A. Proofs

A.1. Proving Theorem 3.1 (Main Theorem)

Before we proceed to proving Theorem 3.1, we first establish a few preliminary results. Let $\mathcal{T} = (T_1, \dots, T_k)$ be our target clustering and let $T_r^{(z)}$ be the subset of points of a cluster T_r on device z. For any point, $A_i^{(z)}$ on device z, let $c(A_i^{(z)})$ denote the index of the cluster it belongs to. That is,

$$A_i^{(z)} \in T^{(z)}_{c(A_i^{(z)})} \subseteq T_{c(A_i^{(z)})}.$$

Also recall the definition of matrix C, the matrix of means. Here the i-th row of C contains the mean of the cluster which contains data points A_i, i.e. $C_i = \mu(T_{c(A_i)})$. Our first lemma bounds how far the 'local' cluster mean $\mu(T_r^{(z)})$ can deviate from $\mu(T_r)$.

Lemma 2 (Lemma 5.2 in Kumar & Kannan (2010))**.** *Let $T_r^{(z)}$ be a subset of T_r on device z. Let $\mu(T_r^{(z)})$ denote the mean of the points indexed by $T_r^{(z)}$. Then,*

$$\|\mu(T_r^{(z)}) - \mu(T_r)\|_2 \leqslant \frac{\|A - C\|}{\sqrt{|T_r^{(z)}|}}.$$

Proof. Let $A^{(z)}$ be the sub-matrix of A on device z and let $\tilde{C}^{(z)}$ be the corresponding sub-matrix of our matrix of means C. Let u be an indicator vector for points in $T_r^{(z)}$. Observe that,

$$\begin{aligned} \left\| \, |T_r^{(z)}|(\mu(T_r^{(z)}) - \mu_r) \, \right\|_2 &= \|(A^{(z)} - \tilde{C}^{(z)}) \cdot u\|_2 \\ &\leqslant \|A^{(z)} - \tilde{C}^{(z)}\| \|u\|_2 \\ &\leqslant \|A - C\| \sqrt{\left|T_r^{(z)}\right|}. \end{aligned}$$

Here, for the last inequality, we note that $(A^{(z)} - \tilde{C}^{(z)})$ contains a subset of rows of $(A - C)$, and therefore $\|A^{(z)} - \tilde{C}^{(z)}\| \leqslant \|A - C\|$. □

Now consider the local clustering problem on each device z. The device has a data matrix $A^{(z)}$, whose rows are a subset of A. Let $T_1^{(z)}, T_2^{(z)}, \dots, T_k^{(z)}$ be subsets of $T_1, T_2, \dots, T_k$ on this device, such that no more than k' of them are non-empty. Construct a matrix $C^{(z)}$, of the same dimensions as $A^{(z)}$ where for each row of $A^{(z)}$, the corresponding row of $C^{(z)}$ contains the mean of the local cluster the point belongs to. That is, the i-th row of $C^{(z)}$ contains $\mu(T^{(z)}_{c(A_i^{(z)})})$. Using this next lemma, we bound the operator norm of the matrix $(A^{(z)} - C^{(z)})$, in terms of $(A - C)$.

Lemma 3. *Let $T_1^{(z)}, T_2^{(z)}, \dots T_k^{(z)}$ be subsets of target cluster that reside on a device such that k' of them are non-empty. Let $A^{(z)}$ be the corresponding $n^{(z)} \times d$ data matrix. Let $C^{(z)}$ be the corresponding matrix of means; that is each row $C_i^{(z)} = \mu(T^z_{c(A_i^z)})$. Then,*

$$\|A^{(z)} - C^{(z)}\| \leqslant 2\sqrt{k'}\|A - C\| \, .$$

Proof. Let $\tilde{C}^{(z)}$ be an $n^{(z)} \times d$ matrix where $\tilde{C}_i^{(z)} = \mu(T_{c(A_i^{(z)})})$. First, consider a unit vector u along the top singular direction and observe that:

$$\begin{aligned} \|\tilde{C}^{(z)} - C^{(z)}\|^2 &= \sum_{r=1}^{k} |T_r^{(z)}| \Big(\big(\mu(T_r^{(z)}) - \mu(T_r)\big) \cdot u \Big)^2 \\ &\leqslant \sum_{r=1}^{k} |T_r^{(z)}| \Big\| \mu(T_r^{(z)}) - \mu(T_r) \Big\|_2^2 \\ &\leqslant_{(a)} k' \|A - C\|^2. \end{aligned}$$

