AISTATS 2020 Supplementary Material

Sequential no-Substitution k-Median-Clustering

Tom Hess and Sivan Sabato

A Bernstein and empirical Bernstein inequalities

Proof of Lemma 3.2. We use the Empirical Bernstein inequality of (Maurer and Pontil, 2009). This inequality states that for $\hat{\sigma}^2 := \frac{1}{2n(n-1)}\sum_{i,j\in[n],i\neq j}(Y_i - Y_j)^2$, with a probability at least $1-\delta$, we have

$$\hat{\mu} - \mu \leq \frac{7\ln(\frac{2}{\delta})}{3(n-1)} + \sqrt{\frac{2\hat{\sigma}^2\ln(\frac{2}{\delta})}{n}}.$$

We have $\hat{\sigma}^2 \leq \frac{n}{2(n-1)}\frac{1}{n^2}\sum_{i,j\in[n]}(Y_i - Y_j)^2 = \frac{n}{2(n-1)}\mathbb{E}[(Y-Y')^2]$, where Y, Y' are drawn independently and uniformly from the fixed sample $Y_1, \ldots, Y_n$. Since $\mathbb{E}[(Y-Y')^2] \leq 2\mathbb{E}[Y^2]$, and $Y \in [0,1]$, we have $\hat{\sigma}^2 \leq \frac{n}{n-1}\mathbb{E}[Y] \equiv \frac{n}{n-1}\hat{\mu}$. Therefore,

$$\hat{\mu} - \mu \leq \frac{7\ln(\frac{2}{\delta})}{3(n-1)} + \sqrt{\frac{2\hat{\mu}\ln(\frac{2}{\delta})}{n-1}}.$$

If $\hat{\mu} = a\ln(\frac{2}{\delta})/(n-1)$ for $a \geq 16$, then the RHS is at most

$$(7/3+\sqrt{2a})\ln(\frac{2}{\delta})/(n-1) \leq a/2\cdot\ln(\frac{2}{\delta})/(n-1) \leq \hat{\mu}/2.$$

□

Proof of Lemma 4.3. Let $\sigma^2 = \mathrm{Var}[Y_i]$. By Bernstein's inequality (Hoeffding, 1963) (See, e.g., Maurer and Pontil 2009 for the formulation below),

$$\mu - \hat{\mu} \leq \frac{\ln(\frac{1}{\delta})}{3n} + \sqrt{\frac{2\sigma^2\ln(\frac{1}{\delta})}{n}}.$$

Since Y_i are supported on $[0,1]$, we have $\sigma^2 \leq \mu$. Since $\mu = a\ln(\frac{1}{\delta})/n$ for $a \geq 10$, we have that the RHS is equal to $(1/3+\sqrt{2a})\ln(\frac{1}{\delta})/n \leq a/2\cdot\ln(\frac{1}{\delta})/n \leq \mu/2$. The statement of the lemma follows. □

B Tightness of multiplicative factor of SKM

Proof of Theorem 3.6. We define a weighted undirected graph $G = (V, E, W)$, and let $(\mathcal{X}, \rho)$ be a metric space such that $\mathcal{X} = V$ and $\rho(u,v)$ is the length of the shortest path in the graph between u and v. G, which is illustrated in Figure 2, is formally defined as follows. The set of nodes is $V := U \cup Y \cup \{o, v\}$, where $U := [0,1]$ and $Y := [3,4]$. The set of edges is

$$E := \{\{u,o\} \mid u \in U\} \cup \{\{v,y\} \mid y \in Y\} \cup \{\{o,v\}\}.$$

Denote $m_1 := m/2$, and let $\eta := 1/(4m_1)$. The weight function W assigns a weight of 1 to all edges except for those that have a node in Y as an endpoint, which are assigned a weight of $2-\eta$.

