On the Optimal Number of Grids for Differentially Private Non-Interactive K-Means Clustering

Gokularam Muthukrishnan, Anshoo Tandon
Center of Data for Public Good, FSID,
Indian Institute of Science, Bengaluru - 560012, India.
e-mail: gokularam.m@datakaveri.org, anshoo.tandon@gmail.com

***Abstract*—Differentially private K-means clustering enables releasing cluster centers derived from a dataset while protecting the privacy of the individuals. Non-interactive clustering techniques based on privatized histograms are attractive because the released data synopsis can be reused for other downstream tasks without additional privacy loss. The choice of the number of grids for discretizing the data points is crucial, as it directly controls the quantization bias and the amount of noise injected to preserve privacy. The widely adopted strategy selects a grid size that is independent of the number of clusters and also relies on empirical tuning. In this work, we revisit this choice and propose a refined grid-size selection rule derived by minimizing an upper bound on the expected deviation in the K-means objective function, leading to a more principled discretization strategy for non-interactive private clustering. Compared to prior work, our grid resolution differs both in its dependence on the number of clusters and in the scaling with dataset size and privacy budget. Extensive numerical results elucidate that the proposed strategy results in accurate clustering compared to the state-of-the-art techniques, even under tight privacy budgets.**

***Index Terms*—Differential privacy, K-means clustering, Non-interactive data analysis, Grid-size selection, Laplace mechanism.**

I. Introduction

Data clustering [1] is a fundamental unsupervised machine learning technique that, when applied to datasets with attributes from several users, identifies meaningful user groups and latent structures within the data. The *K-means clustering* is a widely adopted clustering method. However, the user attributes often contain sensitive information, and differential privacy (DP) [2] can be employed for preserving user privacy. Over the years, several DP clustering algorithms have been developed [3]–[12]. Ghazi *et al.* [13] provided the approximation guarantee for DP clustering. Broadly speaking, private clustering algorithms can be classified into two categories: *interactive* and *non-interactive*. Interactive algorithms deplete the entire privacy budget for the clustering task, and the data cannot be reused for any other purpose without further degradation of privacy guarantee. In contrast, non-interactive algorithms first generate a private summary of the data distribution, which can be reused for multiple queries or downstream tasks without additional privacy leakage.

DPLloyd [3], the private version of Lloyd algorithm, is the prominent interactive scheme. It requires careful distribution of the available privacy budget across iterations: too few iterations prevent convergence, while too many increase the amount of noise injected to preserve privacy. IBM's `diffprivlib` [14] implements DPLloyd with a budget allocation strategy from [4], and decreasing exponential budget distribution is used in [5]. Another private version of Lloyd algorithm has been proposed in [6], where the centroids are resampled randomly in each iteration and the convergence is guaranteed.

Chang *et al.* [15] proposed a non-interactive, *coreset*-based scheme for the local model of DP (later extended to the central model [16]), where the dataset is privately compressed into a smaller, weighted set of points and a classical, non-private clustering algorithm is executed on them. This method scales well for massive datasets, but requires several hyperparameters to be set *a priori*. Uniform grid non-interactive methods partition the data domain into equal-sized grids, providing a simple and efficient structure for releasing private histograms [4], [17]. Among these, the *Extended Uniform Grid K-Means* (EUGkM) algorithm [4] is the most prominent and has been shown to outperform interactive schemes. However, its construct is not directly motivated by the clustering objective, rather adapted from [17], which focused on range queries; consequently, the selected grid size is independent of the number of clusters. Furthermore, the grid size depends on a hyperparameter obtained through extensive tuning, and may fail to maintain consistent performance across configurations.

We derive the optimal grid size for uniform grid non-interactive clustering that effectively balances the discretization bias and the amount of noise injected for privacy: (i) We determine the optimal grid size that minimizes an upper bound on the expected deviation in the K-means objective function. While some existing works [18], [19] consider objective perturbation to guarantee privacy, we *control* the perturbation in the objective function. (ii) Our grid size is dictated by the number of clusters, and it also differs from that of EUGkM in the exponents governing the dataset size and the privacy budget. (iii) Extensive numerical results demonstrate that our solution outperforms state-of-the-art techniques, including leading open-source DP clustering implementations.

Detailed derivation, proofs, and additional empirical results are provided in the full version of this paper [20].

II. Background and Motivation

A. Differential Privacy

Let $\mathcal{D} \in \mathcal{X}^N$ denote the dataset that contains data records $\mathbf{x}_l \in \mathcal{X}$, $l \in \mathbb{N}_N$, collected from N individuals. We term a pair

of datasets $(\mathcal{D}, \breve{\mathcal{D}})$ differing by only a single data record as neighboring datasets, and we denote $\mathcal{D} \asymp \breve{\mathcal{D}}$; we consider add-remove model [21], where $\breve{\mathcal{D}}$ can be constructed by adding or removing a single data record from $\mathcal{D}$.