Here for inequality (a) we invoke Lemma 2. Also, noting that $\|A^{(z)} - \tilde{C}^{(z)}\| \leqslant \|A - C\|$, we get,

$$\begin{aligned}\|A^{(z)} - C^{(z)}\| &\leqslant \|A^{(z)} - \tilde{C}^{(z)}\| + \|\tilde{C}^{(z)} - C^{(z)}\| \\ &\leqslant (1 + \sqrt{k'})\|A - C\| \leqslant 2\sqrt{k'}\|A - C\|.\end{aligned}$$

□

We prove Theorem 3.1 in four parts:

1. In the first step we show that satisfying the active separation condition is sufficient to satisfy the Awasthi-Sheffet separation condition required for Lemma 1 (Lemma 4).

2. Next we use Lemma 4 to show that the first step of k-FED (Algorithm-1) will find local centers $\theta_r^{(z)}$ that are close to true centers $\mu(T_r^{(z)})$ on device z. We state and prove this in Lemma 5.

3. In next step, we show that the process of picking k initial centers in steps 2-6 of k-FED picks exactly one local cluster center $\theta_r^{(z)}$ for each cluster r. That is, we pick k local centers one corresponding to each target cluster. (Lemma 6)

4. Finally, we argue that with this initialization, the clustering of local cluster centers produced $(\tau_1, \dots, \tau_k)$ has the property that, all local cluster centers corresponding the to the same cluster (say T_r) will be in the same set (say τ_r). Moreover, no local cluster center corresponding to any T_s, $s \neq r$ will be in τ_r. As we argue later, this is sufficient for the induced clustering produced by $(\tau_1, \dots, \tau_k)$ to agree with our target clustering $\mathcal{T} = (T_1, T_2, \dots)$ up to permutation of labels and missclassifications incurred at the local clustering stage.

Lemma 4. *Let (T_r, T_s) be cluster pairs such that, $\|\mu_r - \mu_s\|_2 \geqslant 2c\sqrt{m_0}(\Delta_r + \Delta_s)$. Let $T_r^z \subseteq T_r$ and $T_s^z \subseteq T_s$ be large subsets on device z. Then,*

$$\|\mu_r^{(z)} - \mu_s^{(z)}\|_2 \geqslant c\sqrt{k'}\left(\frac{\|A^{(z)} - C^{(z)}\|}{\sqrt{n_r^{(z)}}} + \frac{\|A^{(z)} - C^{(z)}\|}{\sqrt{n_s^{(z)}}}\right).$$

Proof. (Lemma 4) Using the triangle inequality, we have

$$\begin{aligned}\|\mu_r^{(z)} - \mu_s^{(z)}\|_2 &\geqslant \|\mu_r - \mu_s\|_2 - \|\mu_r^{(z)} - \mu_r\|_2 - \|\mu_s - \mu_s^{(z)}\|_2 \\ &\geqslant 2c\sqrt{m_0}(\Delta_r + \Delta_s) - \frac{\|A - C\|}{\sqrt{n_r^{(z)}}} - \frac{\|A - C\|}{\sqrt{n_s^{(z)}}}\end{aligned} \tag{5}$$

using the active separation assumption. Now, expanding the terms can write the left hand side as

$$\begin{aligned}\|\mu_r^{(z)} - \mu_s^{(z)}\|_2 &\geqslant 2c\sqrt{m_0}\left(k'\frac{\|A - C\|}{\sqrt{n_r}} + k'\frac{\|A - C\|}{\sqrt{n_s}}\right) - \frac{\|A - C\|}{\sqrt{n_r^{(z)}}} - \frac{\|A - C\|}{\sqrt{n_s^{(z)}}} \\ &\geqslant \underbrace{\left(2\sqrt{\frac{m_0 n_r^{(z)}}{n_r}} - \frac{1}{ck'}\right)ck'\frac{\|A - C\|}{\sqrt{n_r^{(z)}}}}_{(i)} + \underbrace{\left(2\sqrt{\frac{m_0 n_s^{(z)}}{n_s}} - \frac{1}{ck'}\right)ck'\frac{\|A - C\|}{\sqrt{n_s^{(z)}}}}_{(ii)}.\end{aligned}$$

