Towards Multi-level Fairness and Robustness on Federated Learning

Fengda Zhang [1] **Kun Kuang** [1] **Yuxuan Liu** [2] **Long Chen** [1] **Jiaxun Lu** [3] **Yunfeng Shao** [3] **Fei Wu** [1] **Chao Wu** [2] **Jun Xiao** [1]

Abstract

Federated learning (FL) has emerged as an important machine learning paradigm where a global model is trained based on the private data from distributed clients. However, federated model can be biased due to the spurious correlation or distribution shift over subpopulations, and it may disproportionately advantage or disadvantage some of the subpopulations, leading to the problem of unfairness and non-robustness. In this paper, we formulate the problem of multi-level fairness and robustness on FL to train a global model performing well on existing clients, different subgroups formed by sensitive attribute(s), and newly added clients at the same time. To solve this problem, we propose a unified optimization objective from the view of federated uncertainty set with theoretical analyses. We also develop an efficient federated optimization algorithm named Federated Mirror Descent Ascent with Momentum Acceleration (FMDA-M) with convergence guarantee. Extensive experimental results show that FMDA-M outperforms the existing FL algorithms on multi-level fairness and robustness.

1. Introduction

Federated learning (FL) has emerged as an important machine learning paradigm where distributed clients (e.g., a large number of mobile devices or several organizations) collaboratively train a shared global model while keeping private data on clients (McMahan et al., 2017). However, federated model can be biased because of possible spurious correlation and distribution shift over data subpopulations. As a result, the model performance may degrade significantly on some data subpopulations and bring the problem

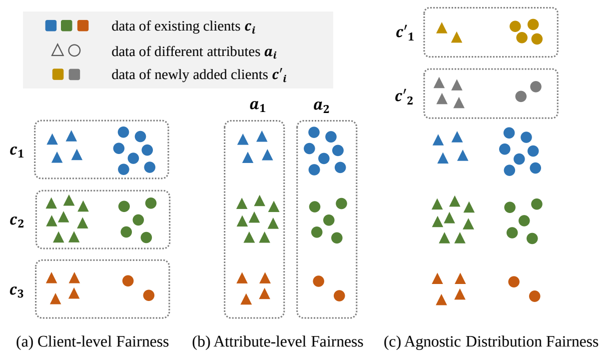


Figure 1. Illustration of three levels in fair FL scenario. (a) client-level fairness: $P(Y = \hat{Y}|c_i) = P(Y = \hat{Y}|c_j)$ for $\forall\ i,\ j$; (b) attribute-level fairness: $P(Y = \hat{Y}|a_i) = P(Y = \hat{Y}|a_j)$ for $\forall\ i, j$; (c) agnostic distribution fairness: $P(Y = \hat{Y}|c'_i) = P(Y = \hat{Y}|c'_j)$ for $\forall\ i,\ j$. The groups are formed by existing clients index, sensitive attribute(s), and index of newly added clients with unknown distribution, respectively.

of unfairness and non-robustness, which becomes an increasing concern, especially in some high-stakes scenarios such as loan approvals, healthcare, etc (Kairouz et al., 2019). How to train an unbiased federated model with fair and robust performance is of paramount importance, and has become an important research theme in recent years (Mohri et al., 2019; Li et al., 2019a; Wang et al., 2021).

It is unfair that the federated model disproportionately advantages or disadvantages some of the clients, since the purpose of clients participating in FL is to get a better model (***client-level fairness***, Fig. 1(a)). Motivated by it, recent research mainly focus on encouraging the federated model to have similar performance over different clients (Mohri et al., 2019; Li et al., 2019a; Wang et al., 2021). However, a federated model trained by client-level method may still suffer from ethical issues in real applications due to neglect of fairness at other levels. For example, consider a scenario that several banks (clients) participate in FL to collaboratively train a loan approval model. Different banks will have different customer demographic compositions (formed by sensitive attribute(s), such as gender). Consider a federated model trained by client-level method, which can treat different banks fairly. Although the model may perform well on male subpopulation, it is also unfair that the model cannot make accurate decision for female subpopulation

[1]College of Computer Science and Technology, Zhejiang University, China [2]School of Public Affairs, Zhejiang University, China [3]Huawei Noah's Ark Lab, China. Correspondence to: Kun Kuang <kunkuang@zju.edu.cn>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA.