A trusted curator holds the sensitive dataset $\mathcal{D}$ and is responsible for securely releasing the result of the query $f : \mathcal{X}^N \to \mathcal{Y}$ on dataset $\mathcal{D}$. DP ensures that it is difficult to infer any individual's presence in $\mathcal{D}$ from the response by *randomizing* it; the algorithm $\mathcal{M}$ that provides randomized output to the query referred to as the *private mechanism*.

Definition 1 (ϵ-Differential Privacy [2])**.** A randomized mechanism $\mathcal{M} : \mathcal{X}^N \to \mathcal{Y}$ is said to guarantee ϵ-differential privacy[1] (ϵ-DP) if for every measurable set $\mathcal{E} \subseteq \mathcal{Y}$ and every pair of neighboring datasets $\mathcal{D} \asymp \breve{\mathcal{D}}$, $\mathbb{P}\{\mathcal{M}(\mathcal{D}) \in \mathcal{E}\} \leq e^{\epsilon}\, \mathbb{P}\{\mathcal{M}(\breve{\mathcal{D}}) \in \mathcal{E}\}$, where $\epsilon \in \mathbb{R}_+$ is the *privacy budget*.

When the query output is a numeric vector, $f(\mathcal{D}) \in \mathbb{R}^M$, *additive noise mechanism* achieves DP by perturbing it as $\mathcal{M}(\mathcal{D}) = f(\mathcal{D}) + \mathbf{Z}$, where $\mathbf{Z} \in \mathbb{R}^M$ is the noise sampled from a known distribution. The *Laplace mechanism* achieves DP by adding noise sampled from a Laplace distribution, with appropriately calibrated scale, to each coordinate of the response. Because of the exponential tails of the noise distribution, it can ensure ϵ-DP with *bounded privacy loss* [22].

B. Private Histogram Query

Consider the setting where the data domain $\mathcal{X}$ is partitioned into M disjoint bins. The *histogram query* [2, E.g., 3.2] requests the number of data entries in the dataset that fall within each of the bins: the j-th coordinate of the response $f(\mathcal{D}) \in \mathbb{Z}_+^M$ represents the number of data entries in $\mathcal{D}$ of *type* $j \in \mathbb{N}_M$. Here, the addition or deletion of a single data record can affect the count in exactly one bin. The privatization of histogram queries is primarily achieved through the Laplace Mechanism.[2] Specifically, to guarantee ϵ-DP, independent noise sampled from $\mathcal{L}\big(0, \frac{1}{\epsilon}\big)$ is added to each bin count.

Remark 1. Due to inherent structural disjointness of the histogram query, the scale of Laplace noise required for ϵ-DP is independent of the number of bins as well as the size of the dataset. Consequently, the Laplace mechanism offers fairly high accuracy: The probability that any bin count deviates larger than $\frac{1}{\epsilon}\log\big(\frac{M}{\tau}\big)$ is at most τ [2, Thm. 3.8].

Histograms provide a granular summary of the data distribution, and, once privatized, they can be reused arbitrarily many times without further privacy leakage. Hence, privatized histograms facilitate *non-interactive analyses*, where the released private data synopses are reused for subsequent queries on the dataset and other downstream tasks (e.g., differentially private selection and synthetic data generation [2]). In this work, we focus on clustering data points based on their noisy histograms, specifically addressing the optimal grid resolution needed to balance bin count error due to injected noise (see Remark 1) and discretization (or quantization) error.

C. Private K-Means Clustering

Given the dataset $\mathcal{D} = \{\mathbf{x}_l \in \mathbb{R}^d, l \in \mathbb{N}_N\}$, the *$K$-means clustering* [1] groups the N data points in $\mathcal{D}$ into $K \leq N$ clusters, $\{\mathcal{C}_i \subset \mathbb{N}_N\}_{i=1}^K$, so as to minimize the average within-cluster sum of squares (WCSS), i.e., $\min\limits_{\mathcal{C}_1,\ldots,\mathcal{C}_K} \frac{1}{K}\sum\limits_{i=1}^{K} J_i$, where $J_i = \sum_{\mathbf{x}_l \in \mathcal{C}_i} \|\mathbf{x}_l - \boldsymbol{\mu}_i\|^2$ and $\boldsymbol{\mu}_i = \frac{1}{|\mathcal{C}_i|}\sum_{\mathbf{x}_l \in \mathcal{C}_i} \mathbf{x}_l$ are respectively the WCSS and centroid corresponding to the i-th cluster $\mathcal{C}_i$, $i \in \mathbb{N}_K$. *Lloyd's algorithm* is the common technique for solving the K-means problem. When initialized with a set of K centroids, it iteratively assigns each data point to its nearest centroid and then updates each cluster's centroid as the mean of the points assigned to it. For DP K-means clustering, we consider bounded data, i.e., $\mathbf{x}_l \in [-r, r]^d \ \forall\, l \in \mathbb{N}_N$, and the goal is to privately release the K cluster centroids.

