Appendix for "A Convergent Federated Clustering Algorithm without Initial Condition"

A Computation and Communication Complexity of SR-FCA

Note that the complexity of the REFINE step is the same as that of IFCA in terms of both computation time and communication since in each case, we need to find the loss of every cluster model on every client's data. The main blowup of $\mathcal{O}(m^2)$ is incurred during ONE_SHOT, which is unavoidable if an initial clustering is not known (eg. see KMeans Lloyd (1982) v/s DBSCAN Ester et al. (1996) or Ward's algorithm where without the initial clustering, we need to perform all pairwise comparisons to check which clients can be clustered together).

B The Trimmed Mean Algorithm

Algorithm 4: TrimmedMeanGD()

```
Input: $0 \leq \beta < \frac{1}{2}$, Clustering $\mathcal{C}_r$
Output: Cluster models $\{\omega_{c,T}\}_{c \in \mathrm{rg}(\mathcal{C}_r)}$
for all clusters $c \in \mathrm{rg}(\mathcal{C}_r)$ in parallel do
    $w_{c,0} \leftarrow w_0$
    for $t = 0$ to $T-1$ do
        $g(w_{c,t}) \leftarrow \mathrm{TrMean}_\beta(\{\nabla f_i(w_{c,t}), \mathcal{C}_r(i) = c\})$
        $w_{c,t+1} \leftarrow proj_{\mathcal{W}}\{w_{c,t} - \eta g_t\}$
    end for
    Return $\{\omega_{c,T}\}_{c \in \mathrm{rg}(\mathcal{C}_r)}$
end for
```

C Proof of proposition 4.1

According to the proposition, for two users i and j, the data is generated by first sampling each coordinate of $x \in \mathbb{R}^d$ from $\mathcal{N}(0,1)$ iid and then computing y as –

$$y_i = \langle x, w_i^\star \rangle + \epsilon_i$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$. Then, the distribution of $y_i|x$ is $\mathcal{N}(\langle x, w_i^\star \rangle, \sigma^2)$. Therefore, the KL divergence between $y_i|x$ and $y_j|x$ is given by

$$KL(p(y_i|x)||p(y_j|x)) = \frac{\langle w_i^\star - w_j^\star, x \rangle^2}{2\sigma^2}$$

Therefore, if we take expectation wrt x, we have

$$\mathbb{E}_x[KL(p(y_i|x)||p(y_j|x))] = \frac{d\|w_i^\star - w_j^\star\|^2}{2\sigma^2}$$

D Proof of lemma 4.6

In ONE_SHOT(), $\mathcal{C}_0 = \mathcal{C}^\star$, if all the edges formed in the graph are correct. This means that if i,j are in the same cluster in $\mathcal{C}^\star$, then $\|w_{i,T} - w_{j,T}\| \leq \lambda$ and if i,j are in different clusters, $\|w_{i,T} - w_{j,T}\| > \lambda$. Note that,

$$w_{i,T} - w_{j,T} = (w_i^\star - w_j^\star) + (w_{i,T} - w_i^\star) - (w_{j,T} - w_j^\star)$$

Now, if we apply triangle inequality, we obtain

$$\mathsf{dist}(w_{i,T}, w_{j,T}) \geq \mathsf{dist}(w_i^\star, w_j^\star) - \Xi_{i,j}, \quad \mathsf{dist}(w_{i,T}, w_{j,T}) \leq \mathsf{dist}(w_i^\star, w_j^\star) + \Xi_{i,j}$$

where $\Xi_{i,j} = \sum_{k=i,j} \mathsf{dist}(w_{k,T}, w_k^\star)$. This decomposition forms the key motivation for our algorithm.

Therefore, if i,j are in the same cluster, then a sufficient condition for edge (i,j) to be incorrect is

$$\lambda \leq \mathsf{dist}(w_i^\star, w_j^\star) + \Xi_{i,j} \implies \Xi_{i,j} \geq \lambda - \epsilon_1$$

Similarly, if i,j are in different clusters, then a sufficient condition for edge (i,j) to be incorrect is

$$\lambda \geq \mathsf{dist}(w_i^\star, w_j^\star) - \Xi_{i,j} \implies \Xi_{i,j} \geq \epsilon_2 - \lambda$$

Therefore, we can set $\Delta_\lambda = \min\{\epsilon_2 - \lambda, \lambda - \epsilon_1\}$, and then a sufficient condition for any edge to be incorrect is $\max_{i,j} \Xi_{i,j} \geq \Delta_\lambda$.
Thus,

$$\begin{aligned} \Pr[\mathcal{C}^\star \neq \mathcal{C}_0] &\leq \Pr[\text{at least 1 edge is incorrect}] \\ &\leq \Pr[\max_{i,j} \Xi_{i,j} \geq \Delta_\lambda] \\ &\leq \Pr[\max_{i,j} \sum_{k=i,j} \|w_{k,T} - w_k^\star\| \geq \Delta_\lambda] \\ &\leq \Pr[\max_{i,j} \max_{k=i,j} (\|w_{k,T} - w_k^\star\| \geq \frac{\Delta_\lambda}{2}] \\ &\leq \Pr[\max_{i \in [m]} \|w_{i,T} - w_i^\star\| \geq \frac{\Delta_\lambda}{2}] \end{aligned} \tag{3}$$

The second and third inequalities are obtained by expanding the terms. The fourth inequality is obtained by $\Pr[a+b \geq c] \leq \Pr[\max\{a,b\} \geq c/2]$. For the fifth inequality, we merge $\max_{i,j} \max_{k=i,j}$ into $\max_{i \in [m]}$. As we can see in Equation (3), we need to bound $\|w_{i,T} - w_i^\star\|$ for each node i. The subsequent Lemma allow us to bound this quantities.

Lemma D.1 (Convergence of $w_{i,T}$)**.** *Let $\frac{n^{2/3}\Delta^{4/3}}{D^{2/3}\hat{L}^{2/3}} \lesssim b_1 d$, for some constant $b_1 > 0$. Then, after running* `ONE_SHOT()` *with $\eta \leq \frac{1}{L}$, for some constant $b_2 > 0$, under assumption 4.3 ,assumption 4.4 and assumption 4.5, we have*

$$\Pr[\|w_{i,T} - w_i^\star\| \geq \frac{\epsilon_2 - \epsilon_1}{4}] \leq d \ \exp(-n \frac{b_2 \Delta}{\hat{L}\sqrt{d}}),$$

where $\Delta = \frac{\mu}{2}(\frac{\Delta_\lambda}{2} - (1 - \frac{\mu}{L})^{T/2} D)$ and $n = \min_{i \in [m]} n_i$.

This lemma follows from Yin et al. (2018). The complete proof of this Lemma is present in appendix D.1.
Now, we can apply lemma D.1 in Eq (3).

$$\begin{aligned} \Pr[\mathcal{C}_0 \neq \mathcal{C}^\star] &\leq \Pr[\max_{i \in [m]} \|w_{i,T} - w_i^\star\| \geq \frac{\Delta_\lambda}{2}] \\ &\leq m \max_{i \in [m]} \Pr[\|w_{i,T} - w_i^\star\| \geq \frac{\Delta_\lambda}{2}] \\ &\leq md \ \exp(-n \frac{b_2 \Delta}{\hat{L}\sqrt{d}}) \end{aligned}$$

For the second inequality, we use $\Pr[\max_{i \in [m]} a_i \geq c] \leq \sum_{i \in [m]} \Pr[a_i \geq c] \leq m \max_{i \in [m]} \Pr[a_i \geq c]$, which follows from union bound.
Note that for $p < 1$, we need the separation to be order of $\Theta(\sqrt{\frac{\log m}{n}})$.

