Distributed Gradient Clustering: Convergence and the Effect of Initialization

1st Aleksandar Armacki*
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA, USA
aarmacki@andrew.cmu.edu

2nd Himkant Sharma*
Electronics and Electrical Communication Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India
himkant@kgpian.iitkgp.ac.in

3rd Dragana Bajović
Department of Power, Electronics and Computer Engineering
Faculty of Technical Sciences, University of Novi Sad
Novi Sad, Serbia
dbajovic@uns.ac.rs

4th Dušan Jakovetić
Department of Mathematics and Informatics
Faculty of Sciences, University of Novi Sad
Novi Sad, Serbia
dusan.jakovetic@dmi.uns.ac.rs

5th Mrityunjoy Chakraborty
Electronics and Electrical Communication Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India
mrityun@ece.iitkgp.ac.in

6th Soummya Kar
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA, USA
soummyak@andrew.cmu.edu

***Abstract*—We study the effects of center initialization on the performance of a family of distributed gradient-based clustering algorithms introduced in [1], that work over connected networks of users. In the considered scenario, each user contains a local dataset and communicates only with its immediate neighbours, with the aim of finding a global clustering of the joint data. We perform extensive numerical experiments, evaluating the effects of center initialization on the performance of our family of methods, demonstrating that our methods are more resilient to the effects of initialization, compared to centralized gradient clustering [2]. Next, inspired by the K-means++ initialization [3], we propose a novel distributed center initialization scheme, which is shown to improve the performance of our methods, compared to the baseline random initialization.**

***Index Terms*—clustering, distributed, networks, convergence, initialization, K-means++**

The first two authors contributed equally. The work of A. Armacki and S. Kar is supported in part by the National Science Foundation under grant ECCS 2330196. The work of H. Sharma was supported in part by a SPARC grant from the MHRD, Govt. of India. The work of D. Bajović is supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-65/2024-03/200156), by Faculty of Technical Sciences, University of Novi Sad, through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad" (No. 01-3394/1), and by Serbian Ministry of Science, Technological development and Innovation, within the bilateral project Serbia-Slovakia No. 337-00-3/2024-05/16. The work of D. Jakovetić is supported by the Science Fund of Republic of Serbia, project "LASCADO", grant No. 7359, by Provincial Secretariat for Higher Education and Scientific Research, grant No. 142-451- 2593/2021-01/2. The work of D. Jakovetić and D. Bajović is also supported by the European Union's Horizon Europe program under grant agreement No. 101093006. *(Corresponding author: Aleksandar Armacki.)*

I. Introduction

Clustering is an unsupervized learning problem, with the goal of finding groups of similar data, without having any knowledge of the underlying distribution, or even the true number of groups [4], [5]. Depending on the approach, data can be assigned to clusters in a *hard* or *soft* manner, with hard clustering assigning data to exclusively one cluster, while soft clustering provides the probability of a sample belonging to each cluster. In this paper we will be studying the problem of hard clustering. Additionally, we will focus on center-based clustering, e.g., [6], where the goal is to find centers which represent the clusters. Many popular algorithms fall in this category, including the celebrated Lloyd's method [7], its extension to Bregman losses [8], Huber loss clustering [9], as well as the recently proposed gradient based clustering [2].

Clustering has traditionally been studied in the centralized regime, where the methods are run across the entire dataset. Another learning paradigm, which has been attracting significant interest is that of *distributed learning*. *Distributed learning* is a popular learning paradigm, wherein many users collaborate to train a joint model, while keeping their data private. There are many approaches to distributed learning, such as federated learning (FL), e.g., [10]–[13], and peer-to-peer (P2P) distributed learning, e.g., [14]–[17]. In this paper we are interested in the P2P setup, where users communicate directly with one another, while no user can communicate directly with all the others. The communication network is modeled as a connected graph $G = (V, E)$. Clustering in

this setup is very challenging, as the data is stored locally at each user, with users only able (or willing) to exchange local parameter estimates (e.g., local centers) in order to achieve the final goal of obtaining a clustering of the entire, joint dataset.