We first only consider the term (i). According to Lemma 3, $\|A - C\| \geqslant \frac{1}{2\sqrt{k'}}\|A^{(z)} - C^{(z)}\|$. Using this we can bound (i) as

$$\left(2\sqrt{\frac{m_0 n_r^{(z)}}{n_r}} - \frac{1}{ck'}\right)ck'\frac{\|A - C\|}{\sqrt{n_r^{(z)}}} \geqslant \left(2\sqrt{\frac{m_0 n_r^{(z)}}{n_r}} - \frac{1}{ck'}\right)c\sqrt{k'}\frac{\|A^{(z)} - C^{(z)}\|}{\sqrt{n_r^{(z)}}}.$$

Now recall that for large cluster subsets $n_r^{(z)} \geqslant \frac{1}{m_0} n_r$ and thus $2\sqrt{\frac{m_0 n_r^{(z)}}{n_r}} - \frac{1}{ck'} \geqslant 2 - \frac{1}{ck'} \geqslant 1$. This means that we can bound term (i) as,

$$\left(2\sqrt{\frac{m_0 n_r^{(z)}}{n_r}} - \frac{1}{ck'}\right) ck' \frac{\|A - C\|}{\sqrt{n_r^{(z)}}} \geqslant c\sqrt{k'} \frac{\|A^{(z)} - C^{(z)}\|}{\sqrt{n_r^{(z)}}} .$$

We get a symmetric expression for term (ii) as well. Using this in equation 5, we get the desired result:

$$\|\mu_r^{(z)} - \mu_s^{(z)}\|_2 \geqslant c\sqrt{k'}\left(\frac{\|A^{(z)} - C^{(z)}\|}{\sqrt{n_r^{(z)}}} + \frac{\|A^{(z)} - C^{(z)}\|}{\sqrt{n_s^{(z)}}}\right).$$

□

Since Algorithm-1 is run locally on each device, it is unaffected by the inactive separation condition, as by definition, subsets of only active cluster pairs exist on each device. This implies that Algorithm-1 solves the local clustering problem successfully. Specifically on device z containing data from some cluster T_r, θ_r^z is not too far from $\mu(T_r^{(z)})$. Showing this result is our second step and we state this formally in Lemma 5 below.

Lemma 5. *Let $(T_1^{(z)}, \dots, T_k^{(z)})$ be the subsets of $(T_1, \dots, T_k)$ on some device z such that no more than k' of them are non-empty. Moreover, assume all non-empty subsets are large, i.e. $|T_r^{(z)}| \geqslant \frac{1}{m_0}|T_r|$. Finally, assume that the active separation requirement is satisfied for all active cluster pairs on z. Then, on termination of Algorithm-1, for each non-empty $T_r^{(z)}$, we have*

$$\|\theta_r^{(z)} - \mu(T_r^{(z)})\|_2 \leqslant \frac{25}{c}\frac{\|A^{(z)} - C^{(z)}\|}{\sqrt{n_r^z}} \leqslant \frac{50\sqrt{k'}}{c}\frac{\|A - C\|}{\sqrt{n_r^z}} ,$$

and,

$$\|\theta_r^{(z)} - \mu(T_r)\|_2 \leqslant 2\sqrt{m_0 k'}\frac{\|A - C\|}{\sqrt{n_r}} \leqslant 2\sqrt{m_0}\lambda .$$

Proof. First note that the local clustering problem with data matrix $A^{(z)}$ and matrix of centers $C^{(z)}$ satisfies the requirements of Lemma 1. Thus it follows that,

$$\|\theta_r^{(z)} - \mu(T_r^{(z)})\|_2 \leqslant \frac{25}{c}\frac{\|A^{(z)} - C^{(z)}\|}{\sqrt{n_r^z}} .$$

Now applying Lemma 3 gives us the first statement.