D.1 PROOF OF LEMMA D.1

We utilize results from Yin et al. (2018), which hold for `TrimmedMeanGD` to analyze convergence for a single node as they yield stronger guarantees under the given assumptions.

Lemma D.2 (Convergence of $w_{i,T}$)**.** *If assumption 4.3,assumption 4.4,and assumption 4.5 hold, and $\eta \leq \frac{1}{L}$, then*

$$\|w_{i,T} - w_i^\star\| \leq (1 - \kappa^{-1})^{T/2} D + \frac{2}{\mu}\Lambda_i \quad \forall i \in [m] \tag{4}$$

where $\kappa = \frac{L}{\mu}$ and Λ_i is a positive random variable with

$$\Pr[\Lambda_i \geq \sqrt{2d}r + 2\sqrt{2}\delta\hat{L}] \leq 2d(1 + \frac{D}{\delta})^d \exp(-n \min\{\frac{r}{2\hat{L}}, \frac{r^2}{2\hat{L}^2}\}) \tag{5}$$

for some $r, \delta > 0$.

We provide the proof of this lemma in appendix E.8.
Using the above Lemma, we can bound the probability $\Pr[\|w_{i,T} - w_i^\star\| \geq \frac{\Delta_\lambda}{2}]$

$$\begin{aligned}
\Pr[\|w_{i,T} - w_i^\star\| \geq \frac{\Delta_\lambda}{2}] &\leq \Pr[2(1-\kappa^{-1})^{T/2}D + \frac{2}{\mu}\Lambda_i + \geq \frac{\Delta_\lambda}{2}] \\
&\leq \Pr[\Lambda_i \geq \Delta], \quad \text{where } \Delta = \frac{\mu}{2}(\frac{\Delta_\lambda}{2} - (1-\kappa^{-1})^{T/2}D) \\
&\leq \Pr[\sqrt{2d}r + 2\sqrt{2}\delta\hat{L} \geq \Delta] \\
&\leq d\exp(-nb_2\frac{\Delta}{\hat{L}\sqrt{d}})
\end{aligned}$$

for some constants $b_1, b_2, b_3, b_4 > 0$, where we set $r = b_3\hat{L}\max\{\frac{\Delta}{\hat{L}\sqrt{d}}, \sqrt{\frac{\Delta}{\hat{L}\sqrt{d}}}\}$ and $\delta = b_4\frac{\Delta}{\hat{L}}$, and for $b_1 d \leq \frac{n^{2/3}\Delta^{4/3}}{D^{2/3}\hat{L}^{4/3}}$, such that $\sqrt{2d}r + 2\sqrt{2}\delta\hat{L} \geq \Delta$ and $n\min\{\frac{r}{2\hat{L}}, \frac{r^2}{2\hat{L}^2}\} > \frac{Dd}{\delta}$ in lemma D.2.

E PROOF OF THEOREM 4.8

E.1 PRELIMINARIES

First, we define certain random variables and their respective probabilities which we will use throughout this proof. Since the edge based analysis and corresponding clique identification involves a lot of dependent events, we try to decompose the absence/presence of edge into a combination of independent events.
Define,

$$X_{ij} = \begin{cases} 1 & \text{If the edge } (i,j) \text{ in } \mathcal{C}_0 \text{ is incorrect in } \mathcal{C}^\star \\ 0 & \text{Otherwise} \end{cases} \tag{6}$$

An edge (i,j) in $\mathcal{C}_0$ is incorrect in $\mathcal{C}^\star$ if either it is present in $\mathcal{C}^\star$ and absent in $\mathcal{C}_0$ or vice versa. We analyze the probability of this event for the case when $\mathcal{C}^\star$ contains the edge (i,j). The case when $\mathcal{C}^\star$ doesn't contain edge (i,j) and it is present in $\mathcal{C}_0$ has exaclty same probability. When $\|w_i^\star - w_j^\star\| \leq \epsilon_1$, then edge is present is $\mathcal{C}^\star$. If it is absent in $\mathcal{C}_0$, then

$$\begin{aligned}
\Pr[X_{ij} = 1] &\leq \Pr[\Xi_{i,j} \geq \Delta_\lambda] \\
&\leq \Pr[\Lambda_i + \Lambda_j \geq 2\Delta]
\end{aligned}$$

The analysis is similar to the proof of `ONE_SHOT()` in appendix D.
Note that the random variables $\{X_{ij}\}$ are not independent. We now define independent random variables X_i such that

$$X_i = \begin{cases} 1 & \text{If } \Lambda_i \geq \Delta \\ 0 & \text{Otherwise} \end{cases} \tag{7}$$

Thus, we can see that $X_{ij} \leq X_i + X_j$. Additionally,

$$\Pr[X_i = 1] \leq \Pr[\Lambda_i \geq \Delta] \leq \frac{p}{m} \tag{8}$$

This follows from analysis of `ONE_SHOT()` in appendix D.
We can further generalize this notion to the random variables defined as $Y_{i,\gamma}$.

$$Y_{i,\gamma} = \begin{cases} 1 & \text{If } \Lambda_i \geq \gamma\Delta, \gamma \in (0,2) \\ 0 & \text{Otherwise} \end{cases} \tag{9}$$

Then,

$$\Pr[Y_{i,\gamma} = 1] \leq \Pr[\Lambda_i \geq \gamma\Delta] \leq d\exp(-nb_2\frac{\gamma\Delta}{\hat{L}\sqrt{d}}) = (\frac{p}{m})^\gamma$$

Note that the set of random variables $\{Y_{i,\gamma}\}_{i=1}^m$ are mutually independent random variables.

Further, we define the $\omega_c^\star$ for every cluster $c \in \mathrm{rg}(\mathcal{C}_0)$. Let $c' \in \mathcal{C}^\star$ be the cluster label of node c. If $G_c = \{i : i \in [m], \mathcal{C}^\star(i) = c'\}$, which is the set of nodes in c which were from c' in the original clustering, then we can define $\omega_c^\star$ and $F_c(w)$ as

$$\omega_c^\star = \underset{w \in \mathcal{W}}{\mathrm{argmin}} \mathbb{E}[\frac{1}{|G_{c'}|} \sum_{i \in G_{c'}} f_i(w)] \tag{10}$$

$$= \underset{w \in \mathcal{W}}{\mathrm{argmin}} \frac{1}{|G_{c'}|} \sum_{i \in G_{c'}} F_i(w) = \underset{w \in \mathcal{W}}{\mathrm{argmin}} F_c(w) \tag{11}$$

We use this definition of $\omega_c^\star$ in the appendices E.5 and E.6.

E.2 Analysis of `REFINE()`

Our goal is to compute total probability of error for `REFINE()` to fail. If we define this error as $\mathcal{C}_1 \neq \mathcal{C}^\star$, then we can define the main sources of error for this event.

1. **$\exists c \in \mathrm{rg}(\mathcal{C}^\star)$ such that no cluster in $\mathcal{C}_0$ has cluster label c :** If the a cluster $c \in \mathrm{rg}(\mathcal{C}^\star)$ is absent in $\mathcal{C}_0$, then subsequent steps of `REFINE()` will never be able to recover it, as they only involve node reclustering and merging existing clusters. The lemma presented below gives an upper bound on the probability of this event.