Literature review. Distributed clustering has been considered in [18]–[24]. Work [18] proposes approximate K-means algorithms for both P2P and FL setups, providing theoretical guarantees only in the FL setup. Works [19], [22], [24] study distributed soft and hard K-means clustering, with only the method in [19] providing convergence guarantees to a local minima of the centralized K-means problem. In [20], the authors study K-means and K-medians problems and rely on the idea of coresets [25], to design methods with provable constant approximation guarantees. Work [21] studies distributed K-means in the special case where users have a single sample, while in [23] the authors design a parametric family of distributed K-means methods, establishing convergence of centers to local minima of the centralized K-means problem. Finally, we propose a unified framework for distributed clustering in [1], that considers popular clustering methods beyond K-means, such as Huber loss clustering [9].

Contributions. In this work we study the effects of center initialization on the performance of the distributed gradient-based clustering (DGC-$\mathcal{F}_\rho$) method proposed in [1]. To that end, we perform extensive numerical experiments, demonstrating that DGC-$\mathcal{F}_\rho$ is more resilient to center initialization, compared to the centralized gradient clustering (CGC) method from [2]. Inspired by the celebrated K-means++ initialization, we then propose a novel distributed center initialization scheme, dubbed Distributed K-means+Clustering (DKM+C), which combines local K-means++ with multiple communication and local clustering rounds, to produce the initial centers. The proposed scheme is shown to result in better performance of the algorithm, compared to the baseline random center initialization.

Paper organization. The rest of the paper is organized as follows. Section II formally states the problem of distributed center-based clustering, Section III introduces the proposed family of methods, Section IV provides theoretical results, Section V provides numerical results and Section VI concludes the paper. The remainder of this section introduces notation.

Notation. The spaces of real numbers and d-dimensional vectors are denoted by $\mathbb{R}$ and $\mathbb{R}^d$, with $\|\cdot\|$ denoting the Euclidean norm. The set of non-negative integers is denoted by $\mathbb{N}$, with $[M]=\{1,\dots,M\}$, for any $M\in\mathbb{N}$. For a matrix $A\in\mathbb{R}^{d\times d}$, $A^\top$ and $\overline{\lambda}(A)$ denote transposition and the largest eigenvalue of A. Superscripts and subscripts denote iterations and users, while brackets correspond to the particular center or cluster, e.g., $x_i^t(k)$ is center k of user i at iteration t.

II. Problem Formulation

Consider a network of $m>1$ users, communicating over a graph $G=(V,E)$, where $V=[m]$ is the set of vertices (i.e., users), E is the set of undirected edges connecting them, such that $\{i,j\}\in E$ if and only if users i, j communicate. Each user contains a local dataset $\mathcal{D}_i=\{y_{i,1},\dots,y_{i,N_i}\}\subset\mathbb{R}^d$, for some $N_i\geq 1$. The goal is to produce a clustering of the global data $\mathcal{D}=\cup_{i\in[m]}\mathcal{D}_i$, into $K\geq 2$ disjoint clusters. Formally, the problem can be stated as

$$\min_{\substack{\mathbf{x}_i\in\mathbb{R}^{Kd},\, C_i\in\mathcal{C}_{K,\mathcal{D}_i},\, i\in[m]\\ \text{subject to } \mathbf{x}_1=\ldots=\mathbf{x}_m}} \sum_{i\in[m]}\sum_{k\in[K]}\sum_{r\in C_i(k)} f(x_i(k),y_{i,r}), \quad (1)$$

where $\mathbf{x}_i=\begin{bmatrix}x_i(1)^\top & \dots & x_i(K)^\top\end{bmatrix}^\top$ is the vector stacking the K centers $x_i(k)\in\mathbb{R}^d$ of user i, $\mathcal{C}_{K,\mathcal{D}_i}$ is the set of all K-partitions of $\mathcal{D}_i$, i.e., $C_i\in\mathcal{C}_{K,\mathcal{D}_i}$ is a K-tuple $C_i=\big(C_i(1),\dots,C_i(K)\big)$, such that $C_i(k)\subseteq\mathcal{D}_i$,[1] $C_i(k)\cap C_i(l)=\emptyset$ and $\cup_{k\in[K]}C_i(k)=\mathcal{D}_i$. Here $f:\mathbb{R}^d\times\mathbb{R}^d\mapsto[0,\infty)$ is a loss function, e.g., $f(x,y)=\|x-y\|^2$ recovers the distributed K-means problem, with many other possibilities, such as Huber, Logistic or Fair loss, see [1]. In general, (1) is NP-hard, even in the centralized setting, e.g., [26]–[28]. As such, the best one can hope for is reaching stationary points, with various schemes guaranteeing this in both centralized, e.g., [2], [7], [8], [29], and distributed settings, e.g., [1], [19], [23].

