Sequential no-Substitution k-Median-Clustering

Tom Hess
Department of Computer Science
Ben-Gurion University of the Negev
Beer Sheva 8410501, Israel

Sivan Sabato
Department of Computer Science
Ben-Gurion University of the Negev
Beer Sheva 8410501, Israel

Abstract

We study the sample-based k-median clustering objective under a sequential setting without substitutions. In this setting, an i.i.d. sequence of examples is observed. An example can be selected as a center only immediately after it is observed, and it cannot be substituted later. The goal is to select a set of centers with a good k-median cost on the distribution which generated the sequence. We provide an efficient algorithm for this setting, and show that its multiplicative approximation factor is twice the approximation factor of an efficient offline algorithm. In addition, we show that if efficiency requirements are removed, there is an algorithm that can obtain the same approximation factor as the best offline algorithm. We demonstrate in experiments the performance of the efficient algorithm on real data sets. evaluation of the downstream application of face detection. Our code is available at `https://github.com/tomhess/No_Substitution_K_Median`.

1 Introduction

Clustering is an important unsupervised task used for various applications, including, for instance, anomaly detection (Leung and Leckie, 2005), recommender systems (Shepitsen et al., 2008) and image segmentation (Ng et al., 2006). The k-median clustering objective is particularly useful when the partition must be defined using centers from the data, as in some types of image categorization (Dueck and Frey, 2007) and video summarization (Hadi et al., 2006). While clustering has been classically applied to fixed offline data, in recent years clustering on sequential data has become a topic of ongoing research, motivated by various applications where data is observed sequentially, such as detecting communities in social networks (Aggarwal and Yu, 2005), online recommender systems (Nasraoui et al., 2007) and online data summarization (Badanidiyuru et al., 2014). Previous work on clustering sequential data (e.g., Guha et al., 2000; Ailon et al., 2009; Ackermann et al., 2012) has typically focused on cases where the main limitation is memory; the clustering needs to be done on massive amounts of data, and so the data cannot be kept in memory in full. In this work, we study sequential k-median clustering in a new setting, which we call the *no-substitution* setting. In this setting, an i.i.d. sequence of examples is observed. An example can be selected as a center only immediately after it is observed, and it cannot be substituted later. The goal is to select a set of centers with a good k-median cost on the distribution which generated the sequence. This is a natural extension to clustering of the problem of irrevocable item selection from a sequence, which is well-studied in various other settings (see, e.g., Kesselheim and Tönnis 2017; Babaioff et al. 2007, 2008).

The no-substitution setting captures applications of clustering in which the selection of each center involves an immediate and irrevocable action in the real world. For instance, consider selecting a small set of users from those arriving to a shopping website. These users will receive an expensive promotional gift, where the goal is to select the users who will be the most effective in spreading the word about the product. Assuming a budget of k gifts, this can be formalized as a k-median objective, with respect to a metric defined by connections between users, where the selected users are the centers. Offering the gift to a user must be done immediately, before the user leaves the website. The gift also cannot later be reassigned to another user. This is captured by the no-substitution setting. As another example, consider selecting participants for a medical experiment from a stream of patients. The par-

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

ticipants should represent the population, formalized as a k-median objective, and each participant should be selected before leaving the reception desk. These two examples demonstrate the usefulness of the no-substitution setting for real-life applications.

Our contributions. We study the no-substitution setting in a general metric space, assuming that the data sequence is sampled i.i.d. from an unknown distribution, and the goal is to minimize the distribution risk of the selected centers. The focus of this work is obtaining theoretical guarantees for this setting, given a predefined length of stream, a fixed number of centers and a given confidence parameter. We provide a computationally efficient and practical algorithm, called SKM, which uses as a black box a given clustering algorithm which is not restricted to the no-substitution setting. We show that the multiplicative approximation factor obtained by SKM is twice the factor obtained by the black-box algorithm, and that this factor of 2 is tight. We further provide another algorithm, called SKM2, which obtains the same approximation factor as the best possible (though not necessarily efficient) offline algorithm. However, the computational complexity of SKM2 is exponential in k. Whether there exists an efficient no-substitution algorithm with the same approximation factor as the best efficient offline algorithm, is an open question which we leave for future work. Lastly, we demonstrate SKM, the efficient algorithm, on real data sets.

Related Work

We are not aware of previous works which study the no-substitutions setting for k-median clustering defined above.[1] Below we review previous work in related settings. Ben-David (2007) studied sample-based k-median clustering in the offline setting. In this setting, the entire set of sampled data points is observed, and then the k centers are selected from this sample. For the case of a general metric space, Ben-David (2007) provides uniform finite-sample bounds on the convergence of the sample risk to the distribution risk of any choice of centers from the sample.

Algorithms studying clustering on sequential data have mainly assumed a fixed data set and an adversarial ordering, under bounded memory. In this setting, the approximation is with respect to the optimal clustering of the data set. Guha et al. (2000) proposed the first single-pass constant approximation algorithm for the k-median objective with bounded memory. Ailon et al. (2009); Chen (2009); Ackermann et al. (2012) develop algorithms for this setting using coreset constructions. Charikar et al. (2003) design algorithms based on the facility-location objective, using a procedure proposed in Meyerson (2001), which also studies facility location under a random arrival order. Braverman et al. (2016) suggests a space-efficient technique to extend any sample-based offline coreset construction to the streaming (bounded-memory) model. Lang (2018) considers the streaming k-median problem under a random arrival order. Unlike the no-substitution setting, these algorithms can repeatedly change their selection of centers, or simply select a center that has appeared sometime in the past.

Liberty et al. (2016) studies the online k-means objective with an arbitrary arrival order, in a setting where each observed point must either be allocated to an already-defined cluster or start a new cluster. This setting can be seen as a variant of the no-substitution setting, since a chosen center cannot be discarded later. However, the proposed algorithm selects $O(k \log m)$ centers, where m is the sample size, and it is shown that in this adversarial setting, one must select more than k elements to obtain a bounded approximation factor. Lattanzi and Vassilvitskii (2017) propose an online k-median algorithm which minimizes the number of necessary recalculations of a clustering.

The no-substitution setting bears a resemblance to the secretary problem under a cardinality constraint. In this setting, a set of limited cardinality must be selected with no substitutions from a sequence of objects, so as to optimize a given objective. Bateni et al. (2010); Feldman et al. (2011); Kesselheim and Tönnis (2017) study this setting when the objective is monotone and submodular. Badanidiyuru et al. (2014) suggest reformulating the k-median objective as a submodular function. However, this reformulation does not preserve the approximation ratio of the k-median objective. It also requires access to an oracle for function value calculations, which is not readily available in the sample-based sequential clustering setting. Sabato and Hess (2018) study a more general problem of converting an offline algorithm to a no-substitution algorithm in an interactive setting.