To prove the second statement, we start off with the triangle inequality:

$$\begin{aligned}\|\theta_r^{(z)} - \mu(T_r)\|_2 &\leqslant \|\theta_r^{(z)} - \mu(T_r^{(z)})\|_2 + \|\mu(T_r^{(z)}) - \mu(T_r)\|_2 \\ &\leqslant \frac{25}{c}\frac{\|A^{(z)} - C^{(z)}\|}{\sqrt{n_r^z}} + \frac{\|A - C\|}{\sqrt{n_r^z}} .\end{aligned}$$

Here for the last inequality we used Lemma 2. Now applying Lemma 3 and taking take $c \geqslant 100$, we get

$$\begin{aligned}\|\theta_r^{(z)} - \mu(T_r)\|_2 &\leqslant \frac{50\sqrt{k'}}{c}\frac{\|A - C\|}{\sqrt{n_r^z}} + \frac{\|A - C\|}{\sqrt{n_r^z}} \\ &\leqslant \left(\frac{50}{c} + \frac{1}{\sqrt{k'}}\right)\sqrt{k'}\frac{\|A - C\|}{\sqrt{n_r^z}} \\ &\leqslant 2\sqrt{k'}\frac{\|A - C\|}{\sqrt{n_r^z}} \leqslant 2\sqrt{m_0 k'}\frac{\|A - C\|}{\sqrt{n_r}} \\ &\leqslant 2\sqrt{m_0}\lambda .\end{aligned}$$

□

This means that for a fixed r, all the $\theta_r^{(z)}$ received at the central server from devices $z \in [Z]$ are 'close' to $\mu(T_r)$.

The next step is to show that in the k initial centers k-FED picks in steps 2-6, there is exactly one corresponding to each target cluster T_i. We will show later that this is sufficient for the final step of the algorithm to correctly assign local cluster centers to the correct partition.

Lemma 6. *Let $\mathcal{T} = (T_1, \ldots, T_k)$ be our target clustering. Assume all active cluster pairs and inactive cluster pairs satisfy their separation requirements. Further let $n_{\min} \geqslant \frac{4}{c^2 k'} n_{\max}$. Then at the end of step 6 of k-*FED*, for every target cluster T_r, there exists an $\theta_s^{(z)} \in M$ such that $\theta_s^{(z)} = \mu(T_s^{(z)})$ for some $z \in [Z]$.*

Before we proceed to proving this lemma, we state and prove a lower bound on how close a local cluster center $\theta_r^{(z)}$ can be to some cluster mean $\mu(T_s)$ for $s \neq r$:

Lemma 7. *Let $\theta_r^{(z)} := \mu(T_r^{(z)})$. The for any $s \neq r$, $z' \in [Z]$,*

$$\|\theta_r^{(z)} - \theta_s^{(z')}\|_2 \geqslant 6\sqrt{m_0}\lambda\,.$$

Proof. First, from the triangle inequality note that,

$$\|\theta_r^{(z)} - \theta_s^{(z')}\|_2 \geqslant \|\mu_r - \mu_s\|_2 - \|\mu_r - \theta_r^{(z)}\|_2 - \|\mu_s - \theta_s^{(z')}\|_2\,.$$

Using Lemma 5 and our inactive separation assumption we bound the right hand side further as,

$$\begin{aligned}\|\mu_r - \mu_s\|_2 - \|\mu_r - \theta_r^{(z)}\|_2 - \|\mu_s - \theta_s^{(z')}\|_2 &\geqslant 10\sqrt{m_0 k'}\frac{\|A - C\|}{\sqrt{n_{\min}}} - 4\sqrt{m_0 k'}\frac{\|A - C\|}{\sqrt{n_r}}\\ &\geqslant 6\sqrt{m_0 k'}\frac{\|A - C\|}{\sqrt{n_{\min}}} \geqslant 6\sqrt{m_0}\lambda,\end{aligned}$$

as desired. □

Proof. (Lemma 6) Let M_t denote the set M in step 2-6 of k-FED, after picking the first t points $(1 \leqslant t \leqslant k)$. Let us denote the point k-FED selects in iteration t as θ_t. That is,

$$\theta_t \leftarrow \operatorname*{arg\,max}_{z \in [Z], i \in [k]} d_{M_{t-1}}(\theta_i^{(z)})\,.$$