The problem (1) ensures clustering of the joint data is produced, by requiring centers across all users to be the same. As (1) is a constrained problem, a relaxation is proposed in [1], making it amenable to a distributed first-order approach. The relaxed problem is given by

$$\min_{\substack{\mathbf{x}\in\mathbb{R}^{Kmd},\\ C\in\mathcal{C}_{m,K,\mathcal{D}}}} J_\rho(\mathbf{x},C)=\sum_{i\in[m]}\sum_{k\in[K]}\Big[\frac{1}{2}\sum_{j\in\mathcal{N}_i}\|x_i(k)-x_j(k)\|^2 + \frac{1}{\rho}\sum_{r\in C_i(k)} w_{i,r}f(x_i(k),y_{i,r})\Big], \quad (2)$$

where $\mathcal{C}_{m,K,\mathcal{D}}$ is the set of all clusterings of the entire data, i.e., for $C\in\mathcal{C}_{m,K,\mathcal{D}}$, we have $C=(C_1,\dots,C_m)$, with $C_i\in\mathcal{C}_{K,\mathcal{D}_i}$, $\mathcal{N}_i=\{j\in V:\{i,j\}\in E\}$ is the set of neighbours of user i (not including i), while $\rho\geq 1$ is a tunable parameter. The formulation (2) relaxes (1), by considering an unconstrained problem which penalizes the difference of centers among neighbouring users and controls the trade-off between center estimation and proximity, via the parameter ρ.

III. The DGC-$\mathcal{F}_\rho$ Family of Methods

In this section we describe the DGC-$\mathcal{F}_\rho$ family of methods proposed in [1]. We refer to DGC-$\mathcal{F}_\rho$ as a family of methods, as it subsumes several distributed clustering methods, such as K-means, Huber, Logistic and Fair loss-based. In each iteration users maintain their center and cluster estimates. To begin, users choose initial centers $\mathbf{x}_i^0\in\mathbb{R}^{Kd}$, $i\in[m]$. At iteration $t\geq 0$, users first form the clusters locally, by finding a $k\in[K]$ for each data point $r\in\mathcal{D}_i$, such that the k-th center is the closest to the point r, i.e., such that

$$\|x_i^t(k)-y_{i,r}\|\leq\|x_i^t(l)-y_{i,r}\|, \text{ for all } l\neq k, \quad (3)$$

[1] In a slight abuse of notation, we will also use $\mathcal{D}_i$ to denote the set of indices of the data, i.e., $\mathcal{D}_i=[N_i]$.

Algorithm 1 DGC-$\mathcal{F}_\rho$

Require: $\alpha > 0$, $\rho \geq 1$, initial centers $\mathbf{x}_i^0 \in \mathbb{R}^{Kd}$, $i \in [m]$.

1: **for** all users i in parallel, in round t = 0,1,...,T-1 **do**
2: Set $C_i^{t+1}(k) \leftarrow \emptyset$, for all $k \in [K]$;
3: **for** each $r \in [N_i]$ **do**
4: Find k so that $\|x_i^t(k) - y_{i,r}\| \leq \|x_i^t(l) - y_{i,r}\|$, $l \neq k$;
5: Update $C_i^{t+1}(k) \leftarrow C_i^{t+1}(k) \cup \{r\}$;
6: Exchange centers with neighbours $j \in \mathcal{N}_i$;
7: Update $x_i^{t+1}(k)$ by performing (4), for all $k \in [K]$;
8: **Return** $(\mathbf{x}_i^T, C_i^T)$, $i \in [m]$.

and assign $y_{i,r}$ to $C_i^{t+1}(k)$. Next, the centers are updated via

$$x_i^{t+1}(k) = x_i^t(k) - \alpha \sum_{j \in \mathcal{N}_i} \left[x_i^t(k) - x_j^t(k)\right] - \frac{\alpha}{\rho} \sum_{r \in C_i^{t+1}(k)} \nabla_x f\left(x_i^t(k), y_{i,r}\right), \quad (4)$$

where $\alpha > 0$ is a fixed step-size. The procedure is summarized in Algorithm 1.[2] Note that centers can be initialized randomly, providing flexibility in designing initialization algorithms, such as distributed variants of K-means++, e.g., [22], or the method we propose in Section V ahead. The center update is built on the consensus+innovation framework, e.g., [30], [31].