2 Setting and Preliminaries

For an integer i, denote $[i] := \{1, \ldots, i\}$. Let $(\mathcal{X}, \rho)$ be a bounded metric space, and assume $\rho \leq 1$. For $c \in \mathcal{X}$ and $r \geq 0$, let $\mathrm{Ball}(c, r) := \{x \in \mathcal{X} \mid \rho(c, x) \leq r\}$. Assume a probability distribution P over $\mathcal{X}$. Below, we assume $X \sim P$, unless explicitly noted otherwise. For $B \subseteq \mathcal{X}$, denote $\mathbb{P}[B] := \mathbb{P}[X \in B]$. A k-clustering is a set of k points $T = \{t_1, \ldots, t_k\} \subseteq \mathcal{X}$ which represent the centers of the clusters. Given a probabil-

[1] Citation of a follow-up work by other authors, which cites an earlier unpublished version of this work, was removed for the anonymous submission.

ity distribution P, the k-median risk of T on P is $R(P,T) := \mathbb{E}[\min_{i\in[k]} \rho(X,t_i)]$. For a finite set $S \subseteq \mathcal{X}$, $R(S,T)$ is the risk of T on the uniform distribution over S. We will generally assume an i.i.d. sample $S \sim P^m$. For convenience of presentation, we treat S as both a sequence and as a set interchangeably, ignoring the possibility of duplicate examples in the sample. These can be easily handled by using multisets, and taking the necessary precautions when selecting an element from S. When a minimization with respect to ρ is performed, we assume that ties are broken arbitrarily.

Denote by $\mathrm{OPT} \in \operatorname{argmin}_{T\in\mathcal{X}^k} R(P,T)$ a specific optimal solution of the k-median clustering problem, where the minimization is over all possible k-clusterings in $\mathcal{X}$; we assume for simplicity that such an optimizer always exists. Denote by $\mathrm{OPT}_S \in \operatorname{argmin}_{T\in\mathcal{X}^k} R(S,T)$ a solution that minimizes the risk on S using centers from $\mathcal{X}$.

In the no-substitution k-median setting, the algorithm does not know the distribution P. It observes the i.i.d. sample $S \sim P^m$ in a sequence and selects centers from S. Formally, there are m time steps. At time step t, a single example $x_t \sim P$ is observed and can be selected as a center. x_t cannot be selected as a center at a later time step. Moreover, once a center is selected, it cannot be removed or substituted. The algorithm can select k elements from S as centers, to form the k-clustering T. The objective is to obtain a small $R(P,T)$, compared to the optimal $R(P,\mathrm{OPT})$.

An *offline k-median algorithm* $\mathcal{A}$ takes as input a finite set of points S from $\mathcal{X}$ and outputs a k-clustering $T \subseteq S$. We say that $\mathcal{A}$ is a *β-approximation* offline k-median algorithm, for some $\beta \geq 1$, if for all input sets S, $R(S,\mathcal{A}(S)) \leq \beta \cdot R(S,\mathrm{OPT}_S)$. It is well known (e.g., Guha et al., 2000) that for any data set S, $R(S, \operatorname{argmin}_{T\in S^k} R(S,T)) \leq 2R(S,OPT_S)$, and that this upper bound is tight.[2] Therefore, the lowest possible value for β in a general metric space is 2.

For a non-negative function $f(k,m,\delta)$, we denote by $O(f(k,m,\delta))$ a function which is upper-bounded by $C \cdot f(k,m,\delta)$ for some universal constant C, for any integer k, $\delta \in (0,1)$, and sufficiently large m.

3 An Efficient Algorithm: SKM

The first algorithm that we propose is called SKM (Sequential K-Median). SKM works in two phases. In the first phase, the incoming elements are observed and no element is selected. In the second phase, elements are selected based on the information gained in the first phase. SKM receives as input a confidence parameter δ, the number of clusters k, the sequence size m, and access to a black-box offline k-median algorithm $\mathcal{A}$. The main challenge in designing SKM is to define a selection rule for elements from the second phase, based on the information gained in the first phase. This information should have uniform finite-sample convergence properties, so that the error of the solution can be bounded. In addition, the selection rule should guarantee selecting k centers with a high probability. SKM constructs this rule by combining the solution of $\mathcal{A}$, calculated on the examples of the first phase, with estimations on the distribution.

Algorithm 1 SKM

input $k, m \in \mathbb{N}$, $\delta \in (0,1)$, offline k-median algorithm $\mathcal{A}$, sequential access to $S = (x_i)_{i=1}^m \sim P^m$
output A k-clustering $T_{\text{out}} \subseteq S$.

```
q ← 43 ln(2m²/δ)/m ; T_out ← ∅
Get m/2 samples from S; S_1 ← (x_1, ..., x_{m/2}).
Run A on S_1, and set {c_1, ... c_k} ← A(S_1).
for j = m/2 + 1 to m do
    Get the next sample x_j
    if ∃i ∈ [k] such that x_j ∈ qball_{S_1}(c_i, q) and T_out ∩ qball_{S_1}(c_i, q) = ∅ then
        T_out ← T_out ∪ {x_j}.
    end if
end for
return T_out
```

SKM is listed in Alg. 1. It uses the following notation. Denote the elements observed in the first phase by S_1, and those in the second phase by S_2. Elements from S_2 are selected as centers if they are close to the centers calculated by $\mathcal{A}$ for S_1. Importantly, closeness is measured relative to the distribution of distances in S_1: An element is considered close to a center if its distance is smaller than all but at most a q fraction of the points in S_1. Formally, for $x, y \in S$, define $B(x,y) := \mathbb{P}[\mathrm{Ball}(x,\rho(x,y))]$. This is the probability mass of points whose distance from x is at most the distance of y. For a set of points S and $x, y \in S$, let $\hat{B}_S(x,y)$ be the fraction of the points in $S \setminus \{x,y\}$ that are in $\mathrm{Ball}(x,\rho(x,y))$. For $q \in (0,1)$, let $\mathrm{qp}_S(x,q)$ be some point $y \in \operatorname{argmin}\{\rho(x,y) \mid y \in S, \hat{B}_S(x,y) \geq q\}$. Denote the ball in $\mathcal{X}$ with center x and radius determined by $\mathrm{qp}_S(x,q)$ by $\mathrm{qball}_S(x,q) := \mathrm{Ball}(x,\rho(x,\mathrm{qp}_S(x,q)))$.