   Lemma E.1. *Under the conditions of lemma 4.6 and if $t = \Theta(c_{\min})$, then there exists constant $a_1 > 0$ such that*

   $$\Pr[\exists c \in \mathrm{rg}(\mathcal{C}^\star) \textit{ such that no cluster in } \mathcal{C}_0 \textit{ has cluster label } c]$$
   $$\leq \frac{m}{c_{\min}} \exp(-a_1 c_{\min})$$

   The proof of this Lemma is presented in appendix E.3

2. **Each cluster $c \in \mathrm{rg}(C)_0$ should have $< \alpha$ fraction of impurities for some $\frac{1}{2} > \beta > \alpha$:** If some cluster has more than α-fraction of impure nodes, then we cannot expect convergence guarantees for `TrimmedMeanGD`$_\beta$.

   The below lemma bounds the probability of this error as

   Lemma E.2. . *For some constants $0 < \alpha < \beta < \frac{1}{2}, a_2 \geq 0, \gamma_1 \in (1,2)$ and $\alpha t = \Theta(m)$, under the conditions in lemma 4.6, we have*

   $$\Pr[\exists c \in \mathrm{rg}(\mathcal{C}_0) \textit{ which has} > \alpha \textit{ fraction of impurities }]$$
   $$\leq \frac{m}{t} \exp(-a_2 m) + (1-\alpha) m (\frac{p}{m})^{\gamma_1}$$

   The proof of this Lemma is presented in appendix E.4.

3. **`MERGE()` error:** We define this as the error for the `MERGE()` to fail. Even though `MERGE()` operates after `RECLUSTER()`, `RECLUSTER()` does not change the cluster iterates. The goal of `MERGE()` is to ensure that all clusters in $\mathcal{C}_0$ with the same cluster labels are merged. Therefore, we define `MERGE()` error as the event when either two clusters with same cluster label are not merged or two clusters with different cluster labels are merged. The below lemma bounds this probability.

   Lemma E.3. *If $\min\{\frac{n^{2/3}\Delta^{4/3}}{D^{2/3}\hat{L}^{2/3}}, \frac{n^2 \Delta'^2}{\hat{L}^2 \log(c_{\min})}\} \geq u_1 d$ for some constants $u_1 > 0$, then for some constant $a_3' > 0$, where $\Delta' = \Delta - \frac{\mu B}{2} > 0$, where $B = \sqrt{\frac{2\hat{L}\epsilon_1}{\mu}}$, we have*

   $$\Pr[\textit{MERGE () Error}] \leq \frac{4dm}{t} \exp(-a_3' n \frac{\Delta'}{2\hat{L}})$$

   The proof of this Lemma is presented in appendix E.5.

4. **`RECLUSTER()` error:** This event is defined as a node going to the wrong cluster after both `MERGE()` and `REFINE()` operations. After `MERGE()`, each cluster in $\mathcal{C}_0$ corresponds to a single cluster in $\mathcal{C}_1$. Therefore, we incur an error due to the `RECLUSTER()` operation if any node i does not go to the cluster $c \in \mathcal{C}_1$ which has cluster label $\mathcal{C}^\star(i)$. The below lemma provides an upper bound on the probability of this error.

Lemma E.4. *If* $\min\{\frac{n^{2/3}\Delta^{4/3}}{D^{2/3}\hat{L}^{2/3}},\frac{n^2\Delta'^2}{\hat{L}^2\log(c_{\min})}\}\geq u_2 d$ *for some constants* $u_2>0$*, then for some constants* $a_3''>0$ *and* $\gamma_2\in(1,2-\frac{\mu B}{2\Delta})$*, we have*

$$\Pr[\mathit{RECLUSTER\,()\,error}]\leq 4d\frac{m}{t}\exp(-a_3''n\frac{\Delta'}{2\hat{L}})+m(\frac{p}{m})^{\gamma_2} \tag{12}$$

The proof of this Lemma is presented in appendix E.6.

The total probability of error after for a single step of `REFINE()` is the sum of probability of errors for these 4 events by the union bound. Therefore,

$$\begin{aligned}\Pr[\mathcal{C}_1\neq\mathcal{C}^\star]\leq&\frac{m}{c_{\min}}\exp(-a_1c_{\min})+\frac{m}{t}\exp(-a_2m)\\&+(1-\beta)m(\frac{p}{m})^{\gamma_1}+8d\frac{m}{t}\exp(-a_3n\frac{\Delta'}{2\hat{L}})+m(\frac{p}{m})^{\gamma_2}\end{aligned}$$

where we set $a_3=\min\{a_3',a_3''\}$.
For some small constants $\rho_1>0, \rho_2\in(0,1)$, we can choose $\gamma_1\in(1,2), \beta\in(0,\frac{1}{2})$ and $\gamma_2\in(1,2-\frac{\mu B}{2\Delta})$ such that $(1-\beta)(\frac{p}{m})^{\gamma_1-1}+(\frac{p}{m})^{\gamma_2-1}\leq\frac{\rho_1}{2m^{1-\rho_2}}$ and for large enough m,Δ' and n, $\frac{m}{c_{\min}}\exp(-a_1c_{\min})+\frac{m}{t}\exp(-a_2m)+8d\frac{m}{t}\exp(-a_3n\frac{\Delta'}{2\hat{L}})\leq\frac{\rho_1}{2m^{1-\rho_2}}p$. This happens because we have terms of $\exp(-m)$,$\exp(-c_{\min})$ and $\exp(-n\Delta')$, which decrease much faster than $\frac{p}{m}$ which has terms of $\mathcal{O}(m\exp(-n\Delta))$, where Δ and Δ' are of the same order. Therefore, the total probability of error can be bounded by

$$\Pr[\mathcal{C}_1\neq\mathcal{C}^\star]\leq\frac{\rho_1}{m^{1-\rho_2}}p \tag{13}$$

E.3 Proof of lemma E.1

$$\begin{aligned}&\Pr[\exists c\in\mathrm{rg}(\mathcal{C}^\star)\text{ such that no cluster in }\mathcal{C}_0\text{ has cluster label }c]\\&\leq\sum_{c\in\mathcal{C}^\star}\Pr[\text{No cluster in }\mathcal{C}_0\text{ has cluster label }c]\end{aligned} \tag{14}$$

Here, we use union bound over the clusters for the second inequality. Now, we analyze the probability that no cluster in $\mathrm{rg}(\mathcal{C}_0)$ has cluster label c for some $c\in\mathrm{rg}(\mathcal{C}^\star)$. Consider a cluster in $\mathrm{rg}(\mathcal{C}_0)$. This cluster has cluster label c if a majority of its nodes are from cluster $c\in\mathrm{rg}(\mathcal{C}^\star)$. Since the size of each cluster in $\mathrm{rg}(\mathcal{C}_0)$ is atleast t and there are C clusters in $\mathrm{rg}(\mathcal{C}^\star)$, if all clusters in $\mathrm{rg}(\mathcal{C}_0)$ have $\leq\frac{t}{C}$ nodes from cluster c, then no cluster will have cluster label c.
Assume that the clique formed by nodes from cluster c has r nodes. Then, every node i in cluster c, must have S_c-r edges absent, which correspond to the edges between a node of the clique and those outside it. Thus, we obtain,

$$\begin{aligned}\Pr[\text{No cluster in }\mathcal{C}_0\text{ has cluster label }c]\leq&\Pr[\bigcap_{\mathcal{C}^\star(i)=c}\{\sum_{j\neq i,\mathcal{C}^\star(i)=c}X_{ij}>S_c-\frac{t}{C}\}]\\\leq&\Pr[\sum\sum_{\mathcal{C}^\star(i)=\mathcal{C}^\star(j)=c}X_{ij}>S_c(S_c-\frac{t}{C})]\\\leq&\Pr[\sum\sum_{\mathcal{C}^\star(i)=\mathcal{C}^\star(j)=c}(X_i+X_j)>S_c(S_c-\frac{t}{C})]\\\leq&\Pr[\frac{1}{S_c}\sum_{\mathcal{C}^\star(i)=c}X_i>1-\frac{t}{CS_c})]\\\leq&\exp(-\left(1-\frac{t}{CS_c}-\frac{p}{m}\right)^2S_c)\\\leq&\exp(-a_1c_{\min})\end{aligned}$$