We will show that the set M_t contains t points corresponding to t different target clusters at every iteration t. This invariant holds trivially at $t = 1$. Assume the statement first became false at some $1 < t' \leqslant k$. Let the point $\theta_{t'}$ correspond to a local cluster mean from cluster T_r. Then there must exist some $1 \leqslant t'' < t'$ such that $\theta_{t''}$ also correspond to a local cluster mean from T_r. Further, there must exist some cluster $s \neq r$ such that $\theta_s^{(z)} \notin M_{t'}$ for any $z \in [Z]$.

Now by definition of $d_{M_{t-1}}(\theta_{t'})$, we have

$$\begin{aligned}d_{M_{t-1}}(\theta_{t'}) &= \min_{\theta \in M_{t-1}} \|\theta_{t'} - \theta\|_2\\ &\leqslant \|\theta_{t'} - \theta_{t''}\|_2\\ &\leqslant_{(a)} \|\theta_{t'} - \mu(T_r)\|_2 + \|\mu(T_r) - \theta_{t''}\|_2\\ &\leqslant_{(b)} 4\sqrt{m_0 k'}\frac{\|A - C\|}{\sqrt{n_r}} \leqslant 4\sqrt{m_0}\lambda\,. \end{aligned} \tag{6}$$

Here inequality (a) follows from the triangle inequality and (b) follows from Lemma 5.

Now consider $\theta_s^{(z)}$ for any z. Since no other local cluster center from T_s is contained in M_t, from Lemma 7 we conclude that for every $\theta \in M_{t-1}$,

$$\|\theta_s^{(z)} - \theta\|_2 \geqslant 6\sqrt{m_0}\lambda\,.$$

But this means that $d_{M_{t-1}}(\theta_{t'}) \leqslant 4\sqrt{m_0}\lambda \leqslant 6\sqrt{m_0}\lambda \leqslant d_{M_{t-1}}(\theta_s^{(z)})$ leading to a contradiction based on the definition of $\theta_{t'}$. This completes our argument. □

Now we are ready to prove our main Theorem 3.1.

Proof. From Lemma 6, we know that the set M at the end of step 6 of k-FED contains exactly one center corresponding to each target clustering. Let the local cluster center $\tilde{\theta}_r \in M$ correspond to the cluster T_r. Observe that for any $z \in [Z]$,

$$\begin{aligned}\|\theta_r^z - \tilde{\theta}_r\|_2 &\leqslant \|\theta_r^z - \mu_r\|_2 + \|\mu_r - \tilde{\theta}_r\|_2 \\ &\leqslant 4\sqrt{m_0}\lambda,\end{aligned}$$

using Lemma 7. Further, for any $s \neq r$,

$$\|\theta_s^z - \tilde{\theta}_r\|_2 \geqslant 6\sqrt{m_0}\lambda.$$

This means that for every r and $z \in [Z]$, θ_r^z is closer to the corresponding initial center $\tilde{\theta}_r$ than to any other initial center $\tilde{\theta}_s$, $s \neq r$. Let τ_r be the set of local cluster centers assigned to $\tilde{\theta}_r$. Then it can be seen that τ_r only contains local cluster centers $\theta_r^{(z)}$ for all devices z, i.e. τ_r contains all the device cluster centers corresponding to target cluster T_r.