The computational complexity of SKM is $O(km\log(m))$ plus the complexity of the black-box algorithm $\mathcal{A}$. In a memory-restricted online setting, a small variant of SKM can be used, which calculates the clustering of $\mathcal{A}$ on half of S_1 and finds the ball radii based on q using the second half of S_1,

[2] The tightness can be observed by considering a star graph where the metric is the shortest path between vertices, and the center of the star is in $\mathcal{X} \setminus S$.

and a memory of $O(k\log(m/\delta))$ examples. Combined with a black-box $\mathcal{A}$ which is itself online and memory-restricted, the result is an online memory-restricted no-substitution algorithm.

3.1 Risk upper bound for SKM

The following theorem provide the guarantee for SKM.

Theorem 3.1. *Suppose that SKM is run with inputs $k \in \mathbb{N}$, $m \geq \max(2k, 24)$, $\delta \in (0,1)$ and $\mathcal{A}$, where $\mathcal{A}$ is a β-approximation offline k-median algorithm. For any $\gamma \in (0, \frac{1}{2})$ and any distribution P over $\mathcal{X}$, with a probability at least $1-\delta$,*

$$\begin{aligned} R(P, T_{\mathrm{out}}) &\leq (2+2\gamma)\beta R(P, \mathrm{OPT}) \\ &\quad + ((2+2\gamma)\beta + 1)\sqrt{(k\ln(m/2) + \ln(4/\delta))/m} \\ &\quad + (1/\gamma)\cdot(44k\log(2m^2/\delta))/m. \end{aligned}$$

Theorem 3.1 gives a range of trade-offs between additive and multiplicative errors, depending on the value of γ. In particular, by setting $\gamma = \sqrt{(k\log(m/\delta))/m}$ and noting that $R(P, \mathrm{OPT}) \leq 1$, we get

$$R(P, T_{\mathrm{out}}) \leq 2\beta R(P, \mathrm{OPT}) + \beta \cdot O(\sqrt{\frac{k\log(\frac{m}{\delta})}{m}}). \quad (1)$$

This guarantee can be compared to the guarantee of an offline algorithm that uses the same k-median algorithm $\mathcal{A}$ as a black box. As shown in Ben-David (2007), for $S \sim P^m$, with a probability at least $1-\delta$, for every k-clustering $T \subseteq S$ and for $T = \mathrm{OPT}$,

$$|R(P,T) - R(S,T)| \leq O(\sqrt{\frac{k\ln m + \ln(\frac{1}{\delta})}{m}}). \quad (2)$$

Denote the RHS by $O(f(m,k,\delta))$. Therefore,

$$\begin{aligned} R(S, \mathcal{A}(S)) &\leq \beta R(S, \mathrm{OPT}_S) \leq \beta R(S, \mathrm{OPT}) \\ &\leq \beta R(P, \mathrm{OPT}) + \beta \cdot O(f(m,k,\delta)). \end{aligned}$$

Since Eq. (2) holds also for $T = \mathcal{A}(S)$, it follows that $R(P, \mathcal{A}(S)) \leq \beta R(P, \mathrm{OPT}) + \beta \cdot O(f(m,k,\delta))$. Therefore, the additive errors of this guarantee and that of Theorem 3.1 have a similar dependence on m, k, δ and β. When $m \to \infty$, the additive errors go to zero, and there remains the approximation factor of 2β for SKM, instead of β for the offline algorithm. We show in Section 3.2 that the 2β approximation factor is tight.

To prove Theorem 3.1, we first prove that with a high probability, SKM succeeds in selecting k centers from S_2. This requires showing that the estimate of the mass of $\mathrm{qball}_{S_1}(c_i, q)$ using S_1 is close to its true mass on the distribution. We use the following lemma, proved in the supplementary material using the empirical Bernstein's inequality of Maurer and Pontil (2009):

Lemma 3.2. *Let $Y_1, \ldots, Y_n$ be i.i.d. random variables over $[0,1]$ with mean μ. Let $\hat{\mu} = \frac{1}{n}\sum_{i\in[n]} Y_i$ be their empirical mean. Then, with a probability at least $1-\delta$, $\hat{\mu} \leq \max(16\ln(\frac{2}{\delta})/(n-1), 2\mu)$.*

This result is used in the proof of the following lemma. For readability, we denote the sizes of S_1 and S_2 by m_1, m_2 respectively.

Lemma 3.3. *For every distribution P over $\mathcal{X}$, if $m_1 \geq \max(k, 12)$ then with a probability at least $1-\delta/2$, for every $i \in [k]$, SKM selects a point in $\mathrm{qball}_{S_1}(c_i, q)$ from S_2.*

Proof. For $x, y \in S_1$, denote $\hat{B} := \hat{B}_{S_1}(x,y)$. Apply Lemma 3.2 by letting $Y_1, \ldots, Y_n$ stand for the indicators $\mathbb{I}[z \in B(x,y)]$ for $z \in S_1 \setminus \{x,y\}$, $n = m_1 - 2$, $\hat{\mu} = \hat{B}, \mu = B(x,y)$. It follows that with a probability at least $1-\delta$, if $\hat{B} \geq 16\ln(\frac{2}{\delta})/(m_1-3)$, then $B(x,y) \geq \hat{B}/2$, hence $B(x,y) \geq 8\ln(\frac{2}{\delta})/(m_1-3)$. By a union bound on the pairs in S_1, we have that with a probability of $1-\delta/4$, for all pairs $x, y \in S_1$,

$$\begin{aligned} &\hat{B}_{S_1}(x,y) \geq 16\ln(\frac{8m_1^2}{\delta})/(m_1-3) \\ \implies \quad &B(x,y) \geq \hat{B}_{S_1}(x,y)/2. \end{aligned}$$