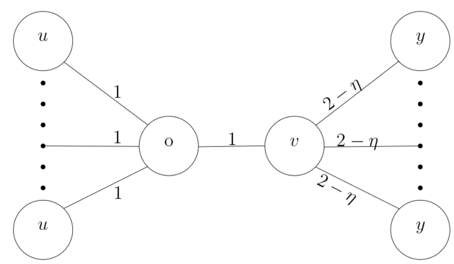

Figure 2: Illustration of the graph G which defines the metric space.

Define the distribution P over $\mathcal{X}$ such that $P(o) = 0$, $P(v) = 1/m_1$, $P(Y) = 2q$, with a uniform conditional distribution over Y. Lastly, $P(U) = 1 - 2q - \frac{1}{m_1}$, with a uniform conditional distribution over U. Note that the latter is positive for a large enough m_1, since $q(m_1) \to 0$.

Let $S \sim P^m$ be the i.i.d. sample used as an input sequence to SKM, and set $k = 1$. Let S_1 be the sample observed in the first phase of SKM, of size m_1. Define the following events:

1. $E_1 := \{o \notin S_1\}$.
2. $E_2 := \{v \text{ appears in } S_1\}$.
3. $E_3 := \{\text{at least } qm_1 \text{ of the samples in } S_1 \text{ are from } Y\}$.

First, observe that all these events occur together with a positive probability, as follows. $\mathbb{P}[E_1] = 1$ since $P(o) = 0$. For E_2, we have

$$\mathbb{P}[E_2] > 1 - (1 - \frac{1}{m_1})^{m_1} > \frac{1}{2}.$$

For E_3, note that the probability mass of Y is $2q$, Apply Lemma 4.3 with $\mu = 2q$, $n = m_1$ and a confidence value of $1/4$. By the assumption of the theorem, for sufficiently small δ, we have $q \geq 5\log(4)/m_1$. Therefore, Lemma 4.3 implies that $P[E_3] \geq 3/4$. It follows that $\mathbb{P}[E_1 \wedge E_2 \wedge E_3] \geq 1/4$.

Now, assume that all the events above hold. By E_1, o does not appear in S_1, and by E_2, v appears in S_1. We show that out of the points in S_1, the 1-clustering $\{v\}$ has the best empirical risk. The only other options in S_1 are centers from Y or from U. For a center $u \in U$ from S_1, note that with a probability 1, it does not have additional copies in S_1. Its distance from all other $u' \in U$ is the same as that of v, while its distance from points in Y and from v is larger. Thus, $R(S_1, \{u\}) > R(S_1, \{v\})$. For a center $y \in Y$, it too does not have additional copies in S_1. Its distance to all other points is larger than that of v. Thus, $R(S_1, \{y\}) > R(S_1, \{v\})$. Therefore, v has the best empirical risk on S_1. Thus, $\mathcal{A}(S_1)$ returns the 1-clustering $\{v\}$.

By E_3, the number of instances of vertices from Y is at least qm_1. Since the points in Y are the closest to v in S_1, we have $y' := \mathrm{qp}_{S_1}(v, q) \in Y$. Therefore, $\mathrm{qball}(v, y') = \{v\} \cup Y$. It follows that SKM selects as a center the first element from $\{v\} \cup Y$ that it observes in the second phase. With a probability $\frac{2q}{2q+\frac{1}{m_1}}$, the first element that SKM observes from $\{v\} \cup Y$ is in Y. Since $q \geq 1/m_1$, this probability is at least $2/3$. Thus, the output center of SKM is from Y with a constant probability.

However, the risk of this clustering is large:

$$\begin{aligned} &R(P, \{y\}) \\ &= (4-\eta)(1-2q-\frac{1}{m_1}) + (2-\eta)\frac{1}{m_1} + (4-\eta)2q. \end{aligned}$$

For large m, we have $m_1 \to \infty$. In addition, $q, \eta \to 0$. Hence, $R(P, \{y\}) \to 4$. In contrast, the risk using o as a center is small:

$$R(P, \{o\}) = (1-2q-\frac{1}{m_1}) + \frac{1}{m_1} + (3-\eta)2q.$$

This approaches 1 for large m. Therefore, for $m \to \infty$, $R(P, \{y\})/R(P, \{o\}) \to 4$. Since $\{y\}$ is the output of SKM with a constant probability, the multiplicative factor obtained by SKM cannot be smaller than $4 = 2\beta$ in this case. □