In the first step, we require each node i to have $S_c-\frac{t}{C}$ wrong edges. For the second inequality, we remove the intersection and thus, the total number of incorrect edges has to be $S_c(S_c-\frac{t}{C})$, since each node has $S_c-\frac{t}{C}$ incorrect edges. For the third inequality, we use $X_{ij}\leq X_i+X_j$ and collect the terms

of X_i for the fourth inequality. In the fifth inequality, we obtain a condition on the sum of independent Bernoulli random variables each with mean $\frac{p}{m}$. Therefore, we can apply Chernoff bound for their sum to obtain the fifth inequality.
A necessary condition for us is $1-\frac{t}{CS_c}-\frac{p}{m}>0$ which translates to $t<CS_c(1-\frac{p}{m})$. If we select $t\leq c_{min}-1$, this inequality is always satisfied. Note that we want the term $\left(1-\frac{t}{CS_c}-\frac{p}{m}\right)^2>a_1$, for some positive constant a_1. If we choose $t=\Theta(m)$, which is possible if $t=\Theta(c_{\min})$ as we assume $c_{\min}=\Theta(m)$, then this is satisfied. We use the lower bound a_1 and $S_c\geq c_{\min}$ to obtain the final inequality. Plugging this in Eq (14), we obtain our result.

E.4 Proof of lemma E.2

$$\begin{aligned}&\Pr[\exists c\in\mathrm{rg}(\mathcal{C}_0)\text{ which has}\geq\alpha\text{ fraction of impurities}]\\&\leq\sum_{c\in\mathrm{rg}(\mathcal{C}_0)}\Pr[\text{cluster }c\text{ has}\geq\alpha\text{ fraction of wrong nodes}]\end{aligned}\tag{15}$$

We use a simple union bound on clusters in $\mathcal{C}_0$ for the above inequality. Let the set of nodes in the cluster c which are from same cluster of $\mathcal{C}^\star$ as the cluster label of c, i.e., which are not impurities, be R_c. Then let $Q_c=|R_c|$. Let Q'_c denote the number of impurities in cluster c.

$$\begin{aligned}\Pr[\text{cluster }c\text{ has}\geq\alpha\text{ fraction of wrong nodes}]\leq&\Pr[Q'_c\geq\frac{\alpha}{1-\alpha}Q_c]\\&\Pr[Q'_c\geq\alpha t]\end{aligned}$$

We use the fact that $Q_c+Q'_c\geq t$, which is the minimum size of any cluster, for the second inequality.
Now, we analyze the probability of a single node to be incorrect. A node is an impurity in cluster c if it has an edge to each of nodes in R_c.

$$\begin{aligned}\Pr[\text{Node }i\text{ is an impurity in cluster c}]\leq&\Pr[\min_{j\in R_c}\|w_{i,T}-w_{j,T}\|\leq\lambda]\\\leq&\Pr[\min_{j\in R_c}(\|w^\star_i-w^\star_j\|-\Xi_{i,j})\leq\lambda]\\\leq&\Pr[\Lambda_i+\max_{j\in R_c}\Lambda_j\geq2\Delta]\end{aligned}\tag{16}$$

Now, if $\max_{j\in R_c}\Lambda_j\leq\gamma_1\Delta$, for $\gamma_1\in(1,2)$, then we need $\Lambda_i\geq(2-\gamma_1)\Delta$ for error.
Using the definition of random variables in appendix E.1

$$\begin{aligned}\Pr[Q'_c\geq\alpha t]\leq&\Pr[Q'_c\geq\alpha t|\max_{j\in R_c}\Lambda_j\leq\gamma_1\Delta]+\Pr[\max_{j\in R_c}\Lambda_j\geq\gamma_1\Delta]\\\leq&\Pr[\sum_{i=1}^mY_{i,2-\gamma_1}\geq\alpha t]+\Pr[\max_{j\in R_c}\Lambda_j\geq\gamma_1\Delta]\end{aligned}$$

For the first inequality, we use union bound over the value of $\max_{j\in R_c}\Lambda_j$ and for the second inequality, we need atleast αt impurities, so atleast αt of all $Y_{i,2-\gamma_1}$ should be 1.
We now bound the two terms in the final inequality separately.
For the second term, if $\max_{j\in R_c}\Lambda_j\geq\gamma_1\Delta$.

$$\Pr[\max_{j\in R_c}\Lambda_j\geq\gamma_1\Delta]\leq Q_c\Pr[Y_{j,\gamma_1}=1]\leq Q_c(\frac{p}{m})^{\gamma_1}$$

Here, we use union bound over all elements in R_c for the first inequality and the second inequality is plugging in the value of $\Pr[Y_{j,\gamma_1}=1]$, which we have already computed.
Now, we need to provide a bound on Q_c. Note that if Q_c denotes the correct number of nodes, which corresponds to the majority of nodes, then $Q_c\leq(1-\alpha)S_c$, where S_c is the size of the cluster c.
For the first term, we can use Chernoff bound as $Y_{i,2-\gamma_1}$ are independent random variables with expectation $\frac{p}{m}$

$$\Pr[\frac{1}{m}\sum_{i=1}^mY_{i,2-\gamma_1}\geq\alpha\frac{t}{m}]\leq\exp(-(\alpha\frac{t}{m}-\mathbb{E}[Y_{i,2-\gamma_1}])^2m)\leq\exp(-a_2m)$$

We need $\alpha\frac{t}{m}\geq\mathbb{E}[Y_{i,2-\gamma_1}]$,which implies $\alpha t\geq1$, since $Y_{i,2-\gamma_1}$ is a bernoulli random variable. Further, we require $\alpha t=\Theta(m)$, so that we can bound the probability using a constant $a_2\geq0$. If we choose γ_1 as a constant independent of m, then we are done.

Now, plugging all these inequalities into Eq (15), we get

$$\begin{aligned}
&\Pr[\exists c \in \mathrm{rg}(\mathcal{C}_0) \text{ which has} \geq \alpha \text{ fraction of wrong nodes}]\\
&\quad \leq \mathrm{rg}(\mathcal{C}_0)\exp(-a_2 m) + \sum_{c\in \mathrm{rg}(\mathcal{C}_0)} (1-\alpha) S_c (\frac{p}{m})^{\gamma_1}\\
&\quad \leq |\mathrm{rg}(\mathcal{C}_0)|\exp(-a_2 m) + (1-\alpha) m (\frac{p}{m})^{\gamma_1}\\
&\quad \leq \frac{m}{t}\exp(-a_2 m) + (1-\alpha) m (\frac{p}{m})^{\gamma_1}
\end{aligned}$$

For the second inequality, we use $\sum_{c\in\mathcal{C}_0} S_c = m$ and for the third inequality, we use $|\mathrm{rg}(\mathcal{C}_0)| t \leq m$.