In particular, this holds for $x = c_i$ and $y = y_i := \mathrm{qp}_{S_1}(c_i, q)$, where $c_1, \ldots, c_k$ are the centers returned by $\mathcal{A}$ in SKM. Denote $\hat{B}_i = \hat{B}_{S_1}(c_i, y_i)$. By definition of y_i, for all $i \in [k]$, $\hat{B}_i \geq q$. In addition, by definition of $B(\cdot,\cdot)$ and qball, we have that $\mathbb{P}[\mathrm{qball}_{S_1}(c_i,q)] = B(c_i, y_i)$. Since $m_1 \geq 12$, we have $m_1 - 3 \geq 3m_1/4$. Therefore, $\hat{B}_i \geq q = 43\ln(2m^2/\delta)/m \geq 16\ln(8m_1^2/\delta)/(m_1-3)$. Therefore, with a probability at least $1-\delta/4$, S_1 satisfies that for all $i \in [k], \mathbb{P}[\mathrm{qball}_{S_1}(c_i,q)] \geq q/2 \geq \frac{\ln(\frac{4k}{\delta})}{m_2} =: \eta$, where we used $m_1 \geq k$. If this event holds for S_1, then the probability over $S_2 \sim P^{m_2}$ that $S_2 \cap \mathrm{qball}_{S_1}(c_i,q) = \emptyset$ is at most $(1-\eta)^{m_2} \leq \exp(-m_2\eta)$. By a union bound, the probability that for some c_i a center is not found in S_2 is at most $k\exp(-m_2\eta) \leq \delta/4$. Combining the two events, we conclude that the probability that a point is found in S_2 for all centers is at least $1-\delta/2$. $\square$

We now bound the risk of the output of SKM, under the assumption that indeed all centers have been successfully selected. The condition in step 6 of the algorithm guarantees that all the selected centers are in the qball around the centers returned by $\mathcal{A}$. The following two lemmas bound the risk that the selected centers induce compared to the original centers. The lemmas are formulated more generally to apply to a general distribution. The first lemma considers a single center. For a distribution Q over $\mathcal{X}$ and $c, t \in \mathcal{X}$, denote $B^o_Q(c,t) := \mathbb{P}_{X\sim Q}[\rho(X,c) < \rho(t,c)]$.

Lemma 3.4. *Let $\tau \in (0,1)$. Let Q be a distribution over $\mathcal{X}$. Let $c \in \mathcal{X}$, $t \in \mathcal{X}$ such that $B^o_Q(c,t) \leq \tau$. Then $R(Q,\{t\}) \leq (1 + 1/(1-\tau))R(Q,\{c\})$.*

Proof. Denote $r := \rho(t,c)$. Using the triangle inequality, and letting $X \sim Q$, we have

$$\begin{aligned} R(Q,\{t\}) &= \mathbb{E}[\rho(X,t)] \leq \mathbb{E}[\rho(X,c) + \rho(t,c)] \\ &= R(Q,\{c\}) + r. \end{aligned}$$

To upper-bound r, note that by the conditions on t, $\mathbb{P}[\rho(X,c) \geq r] \geq 1-\tau$. Therefore, $R(Q,\{c\}) \geq r \cdot \mathbb{P}[\rho(X,c) \geq r] \geq (1-\tau)r$. It follows that $r \leq R(Q,\{c\})/(1-\tau)$, which completes the proof. □

The lemma above provides a multiplicative upper bound on the risk obtained when replacing a center c_i with another center t_i. However, this upper bound is only useful if τ is small. In the general case, an additive error term cannot be avoided. For instance, suppose that the optimal clustering has a risk of zero, and there is at least one very small cluster. In this case, the algorithm might not succeed in choosing a good center for this cluster, and some additive error will ensue. The following lemma bounds the overall risk of the clustering when all centers are replaced.

Lemma 3.5. *Let $\tau \in (0,1)$ and let Q be a distribution over $\mathcal{X}$. Let $O = \{c_1,\ldots,c_k\} \subseteq \mathcal{X}$, and $T = \{t_1,\ldots,t_k\} \subseteq \mathcal{X}$ such that $B^o_Q(c_i,t_i) \leq \tau$. Then for any $\gamma \in (0,\frac{1}{2})$,*

$$R(Q,T) \leq (2+2\gamma)R(Q,O) + k\tau/\gamma.$$

Proof. Let $C_i := \{x \in \mathcal{X} \mid i = \mathrm{argmin}_{j\in[k]}\,\rho(c_j,x)\}$ and $\beta_i := \mathbb{P}_{X\sim Q}[X \in C_i]$. Let $q_i := B^o_Q(c_i,t_i)/\beta_i$, and let Q_i be the conditional distribution of $X \sim Q$ given $X \in C_i$. Distinguish between two types of clusters. If $q_i \geq \gamma$, then $\gamma \leq q_i \leq \tau/\beta_i$, where the second inequality follows from the assumption on t_i. Thus $\beta_i \leq \tau/\gamma$. Since $\rho \leq 1$, $R(Q_i,\{t_i\}) \leq 1$. Therefore, $\sum_{i:q_i\geq\gamma} \beta_i \cdot R(Q_i,\{t_i\}) \leq k\tau/\gamma$. On the other hand, if $q_i < \gamma$, then

$$\begin{aligned} B^o_{Q_i}(c_i,t_i) &= \mathbb{P}_{X\sim Q}[\rho(X,c_i) < \rho(t_i,c_i) \mid X \in C_i] \\ &\leq B^o_Q(c_i,t_i)/\beta_i = q_i < \gamma. \end{aligned}$$

Thus, Lemma 3.4 holds for $\tau := \gamma$, $Q := Q_i$, $t := t_i$ and $c := c_i$, hence $R(Q_i,\{t_i\}) \leq (1+\frac{1}{1-\gamma})R(Q_i,\{c_i\})$. Since $\gamma \in (0,\frac{1}{2})$, we have $1+\frac{1}{1-\gamma} \leq 2+2\gamma$. Therefore,

$$\begin{aligned} \sum_{i:q_i<\gamma} \beta_i \cdot R(Q_i,\{t_i\}) &\leq (2+2\gamma)\sum_{i:q_i<\gamma} \beta_i \cdot R(Q_i,\{c_i\}) \\ &\leq (2+2\gamma)\cdot R(Q,O). \end{aligned}$$

We thus have

$$\begin{aligned} R(Q,T) &\leq \sum_{i\in[k]} \beta_i \cdot R(Q_i,\{t_i\}) \\ &= \sum_{i:q_i<\gamma} \beta_i \cdot R(Q_i,\{t_i\}) + \sum_{i:q_i\geq\gamma} \beta_i \cdot R(Q_i,\{t_i\}) \\ &\leq (2+2\gamma)\cdot R(Q,O) + k\tau/\gamma, \end{aligned}$$

which completes the proof. □

Using the results above, Theorem 3.1 can now be proved.