IV. Convergence Guarantees

We start by defining the notion of points to which DGC-$\mathcal{F}_\rho$ converges to, referred to as *fixed points*.

Definition 1: Let $\mathbf{x} \in \mathbb{R}^{Kmd}$ be cluster centers. We say that $U_{\mathbf{x}} \subset \mathcal{C}_{m,K,\mathcal{D}}$ is the set of optimal clusterings with respect to $\mathbf{x}$, if condition (3) is satisfied for all clusterings $C \in U_{\mathbf{x}}$.

Definition 2: The pair $(\mathbf{x}^\star, C^\star) \in \mathbb{R}^{Kmd} \times \mathcal{C}_{m,K,\mathcal{D}}$ is a fixed point of DGC-$\mathcal{F}_\rho$, if 1) $C^\star \in U_{\mathbf{x}^\star}$; 2) $\nabla J_\rho(\mathbf{x}^\star, C^\star) = 0$.

Definition 3: $\overline{U}_{\mathbf{x}} \subset \mathcal{C}_{m,K,\mathcal{D}}$ is the set of clusterings, such that 1) $\overline{U}_{\mathbf{x}} \subseteq U_{\mathbf{x}}$; 2) $\nabla J_\rho(\mathbf{x}, C) = 0$, for all $C \in \overline{U}_{\mathbf{x}}$.

Note that Definition 2 requires $(\mathbf{x}^\star, C^\star)$ to be a stationary point of J_ρ, in the sense that clusters $C^\star$ are optimal for fixed centers $\mathbf{x}^\star$ and centers $\mathbf{x}^\star$ are optimal for fixed clusters $C^\star$. As such, it is not possible to further improve the clusters, nor the centers at a fixed point. By Definitions 1-3, a point $\mathbf{x}$ is a fixed point if and only if $\overline{U}_{\mathbf{x}} \neq \emptyset$. As such, we will call a point $\mathbf{x}$ a fixed point if $\overline{U}_{\mathbf{x}} \neq \emptyset$. We next state our assumptions.

Assumption 1: The full data has at least K distinct samples.

Assumption 2: The graph $G = (V, E)$ is connected.

Assumption 3: The loss f is coercive, convex and β-smooth with respect to the first argument and preserves the ordering with respect to Euclidean distance, i.e., for each $x, y, z \in \mathbb{R}^d$ 1) $\lim_{\|x\| \to \infty} f(x, y) = \infty$; 2) $0 \leq f(x, y) - f(z, y) - \langle \nabla_z f(z, y), x - z \rangle \leq \frac{\beta}{2}\|x - z\|^2$; 3) $f(x, y) < f(z, y)$ if $\|x - y\| < \|z - y\|$ and $f(x, y) = f(z, y)$ if $\|x - y\| = \|z - y\|$.

Assumptions 1-3 are mild assumptions on the global data, communication graph and loss function. Assumption 1 requires the full data to have K distinct points, while placing no requirements on the local datasets. Assumption 2 requires the communication graph to be connected, which allows for (1) to be solved in a distributed manner. Finally, Assumption 3 is a mild assumption on the behaviour of the loss, which is an intrinsic property of the loss, independent of the data that we wish to cluster. It is shown to be satisfied by a broad class of popular clustering losses, including K-means (i.e., squared Euclidean), Huber, Logsitic and Fair loss, see [1] for details. We are now ready to state the convergence result from [1].

[2]The method in [1] is more general, in that it allows for distance metrics beyond Euclidean and multiple center updates per iteration, see [1] for details.

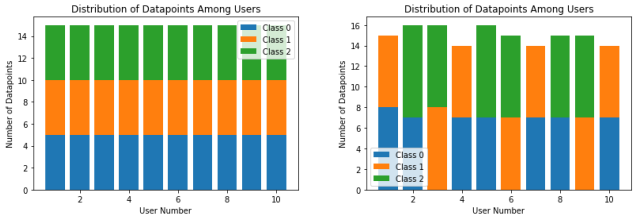

Fig. 1. Homogeneous and heterogeneous data distributions across users.