Now consider the definition of k-FED induced clustering (Definition 3.3), where we define

$$T_r' = \{i : A_i^{(z)} \in U_s^{(z)} \text{ and } \theta_s^{(z)} \in \tau_r\}.$$

In this case, we know that only local cluster centers corresponding to cluster T_r is contained in τ_r. Thus our induced cluster T_r' becomes,

$$T_r' = \{i : A_i^{(z)} \in U_r^{(z)}\}.$$

Now from Lemma 1 we know that on each device the sets $(U_1^{(z)}, \ldots, U_{k'}^{(z)})$ and $(T_1^{(z)}, \ldots, T_{k'}^{(z)})$ only differ on at most $O(\frac{1}{c^2})n^{(z)}$. Summing this error over all devices z, we see that our induced clustering $(T_1', \ldots, T_k')$ and the target clustering $(T_1, \ldots, T_k)$ differ only on $O(\frac{1}{c^2})n$ points. Finally, if all the local points satisfy their respective proximity condition (Definition 3.1), then no points are missclassified. This concludes our proof. □

A.2. Running Time of *k*-FED and Handling New Devices

We now analyze the running time of k-FED steps 2-8. Since step 1 is running Algorithm-1 on individual devices, we do not include the running time of this step as part of our analysis. Note that with the separation assumptions in place, Algorithm-1 will converge in polynomial time. However, as observed in practise, Lloyd like methods typically only take a few iterations to terminate.

Theorem A.1. *Steps 2-8 of* k-FED *takes* $O(Zk' \cdot k^2)$ *pairwise distance computations to terminate. Further, after the set* M *in step 6 has been computed, new local cluster centers* $\Theta^{(z)}$ *from a yet unseen device* z *can be correctly assigned in* $O(k' \cdot k)$ *distance computations.*

Proof. (Theorem 3.2) The proof of the first part follows from a simple step by step analysis. Step 1 can be performed in $O(1)$. Step 2-6 executes exactly k times. At each iteration t, $(1 \leqslant t \leqslant k)$, we compute the distance of all device cluster centers, of which there are most Zk', to the points in M_{t-1}. Thus at iteration t, this can be implemented with $Zk' \cdot t$ distance computations. Summing over all t, we see that steps 2-6 can run in $O(Zk' \cdot k^2)$ distance computations. Finally, step 7 requires us to assign all the Zk' device cluster centers to one of the k initial points in M. This can be implemented in $O(Zk' \cdot k)$ distance computations. Thus the overall complexity in terms of pairwise distance computations is $O(Zk' \cdot k^2)$.

The second part of the statement follows from noting that for each $\theta_r^{(z)} \in \Theta^{(z)}$, the nearest point in set M must be the initial center $\tilde{\theta}_r$ we picked as was demonstrated in the proof of Theorem 3.1. Thus every $\theta_r^{(z)} \in \Theta^{(z)}$ is assigned to the correct partition τ_r as required. □

A.3. Separating Data from Mixture of Gaussian

We now prove Theorem 4.1. Recall that we are working in the setting where $k' \leqslant \sqrt{k}$. Our proof builds on results from Lemma 6.3, Kumar & Kannan (2010).

Proof. First consider an active cluster pair r, s. Based on our separation requirement, we have:

$$\begin{aligned}\|\mu_r - \mu_s\|_2 &\geqslant \frac{2c\sqrt{km_0}\sigma_{\max}}{\sqrt{w_{\min}}}\text{polylog}\left(\frac{d}{w_{\min}}\right) \\ &\geqslant 2c\sqrt{km_0}\frac{\sigma_{\max}\sqrt{n}}{\sqrt{w_{\min}n}}\text{polylog}\left(\frac{d}{w_{\min}}\right) .\end{aligned}$$

We further simplify the right hand to get,

$$\|\mu_r - \mu_s\|_2 \geqslant c\sqrt{km_0}\sigma_{\max}\sqrt{n}(\frac{1}{\sqrt{w_r n}} + \frac{1}{\sqrt{w_s n}})\text{polylog}\left(\frac{d}{w_{\min}}\right) .$$

Now note the number of points from each component F_r is very close to $w_r n_r$ with very high probability. Here w_r is the mixing weight of component r and n_r is the number of data points. Using this, with high probability we have

$$\|\mu_r - \mu_s\|_2 \geqslant c\sqrt{km_0}\sigma_{\max}\sqrt{n}(\frac{1}{\sqrt{n_r}} + \frac{1}{\sqrt{n_s}})\text{polylog}\left(\frac{d}{w_{\min}}\right) .$$