Proof of Theorem 3.1. Recall that S_1, S_2 are independent i.i.d. samples of size m_1, m_2 drawn from P. By Hoeffding's inequality and the fact that $\rho \leq 1$ we have that for any fixed k-clustering T, $\mathbb{P}[|R(P,T) - R(S_1,T)| \geq \epsilon] \leq 2e^{-2\epsilon^2 m_1}$. By a union bound on all the k-clusterings in S_2 and on $T = \mathrm{OPT}$, we get that with a probability $1-\delta/2$, all such clusterings T satisfy

$$\begin{aligned} |R(P,T) - R(S_1,T)| &\leq \sqrt{(k\ln(m_2) + \ln(4/\delta))/(2m_1)} \\ &= \sqrt{(k\ln(m/2) + \ln(4/\delta))/m} =: \epsilon_1\ , \end{aligned} \tag{3}$$

where we used $m_1 = m_2 = m/2$.

In addition, by Lemma 3.3, with a probability at least $1-\delta/2$, SKM selects k centers from S_2. The two events thus hold simultaneously with a probability at least $1-\delta$. Condition below on these events and let $t_1,\ldots,t_k$ be the selected centers, ordered so that $t_i \in \mathrm{qball}_{S_1}(c_i,q)$. Denote $N_i = |\{z \in S_1 \mid \rho(c_i,z) < \rho(c_i,t_i)\}|$. Since $t_i \in \mathrm{qball}_{S_1}(c_i,q)$, we have by definition of qball that $N_i/|S_i| \leq ((m_1-2)q+1)/m_1 \leq q + 1/m_1$. Therefore, Lemma 3.5 holds with Q set to the uniform distribution on S_1, $O := \mathcal{A}(S_1)$, and $\tau := q + 1/m_1$. Hence,

$$R(S_1,T_{\mathrm{out}}) \leq (2+2\gamma)R(S_1,\mathcal{A}(S_1)) + k(q+1/m_1)/\gamma.$$

By the assumptions on $\mathcal{A}$ and by Eq. (3),

$$\begin{aligned} R(S_1,\mathcal{A}(S_1)) &\leq \beta R(S_1,\mathrm{OPT}_{S_1}) \leq \beta R(S_1,\mathrm{OPT}) \\ &\leq \beta(R(P,\mathrm{OPT}) + \epsilon_1). \end{aligned}$$

In addition, $R(P,T_{\mathrm{out}}) \leq R(S_1,T_{\mathrm{out}}) + \epsilon_1$. Combining the inequalities and noting that $m_1 = m/2$, we get

$$\begin{aligned} R(P,T_{\mathrm{out}}) \leq\ &(2+2\gamma)\beta(R(P,\mathrm{OPT}) + \epsilon_1) \\ &+ k(q+2/m)/\gamma + \epsilon_1. \end{aligned}$$

The theorem follows by setting q as in SKM. □

We have thus shown that SKM obtains an approximation factor at most twice that of the offline algorithm. In the next section, we show that this upper bound on the multiplicative factor is tight.

3.2 Tightness of the multiplicative factor

In this section we show that the multiplicative approximation factor of 2β given in Eq. (1) is tight for SKM. Let $\mathcal{A}$ be an offline k-median algorithm, which for every sample S returns a k-clustering $T \subseteq S$ that minimizes $R(S,T)$. As discussed in Section 2, $\mathcal{A}$ is a 2-approximation offline k-median algorithm. Thus, $\beta = 2$ in Eq. (1). We now show that SKM in this case cannot have a multiplicative factor of less than 4, thus showing that the approximation factor is tight. Moreover, this holds for any setting of q, not necessarily the one used in Alg. 1. Note that if the probability mass of the q-ball set by SKM is smaller than $\log(1/\delta)/m$, then the probability of finding a center in the second phase is less than $1-\delta$. Therefore, one must have $q \geq \log(1/\delta)/m$. In addition, one must have $q \equiv q(m) \to 0$ when $m \to \infty$, otherwise the additive error would not vanish for large m.[3] The following theorem shows that for any q which satisfies these requirements, the approximation factor of SKM is at least 4. The proof of the theorem is provided in the supplementary material.

Theorem 3.6. *Consider running SKM with any setting of $q = q(m)$ such that $q(m) \to 0$ when $m \to \infty$, and $q(m) \geq \log(1/\delta)/m$ for all m. Then, the multiplicative factor of SKM cannot be smaller than $4 = 2\beta$ for $\mathcal{A}$ as defined above.*

We conclude that the multiplicative factor of 2β for SKM is tight. SKM uses a black-box algorithm $\mathcal{A}$, and it is computationally efficient if $\mathcal{A}$ is computationally efficient. In the next section, we show that if efficiency limitations are removed, there is an algorithm for the no-substitution setting that obtains the same approximation factor as an optimal (possibly also inefficient) offline algorithm.

4 Obtaining the Optimal Approximation Factor: SKM2

If efficiency considerations are ignored, the offline algorithm can use a β-approximation algorithm with the best possible β, which is equal to 2, as discussed above. Using Eq. (2), this gives the following guarantee for the offline algorithm:

$$R(P, \operatorname*{argmin}_{T\in S^k} R(S,T)) \leq 2R(S, \mathrm{OPT}_S) + O\left(\sqrt{(k\log(m) + \log(1/\delta))/m}\right).$$

[3] To see this, consider a case with a very small optimal risk, in which one of the clusters has a probability mass of $q/2$. With a constant probability, the center for this cluster will be selected from another cluster, resulting in an additive error of $\Omega(q)$.

We now give an algorithm for the no-substitution setting, which obtains the same approximation factor of 2, and a similar additive error to that of the offline algorithm. The algorithm, called SKM2, is listed in Alg. 2. It receives as input the confidence parameter δ, the number of clusters k, and the sequence size m. Similarly to SKM, it also works in two phases, where the first phase is used for estimation, and the second phase is used for selecting centers. The first phase is further split to sub-sequences $S_0, S_1, \ldots, S_k$. The second phase is denoted $\bar{S}$.

The main challenge in designing SKM2 is to make sure that elements are selected as centers only if it will later be possible, with a high probability, to select additional centers so that the final risk will be near-optimal. To this end, we define a recursive notion of *goodness*. For a set of size k, we say that it is good if its risk on S_0 is lower than some threshold. For a set of size less than k, it is good if there is a sufficient probability to find another element to add to this set, such that the augmented set is good. The following definition formalizes this.