Theorem 1: Let Assumptions 1-3 hold. For the step-size $\alpha < (\beta/\rho + \overline{\lambda}(L))^{-1}$, any initialization $\mathbf{x}^0 \in \mathbb{R}^{Kmd}$ and $\rho \geq 1$, the sequence of centers $\{\mathbf{x}^t\}_{t \in \mathbb{N}}$ generated by DGC-$\mathcal{F}_\rho$ converges to a fixed point $\mathbf{x}^\star = \mathbf{x}^\star(0, \rho) \in \mathbb{R}^{Kmd}$, such that $\overline{U}_{\mathbf{x}^\star} \neq \emptyset$. Moreover, the clusters converge in finite time, i.e., there exists a $t_0 > 0$ such that $U_{\mathbf{x}^t} \subseteq U_{\mathbf{x}^\star}$, for all $t \geq t_0$.

Theorem 1 shows that the sequence of centers generated by DGC-$\mathcal{F}_\rho$ is guaranteed to converge to a fixed point, for any center initialization. We emphasize that the fixed point to which the sequence of centers converges, $\mathbf{x}^\star = \mathbf{x}^\star(0, \rho)$, depends on center initialization and penalty parameter ρ. In the next section we perform extensive empirical studies of the effect of center initialization on the performance of DGC-$\mathcal{F}_\rho$. A detailed theoretical study on the effect of parameter ρ on fixed points of DGC-$\mathcal{F}_\rho$, as $\rho \to \infty$, is provided in [1].

V. Numerical Results

In this section we study the effect of center initialization of the performance of DGC-$\mathcal{F}_\rho$. In particular, we test the performance of our method using K-means loss, dubbed DGC-KM. All experiments are performed on Iris data [32], which consists of 150 samples, belonging to $K = 3$ classes (50 samples per class), and dimension $d = 4$. We distribute the samples across a ring network of $m = 10$ users. We consider two types of data distributions across users: *homogeneous* and *heterogeneous*. In the homogeneous setup, each user is assigned samples from all three classes, in equal proportion. In the heterogeneous setup, each user is assigned samples from two out of three classes, with different number of samples per class and per user. The data distributions are visualized in Figure 1. We set $\rho = 10$ in the homogeneous setup and $\rho = 100$ in the heterogeneous setup, due to different data distributions across users, to enforce consensus more strongly.

To test the robustness to initialization of our method, we apply our DGC-KM method with two different center initializations: *random* and *local K-means++*. For random initialization, each user selects three centers uniformly at

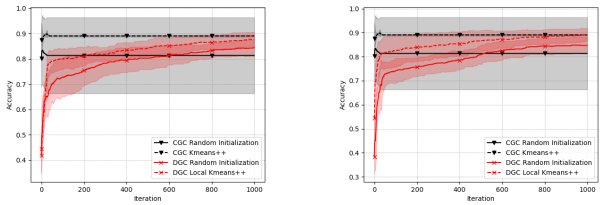

Fig. 2. Performance of methods on homogeneous and heterogeneous data.

random from their local data. For local K-means++, each user initializes their centers using the K-means++ scheme on their local data. To compare the sensitivity of our algorithm, we use the CGC method from [2], also using the K-means cost. CGC is also initialized using random and K-means++ initialization, with the difference being that CGC chooses samples from the entire dataset, as it is a centralized algorithm. We run both methods for $T = 1000$ iterations, on both homogeneous and heterogeneous data.[3] We measure the performance via clustering *accuracy*, i.e., by comparing the true labels to the ones produced by the clustering methods, accounting for label permutation. We repeat the experiments across 5 runs and present the average accuracy. For our distributed method, we additionally average the accuracy across users. The results are presented in Figure 2. The solid lines represent accuracy per iteration using random initialization, while dashed lines represent the performance using K-means++. We can see that our DGC-KM method shows less sensitivity to center initialization compared to CGC-KM, with the gap in performance of our method under different initialization much smaller than that of CGC-KM. This phenomena has previously been observed in [1], [19], where it was noted that distributed clustering algorithms are less sensitive to initialization, compared to their centralized counterparts, as they in essence perform m initialization, one per each user, whereas the centralized algorithms only perform one initialization. Combined with the effects of consensus, these multiple initializations help mitigate the effects of bad initialization at some users and lead the algorithm to a solution of better quality.

