We thank the reviewers for their comments. Below, we first respond to several common questions and then respond to more specific questions raised by each individual reviewer.

Common

Theory All reviewers agree that our theoretical results are solid and well-explained. The only concern (from **R1, R3**) is about our initialization condition. As mentioned in the paper (line 190-192), good initialization is a very standard assumption in the convergence analysis of mixture models (such as clustering; see ref [1, 44]), due to the non-convex optimization landscape of mixture model problems. In fact, in a paper by Jin et al 2016 titled *Local Maxima in the Likelihood of Gaussian Mixture Models: Structural Results and Algorithmic Consequences*, it has been shown that bad local minima provably exists in EM algorithms for Gaussian mixtures. As mentioned in line 58-60, in practice, we *do not* need this assumption as random initialization with restarts works well. **R1** mentioned that with good initialization, *"performing a clustering on the initial models should already give the right clusters"*. We argue that this does not hold in practice: we usually observe that the number of wrongly clustered worker machines is high at the beginning of the algorithm, and keeps decreasing as we run more iterations.

Experiments We observe that the following common questions about experiments were raised by the reviewers.

1) Comparison with other baseline algorithms and more realistic datasets: We conducted comparison with the one-shot clustering algorithm proposed in ref [9]. In Table 1, we present the results on the Federated EMNIST (FEMNIST) dataset which is one of the *realistic* federated learning datasets in the literature (see the paper by Caldas et al. 2018, *LEAF: A Benchmark for Federated Settings*, i.e., ref [2] in our submission). The comparison on other datasets will be added to the revised version.

Table 1: Test accuracy on FEMNIST

IFCA ($k=2$)	IFCA ($k=3$)	one-shot ($k=2$)	one-shot ($k=3$)	global	local
86.88	86.90	86.55	86.64	83.22	73.86

In this experiments, for IFCA and one-shot clustering algorithm, we share the representation layers among all the models, but the last layers for different models are trained based on clustering. As we can see, the results of IFCA are on par with the one-shot clustering algorithm. However, an important goal of FL is to reduce the computational cost at the central server and take full advantage of on-device intelligence. **In the one-shot clustering algorithm, the clustering is done at the center machine, which may lead to much higher computational cost at the center compared to our algorithm.**

2) Knowledge of the number of clusters: Similar to many other clustering algorithms, our algorithm requires a hyper parameter on the number of clusters. In our experiments, we observe that our algorithm is robust to the choice of number of clusters. For example, on Rotated MNIST/CIFAR, when we choose the number of clusters larger than the actual number, the algorithm can quickly identify that one of the cluster contains zero worker machines, and thus the model can be discarded, when we choose the number of clusters smaller than the actual number, several clusters will be classified as a single cluster, and in this case we can still improve over the *global model* and *local model* baselines. Moreover, as we can see in Table 1, for FEMNIST, in which the clusters are more ambiguous, we also observe that our algorithm is robust to the choice of number of clusters. We will provide more detailed discussions on the number of clusters in the revision.

Specific

R1 *"Was a separate dataset for parameter evaluation used?"*: We choose the hyperparameters within a wide range. For each algorithm, we choose the hyperparmeters that produce the best result. This is a common method when running experiments on public datasets.

R1 *"criterion of counting the experiment as successful for the synthetic data is not truly justified"* The success criterion in the synthetic data experiments only needs to be a constant multiple of the standard deviation of the noise. Our results are robust to the choice of the constant, and we will clarify in the revision.

R1 *"convergence rate seems not to address the number of participating workers"* As mentioned in line 154, we present the results for full participation in order to streamline the analysis. Extensions to partial participation is straightforward.

R2 *"whether ... will work in non-linear problem."* It will work for non-linear problems: we prove theoretical results for strongly convex loss, which can be non-linear, and we show experimental results for neural networks.

R2 *"privacy issue"* We did not aim to tackle it in this paper, but privacy is an interesting and important future direction.

R3 *"comparison with ClusteredFL, Sattler et al. 2019"* We will add this comparison in the revision. We emphasize that one important contribution of this work is the rigorous analysis of convergence rates, which Sattler et al. did not fully address. In addition, in our algorithm, the worker machines identify their cluster membership by themselves, whereas in Sattler et al., the clustering was done in a centralized manner, similar to ref [9]. Thus, our algorithm reduces the computational cost of the central server, which is one of the major goals of FL as mentioned above.

R3 *"reduce communication cost"* There are two ways to reduce communication cost: 1) when the cluster identity is stable (which usually happens after running a few iterations), we don't need to send all the models to all the worker machines, and instead we send the model corresponding to the worker's cluster, and 2) we can use weight sharing as mentioned above (e.g., we only need to train k different last layers). We will discuss these points in the revision.

R3 *"mini batch ... to estimate the local cluster"* We use the full local data, and will clarify in the revision.