E.5 Proof of lemma E.3

First, let $i,j \in [m]$ be a node in cluster $c,c' \in \mathrm{rg}(\mathcal{C}_0)$ respectively such that $\mathcal{C}^\star(j)$ and $\mathcal{C}^\star(i)$ are the cluster labels of clusters c and c' respectively. Then, if we repeat our thresholding analysis for `MERGE()` operation, we obtain

$$\begin{aligned}
&\mathsf{dist}(w_i^\star, w_j^\star) - \Psi_{c,c'} \leq \mathsf{dist}(\omega_{c,T}, \omega_{c',T}) \leq \mathsf{dist}(w_i^\star, w_j^\star) + \Psi_{c,c'}\\
&\text{where } \Psi_{c,c'} = \mathsf{dist}(\omega_c^\star, w_i^\star) + \mathsf{dist}(\omega_{c'}^\star, w_j^\star) + \sum_{k=c,c'} \mathsf{dist}(w_{k,T}, w_k^\star)
\end{aligned}$$

We obtain the above equations by a simple application of triangle inequality. Here, $\omega_c^\star$ is as defined in appendix E.1.
To analyze the above quantities, we need to bound $\|\omega_c^\star - \omega_{c,T}\|$ and $\left\|\omega_c^\star - w_j^\star\right\|$ for some $j \in G_c$. The following Lemmas provide these bounds.

Lemma E.5 (Convergence of $\omega_{c,T}$). *If assumption 4.3,assumption 4.4 and assumption 4.5 hold, and $\eta \leq \frac{1}{L}$, then*

$$\|\omega_{c,T} - \omega_c^\star\| \leq (1-\kappa^{-1})^{T/2} D + \frac{2}{\mu}\Lambda_c \quad \forall c \in \mathrm{rg}(\mathcal{C}_0) \tag{17}$$

where $\kappa = \frac{L}{\mu}$ and Λ_c is a positive random variable with

$$\begin{aligned}
&\Pr[\Lambda_c \geq \sqrt{2d}\frac{r+3\beta s}{1-2\beta} + \sqrt{2}\frac{2(1+3\beta)}{1-2\beta}\delta\hat{L}]\\
&\leq 2d(1+\frac{D}{\delta})^d (\exp(-(1-\alpha)S_c n \min\{\frac{r}{2\hat{L}}, \frac{r^2}{2\hat{L}^2}\})\\
&\quad + (1-\alpha)S_c \exp(-n\min\{\frac{s}{2\hat{L}}, \frac{s^2}{2\hat{L}^2}\}))
\end{aligned} \tag{18}$$

for some $r,s,\delta > 0$ where S_c is the size of cluster c.

Proof is presented in appendix E.7

Lemma E.6 (Distance between cluster minima and node minima). *If assumption 4.3 and assumption 4.5 are satisfied then, for all $j \in [m]$, where j is a node in cluster $c \in \mathcal{C}_0$ where $\mathcal{C}^\star(j)$ is the cluster label of node c, we have*

$$\left\|\omega_c^\star - w_j^\star\right\| \leq \sqrt{\frac{2\hat{L}\epsilon_1}{\mu}} := B \tag{19}$$

Proof is presented in appendix E.9.
Now, that we have our required quantities, we are ready to analyze the probability of error after the merge and reclustering operations.
First, we analyze the probabilty of `MERGE()` operation. Note that if correct nodes of c and c' were from the same cluster $\mathcal{C}^\star$ then, $\left\|w_i^\star - w_j^\star\right\| \leq \epsilon_1, \forall i \in G_c, j \in G_{c'}$. If correct nodes of c' and c were from different clusters in $\mathcal{C}^\star$, then, $\left\|w_i^\star - w_j^\star\right\| \geq \epsilon_2, \forall i \in G_c, j \in G_{c'}$. Therefore, the probability of `MERGE()`

error is upper bounded by

$$
\begin{aligned}
\Pr[\texttt{MERGE()}\text{ Error}] &\leq \Pr[\text{at least 1 edge is incorrect}] \\
&\leq \Pr[\max_{c,c'} \Psi_{c,c'} \geq \Delta_\lambda] \\
&\leq \Pr[\max_{c,c'} \sum_{k=c,c'} \frac{2\Lambda_k}{\mu} \geq \Delta_\lambda - 2(1-\kappa^{-1})^{T/2} D - 2B] \\
&\leq \max_{c \in \mathrm{rg}(\mathcal{C}_0)} \Pr[\Lambda_c \geq \frac{\mu}{2}(\frac{\Delta_\lambda}{2} - (1-\kappa^{-1})^{T/2} D - B)] \\
&\leq \max_{c \in \mathrm{rg}(\mathcal{C}_0)} \Pr[\Lambda_c \geq \Delta'] && (20) \\
&\leq \max_{c \in \mathrm{rg}(\mathcal{C}_0)} 4d\exp(-a_3' n \frac{\Delta'}{2\hat{L}}) \\
&\leq \sum_{c \in \mathrm{rg}(\mathcal{C}_0)} 4d\exp(-a_3' n \frac{\Delta'}{2\hat{L}}) \leq \frac{4dm}{t} \exp(-a_3' n \frac{\Delta'}{2\hat{L}}) && (21)
\end{aligned}
$$

For the second inequality, we expand all the terms of $\Phi_{c,c'}$. We set $\Delta' = \frac{\mu}{2}(\frac{\Delta_\lambda}{2} - (1-\kappa^{-1})^{T/2} D - B)$. Then, we set $r = \Theta(\hat{L}\max\{\frac{\Delta'}{S_c\sqrt{d}\hat{L}}, \sqrt{\frac{\Delta'}{S_c\sqrt{d}\hat{L}}}\})$, $s = \Theta(\hat{L}\max\{\frac{\Delta'}{S_c\sqrt{d}\hat{L}} + \frac{2\log(S_c)}{n}, \sqrt{\frac{\Delta'}{S_c\sqrt{d}\hat{L}} + \frac{2\log(S_c)}{n}}\})$ and $\delta = \Theta(\frac{Dd^{3/2}\hat{L}}{n\Delta'})$. Now, if
$d = \Omega(\min\{\frac{n^{2/3}\Delta^{4/3}}{D^{2/3}\hat{L}^{2/3}}, \frac{n^2\Delta'^2}{\hat{L}^2\log(c_{\min})}\})$, such that $\sqrt{2d}\frac{r+3\beta s}{1-2\beta} + \sqrt{2}\frac{2(1+3\beta)}{1-2\beta}\delta\hat{L} \geq \Delta'$, then there exist some constant $a_3' > 0$ such that the second inequality is satisfied by lemma E.5. We then use the union bound, followed by $|\mathrm{rg}(\mathcal{C}_0)| \leq \frac{m}{t}$.

