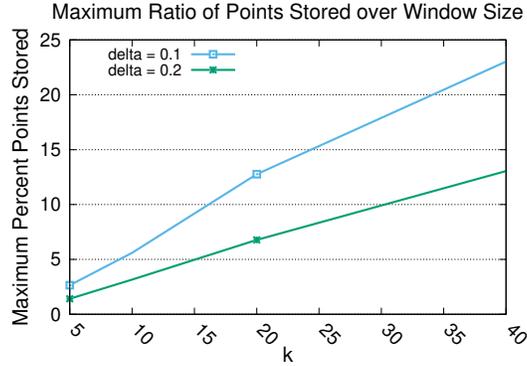


Figure 5: Space usage of our algorithm for different values of k and δ and $W = 40,000$ in COVER-TYPE. Space usage is reported as percent of the window points stored.

E.7 Experiments using a different p

All the experiments we have reported are for k -means ($p = 2$). We observe that our implementation easily adapts to any p by just changing the underlying k -clustering solver. For sake of completeness, we also ran some experiments with the k -median objective, $p = 1$. All the results we observe are in line with that of the k -means case ($p = 2$). For instance, replicating the clustering accuracy experiment using $p = 1$ (See Table 6 for the results for $p = 2$) we obtain the following V-Measure accuracy for the offline algorithm, our stream algorithm and the sampling baseline (same space), 81.8%, 79.9%, 78.5%, respectively. This confirms the same trend observed for k -means. We observe the same trend in terms of speedups and space savings w.r.t. the gold-standard baseline.

F Appendix: Relaxing the assumptions on the input

In the paper we made the following assumptions for the sake of exposition:

- the distances between points are normalized to lie between 1 and Δ ,
- the points are distinct, and
- we have access to a lower bound m and upper bound M of the cost of the optimal solution in any sliding window.

We can relax these assumptions as follows.

F.1 Relaxing the distinct points' assumption.

Notice that we made the assumption that the points in X are distinct only for sake of simplifying the notation of the mapping function from points to centers (we want to define a function from a point). In general, points can be repeated in the stream and without any loss of generality one can define a function (implicitly) on the points assuming their arrival time make them distinct entities (note that we do not need to store the mapping which is not used by the algorithm, this is only for sake of analysis).

F.2 Relaxing the assumption having distances in $[1, \Delta]$ in Meyerson algorithm

Notice that in the Meyerson algorithm (Algorithm 3) we assume that the min distance of two non-coincident points is 1 and we assume to know the maximum distance Δ . The latter value is used to sample points with probability depending on $\log(\Delta)$. In reality we do not need such assumptions. Notice that the $\log(\Delta)$ factor is needed to perform a union bound over the number of exponentially growing balls around a center that starts at distance given by the average cost of a point in a cluster. It is possible to observe that the maximum distance from a point in a cluster C to the center, is at most $|C|$ times the average distance of points to the center in the cluster. So if we assume that the algorithm

is run on a stream of size at most $O(w)$ we can use $\log(w)$ instead of $\log(\Delta)$ in the algorithm, and ignore the assumption on the min distance of points. In the Section F.5 we show that we never need to compute the sketches on streams of length larger than $2w$ so this we can replace $\log(\Delta)$ with $\log(2w)$ in Algorithm 3.

F.3 Estimating the min cost m

The final assumption is knowing a lower bound on the minimum cost m and an upper bound on cost, M for defining the grid containing the guesses of the optimum. To have a lower bound on m , we maintain the most recent $k + 1$ distinct points at all times in a set S . Whenever a new point x arrives, we check to see if we can replace one of the $k + 1$ points in S with x . For instance if $x \notin S$, we add x to S and remove the most stale (oldest) point in S . If an identical point $y = x$ already exists in set S , we just replace y with this new copy x . This way, we always keep the most recent distinct $k + 1$ points in S . If some of the points in S lie outside the current active window W , the total number of distinct points in the active window does not exceed k . In this case, the optimum clustering has cost zero and we have access to the $\leq k$ centers yielding this optimum clustering. Otherwise, all points in S belong to the active window W . Therefore the minimum cost of clustering S lower bounds the optimum clustering cost of the whole window W . Notice that the minimum cost of clustering S is simply the minimum pairwise distance between points in S . This is a valid lower bound on m .