C Full results of experiments

The results of the experiments for large stream sizes with the k-medoids as the black box are reported in Figure 3. The results for the BIRCH black-box are reported in Figure 4 and in Figure 5. For the k-medoids black box, the risk ratios for large stream sizes are in the following ranges: `MNIST` $1.02-1.04$, `Covertype` $1.04-1.08$, `Census` $1-1.04$. For the BIRCH black box, the risk ratios for large stream sizes are in the following ranges: `MNIST` $1.03-1.04$, `Covertype` $1.05-1.1$, `Census` $1-1.02$. Thus, the risk ratio converges to a ratio very close to 1.

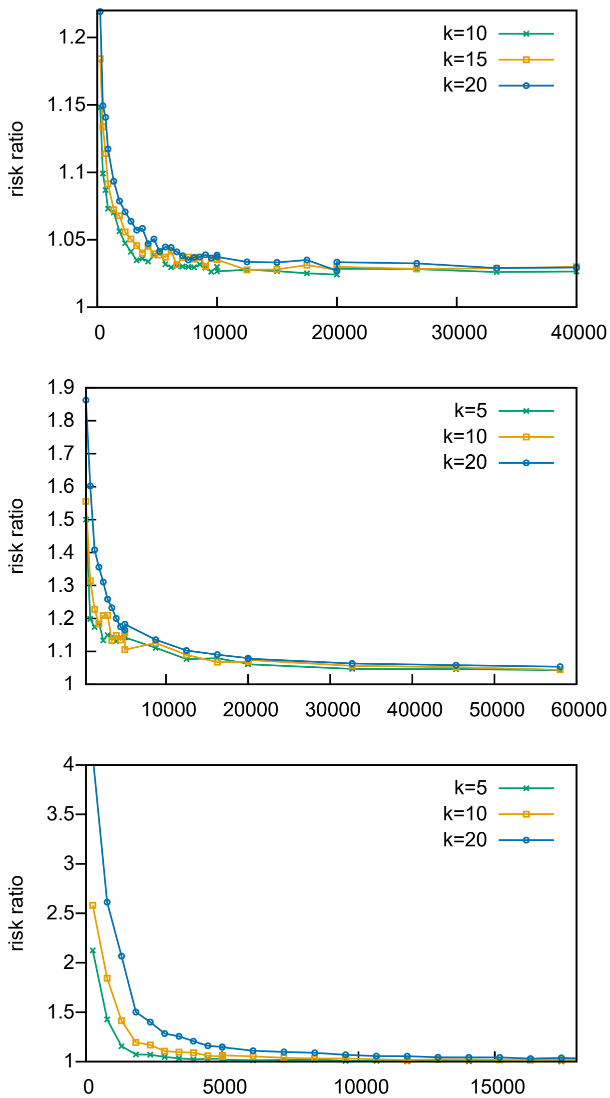

Figure 3: Risk ratio between SKM with k-medoids and offline k-medoids for large stream sizes, as a function of the stream size, for various values of k. Top to bottom: `MNIST, Covertype, Census`.