E.6 Proof of lemma E.4

We can apply our thresholding analysis to
$\|\omega_{c,T} - w_{i,T}\|$ for $c \in \mathrm{rg}(\mathcal{C}_0)$. First, let j be a node in cluster c such that $\mathcal{C}^\star(j)$ is the cluster label of c.

$$
\begin{aligned}
\mathsf{dist}(w_j^\star, w_i^\star) + \Phi_{c,i} \leq \mathsf{dist}(\omega_{c,T}, w_{i,T}) \leq \mathsf{dist}(w_j^\star, w_i^\star) + \Phi_{c,i} \\
\text{where } \Phi_{c,i} = \mathsf{dist}(\omega_{c,T}, \omega_c^\star) + \mathsf{dist}(\omega_c^\star, w_j^\star) + \mathsf{dist}(w_{i,T}, w_i^\star)
\end{aligned}
$$

From appendix D and appendix E.5, we have bounds for all the terms involved. Note that after merging, each cluster in $\mathcal{C}^\star$ should have only 1 cluster in $\mathcal{C}_1$. Therefore, after we recluster according to $\|\omega_{c,T} - w_{i,T}\|$, we incur an error if i goes to the wrong cluster. Suppose that the c corresponds to the correct cluster for i and c' is the cluster to which it is assigned , with $c, c' \in \mathrm{rg}(\mathcal{C}_1), c \neq c'$. Then,

$$
\begin{aligned}
\Pr[\text{Reclustering Error}] &\leq \Pr[\max_{i \in [m]} \max_{c' \neq c} \|\omega_{c',T} - w_{i,T}\| \leq \|\omega_{c,T} - w_{i,T}\|] \\
&\leq \Pr[\max_{i \in [m]} \max_{c' \neq c} \epsilon_2 - \Phi_{c',i} \leq \epsilon_1 + \Phi_{c,i}] \\
&\leq \Pr[\max_{i \in [m]} \max_{c' \in \mathcal{C}_0'} \Phi_{c,i} \geq \frac{\epsilon_2 - \epsilon_1}{2}] \\
&\leq \Pr[\max_{i \in [m]} \max_{c' \in \mathcal{C}_0'} (\Lambda_c + \Lambda_i) \geq \Delta + \Delta'] && (22) \\
&\leq \Pr[\max_{c \in \mathcal{C}_0'} \Lambda_c \geq \Delta' - (\gamma_2 - 1)\Delta] + \Pr[\max_{i \in [m]} \Lambda_i \geq \gamma_2 \Delta] \\
&\leq \max_{c \in \mathrm{rg}(\mathcal{C}_0)'} \Pr[\Lambda_c \geq \Delta''] + \max_{i \in m} \Pr[\Lambda_i \geq \gamma_2 \Delta] && (23)
\end{aligned}
$$

For the second inequality, we use the thresholding analysis on $\|\omega_{c,T} - w_{i,T}\|$. For the third inequality, we rearrange the terms and combine max over $c' \neq c$ with c, and use. For the fourth inequality, we expand the terms of $\Phi_{c,T}$ and substitute the values of Δ and Δ', using the inequality $\Delta_\lambda \leq \frac{\epsilon_2 - \epsilon_1}{2}$. For the fifth inequality, we use consider some $\gamma_2 \in (1, 2 - \frac{\mu B}{2\Delta})$ and break the terms using union bound such that $\Delta'' = \Delta' - (\gamma_2 - 1)\Delta \geq 0$. Finally, we use the union bound on $c \in \mathrm{rg}(\mathcal{C}_0)'$ and $i \in [m]$.
Now, we bound the two terms in Eq (23) separately. The second term can be bounded in terms of Y_{i,γ_2}. Thus,

$$
\max_{i \in [m]} \Pr[\Lambda_i \geq \gamma_2 \Delta] = \max_{i \in [m]} \Pr[Y_{i,\gamma_2} = 1] \leq m(\frac{p}{m})^{\gamma_2} \qquad (24)
$$

We use expectation of Y_{i,γ_2} calculated in appendix E.4 and then bound max by sum.
For the first term, our analysis is similar to that of `MERGE()` error. Assume that there is some constant $u_2 > 1$ such that $\Delta'' \geq u_2 \Delta'$. We set $\delta = \Theta(\frac{Dd^{3/2}\hat{L}}{n\Delta'})$, $r = \Theta(\hat{L}\max\{\frac{\Delta'}{S_c\sqrt{d}\hat{L}}, \sqrt{\frac{\Delta'}{S_c\sqrt{d}\hat{L}}}\})$, $s = \Theta(\hat{L}\max\{\frac{\Delta'}{S_c\sqrt{d}\hat{L}} + \frac{2\log(S_c)}{n}, \sqrt{\frac{\Delta'}{S_c\sqrt{d}\hat{L}} + \frac{2\log(S_c)}{n}}\})$, and if $d = \Omega(\min\{\frac{n^{2/3}\Delta^{4/3}}{D^{2/3}\hat{L}^{2/3}}, \frac{n^2\Delta'^2}{\hat{L}^2\log(c_{\min})}\})$, such that $\sqrt{2d}\frac{r+3\beta s}{1-2\beta} + \sqrt{2}\frac{2(1+3\beta)}{1-2\beta}\delta\hat{L} \geq \Delta'$, then there exist some constant $a_3'' > 0$ such that the second inequality is satisfied by lemma E.5. We then use the union bound, followed by $|\mathrm{rg}(\mathcal{C}_0)| \leq \frac{m}{t}$.

$$\max_{c\in\mathrm{rg}(\mathcal{C}_0)'} \Pr[\Lambda_c \geq \Delta''] \leq \max_{c\in\mathrm{rg}(\mathcal{C}_0)'} 4d\exp(-a_3'' n\frac{\Delta'}{2\hat{L}}) \tag{25}$$

$$\leq \sum_{c\in\mathrm{rg}(\mathcal{C}_0)'} 4d\exp(-a_3'' n\frac{\Delta'}{2\hat{L}}) \tag{26}$$

$$\leq \frac{4dm}{t}\exp(-a_3'' n\frac{\Delta'}{2\hat{L}}) \tag{27}$$

E.7 Proof of lemma E.5

First, we use an intermediate Lemma from Yin et al. (2018). This characterizes the behavior of $TrimmedMean_\beta$ gradient estimator.

Lemma E.7 (TrimmedMean Estimator Variance). *Let $g_c(w)$ be the output of* TrMean_β *estimator for cluster $c \in \mathcal{C}_0$ with size of cluster S_c. If assumption 4.5 holds, then*

$$\begin{aligned}
&\|g_c(w) - \nabla F_c(w)\| \leq \Lambda \\
&\textit{where } \Pr[\Lambda \geq \sqrt{2d}\frac{r+3\beta s}{1-2\beta} + \sqrt{2}\frac{2(1+3\beta)}{1-2\beta}\delta\hat{L}] \\
&\leq 2d(1+\frac{D}{\delta})^d \Bigg(\exp(-(1-\alpha)S_c n \min\{\frac{r}{2\hat{L}}, \frac{r^2}{2\hat{L}^2}\}) \\
&\quad + (1-\alpha)S_c\exp(-n\min\{\frac{s}{2\hat{L}}, \frac{s^2}{2\hat{L}^2}\})\Bigg)
\end{aligned} \tag{28}$$

for some $r,s,\delta > 0$.

Proof. The proof of this Lemma follows from coordinate-wise sub-exponential distribution of ∇F_c. Since loss per sample $f(w,z)$ is Lipschitz in each of its coordinates with Lipschitz constant L_k for $k \in [d]$. Thus, $F_c(w)$ is also L_k-Lipschitz for each coordinate $k \in [d]$ from corollary G.6. Now, every subgaussian variable with variance σ^2 is σ-sub exponential. Thus, each coordinate of $\nabla_w f(w,z)$ is $\hat{L}$-sub-exponential, since $\hat{L} > L_k, \forall k \in [d]$. The remainder of proof can be found in Appendix E.1 in Yin et al. (2018). □

Now, using the above Lemma, we can bound the iterate error for a cluster $c \in \mathcal{C}_0$. Consider $\|\omega_{c,t+1} - \omega_c^\star\|^2$,

$$\begin{aligned}
\|\omega_{c,t+1} - \omega_c^\star\| &\leq \|proj_{\mathcal{W}}\{\omega_{c,t} - \eta\nabla g(\omega_{c,t})\} - \omega_c^\star\| \\
&\leq \|\omega_{c,t} - \eta\nabla g(\omega_{c,t}) - \omega_c^\star\| \\
&\leq \|\omega_{c,t} - \eta\nabla F(\omega_{c,t}) - \omega_c^\star\| + \eta\|g(\omega_{c,t}) - \nabla F(\omega_{c,t})\| \\
&\leq \|\omega_{c,t} - \eta\nabla F(\omega_{c,t}) - \omega_c^\star\| + \eta\Lambda
\end{aligned}$$

