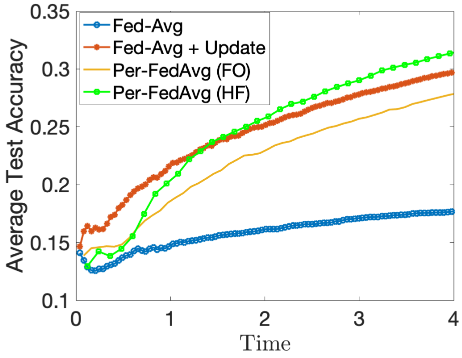
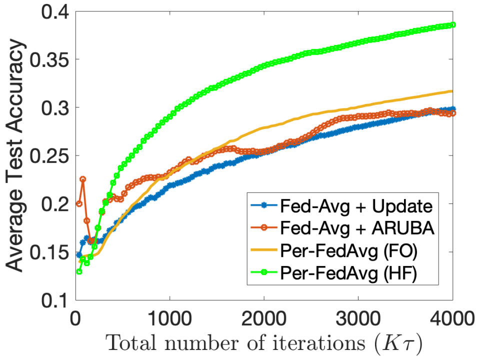


(a) Fig. 1



(b) Fig. 2

We thank the reviewers for their careful consideration and constructive feedback. Below, please find our responses.

Reviewer 1. **Possibility of using a decreasing stepsize:** Indeed, it is possible to achieve the same complexity bound using a diminishing stepsize. In particular, by using β_k as the stepsize at iteration k, eq. (109) holds with $\beta = \beta_k$. Hence, summing up this equation for $k = 0, ..., K-1$, we recover the same complexity bounds using $\beta_k = \mathcal{O}(1/\sqrt{\tau k})$. We'll mention this point as a remark in the revised paper. **Concerning the term $\alpha^2\sigma_G^2/D$ & comparison with [37]:** Note that the term $\alpha^2 L^2\sigma_G^2/D$ appears in the upper bound due to the fact that $\tilde{\nabla} F_i(w)$ is a *biased* estimator of $\nabla F_i(w)$. In particular, as shown in Lemma 4.3, the bias is bounded by $\alpha L\sigma_G/\sqrt{D}$. This bias term will be eliminated if we assume that we have access to the exact gradients at training time (see the discussion after Lemma 4.3), which is the case in [37] where the authors focus on the deterministic case. We'll make the differences with [37] clearer in the revised version. Thank you for your suggestion. **Questions regarding the numerical experiments:** Please see Fig. 1 which illustrates the average test accuracy of all studied algorithms with respect to time. We will include this in the final version of the paper. **Regarding lines 266-268:** We will clarify the data distribution using a figure.

Reviewer 2. **Novelty of the paper and comparison with other theoretical results:** We'd like to emphasize that the main contribution of our work is to provide the first convergence guarantees for meta-federated learning algorithms in the model-agnostic meta-learning regime and for non-convex functions. In particular, [32], mentioned by the reviewer, focuses on the analysis of MAML for centralized settings, and hence, it does not include the local updates on each node ($\tau > 1$) which is one of the main challenges in the analysis of FL algorithms in general. In fact, Proposition F.1 is stated to deal with this challenge which does not exist in the centralized setting at all. In addition, we'd like to add that our analysis is totally different from other meta-federated learning works, such as [38], since they consider different meta-learning regimes. Moreover, it is worth noting that [38] focuses on strongly-convex functions while we study non-convex objective functions. **Regarding comparison with other algorithms such as [38]:** Following your suggestion, we also compare our method with ARUBA. To do so, we also report the output of FedAvg+ARUBA after refinement for each user. In particular, we consider $\tau = 4$ and $K = 1000$, and also tune hyper-parameters of ARUBA for a fair comparison. The final accuracy of all algorithms is as follows: Per-FedAvg(FO): 34.1 ± 0.08, Fed-Avg+ARUBA (with refinement): 36.74 ± 0.1, Per-FedAvg(HF): 43.71 ± 0.12. In Fig. 2, we have also depicted *one realization* of training path, just to provide intuition on the convergence speed of these methods. **Regrading ϵ-stationary definition:** The reviewer is right that there is an inconsistency here. We'll update our definition as $E(\|\nabla f(w)\|^2) \leq \epsilon$ to make it consistent with the result of Corollary 4.6. Thanks for catching this typo. **Dependence of Wasserstein distance on dimension:** The reviewer is right that the convergence speed of Wasserstein distance in d dimension is exponentially slow. Our main goal was to elaborate on the dependence of Wasserstein distance on the number of samples. We will clarify this matter. In addition, we'd like to highlight that our result on TV distance does not suffer from the same issue. Thanks for raising this point.

Reviewer 3. **NLP Example**: NLP example is mainly mentioned in the introduction to highlight the role of data heterogeneity. Indeed the same story holds for images stored on users' devices, which is more consistent with our experiment. We thank the reviewer for this suggestion, and to complete the story, we will add a language model experiment as well. **Regarding the clarity of the paper:** We will provide more details about the setup of our experiments and also clarify the points that the reviewers brought to our attention. **Regarding details of the experiment:** The reviewer is right about the experiment setup (distribution of images). We will clarify our setting, and will also add a figure to explain the distribution of images better. We have also provided the code and will include the updated code in the final version as well. We have used a fully connected neural network with two hidden layers in this experiment. Thanks for your feedback. **$D_t^i, D_t'^i, D_t''^i$ in Eq. 8:** For the sake of analysis, we need these datasets to be independent. That's why we use different datasets. We'll highlight this point. **Experiments:** Please see Figure 1 for the performance of Fed-Avg with and without update, Per-FedAvg (FO), and Per-FedAvg (HF) with respect to time.