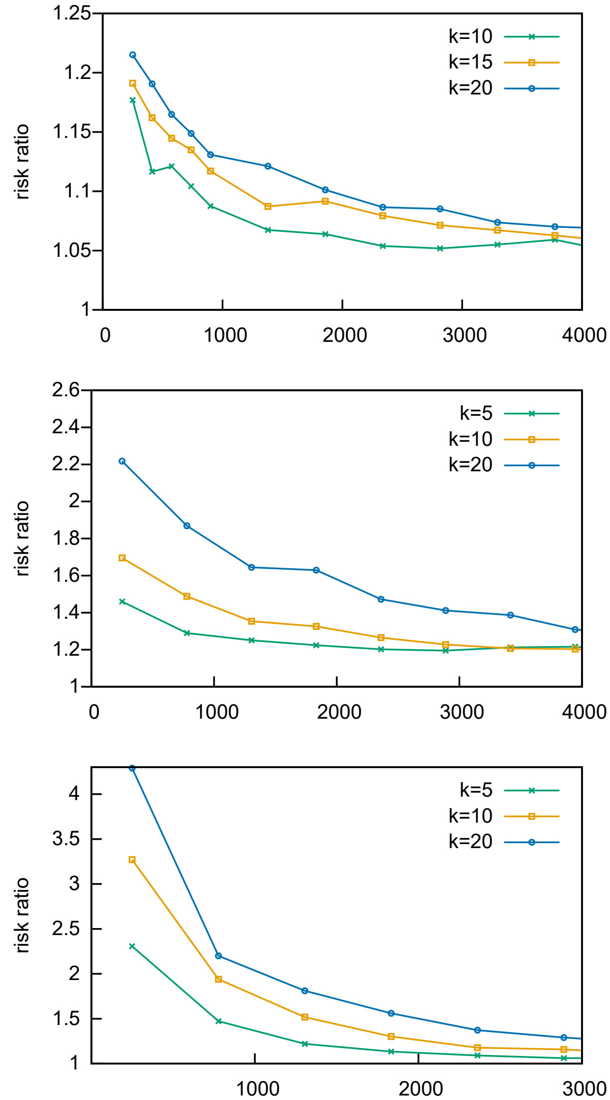

Figure 4: Risk ratio between SKM with BIRCH and offline BIRCH as a function of the stream size, for various values of k. Top to bottom: `MNIST, Covertype, Census`.

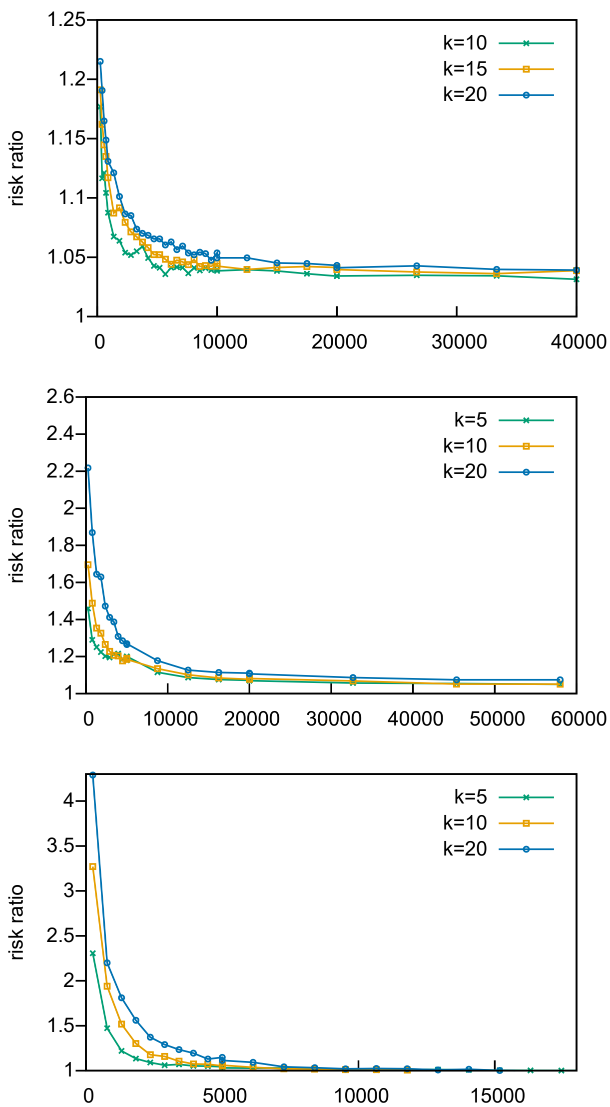

Figure 5: Risk ratio between SKM with BIRCH and offline BIRCH for large stream sizes, as a function of the stream size, for various values of k. Top to bottom: `MNIST`, `Covertype`, `Census`.