1) DPLloyd: DPLloyd [3] is a popular interactive K-means clustering approach based on the standard Lloyd's algorithm. In each iteration, it injects noise into the counts and the sums of data points assigned to each cluster before updating the centroids. The main drawback is that the privacy budget ϵ must be split across a fixed number of iterations; thus, the noise added in each step increases with the number of iterations, leading to performance degradation without achieving convergence [4].

2) Non-Interactive DP K-Means Clustering: In this work, we consider non-interactive approach to private K-means clustering. The data domain $[-r, r]^d$ is partitioned into M equal-sized grids, and the histogram comprising the number of data points falling into each grid is obtained. Once the grid counts are privatized through Laplace mechanism,[2] weighted Lloyd's algorithm is executed on the grid centres using the corresponding noisy counts as weights. The choice of M is critical: A larger M provides a finer resolution, thereby reducing the *bias* in the centroid estimates but increases the *variance* as the dimension of the noise vector increases.

The EUGkM algorithm [4] selects the number of grids as $M = \big(\frac{N\epsilon}{10}\big)^{\frac{2d}{2+d}}$, generalizing the uniform gridding framework developed for two-dimensional geospatial data [17]. While this offers a structured approach for private histogram release, the derivation was primarily focused on optimizing range query's accuracy and relies on hyperparameter tuning rather than being directly informed by the cluster structure underlying the data. Consequently, this choice of M remains independent of the number of clusters K, which can lead to suboptimal or even degenerate results in many scenarios, such as when the grid is too coarse to distinguish K distinct centroids (i.e., $M < K$).

We propose a refined grid-size selection rule that explicitly accounts for the number of clusters K. Specifically, we determine the optimal number of grids by minimizing an upper bound on the expected deviation in the K-means objective

[1]While this work focuses on ϵ- DP (or *pure* DP), a relaxed notion, namely (ϵ, δ)-DP (or *approximate* DP), allows the indistinguishability bound to hold up to an additive slack $\delta \in [0, 1)$.

[2]While the discrete Laplace mechanism is the optimal ϵ-DP mechanism for histogram queries [23], which also ensures that perturbed outputs are integer-valued and provides significant utility gains over Laplace in the high-privacy regime (i.e., as $\epsilon \to 0$), we consider the Laplace mechanism in this article to maintain consistency with [4]: Extending our analysis to the discrete counterpart will be investigated as a part of future work.

$$T_1 \triangleq \mathbb{E}\Bigg[\Bigg|\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i} Z_{ij}\,\|\mathbf{Y}_{ij}\|^2\Bigg|\Bigg],\quad T_2 \triangleq \mathbb{E}\Bigg[\Bigg|\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i}\sum_{\mathbf{x}_q\in g_{ij}} (\mathbf{b}_{q,ij})^\top \mathbf{Y}_{ij}\Bigg|\Bigg],\quad \text{and}\quad T_3 \triangleq \sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i}\sum_{\mathbf{x}_q\in g_{ij}} \mathbb{E}\big[\|\mathbf{b}_{q,ij}\|^2\big] \tag{1}$$

$$\mathbb{E}\Bigg[\Bigg|\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i}\sum_{\mathbf{x}_q\in g_{ij}} (\mathbf{b}_{q,ij})^\top \mathbf{Y}_{ij}\Bigg|\Bigg] \le \sqrt{\mathbb{E}\Bigg[\Bigg(\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i}\sum_{\mathbf{x}_q\in g_{ij}} (\mathbf{b}_{q,ij})^\top \mathbf{Y}_{ij}\Bigg)^2\Bigg]} \approx \sqrt{\mathbb{E}\Bigg[\Bigg(\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i}\sum_{\mathbf{x}_q\in g_{ij}} (\mathbf{b}_{q,ij})^\top \mathbf{Y}_{ij}^{(0)}\Bigg)^2\Bigg]} \tag{2}$$

function arising from both discretization error and the noise introduced to ensure privacy. In particular, we show that the optimal grid resolution scales as $M = \Theta\Big((N\epsilon)^{\frac{2d}{4+d}}\,K^{\frac{4}{4+d}}\Big)$, differing from the EUGkM rule in both its explicit dependence on K and the exponents that govern N and ϵ. We term this grid selection strategy as RUGNIK (*Refined Uniform Gridding for Non-Interactive K-means clustering*). The derivation follows.