Another interesting observation from Figure 2 is that the performance of DGC-KM is improved under local K-means++ initialization, even when the data across users is heterogeneous. This leads to a natural question of can the benefits of K-means++ initialization be further exploited in the distributed setup, if users are allowed to collaborate during the initialization phase? To answer this question, we design a novel distributed center initialization algorithm, dubbed Distributed K-Means+Clustering (DKM+C). The algorithm runs for a user-specified number of communication rounds $L \geq 0$, with $L = 0$ corresponding to each user initializing their centers by performing K-means++ on their local data. If $L \geq 1$, then in round $l = 1, \ldots, L$, users first exchange their centers from round $l-1$ with their neighbours. Next, in order

[3]Note that the distinction between homogeneous and heterogeneous data is irrelevant for CGC, as it is a centralized algorithm with access to all data.

Algorithm 2 DKM+C

Require: Number of centers K and communication rounds $L \geq 0$.

1: **for** all users i in parallel, in communication round $l = 0, 1, \ldots, L$ **do**
2: **if** l = 0 **then**
3: Produce $\mathbf{x}_i^l$ by performing K-means++ on local data;
4: Exchange centers $\mathbf{x}_i^{l-1}, \mathbf{x}_j^{l-1}$ with neighbours $j \in \mathcal{N}_i$;
5: Produce new centers $\mathbf{x}_i^l$, by running K-means on $\{x_i^{l-1}(k),\ x_j^{l-1}(k) :\ k \in [K],\ j \in \mathcal{N}_i\}$;
6: Initialize centers via $\mathbf{x}_i^0 \leftarrow \mathbf{x}_i^L$, for all $i \in [m]$.

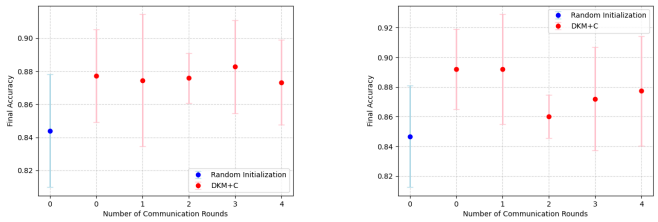

Fig. 3. Performance of methods on homogeneous and heterogeneous data.

to account for potential label mismatch among different users, each user locally runs the K-means clustering algorithm on their own and their neighbours centers, i.e., on the dataset $\{x_i^{l-1}(k),\ x_j^{l-1}(k) :\ k \in [K],\ j \in \mathcal{N}_i\} \subset \mathbb{R}^d$, of size $(|\mathcal{N}_i|+1)K$. The new centers $\mathbf{x}_i^l$ are the centers returned by K-means. The steps are then repeated, until all L communication rounds are performed. The proposed initialization scheme is summarized in Algorithm 2. The idea behind the proposed scheme is to combine the power of K-means++ with local communications, to produce center initializations carrying more information than purely local initialization. Compared to some existing distributed K-means schemes, e.g., [22], where users are required to run the max-consensus algorithm until convergence a total of $2k$ times, we provide a communication-efficient algorithm, that only communicates for a fixed number of rounds, while performing center inference locally, trading communication for computation.

We test the impact of our new initialization scheme, by running DGC-KM for $T = 1000$ iterations, using random initialization and DKM+C with $L = \{0, 1, 2, 3, 4\}$, on both homogeneous and heterogeneous data. We then report the final accuracy of our method, averaged across 5 runs. We again use $\rho = 10$ for homogeneous and $\rho = 100$ for heterogeneous data. The results are presented in Figure 3. The x axis represents the number of communication rounds, while the y-axis represents the final accuracy obtained by DGC-KM. We can see that increasing the number of communication rounds benefits the initialization in both homogeneous ($L = 3$) and heterogeneous ($L = 1$) setup. More importantly, the initialization scheme outperforms the random initalization in both setups, showing clear improvements.

VI. Conclusion

In this work we studied the effects of initialization on the performance of distributed clustering method originally proposed in [1]. We demonstrated through experiments on real data that the performance of DGC-$\mathcal{F}_\rho$ is more robust to initialization compared to centralized gradient clustering method from [2]. Next, we propose an initialization scheme, dubbed DKM+C, inspired by K-means++, which is shown to improve the performance of DGC-$\mathcal{F}_\rho$ compared to the baseline random initialization. Future work includes a rigorous theoretical analysis of the benefits of the proposed initialization scheme, as well as studying a version of the DGC-$\mathcal{F}_\rho$ method with a time-varying penalty ρ_t, such that $\rho_t \to \infty$, as $t \to \infty$.

References

[1] A. Armacki, D. Bajović, D. Jakovetić, and S. Kar, "A unified framework for gradient-based clustering of distributed data," *arXiv preprint arXiv:2402.01302*, 2024.