Now, we bound $\|\omega_{c,t} - \eta\nabla F(\omega_{c,t}) - \omega_c^\star\|^2$ using μ-strong convexity and L-smoothness of F_c. The analysis is similar to the convergence analysis in appendix D.1. Thus, for $\eta \leq \frac{1}{L}$

$$\|\omega_{c,t} - \eta\nabla F(\omega_{c,t}) - \omega_c^\star\|^2 \leq (1-\eta\mu)\|\omega_{c,t} - \omega_c^\star\|^2$$

Using this bound we can analyze the original term with $\|\omega_{c,t+1}-\omega_c^\star\|$.

$$
\begin{aligned}
\|\omega_{c,t+1}-\omega_c^\star\| &\leq \sqrt{1-\eta\mu}\|\omega_{c,t}-\omega_c^\star\|+\eta\Lambda\\
\|\omega_{c,T}-\omega_c^\star\| &\leq (1-\eta\mu)^{T/2}\|\omega_{c,0}-\omega_c^\star\|+\eta\Lambda(\sum_{t=0}^{T-1}(1-\eta\mu)^{t/2})\\
&\leq (1-\kappa^{-1})^{T/2}\|\omega_{c,0}-\omega_c^\star\|+\eta\Lambda(\sum_{t=0}^{\infty}(1-\frac{\eta\mu}{2})^t)\\
&\leq (1-\kappa^{-1})^{T/2}D+\frac{2}{\mu}\Lambda
\end{aligned}
$$

For the second inequality, we use $\kappa=\frac{L}{\mu}$ and unroll the recursion for T steps. For the third inequality, we use $\sqrt{1-x}\leq 1-\frac{x}{2}$ and upper bound the finite geometric sum by its infinite counterpart. Finally we use the boundedness of $\mathcal{W}$ and the sum of the geometric series to get our result.

E.8 Proof of lemma D.2

We present the proof for this lemma here as it is a corollary of lemma E.5.
We utilize the intermediate lemma E.7. Now, if we set $\alpha=\beta=0$ and $S_c=1$, we obtain the generalization guarantee for GD on a single node $i\in[m]$. Further, we do not need the terms of s as they appear with β, and thus, we can choose s very large, so that we can ignore its contribution to error probability. The remainder of the proof follows that of lemma E.5.

E.9 Proof of lemma E.6

Since F_c is $\hat{L}$-Lipshchitz and μ-strongly convex with minima $\omega_c^\star$,

$$
\begin{aligned}
F_c(w_i^\star)-F_c(\omega_c^\star) &= \frac{F_i(w_i^\star)-F_i(\omega_c^\star)}{Q_c}+\sum_{j\neq i,\mathcal{C}_0(j)=c}\frac{F_j(w_i^\star)-F_j(\omega_c^\star)}{Q_c}\\
&\leq \frac{F_i(w_i^\star)-F_i(\omega_c^\star)}{Q_c}+\sum_{j\neq i,\mathcal{C}_0(j)=c}\frac{F_j(w_i^\star)-F_j(w_j^\star)}{Q_c}\\
&\leq -\frac{\mu\|w_i^\star-\omega_c^\star\|^2}{2Q_c}+\sum_{j\neq i,\mathcal{C}_0(j)=c}\frac{\hat{L}\|w_i^\star-w_j^\star\|}{Q_c}\\
\frac{\mu}{2}\|w_i^\star-\omega_c^\star\|^2 &\leq -\frac{\mu\|w_i^\star-\omega_c^\star\|^2}{2Q_c}+\frac{(Q_c-1)\hat{L}\epsilon_1}{Q_c}\\
\frac{\mu}{2}\|w_i^\star-\omega_c^\star\|^2 &\leq -\frac{\mu\|w_i^\star-\omega_c^\star\|^2}{2Q_c}+\frac{(Q_c-1)\hat{L}\epsilon_1}{Q_c}\\
\|w_i^\star-\omega_c^\star\|^2 &\leq \frac{2\hat{L}\epsilon_1}{\mu}\\
\|w_i^\star-\omega_c^\star\| &\leq \sqrt{\frac{2\hat{L}\epsilon_1}{\mu}}
\end{aligned}
$$

For the first equation, we expand F_c into its component terms, where Q_c denotes the number of correct nodes in cluster c. For the second inequality, we use the fact that $w_j^\star=\text{argmin}_{w\in\mathcal{W}}F_j(w)$. For the third inequality, we use strong-convexity of F_i and $\hat{L}$-Lipschitzness for $F_j, j\neq i$. For the fourth inequality, we use a lower bound on $F_c(w_i^\star)-F_c(\omega_c^\star)$ using μ-strong convexity of F_c. Finally, we manipulate the remaining terms to obtain the final bound.

F Proof of theorem 4.14

By theorem 4.8, $\mathcal{C}_R\neq\mathcal{C}^\star$, with probability $\left(\frac{\rho_2}{m^{(1-\rho_1)}}p\right)^R$. For the $(R+1)^{th}$ step, we bound probability of error by 1. Therefore, with probability $1-\exp(-\frac{5}{8}R)p$. For the $(R+1)^{th}$ step, we optimize the cluster iterates from `TrimmedMeanGD()` to improve convergence instead of clustering error. Since $\mathcal{C}_{R+1}=\mathcal{C}_R$, each cluster in $\mathcal{C}_{R+1}$ maps to some cluster in $\mathcal{C}^\star$. Without loss of generality, assume that cluster $c \in \mathrm{rg}(\mathcal{C}_{R+1})$ maps to the same cluster $c \in \mathcal{C}$. Now, if $\{c_1, c_2, ..., c_l\}$ are the clusters in $\mathcal{C}_R$ which merged to form cluster $c \in \mathrm{rg}(\mathcal{C}_{R+1})$. Then, we can write

$$\begin{aligned}
\|\omega_{c,T} - \omega_c^\star\| &= \left\| \frac{1}{l} \sum_{j=1}^{l} (\omega_{c_j,T} - \omega_c^\star) \right\| \\
&\leq \frac{1}{l} \sum_{j=1}^{l} \|\omega_{c_j,T} - \omega_c^\star\| \\
&\leq \frac{1}{l} \sum_{j=1}^{l} (\left\|\omega_{c_j,T} - \omega_{c_j}^\star\right\| + \left\|\omega_{c_j}^\star - \omega_c^\star\right\|)
\end{aligned}$$

For the first inequality, we used the definition of $\omega_{c,T}$ from `MERGE()`. For the second inequality, we used the triangle inequality for the l elements. The third inequality is obtained by using triangle inequality and adding and subtracting $\omega_{c_j}^\star$ as defined in appendix E.1.