III. RUGNIK Derivation

Consider a cluster $\mathcal{C}_i$, $i \in \mathbb{N}_K$, with centroid $\boldsymbol{\mu}_i$ and the corresponding WCSS J_i. The data domain $[-r, r]^d$ is partitioned uniformly into $M = m^d$ grids. Let $\mathcal{G}_i$ denote the set of all grids associated[3] to cluster $\mathcal{C}_i$, and let $g_{ij} \in \mathcal{G}_i$, $j \in \mathbb{N}_{|\mathcal{G}_i|}$, denote the j-th grid, with centre $\mathbf{s}_{ij} \in \mathbb{R}^d$, that is associated to cluster $\mathcal{C}_i$. Also, let t_{ij} denote the number of data points falling within the grid g_{ij}, $t_{ij} = |\{\mathbf{x}_q \in g_{ij}\}|$, and let n_{ij} be the corresponding noisy grid counts, $n_{ij} = t_{ij} + Z_{ij}$, where $Z_{ij} \sim \mathcal{L}\big(0, \frac{1}{\epsilon}\big)$. Thus, the *observed* WCSS for cluster $\mathcal{C}_i$ is $\widetilde{J}_i = \sum_{g_{ij}\in\mathcal{G}_i} n_{ij}\,\|\mathbf{s}_{ij} - \boldsymbol{\mu}_i\|^2 = \sum_{g_{ij}\in\mathcal{G}_i} (t_{ij} + Z_{ij})\,\|\mathbf{s}_{ij} - \boldsymbol{\mu}_i\|^2 \implies \widetilde{J}_i = \sum_{g_{ij}\in\mathcal{G}_i}\sum_{\mathbf{x}_q\in g_{ij}} \|\mathbf{s}_{ij} - \boldsymbol{\mu}_i\|^2 + \sum_{g_{ij}\in\mathcal{G}_i} Z_{ij}\,\|\mathbf{s}_{ij} - \boldsymbol{\mu}_i\|^2$.

Expanding J_i as $\sum_{g_{ij}\in\mathcal{G}_i}\sum_{\mathbf{x}_q\in g_{ij}} \|\mathbf{x}_q - \boldsymbol{\mu}_i\|^2$, the deviation in the WCSS of the cluster $\mathcal{C}_i$ can be expressed as $\widetilde{J}_i - J_i = \sum_{g_{ij}\in\mathcal{G}_i} Z_{ij}\,\|\mathbf{s}_{ij} - \boldsymbol{\mu}_i\|^2 - \sum_{g_{ij}\in\mathcal{G}_i}\sum_{\mathbf{x}_q\in g_{ij}} \big(\|\mathbf{x}_q - \boldsymbol{\mu}_i\|^2 - \|\mathbf{s}_{ij} - \boldsymbol{\mu}_i\|^2\big)$, where the second term corresponds to the deviation due to quantization. Let $\mathbf{b}_{q,ij} \triangleq \mathbf{x}_q - \mathbf{s}_{ij}$ and $\mathbf{Y}_{ij} \triangleq \mathbf{s}_{ij} - \boldsymbol{\mu}_i$. Thus, $\|\mathbf{x}_q\|^2 = \|\mathbf{s}_{ij} + \mathbf{b}_{q,ij}\|^2 = \|\mathbf{s}_{ij}\|^2 + 2\,(\mathbf{b}_{q,ij})^\top \mathbf{s}_{ij} + \|\mathbf{b}_{q,ij}\|^2 \implies \|\mathbf{x}_q - \boldsymbol{\mu}_i\|^2 - \|\mathbf{s}_{ij} - \boldsymbol{\mu}_i\|^2 = \|\mathbf{x}_q\|^2 - \|\mathbf{s}_{ij}\|^2 - 2\,(\mathbf{x}_q - \mathbf{s}_{ij})^\top \boldsymbol{\mu}_i = 2\,(\mathbf{b}_{q,ij})^\top \mathbf{Y}_{ij} + \|\mathbf{b}_{q,ij}\|^2$. The space $[-r, r]^d$ contains K clusters. Assuming equally sized and uniformly spread clusters, each cluster is (roughly) contained in a hypercube of side $\frac{2r}{K^{1/d}}$. Hence, $\|\mathbf{Y}_{ij}\|_\infty \lesssim \frac{r}{K^{1/d}} \implies \|\mathbf{Y}_{ij}\|_2 \lesssim \frac{r}{K^{1/d}}\sqrt{d}$.

Now, the deviation in the objective function of the K-means clustering, $\nu = \frac{1}{K}\big|\sum_{i=1}^{K}\widetilde{J}_i - \sum_{i=1}^{K} J_i\big|$, can be expressed as

$$\nu = \frac{1}{K}\Bigg|\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i} Z_{ij}\,\|\mathbf{Y}_{ij}\|^2 - 2\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i}\sum_{\mathbf{x}_q\in g_{ij}} (\mathbf{b}_{q,ij})^\top \mathbf{Y}_{ij} - \sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i}\sum_{\mathbf{x}_q\in g_{ij}} \|\mathbf{b}_{q,ij}\|^2\Bigg|.$$

Therefore, $\mathbb{E}[\nu] \le \frac{1}{K}\big(T_1 + 2\,T_2 + T_3\big)$, where T_1, T_2, and T_3 are given in (1) and the expectation is with respect to noise and data distribution. We now bound each term in the above summation in terms of m, N, ϵ, K, and d. Noting that $\Big|\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i} Z_{ij}\,\|\mathbf{Y}_{ij}\|^2\Big| \lesssim \frac{dr^2}{K^{2/d}}\Big|\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i} Z_{ij}\Big|$ and $\mathbb{E}\Big[\Big|\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i} Z_{ij}\Big|\Big] \le \sqrt{\mathbb{E}\Big[\Big(\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i} Z_{ij}\Big)^2\Big]} = \frac{\sqrt{2M}}{\epsilon}$ as a consequence of Jensen's inequality, we have $T_1 \lesssim \frac{dr^2}{K^{2/d}}\frac{\sqrt{2M}}{\epsilon}$.