Definition 4.1. *Let $Z \subseteq \mathcal{X}$ of size at most k. Let $r > 0$ and $q \in (0,1)$. The predicate (r,q)-good is defined as follows, with respect to the sub-samples $S_0, S_1, \ldots, S_k \subseteq \mathcal{X}$.*

- *For Z of size k, Z is (r,q)-good (or simply r-good) if $R(S_0, Z) \leq r$.*
- *For Z of size $j \in \{0, \ldots, k-1\}$, define $\hat{\psi}_{r,q}(Z) := \mathbb{P}_{X\sim S_{j+1}}[Z \cup \{X\}$ is (r,q)-good$]$. Z is (r,q)-good if $\hat{\psi}_{r,q}(Z) \geq 2q$.*

SKM2 sets the value of q depending on the input parameters, and finds a value for r such that $\emptyset$ is (r,q)-good. It then iteratively gets the examples, and adds the observed example as a center if the addition preserves the goodness of the solution collected so far. We show below that if $\emptyset$ is (r,q)-good for q as defined in Alg. 2, then with a high probability SKM2 will succeed in selecting k centers with a risk at most r on S_0, and that this will result in a near-optimal k-clustering. SKM2 has a computational complexity exponential in k, since it considers recursively all the elements of $S_1 \times S_2 \times \ldots \times S_k$. We prove the following result for SKM2.

Theorem 4.2. *Suppose that SKM2 is run with inputs $k, m \in \mathbb{N}$ and $\delta \in (0,1)$. For any $\gamma \in (0, \frac{1}{2})$ and distribution P over $\mathcal{X}$, with a probability at least $1-\delta$,*

$$\begin{aligned} R(P, T_{\mathrm{out}}) \leq{}& (2+2\gamma)R(P, \mathrm{OPT}) \\ &+ \frac{1}{\gamma}\cdot O\left((k^3\log(m) + k^2\log(1/\delta))/m\right) \\ &+ O\left((k\log(m) + \log(1/\delta))/m\right). \end{aligned}$$

Algorithm 2 SKM2

input $k, m \in \mathbb{N}$, $\delta \in (0,1)$, sequential access to $S = (x_1, \dots, x_m) \sim P^m$.

output A k-clustering $T_{\text{out}} \subseteq S$

1: $q \leftarrow (32k^2 \log(8m) + 32k \log(8/\delta))/m$; $T_{\text{out}} \leftarrow \emptyset$

2: Get $m/2$ examples from S. Set $m/4$ examples as S_0, and split the rest of the examples equally between $S_1, \dots, S_k$.

3: Set $\beta_m := 1/\sqrt{m}$. Let $r \leftarrow \min\{r = \beta_m(1+\beta_m)^n \mid n \in \mathbb{N},$ and $\emptyset$ is (r,q)-good$\}$.

4: **for** $j = m/2 + 1$ to m **do**

5: Get the next sample x_j.

6: **If** $|T_{\text{out}}| < k$ and $T_{\text{out}} \cup \{x_j\}$ is (r,q)-good **then** $T_{\text{out}} \leftarrow T_{\text{out}} \cup \{x_j\}$.

7: **end for**

8: **return** T_{out}

By setting $\gamma = \sqrt{(k^3 \log(m) + k^2 \log(1/\delta))/m}$ and noting the $R(P, \text{OPT}) \leq 1$, we get

$$R(P, T_{\text{out}}) \leq 2R(P, \text{OPT}) + O\left(\sqrt{(k^3 \log(m) + k^2 \log(1/\delta))/m}\right).$$

As discussed above, this is the same multiplicative approximation factor as the optimal offline algorithm. The additive error is larger by a factor of k.

We now prove Theorem 4.2. Note that by the definition of goodness for Z of size k, it follows that if SKM2 succeeds in selecting k centers, then the solution it finds has a risk of at most r on S_0. We thus need to show that indeed k centers are selected with a high probability, that r is close to the optimal achievable risk, and that the risk on S_0 is close to the risk on P. We use the following lemma, proved in the supplementary material based on Bernstein's inequality.

Lemma 4.3. *Let $Y_1, \dots, Y_n$ be i.i.d. random variables in $[0,1]$ with mean $\mu \geq 10\ln(\frac{1}{\delta})/n$. Let $\hat{\mu} = \frac{1}{n}\sum_{i \in [n]} Y_i$ be the empirical mean. Then, with a probability at least $1-\delta$, $\hat{\mu} \geq \mu/2$.*

Denote the sizes of $S_0, S_1, \dots, S_k, \bar{S}$ by $m_0, m_1, \dots, m_k, \bar{m}$ respectively. First, we show that SKM2 selects k centers with a high probability.

Lemma 4.4. *With a probability at least $1-\delta/2$, by the end of the run SKM2 has collected k centers.*

Proof. Let $\Upsilon = \{\beta_m(1+\beta_m)^n \mid n \in \mathbb{N}\} \cap (0,1)$ be the possible values of r examined by the algorithm which are smaller than 1. Note that since $\exp(x/2) \leq 1 + x$ for $x \in (0,1)$, the largest n such that $\beta_m(1+\beta_m)^n < 1$ satisfies $\beta_m \exp(n\beta_m/2) < 1$. Therefore, $|\Upsilon| \leq \frac{2}{\beta_m}\log(1/\beta_m) = \sqrt{m}\log(m) \leq m$. By Lemma 3.2 and a union bound, with a probability at least $1-\delta/4$, for any $r \in \Upsilon$, $j \in \{0, \dots, k-1\}$, and $T \subseteq \bar{S}$ of size j, $\hat{\psi}_{r,q}(T) \geq 16\ln(8\bar{m}^k/\delta)/(m_{j+1}-1) \Rightarrow \mathbb{P}[T \cup \{X\}$ is (r,q)-good$] \geq \hat{\psi}_{r,q}(T)/2$. Condition below on this event. Let r be the value selected by SKM2, let T_i be the set of points collected by the algorithm until iteration i, and let $j = |T_i| < k$. If $T_i = \emptyset$, then it is (r,q)-good by the definition of r. Otherwise, it is (r,q)-good by the condition on line 6. Therefore, by definition, $\hat{\psi}_{r,q}(T_i) \geq 2q$. This implies the LHS of the implication above, hence $\mathbb{P}[T_i \cup \{X\}$ is (r,q)-good$] \geq q$.