[2] A. Armacki, D. Bajovic, D. Jakovetic, and S. Kar, "Gradient based clustering," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 929–947, PMLR, 17–23 Jul 2022.

[3] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, (New Orleans, Louisiana), p. 1027–1035, SIAM, 2007.

[4] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[5] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[6] P. Awasthi and M.-F. Balcan, "Foundations for center-based clustering: worst-case approximations and modern developments," in *Handbook of cluster analysis* (C. Henning, M. Meila, F. Murtagh, and R. Rocci, eds.), pp. 67–100, Chapman and Hall/CRC, 1st ed., 2016.

[7] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[8] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *Journal of Machine Learning Research*, vol. 6, no. 58, pp. 1705–1749, 2005.

[9] A. K. Pediredla and C. S. Seelamantula, "A Huber-loss-driven clustering technique and its application to robust cell detection in confocal microscopy images," in *2011 7th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 501–506, 2011.

[10] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (A. Singh and J. Zhu, eds.), vol. 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, PMLR, 20–22 Apr 2017.

[11] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, H. Eichner, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[12] A. Armacki, D. Bajović, D. Jakovetić, and S. Kar, "A one-shot framework for distributed clustered learning in heterogeneous environments," *IEEE Transactions on Signal Processing*, vol. 72, pp. 636–651, 2024.

[13] A. Armacki, D. Bajovic, D. Jakovetic, and S. Kar, "Personalized federated learning via convex clustering," in *2022 IEEE International Smart Cities Conference (ISC2)*, pp. 1–7, 2022.

[14] S. Vlaski, S. Kar, A. H. Sayed, and J. M. Moura, "Networked signal and information processing: Learning by multiagent systems," *IEEE Signal Processing Magazine*, vol. 40, no. 5, pp. 92–105, 2023.

[15] A. H. Sayed, "Adaptive networks," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 460–497, 2014.

[16] D. Jakovetić, J. Xavier, and J. M. F. Moura, "Fast distributed gradient methods," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1131–1146, 2014.

[17] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[18] S. Datta, C. Giannella, and H. Kargupta, "Approximate distributed k-means clustering over a peer-to-peer network," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 10, pp. 1372–1388, 2009.

[19] P. A. Forero, A. Cano, and G. B. Giannakis, "Distributed clustering using wireless sensor networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 707–724, 2011.

[20] M.-F. F. Balcan, S. Ehrlich, and Y. Liang, "Distributed k-means and k-median clustering on general topologies," in *Advances in Neural Information Processing Systems* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds.), vol. 26, Curran Associates, Inc., 2013.

[21] G. Oliva, R. Setola, and C. N. Hadjicostis, "Distributed k-means algorithm," *arXiv preprint arXiv:1312.4176*, 2013.

[22] J. Qin, W. Fu, H. Gao, and W. X. Zheng, "Distributed k -means algorithm and fuzzy c -means algorithm for sensor networks based on multiagent consensus theory," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 772–783, 2017.

[23] S. Kar and B. Swenson, "Clustering with distributed data," *arXiv preprint arXiv:1901.00214*, 2019.

[24] H. Yu, H. Chen, S. Zhao, and Q. Shi, "Distributed soft clustering algorithm for iot based on finite time average consensus," *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 16096–16107, 2021.

[25] S. Har-Peled and S. Mazumdar, "On coresets for k-means and k-median clustering," in *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '04, (New York, NY, USA), p. 291–300, Association for Computing Machinery, 2004.

[26] S. Z. Selim and M. A. Ismail, "K-means-type algorithms: A generalized convergence theorem and characterization of local optimality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 1, pp. 81–87, 1984.

[27] A. Vattani, "The hardness of k-means clustering in the plane." https://cseweb.ucsd.edu/ avattani/papers/kmeans_hardness.pdf, 2009.

[28] P. Awasthi, M. Charikar, R. Krishnaswamy, and A. K. Sinop, "The hardness of approximation of euclidean k-means," *arXiv preprint arXiv:1502.03316*, 2015.

[29] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *5-th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, University of California Press, 1967.

[30] S. Kar, J. M. F. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3575–3605, 2012.

[31] S. Kar and J. M. Moura, "Consensus + innovations distributed inference over networks: cooperation and sensing in networked systems," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 99–109, 2013.

[32] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.