Before proceeding further with the simplification of the remaining terms, we make a few assumptions on the data distribution. We assume that the data points $\{\mathbf{x}_l\}_{l=1}^{N}$ are independent, and within each grid, the points falling in that grid are roughly uniformly distributed. Furthermore, the i-th cluster centroid $\mu_i \approx \mu_i^{(0)}$, where $\mu_i^{(0)}$ is the ground truth centroid. Hence, $[\mathbf{b}_{q,ij}]_l \sim \mathcal{U}\big(-\frac{r}{m}, \frac{r}{m}\big)$, and $\mathbf{Y}_{ij} \approx \mathbf{Y}_{ij}^{(0)} \triangleq \mathbf{s}_{ij} - \boldsymbol{\mu}_i^{(0)}$.

Under these considerations, we deduce (2), and hence, we have $T_2 \lesssim \left(\mathbb{E}\left[\left(\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i}\sum_{\mathbf{x}_q\in g_{ij}} (\mathbf{b}_{q,ij})^\top \mathbf{Y}_{ij}^{(0)}\right)^2\right]\right)^{1/2} = \left(\sum_{i=1}^{K}\sum_{g_{ij}\in\mathcal{G}_i}\sum_{\mathbf{x}_q\in g_{ij}} \mathbb{E}\left[\left((\mathbf{b}_{q,ij})^\top \mathbf{Y}_{ij}^{(0)}\right)^2\right]\right)^{1/2}$. Furthermore, $\mathbb{E}\left[\left((\mathbf{b}_{q,ij})^\top \mathbf{Y}_{ij}^{(0)}\right)^2\right] \le \mathbb{E}\Big[\|\mathbf{b}_{q,ij}\|^2\|\mathbf{Y}_{ij}^{(0)}\|^2\Big] \lesssim \frac{dr^2}{K^{2/d}}\frac{dr^2}{3m^2}$. Thus, we have $T_2 \lesssim \sqrt{\frac{N}{3}}\frac{dr^2}{mK^{1/d}}$.

Additionally, we have $\mathbb{E}\big[\|\mathbf{b}_{q,ij}\|^2\big] = \frac{dr^2}{3m^2} \implies T_3 = N\frac{dr^2}{3m^2}$. Combining the bounds on T_1, T_2, and T_3, we have $\mathbb{E}[\nu] \lesssim \frac{dr^2}{K}h(m)$, where $h(m) = \frac{\sqrt{2}m^{d/2}}{\epsilon K^{2/d}} + \frac{2}{mK^{1/d}}\sqrt{\frac{N}{3}} + \frac{N}{3m^2}$. Since $h(m)$ is a convex function[4] in m, the solution to $h'(m) = 0$ minimizes the upper bound on the expected deviation in the objective function of the K-means clustering. Moreover, $h'(m) = \frac{d}{\sqrt{2}\epsilon K^{2/d}}\frac{1}{m^3}\xi(m)$, where

$$\xi(m) = m^{d/2+2} - \frac{\epsilon}{d}\sqrt{\frac{8N}{3}}K^{1/d}m - \frac{\epsilon}{d}\frac{\sqrt{8}N}{3}K^{2/d}, \tag{3}$$

and hence, the optimal m solves $\xi(m) = 0$. We cannot solve this in closed form. However, from the following proposition, we know that $\xi(m)$ has a unique root.

Proposition 1. ξ has a unique root on $\mathbb{R}_{++}$.

Proof. We observe that ξ is strictly convex on $\mathbb{R}_+$; hence, ξ' is a strictly increasing function on $\mathbb{R}_+$. Moreover, since $\xi(0) < 0$ and $\lim_{m\to\infty}\xi(m) = \infty$, it follows that there exists a unique $\tau \ge 0$ such that $\xi(\tau) = \xi(0) < 0$ and $\xi'(m) > 0\ \ \forall\, m > \tau$ (take

[3] We say the grid g with centre $\mathbf{s} \in \mathbb{R}^d$ is associated to cluster $\mathcal{C}_i$ if $\|\mathbf{s} - \boldsymbol{\mu}_i\| < \|\mathbf{s} - \boldsymbol{\mu}_k\|\ \ \forall\, k \in \mathbb{N}_K \setminus \{i\}$.