Therefore, conditioned on the event above, the probability that the next sample x_j satisfies that $T_i \cup \{x_j\}$ is (r,q)-good is at least q. Since this holds for all iterations until there are k centers in T_i, the probability that the algorithm collects less than k centers is at most the probability of obtaining less than k successes in $\bar{m} = m/2$ independent experiments with a probability of success q. Let $\hat{s}$ be the empirical fraction of successes on $m/2$ experiments. By Lemma 4.3, since $q \geq 10\log(4/\delta)/(m/2)$, with a probability $1-\delta/4$, $\hat{s} \geq q/2$. Since $q \geq 2k/m$, we have $\hat{s} \geq k/m$. Therefore, taking a union bound, with a probability of at least $1-\delta/2$, the algorithm selects k centers. □

We now show that the value of r selected by SKM2 is close to the optimal risk. By Hoeffdings's inequality and a union bound over the possible choices of T, for all $T \subseteq S \setminus S_0$ of size k, with a probability $1-\delta/4$, $|R(P,T) - R(S_0,T)| \leq \sqrt{(2k\ln(m) + 2\ln(\frac{8}{\delta}))/m}$. Call this event E_0 and denote the RHS by ϵ_2.

Lemma 4.5. *Let $\gamma \in (0, \frac{1}{2})$, and define the value $r_0 := ((2+2\gamma)R(P, \text{OPT}) + 4qk/\gamma + \epsilon_2)$. With a probability of $1-\delta/4$, E_0 implies that the value of r set by SKM2 satisfies $r \leq (1+\beta_m)r_0$.*

Proof. Let $j \in \{0, \dots, k\}$. For sets $D_1, \dots, D_j$, denote by $\bar{D}_j$ the collection of all sets of size j that include exactly one element from each of $D_1, \dots, D_j$. We start by showing that with a high probability, there exist sets $D_1, \dots, D_k$ such that for all $i \in [k]$, $D_i \subseteq S_i$, $|D_i| \geq 2qm_i$, and $\max_{Z \in \bar{D}_k} R(S_0, Z) \leq r_0$. Let $\text{OPT} = \{o_1, \dots, o_k\} \subseteq \mathcal{X}$ be an optimal k-clustering for P. For $i \in [k]$, let $\alpha_i \geq 0$ such that $\mathbb{P}[\text{Ball}(o_i, \alpha_i)] \geq 4q$ and $\mathbb{P}[\rho(X, o_i) < \alpha_i] \leq 4q$. Let $D_i = \text{Ball}(o_i, \alpha_i) \cap S_i$. Denote $B_i = \mathbb{P}[\text{Ball}(o_i, \alpha_i)]$. By Lemma 4.3, since $B_i \geq 4q \geq 10\ln(4k/\delta)/m_i$, we have that with a probability at least $1-\delta/4$, for all $i \in [k]$, $|D_i|/|S_i| \geq 2q$, as required.

We now show that $\max_{Z \in \bar{D}_k} R(S_0, Z) \leq r_0$. By the definition of α_i, for any $d_i \in D_i$ we have $B_P^o(c_i, d_i) \leq 4q$, where B^o is defined above Lemma 3.4. Therefore, the conditions of Lemma 3.5 hold with $Q := P$, $O :=$

OPT, $T := Z$ and $\tau := 4q$. Hence, for $\gamma \in (0, \frac{1}{2})$,

$$R(P, Z) \leq (2 + 2\gamma)R(P, \mathrm{OPT}) + 4qk/\gamma.$$

Under E_0, we get that for all $Z \in \bar{D}_k$, $R(S_0, Z) \leq (2 + 2\gamma)R(P, \mathrm{OPT}) + 4qk/\gamma + \epsilon_2 \equiv r_0$.

Lastly, we show that the existence of $D_1, \ldots, D_k$ implies an upper bound on the value of r set by the algorithm. First, we show that $\emptyset$ is (r_0, q)-good. This can be seen by induction on the definition of goodness: For $|Z| = k$, all $Z \in \bar{D}_k$ are (r_0, q)-good since $R(S_0, Z) \leq r_0$. Now, suppose that all sets $Z \in \bar{D}_j$ for some $j \in [k]$ are (r_0, q)-good, and let $Z' \in \bar{D}_{j-1}$. Then, since for all $x \in D_j$ we have $Z' \cup \{x\} \in \bar{D}_j$, it follows that $\hat{\psi}_{r_0,q}(Z') = \mathbb{P}_{X \sim S_j}[Z' \cup \{X\} \text{ is } (r_0, q)\text{-good}] \geq |D_j|/|S_j| \geq 2q$. Therefore, by definition, Z' is (r_0, q)-good. By induction, we conclude that $\emptyset \in \bar{D}_0$ is also (r_0, q)-good. Clearly, $\emptyset$ is also (r_1, q)-good for any $r_1 \geq r_0$. Since the value r selected by SKM2 is set to the smallest value $\beta_m(1 + \beta_m)^n$ such that n is natural and $\emptyset$ is (r, q)-good, and since $r_0 \geq \epsilon_2 \geq \beta_m$, we conclude that $r \leq r_0(1 + \beta_m)$, as required. $\square$

The proof of Theorem 4.2 can now be provided.

Proof of Theorem 4.2. Assume that E_0 holds, as well as the events of Lemma 4.4 and Lemma 4.5. This occurs with a probability at least $1 - \delta$. By Lemma 4.4 the algorithm selects T_{out} which is of size k and is (r, q)-good. Thus, by the definition of goodness, $R(S_0, T_{\text{out}}) \leq r$. By E_0, $R(P, T_{\text{out}}) \leq r + \epsilon_2$. By Lemma 4.5, $r \leq (1+\beta_m)((2+2\gamma)R(P, \mathrm{OPT})+4qk/\gamma+\epsilon_2)$. The theorem follows by plugging in the values of β_m, q, ϵ_2 and simplifying. $\square$

5 Experiments

We demonstrate SKM [4] on 3 datasets: `MNIST` (LeCun et al., 1998), `Covertype`[5] and `Census 1990` (Dua and Graff, 2017). While Alg. 1 uses $q = 43 \ln(2m^2/\delta)/m$, it can be seen from the proof of Theorem 3.1 that except for very small values of m, the guarantees of SKM hold also with significantly smaller values. In the experiments we used $q = 9 \ln(2m^2/\delta)/m$. In all experiments, the features were normalized, and PCA was used to reduce the dimension, so that 95% of the signal was retained. As black-box k-median algorithms, we used the implementation of k-medoids in Novikov (2019), and the BIRCH algorithm Zhang et al. (1997), implemented in Pedregosa et al. (2011). All risks were estimated on the same holdout set, and averaged over 20 runs. Figure 1 reports the ratio between the clustering risk obtained by SKM and the risk of the offline algorithm, for k-medoids. Results for BIRCH are reported in the supplementary material. It can be seen that in practice, the risk ratio obtained by SKM is usually close to 1. The results for large stream sizes, provided in the supplementary material, show a convergence to values very close to 1. As expected, the convergence is slower for larger values of k.