As the sliding window shifts, the lower bound on m might increase or decrease. Whenever, the lower bound increases, in algorithm $SimpleMeyerson(X_t)$, we need to discard some runs of $ComputeMeyerson(X_t, L^p)$. These are instances where L^p is smaller than the new lower bound on m and therefore not relevant anymore. Discarding those runs is easy to do.

However, when the lower bound on m decreases, we need to instantiate new runs of $ComputeMeyerson(X_t, L^p)$ for some smaller values of L^p . But these new runs are only valid for sliding windows that consist of only the current set of points in S . Older versions of S were providing higher lower bounds on m and therefore we do not need to do anything for those older sliding windows. So the only sliding window for which we need to run new instances of $ComputeMeyerson(X_t, L^p)$ is the sliding window starting at time when S begins, which consists of only distinct points in S so far. Notice that we can easily keep the multiplicities of the coincident points for each point in S . Therefore we have all the information we need to run the new instances of $ComputeMeyerson(X_t, L^p)$ for the new smaller values of L^p simulating the algorithm on an arbitrary ordering of points (with multiplicities) from S .

F.4 Estimating the max cost M

Now we will talk about maintaining an upper bound M on the optimum cost. The number of points in the window is w , and the optimum cost cannot be more than w times the diameter of the points in the sliding window. Cohen-Addad et al. [26] design an algorithm that approximates the diameter of points in a sliding window with a constant $3 + \epsilon$ approximation guarantee and using $O(\log(\Delta)/\epsilon)$ memory. Therefore, we can always use this dynamically changing upper bound on M . Whenever our upper bound decreases, we need to discard some runs of $ComputeMeyerson(X_t, L^p)$ for values of L^p that are greater than the new upper bound on M which is easy to do.

However, when the upper bound on M increases, there are new values of L^p for which we need to run $ComputeMeyerson(X_t, L^p)$. The challenge is to run these sketches retrospectively. We do not have access to the whole active window W . Let L^{old} be the highest value of L^p we had before we increased our upper bound on M . For any new value of L^p that exceeds our previous upper bound on M (let's call this new value L^{new}), we initialize the sketch with the sketch we have maintained so far for L^{old} . In other words, we use the closest value of L^p to initialize all new sketches.

In the sampling procedure, L^p is the denominator in the sampling probability. So reusing the sketch with lower L^p means that the previous points have been over-sampled. The main properties of the sketch proved in Lemma B.2 are about upper bounding the cost of Meyerson sketch and its size. The cost is upper bounded in the same way and the proof is intact as the sampling probabilities only increased.

The upper bound on the size can be adapted as follows. We divide the stream into chunks using the times we increased our upper bound on M as pivot points. Notice that we do not need to update L^p

each time M changes but only when M increases by a constant factor. The sketch L^{new} consists of the points we sample in each of these chunks. So we upper bound the size of the sketch in each chunk. The last chunk is similar to any other Meyerson sketch with a set of points being pre-selected in previous chunks. So the proof of Lemma B.2 works here as well. For any other chunk S' , we are running the sketch with the maximum L^p at the time the points in S' are arriving. So this L^p is at least a constant fraction of the optimum cost of S' . This suffices for the argument to upper bound the size in Lemma B.2. To summarize, we have a new overall upper bound on the size of the sketch with an extra $\log(\Delta)$ as the number of chunks is related to logarithm of how much diameter grows (as we start a new chunks only when M increases by a constant factor).

F.5 Ensuring that the each sketch is computed on at most $O(w)$ points

Finally, we now show that we never need to compute any streaming algorithm on a too long stream. This can be obtained by at most doubling the time and memory requirement of the algorithm with the following technique. We keep at most 2 independent instances of our sliding window algorithm (Algorithm 2) the following way. Every w steps we start a new algorithm instance that receives points from this moment on. When an algorithm instance has consumed $2w$ points that algorithm is discarded, meanwhile we update the (at most 2 active algorithms) with each point received. Any time a solution is needed we use the active algorithm that has started earlier. It is easy to see that for such algorithm all the properties needed by the theorem 3.4 are true as the algorithm has consumed the entire active window, so the solution is good. Notice that this ensures that no sketch consumes more than $2w$ points. Also notice that the running time and the memory is doubled at most.