[4] Since $h''(m) = \Big(\frac{d}{2} - 1\Big)\frac{dm^{d/2-2}}{\sqrt{2}\epsilon K^{2/d}} + \frac{4}{m^3 K^{1/d}}\sqrt{\frac{N}{3}} + \frac{2N}{m^4} \ge 0\ \ \forall\, d \ge 2$.

$\tau = 0$ if $\xi'(0) \geq 0$). Thus, ξ is strictly increasing on (τ, ∞), and since $\xi(m) < 0 \ \forall\, m \in [0, \tau]$, the result follows. $\square$

The following result presents an upper and a lower bound on the optimal M under mild technical conditions on ϵ.

Theorem 2. The optimal number of grids is bounded as follows.

(i) If $\epsilon \geq \epsilon_0 \triangleq d\sqrt{\frac{3K}{8N}}\left(1+\sqrt{\frac{N}{3}}\right)^{-1}$, $M \geq \left\lfloor \left(\eta K^{2/d}\right)^{\frac{2d}{4+d}} \right\rfloor$, and

(ii) if $\epsilon \leq \epsilon_1 \triangleq \frac{d\sqrt{2K}}{3}\left(\frac{4N}{3}\right)^{d/4}$, $M \leq \left\lceil \left(\gamma\sqrt{3N}\eta K^{2/d}\right)^{\frac{2d}{4+d}} \right\rceil$,

where $\eta \triangleq \frac{\epsilon}{d}\sqrt{\frac{8N}{3}}\left(1+\sqrt{\frac{N}{3}}\right)$ and $\gamma \triangleq \left(1+\sqrt{\frac{N}{3}}\right)^{-1}$.

Proof. Note that $\xi(m) = m^{d/2+2} - \eta\gamma K^{1/d} m - \eta(1-\gamma)K^{2/d}$. Let $m_l = \left(\eta K^{2/d}\right)^{\frac{2}{4+d}}$; observing that $\epsilon \geq \epsilon_0 \iff \eta \geq \sqrt{K}$, we deduce $\xi(m_l) \leq 0$. Also, taking $m_u = \left(\gamma\sqrt{3N}\eta K^{2/d}\right)^{\frac{2}{4+d}}$ and noting $\epsilon \leq \epsilon_1 \iff \eta \leq \frac{2\sqrt{K}}{3\gamma}\left(\frac{4N}{3}\right)^{d/4+1/2}$, we have $\xi(m_u) \geq 0$. Following Proposition 1, we prove the result. $\square$

Table I provides the ranges of ϵ where these bounds hold. Clearly, the optimal grid size is $M = \Theta\left((N\epsilon)^{\frac{2d}{4+d}} K^{\frac{4}{4+d}}\right)$ for acceptable values of ϵ [24]. Unlike EUGkM, this grid size depends explicitly on K and have different exponents for N and ϵ. Also, note that the grid size is independent of r.

TABLE I
RANGES OF ϵ FOR WHICH THE BOUNDS ON M IN THEOREM 2 HOLD.

		$K=2$		$K=4$		$K=8$	
	N	**100**	**200**	**200**	**400**	**400**	**800**
$d=2$	ϵ_0	0.026	0.014	0.019	0.010	0.014	0.008
	ϵ_1	15.396	21.773	30.792	43.546	61.584	87.092
$d=3$	ϵ_0	0.039	0.021	0.029	0.015	0.021	0.011
	ϵ_1	78.475	131.979	186.647	313.901	443.924	746.588

IV. SIMULATION RESULTS

We evaluate the performance of our proposed scheme, RUGNIK, against several state-of-the-art DP K-means clustering methods, namely, EUGkM [4], DPLloyd implementation in IBM's `diffprivlib` [14], and Google's coreset-based clustering (GCC) [16]. All methods are tested on synthetic datasets generated as described in [4] for a variety of configurations.

Table II summarizes the average WCSS achieved by different techniques across these settings for a wide range of privacy

TABLE II
WITHIN-CLUSTER SUM OF SQUARES (WCSS) ACHIEVED BY VARIOUS METHODS.