[4] Our code is available at `https://github.com/tomhess/No_Substitution_K_Median`.

[5] Reuse of this database is unlimited with retention of copyright notice for Jock A. Blackard and Colorado State University.

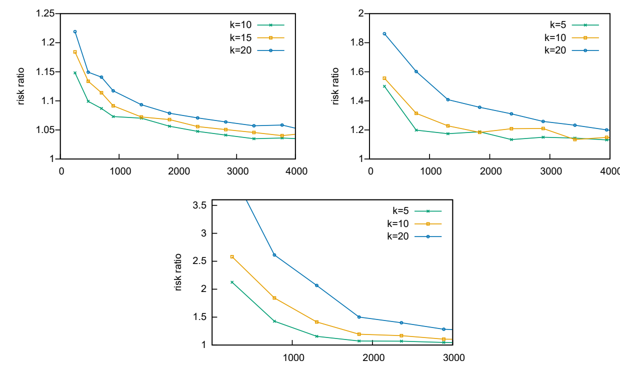

Figure 1: Risk ratio between SKM with k-medoids and offline k-medoids as a function of the stream size, for various values of k. Top left:`MNIST`, Top right: `Covertype`, Bottom: `Census`.

6 Discussion

In this work, we obtained an approximation factor which is twice that of the sample-based offline algorithm in the no-substitution setting. We showed that when disregarding computational considerations, the factor of 2 can be removed. It is an open question whether there is an efficient no-substitution algorithm with the same approximation factor as the best efficient offline algorithm. SKM2 obtains an improved approximation factor by requiring that only centers with many possible choices of other centers are selected. This is related to notions of stability, or robustness, which have been previously studied for clustering algorithms in other contexts (see, e.g., Lange et al., 2004; Ackerman et al., 2013), and more generally for learning algorithms (Bousquet and Elisseeff, 2002). The relationship between stability of algorithms and success in the no-substitution setting is an interesting direction for future research.

Acknowledgements This work was supported in part by the Israel Science Foundation (grant No. 555/15), and by the Lynn and Williams Frankel Center for Computer Science at Ben-Gurion University.

References

M. Ackerman, S. Ben-David, D. Loker, and S. Sabato. Clustering oligarchies. In *Artificial Intelligence and Statistics*, pages 66–74, 2013.

M. R. Ackermann, M. Märtens, C. Raupach, K. Swierkot, C. Lammersen, and C. Sohler. Streamkm++: A clustering algorithm for data streams. *Journal of Experimental Algorithmics (JEA)*, 17:2–4, 2012.

C. C. Aggarwal and P. S. Yu. Online analysis of community evolution in data streams. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 56–67. SIAM, 2005.

N. Ailon, R. Jaiswal, and C. Monteleoni. Streaming k-means approximation. In *Advances in neural information processing systems*, pages 10–18, 2009.

M. Babaioff, N. Immorlica, D. Kempe, and R. Kleinberg. A knapsack secretary problem with applications. In *Approximation, randomization, and combinatorial optimization. Algorithms and techniques*, pages 16–28. Springer, 2007.

M. Babaioff, N. Immorlica, D. Kempe, and R. Kleinberg. Online auctions and generalized secretary problems. *ACM SIGecom Exchanges*, 7(2):7, 2008.

A. Badanidiyuru, B. Mirzasoleiman, A. Karbasi, and A. Krause. Streaming submodular maximization: Massive data summarization on the fly. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 671–680. ACM, 2014.

M. Bateni, M. Hajiaghayi, and M. Zadimoghaddam. Submodular secretary problem and extensions. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 39–52. Springer, 2010.

S. Ben-David. A framework for statistical clustering with constant time approximation algorithms for k-median and k-means clustering. *Machine Learning*, 66(2-3):243–257, 2007.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar): 499–526, 2002.

V. Braverman, D. Feldman, and H. Lang. New frameworks for offline and streaming coreset constructions. *arXiv preprint arXiv:1612.00889*, 2016.

M. Charikar, L. O'Callaghan, and R. Panigrahy. Better streaming algorithms for clustering problems. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 30–39. ACM, 2003.

K. Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

D. Dueck and B. J. Frey. Non-metric affinity propagation for unsupervised image categorization. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

M. Feldman, J. S. Naor, and R. Schwartz. Improved competitive ratios for submodular secretary problems. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 218–229. Springer, 2011.

S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. In *Foundations of computer science, 2000. proceedings. 41st annual symposium on*, pages 359–366. IEEE, 2000.

Y. Hadi, F. Essannouni, and R. O. H. Thami. Video summarization by k-medoid clustering. In *Proceedings of the 2006 ACM symposium on Applied computing*, pages 1400–1401. ACM, 2006.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

T. Kesselheim and A. Tönnis. Submodular secretary problems: Cardinality, matching, and linear constraints. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

H. Lang. Online facility location against at-bounded adversary. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1002–1014. Society for Industrial and Applied Mathematics, 2018.

T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann. Stability-based validation of clustering solutions. *Neural computation*, 16(6):1299–1323, 2004.

S. Lattanzi and S. Vassilvitskii. Consistent k-clustering. In *International Conference on Machine Learning*, pages 1975–1984, 2017.

Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

K. Leung and C. Leckie. Unsupervised anomaly detection in network intrusion detection using clusters. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38*, pages 333–342. Australian Computer Society, Inc., 2005.

E. Liberty, R. Sriharsha, and M. Sviridenko. An algorithm for online k-means clustering. In *2016 Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 81–89. SIAM, 2016.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

A. Meyerson. Online facility location. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 426–431. IEEE, 2001.

O. Nasraoui, J. Cerwinske, C. Rojas, and F. Gonzalez. Performance of recommendation systems in dynamic streaming environments. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 569–574. SIAM, 2007.

H. Ng, S. Ong, K. Foong, P. Goh, and W. Nowinski. Medical image segmentation using k-means clustering and improved watershed algorithm. In *2006 IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 61–65. IEEE, 2006.

A. Novikov. PyClustering: Data mining library. *Journal of Open Source Software*, 4(36):1230, apr 2019. doi: 10.21105/joss.01230. URL `https://doi.org/10.21105/joss.01230`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

S. Sabato and T. Hess. Interactive algorithms: Pool, stream and precognitive stream. *Journal of Machine Learning Research*, 18(229):1–39, 2018.

A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 259–266. ACM, 2008.

T. Zhang, R. Ramakrishnan, and M. Livny. Birch: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182, 1997.