Now, consider the set of nodes $\{i_1, i_2, ..., i_l\} \subseteq [m]$, such that $i_j \in c_j \forall j \in [l]$ and $\mathcal{C}^\star(i_j) = c \forall j \in [l]$. Therefore, we can split each term of $\left\|\omega_{c_j}^\star - \omega_c^\star\right\|$ as –

$$\begin{aligned}
\|\omega_{c,T} - \omega_c^\star\| &\leq \frac{1}{l} \sum_{j=1}^{l} (\left\|\omega_{c_j,T} - \omega_{c_j}^\star\right\| + \left\|\omega_{c_j}^\star - w_{i_j}\right\| + \|w_{i_j} - \omega_c^\star\|) \\
&\leq \frac{1}{l} \sum_{j=1}^{l} \left\|\omega_{c_j,T} - \omega_{c_j}^\star\right\| + 2B
\end{aligned}$$

From lemma E.6, since i_j contributes to both clusters c_j and $c^\star$, we can bound the difference from their minima by B. Further, we can use lemma E.5 and the lemma E.7, which is adapted from Theorem 4 in Yin et al. (2018),to bound the convergence of $\left\|\omega_{c_j,T} - \omega_{c_j}^\star\right\|$. If we set $\delta = \frac{1}{nS_{c_j}\hat{L}D}$ and

$$\begin{aligned}
r &= \hat{L} \max\{\frac{8d}{nS_{c_j}} \log(1 + nS_c\hat{L}D), \sqrt{\frac{8d}{nS_{c_j}} \log(1 + nS_c\hat{L}D)}\} \\
s &= \hat{L} \max\{\frac{4d}{n}(d\log(1 + nS_{c_j}\hat{L}D) + \log m), \sqrt{\frac{4d}{n}(d\log(1 + nS_{c_j}\hat{L}D) + \log m)}\}
\end{aligned}$$

where S_{c_j} is the size of cluster c_j, we obtain

$$\|\omega_{c,T} - \omega_c^\star\| \leq (1 - \kappa^{-1})^{T/2} D + \Lambda' + 2B$$

where

$$\Lambda' = \mathcal{O}\left(\frac{\hat{L}d}{1 - 2\beta} \left(\frac{\beta}{\sqrt{n}} + \frac{1}{\sqrt{nc_{\min}}} \right) \sqrt{\log(n \max_{j \in [l]} S_{c_j} \hat{L}D)} \right)$$

We can further upper bound $\max_{j \in [l] S_{c_j}}$ by m. Now, the probability of error for each cluster $c \in \mathrm{rg}(\mathcal{C}_R)$ for given values of r and s is $\frac{4d}{(1 + nc_{\min}\hat{L}D)^d}$, therefore, we can use union bound and multiply this probability of error by $\mathrm{rg}(\mathcal{C}_R) \leq \frac{m}{t}$. Since $t = \Theta(c_{\min})$, we can upper bound this by $\frac{mu''}{c_{\min}}$ for some positive constant $c_{\min}$.

G Additional Definitions and Lemmas

We start with reviewing the standard definitions of strongly convex and smooth functions $f : \mathbb{R}^d \mapsto \mathbb{R}$.

Definition G.1. f is μ-strongly convex if $\forall w, w'$, $f(w') \geq f(w) + \langle \nabla f(w), w' - w \rangle + \frac{\mu}{2} \|w' - w\|^2$.

Definition G.2. f is L-smooth if $\forall w, w'$, $\|\nabla f(w) - \nabla f(w')\| \leq L\|w - w'\|$.

Definition G.3. f is L_k Lipschitz for every coordinate $k \in [d]$ if, $|\partial_k f(w)| \leq L_k$, where $\partial_k f(w)$ denotes the k-th coordinate of $\nabla f(w)$.

Lemma G.4. *If $f,g:\mathbb{R}^d\rightarrow\mathbb{R}$ are two μ-strongly convex functions on a domain $\mathcal{W}$. Then, $\frac{f+g}{2}$ is also μ-strongly convex on the same domain.*

Proof. If f and g are μ-strongly convex on a domain $\mathcal{W}$, then for any $w_1,w_0\in\mathcal{W}$

$$f(w_1)\geq f(w_0)+\langle\nabla f(w_0),w_1-w_0\rangle+\frac{\mu}{2}\|w_1-w_0\|^2$$
$$g(w_1)\geq g(w_0)+\langle\nabla g(w_0),w_1-w_0\rangle+\frac{\mu}{2}\|w_1-w_0\|^2$$

Adding the above equations, we get

$$\frac{f(w_1)+g(w_1)}{2}\geq\frac{f(w_0)+g(w_0)}{2}+\left\langle\frac{\nabla f(w_0)+\nabla g(w_0)}{2},w_1-w_0\right\rangle+\frac{\mu}{2}\|w_1-w_0\|^2$$

Thus, $\frac{f+g}{2}$ is also μ-strongly convex. □

Lemma G.5. *If $f,g:\mathbb{R}^d\rightarrow\mathbb{R}$ are two L-smooth functions on a domain $\mathcal{W}$. Then, $\frac{f+g}{2}$ is also L-smooth on the same domain.*

Corollary G.6. *If $f,g:\mathbb{R}^d\rightarrow\mathbb{R}$ are two L-Lipschitz functions on a domain $\mathcal{W}$. Then, $\frac{f+g}{2}$ is also L-Lipschitz on the same domain.*

Proof. Consider the following term for any $w_1,w_0\in\mathcal{W}$

$$\begin{aligned}&\left\|\frac{\nabla f(w_1)+\nabla g(w_1)}{2}-\frac{\nabla f(w_0)+\nabla g(w_0)}{2}\right\|\\&\leq\frac{1}{2}\|(\nabla f(w_1)-\nabla f(w_0))+(\nabla g(w_1)-\nabla g(w_0))\|\\&\leq\frac{1}{2}(\|\nabla f(w_1)-\nabla f(w_0)\|+\|\nabla g(w_1)-\nabla g(w_0)\|)\\&\leq\frac{1}{2}(L\|w_1-w_0\|+L\|w_1-w_0\|)\\&\leq L\|w_1-w_0\|\end{aligned}$$

In the second inequality, we use the triangle inequality of norms. For the third inequality, we use the L-smoothness of f and g. Thus, $\frac{f+g}{2}$ is also L-smooth The proof of the corollary is same as above, by replacing terms of ∇f and ∇g by f and g respectively. □

Lemma G.7. *If each coordinate of a function $f:\mathbb{R}^d\rightarrow\mathbb{R}$ is L_k-Lipschitz for $k\in[d]$ on the domain $\mathcal{W}$, then f is $\hat{L}=\sqrt{\sum_{k=1}^d L_k^2}$-Lipschitz on the same domain $\mathcal{W}$.*

Proof. Consider $w_1,w_0\in\mathcal{W}$.Define a sequence of variables $\{w[k]=((w_1)_1,(w_1)_2...,(w_1)_k,(w_0)_{k+1},...(w_0)_d)^\intercal\}_{k=0}^d$. Then, $w_1=w[d]$ and $w_0=w[0]$

$$\begin{aligned}|f(w_1)-f(w_0)|&=\left|\sum_{k=1}^d(f(w[k])-f(w[k-1]))\right|\\&=\sum_{k=1}^d L_k|(w_1)_k-(w_0)_k|\end{aligned}$$

The second inequality follows by using triangle rule. Then, $f(w[k])$ and $f(w[k-1])$ differ only in the k^{th} coordinate, so we apply L_k coordinate-wise Lipschitzness. Now, consider a random variable $v\in\mathbb{R}^d$ such that $v_k=L_k\frac{|(w_1)_k-(w_0)_k|}{(w_1)_k-(w_0)_k}$ if $(w_1)_k-(w_0)_k\neq0$, else 0. Then,

$$\begin{aligned}\sum_{k=1}^d L_k|(w_1)_k-(w_0)_k|&=\langle v,w_1-w_0\rangle\\&\leq\|v\|\|w_1-w_0\|\\&\leq\sqrt{\sum_{k=1}^d L_k^2}\|w_1-w_0\|\end{aligned}$$

Here, we use the Cauchy-Schwartz inequality for the second step. Then, note that each coordinate of v is bounded by L_k. □