			$d=2$						$d=3$					
	Privacy budget, ϵ		**0.1**	**0.15**	**0.25**	**0.4**	**0.6**	**1.0**	**0.1**	**0.15**	**0.25**	**0.4**	**0.6**	**1.0**
$K=2$	$N=100$	**IBM**	28.556	28.014	26.775	24.757	22.498	18.500	45.597	44.860	43.051	40.357	36.978	30.918
		GCC	15.035	14.772	14.233	13.434	12.486	10.950	19.990	19.673	19.044	18.092	16.897	14.952
		EUGkM	16.523	16.523	3.968	3.621	3.509	2.479	21.355	21.355	21.355	6.931	6.413	**3.797**
		RUGNIK	**6.114**	**6.002**	**3.824**	**2.417**	**1.862**	**1.537**	**13.091**	**10.604**	**8.214**	**6.286**	**4.651**	3.800
		No priv.	1.249						1.900					
	$N=200$	**IBM**	54.555	52.172	47.163	41.049	33.716	23.594	87.687	84.612	77.367	67.751	57.027	42.012
		GCC	29.165	28.144	26.179	23.632	20.849	16.854	38.866	37.671	35.308	32.112	28.565	23.281
		EUGkM	33.190	7.483	6.977	5.010	4.882	2.927	42.708	14.980	12.883	12.182	7.294	6.885
		RUGNIK	**9.077**	**6.682**	**4.041**	**3.253**	**2.930**	**2.696**	**18.004**	**14.964**	**10.428**	**8.089**	**5.697**	**4.838**
		No priv.	2.458						3.838					
$K=4$	$N=200$	**IBM**	16.669	16.408	15.917	15.092	14.043	12.053	30.666	30.231	29.297	27.748	25.788	22.417
		GCC	18.926	18.436	17.558	16.442	15.302	13.607	22.871	22.414	21.550	20.348	19.042	17.023
		EUGkM	21.808	4.533	4.533	2.660	2.594	1.423	25.145	**8.140**	7.620	7.371	3.981	3.685
		RUGNIK	**4.508**	**3.249**	**2.000**	**1.357**	**1.032**	**0.844**	**8.849**	8.309	**5.920**	**4.493**	**3.030**	**2.184**
		No priv.	0.683						1.319					
	$N=400$	**IBM**	32.278	31.015	29.130	26.104	22.679	17.309	58.917	57.131	53.059	47.888	42.125	33.462
		GCC	36.170	34.569	31.911	28.971	26.022	21.663	43.900	42.285	39.422	35.987	32.513	27.311
		EUGkM	9.133	9.132	5.218	3.012	1.890	1.624	15.708	15.125	8.276	7.443	4.420	4.082
		RUGNIK	**5.259**	**3.319**	**2.297**	**1.797**	**1.654**	**1.515**	**13.693**	**10.488**	**7.124**	**5.099**	**3.884**	**3.345**
		No priv.	1.382						2.662					
$K=8$	$N=400$	**IBM**	8.981	8.903	8.664	8.343	7.898	7.063	20.068	19.835	19.366	18.585	17.603	15.829
		GCC	20.018	19.134	17.668	16.065	14.540	12.479	26.408	25.543	24.034	22.298	20.531	18.073
		EUGkM	6.408	6.408	2.431	2.050	1.331	0.874	8.361	8.352	5.347	5.044	3.187	2.986
		RUGNIK	**2.723**	**1.944**	**1.268**	**0.785**	**0.601**	**0.481**	**6.975**	**6.383**	**4.564**	**3.341**	**2.395**	**1.721**
		No priv.	0.362						0.946					
	$N=800$	**IBM**	17.419	16.989	16.176	14.889	13.410	10.952	38.841	37.896	36.046	33.368	30.134	24.807
		GCC	37.018	34.490	30.835	27.167	23.624	18.502	49.828	47.180	43.113	38.757	34.549	28.448
		EUGkM	4.800	4.783	4.253	1.761	1.280	0.835	16.616	10.258	9.796	5.923	3.607	2.630
		RUGNIK	**3.082**	**2.012**	**1.322**	**0.972**	**0.865**	**0.794**	**10.532**	**8.159**	**5.461**	**3.888**	**3.074**	**2.524**
		No priv.	0.701						1.847					

budgets ϵ. Lower WCSS indicates better clustering quality. The results show that our scheme consistently outperforms all baseline schemes across all considered settings. Furthermore, RUGNIK offers a significant improvement over EUGkM that becomes more pronounced in regimes with stronger privacy constraints (i.e., smaller ϵ) and larger numbers of clusters, where the choice of grid resolution plays a significant role.

Table III compares the number of grids M selected by EUGkM and our proposed scheme. It is evident that EUGkM results in degenerate choices, $M < K$ (marked *), in high-privacy regimes. In contrast, RUGNIK appropriately scales the grid resolution by explicitly accounting for K, ensuring high clustering accuracy even under strong privacy constraints.

V. Conclusions

We studied the problem of non-interactive differentially private K-means clustering, and we derived the optimal grid resolution for uniform gridding. Our selection rule is theoretically grounded and explicitly incorporates the number of clusters in the grid size. The efficacy and robustness of our refined grid resolution are validated using extensive empirical analysis under a wide range of settings, and we demonstrated superior performance over the existing state-of-the-art and benchmark open-source implementations such as `diffprivlib` [14], and Google's coreset-based clustering [16], with significant improvements in high privacy regime. Future work will focus on extending the analysis to adaptive grid structures and on discrete noise mechanisms that will further improve performance in the high-privacy regimes. It will also be interesting to evaluate RUGNIK's performance on real-world datasets.