Further, it can be shown that $\|A - C\|$ is $O\left(\sigma_{\max}\sqrt{n} \cdot \text{polylog}\left(\frac{d}{w_{\min}}\right)\right)$ with high probability (see (Dasgupta et al., 2007)). Thus we conclude that, with high probability

$$\|\mu_r - \mu_s\|_2 \geqslant c\sqrt{km_0}\left(\frac{\|A - C\|}{\sqrt{n_r}} + \frac{\|A - C\|}{\sqrt{n_s}}\right).$$

Thus the active separation requirement is satisfied. The proof for the inactive separation condition is similar. Finally, the proximity condition follows from the concentration properties of Gaussians. □

B. Experimental Details

B.1. Datasets

For all experiments involving real data, we use the EMNIST, FEMNIST, and Shakespeare datasets. These datasets and their corresponding models are available at the LEAF benchmark: `https://leaf.cmu.edu/`. For client selection experiments, we manually partition a subset of FEMNIST (first 10 classes) by assigning 2 classes to each device. There are 500 devices in total. Both the number of training samples across all devices and the number of training samples per class within each device follow a power law. We use the natural partition of Shakespeare where each device corresponds to a speaking role in the plays of William Shakespeare. We randomly sample 109 users from the entire dataset. For personalization experiments, following Ghosh et al. (2020), we use a CNN-based model with one hidden layer and 200 hidden units trained with a learning rate of 0.01 and 10 local updates on each device.

B.2. Choosing k Based on Separation

As mentioned in Section 4.2, to create our *oracle clustering*, we compute the quantity $c_{rs} = \frac{\|\mu_r - \mu_s\|}{2\sqrt{m_0}(\Delta_r + \Delta_s)}$ for each cluster pairs (r, s), for every candidate value of k we are considering. We construct a distribution plot of these c_{rs}. An example of such a plot for the MNIST dataset is provided in Figure 5. As can be seen here, for all values of k, the relative separation is quite small. Thus even for this oracle clustering, the actual separation between cluster means is small. To pick a k for our oracle clustering, we pick a fixed value c_0 (say 0.5) and then pick the value of k which leads to maximum fraction of cluster pairs (r, s) to have $c_{rs} > c_0$.

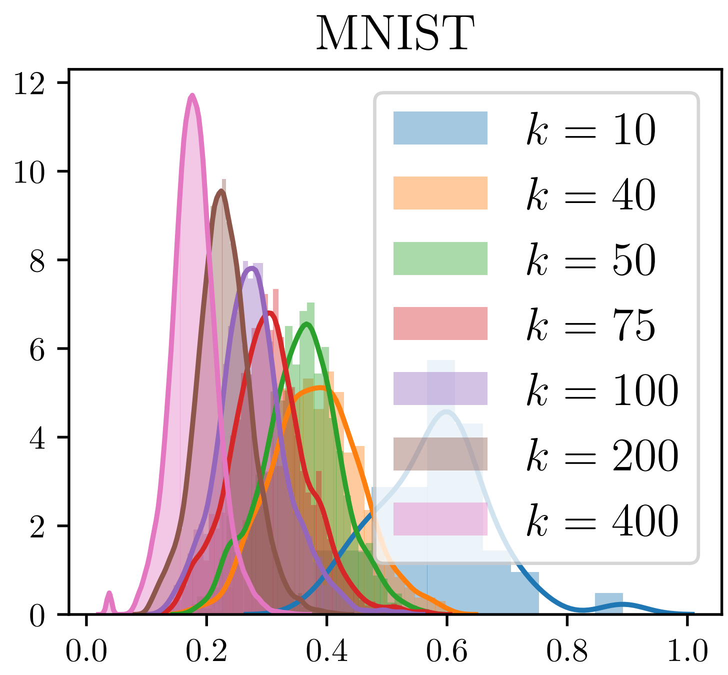

Figure 5. Distribution plot of c_{rs}, for various values of k on the MNIST dataset. As can be seen, $c_{rs} < 1$ for most cluster pairs, indicating that the separation between them is relatively small.