References

[1] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[2] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, 2014.

[3] A. Blum *et al.*, "Practical privacy: the SuLQ framework," in *Proc. ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, 2005, pp. 128–138.

[4] D. Su *et al.*, "Differentially Private K-Means Clustering and a Hybrid Approach to Private Optimization," *ACM Trans. Priv. Secur.*, vol. 20, no. 4, pp. 1–33, 2017.

[5] C. Dwork, "A firm foundation for private data analysis," *Commun. ACM*, vol. 54, no. 1, pp. 86–95, 2011.

[6] Z. Lu and H. Shen, "Differentially Private k-Means Clustering With Convergence Guarantee," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 4, pp. 1541–1552, 2021.

[7] Y. Wang, Y.-X. Wang, and A. Singh, "Differentially private subspace clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015.

[8] M. Jones, H. L. Nguyen, and T. D. Nguyen, "Differentially private clustering via maximum coverage," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 13, 2021, pp. 11 555–11 563.

[9] A. Epasto *et al.*, "k-means clustering with distance-based privacy," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 19 570–19 593.

[10] T. D. Nguyen *et al.*, "Privacy aware k-means clustering with high utility," in *Proc. Pacific-Asia Conf. Knowl. Discov. Data Mining*. Springer, 2016, pp. 388–400.

[11] V. Cohen-Addad *et al.*, "Near-optimal private and scalable k-clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 10 462–10 475.

[12] A. Gilad *et al.*, "Differentially Private Explanations for Clusters," *Proc. ACM Manage. Data*, vol. 3, no. 4, pp. 1–27, 2025.

[13] B. Ghazi, R. Kumar, and P. Manurangsi, "Differentially private clustering: Tight approximation ratios," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4040–4054.

[14] N. Holohan *et al.*, "`Diffprivlib`: The IBM differential privacy library," *2019, arXiv:1907.02444*.

[15] A. Chang *et al.*, "Locally private k-means in one round," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2021, pp. 1441–1451.

[16] A. Chang and P. Kamath. (21 Oct. 2021) Practical Differentially Private Clustering. [Online]. Available: https://research.google/blog/practical-differentially-private-clustering/

[17] W. Qardaji, W. Yang, and N. Li, "Differentially private grids for geospatial data," in *Proc. IEEE Int. Conf. Data Eng.* IEEE, 2013, pp. 757–768.

[18] J. Zhang *et al.*, "Functional mechanism: Regression analysis under differential privacy," in *Proc. VLDB Endowment*, vol. 5, no. 11, 2012.

[19] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *J. Mach. Learn. Res.*, vol. 12, no. 3, 2011.

[20] G. Muthukrishnan and A. Tandon, "On the Optimal Number of Grids for Differentially Private Non-Interactive K-Means Clustering," *arXiv:2603.26963*, 2026.

[21] Y. Zhu and Y.-X. Wang, "Poission subsampled Rényi differential privacy," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2019, pp. 7634–7642.

[22] X. Tian and J. Taylor, "Selective inference with a randomized response," *Ann. Statist.*, vol. 46, no. 2, pp. 679–710, 2018.

[23] Q. Geng and P. Viswanath, "The optimal noise-adding mechanism in differential privacy," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 925–951, 2016.

[24] J. P. Near *et al.*, "Guidelines for Evaluating Differential Privacy Guarantees," National Institute of Standards and Technology, Tech. Rep. NIST SP 800-226, 2025.

TABLE III
NUMBER OF GRIDS SELECTED BY EUGkM AND RUGNIK.

			$d=2$						$d=3$					
	Privacy budget, ϵ		**0.1**	**0.15**	**0.25**	**0.4**	**0.6**	**1.0**	**0.1**	**0.15**	**0.25**	**0.4**	**0.6**	**1.0**
$K=2$	$N=100$	**EUGkM**	1*	1*	4	4	4	9	1*	1*	1*	8	8	27
		RUGNIK	4	9	9	16	16	25	8	8	8	27	27	27
	$N=200$	**EUGkM**	1*	4	4	9	9	16	1*	8	8	8	27	27
		RUGNIK	9	9	16	25	36	49	8	8	27	27	64	64
$K=4$	$N=200$	**EUGkM**	1*	4	4	9	9	16	1*	8	8	8	27	27
		RUGNIK	16	16	25	36	49	64	8	27	27	64	64	125
	$N=400$	**EUGkM**	4	4	9	16	25	36	8	8	27	27	64	64
		RUGNIK	25	25	36	49	81	100	27	27	64	64	125	216
$K=8$	$N=400$	**EUGkM**	4*	4*	9	16	25	36	8	8	27	27	64	64
		RUGNIK	36	36	64	81	121	169	27	64	64	125	125	216
	$N=800$	**EUGkM**	9	9	16	36	49	81	8	27	27	64	125	216
		RUGNIK	49	64	100	121	169	256	64	64	125	216	343	512