(***attribute-level fairness***, Fig. 1(b)). Moreover, the fairness of federated model on those banks that newly participate in FL (***agnostic distribution fairness***, Fig. 1(c)) is also not guaranteed. A model violating any of the above fairness may lead to serious ethical problems, which naturally leads us to ask: *Can we propose a unified FL framework to train a federated model achieving client-level, attribute-level and agnostic distribution fairness* ***at the same time****?*

In fact, a model with good fairness (e.g. a model with similar performance among subpopulations) yet low performance is meaningless. The overall performance is a common metric, but in many real applications (e.g., medical diagnosis and credit evaluations), we are more concerned about FL model robustness, i.e. the worst-performing subpopulation.

In this paper, we focus on the problem of multi-level fairness and robustness on FL to train a federated model performing well on all subpopulations including existing clients, the subgroups formed by sensitive attribute(s), and newly added clients at the same time. To address this problem, we propose a unified risk towards fair and robust FL from the view of federated uncertainty set (Delage & Ye, 2010; Wiesemann et al., 2014; Duchi & Namkoong, 2017). Theoretically, we prove that the proposed unified risk provides an upper bound for both client-level and attribute-level risks, which helps to deal with complex distribution shifts and thus guarantee fairness and robustness at multiple levels simultaneously. We also develop an efficient federated optimization algorithm named Federated Mirror Descent Ascent with Momentum Acceleration (FMDA-M) to optimize the proposed risk with convergence guarantee. Empirically, the advantages of our proposed FMDA-M method, in terms of multi-level fairness and robustness which refers to the worst performance in groups, are demonstrated under different kinds of distribution shift on three real-world datasets.

2. Problem Formulation

2.1. Preliminary on Federated Learning

Suppose that there are N clients in FL and each client $i \in \{1, 2, \dots, N\}$ is associated with a local dataset $D_i^c = \{(x_{i,1}^c, y_{i,1}^c), \dots, (x_{i,n_i^c}^c, y_{i,n_i^c}^c)\}$, where n_i^c is the sample size of client i. Let $D = \{D_1^c, \cdots, D_N^c\}$ be the full dataset with sample size $n = \sum_{i=1}^N n_i^c$. Let P_i^c and P denote the data-generating distribution of each client data D_i^c and whole data D over $\mathcal{X} \times \mathcal{Y}$, respectively. In general, the basic goal of FL is to learn a global model with parameters $\theta \in \Theta$ that performs well on distribution P (in terms of average performance) without accessing the private data of clients.

2.2. Multi-level Fairness and Robustness on FL

In this section, we first define metrics and introduce three levels in FL setting. Then we formulate the problem of multi-level fairness and robustness on FL.

2.2.1. Fairness and Robustness Metrics

Suppose that the full dataset D is divided into M groups: $D = \{D_1^g, D_2^g, \dots, D_M^g\}$. We let P_i^g denote the data-generating distribution of D_i^g. We first define $Disparity$ of a FL model across groups $\{D_i^g | i = 1, 2, \dots, M\}$ as:

$$Disparity = \sqrt{\tfrac{1}{M}\textstyle\sum_{i=1}^{M}(Acc(D_i^g) - Avg\_Acc)^2}, \quad (1)$$

where $Acc(D_i^g)$ is the predictive accuracy on group D_i^g, and $Avg\_Acc = \frac{1}{M}\sum_{i=1}^{M} Acc(D_i^g)$. In this paper, following the *difference principle* on distributive justice and stability (Rawls, 2001), we view the performance of federated model as the resource which is supposed to be allocated into groups fairly. Specifically, we define fairness by $Disparity$, and the smaller $Disparity$, the fairer of a FL model.

Besides, we are also concerned about FL model robustness. $Robustness$ in FL refers to the performance of the worst group with the following definition:

$$Robustness = \min_i Acc(D_i^g). \quad (2)$$

In this paper, we focus on improving both the fairness (in terms of $Disparity$) and $Robustness$ of the FL model.

2.2.2. Multiple Levels in Federated Setting

Suppose that the groups $\{D_1^g, D_2^g, \dots, D_M^g\}$ are formed by a given sensitive variable S. Note that the sensitive variable S can be defined flexibly, and different sensitive variables S correspond to different problems. In this paper, we focus on the following three common cases: 1) ***client level***: S specified as the index of existing clients; 2) ***attribute level***: S specified as sensitive attribute; 3) ***agnostic distribution***: S specified as the index of (potential) newly added clients with agnostic distribution. We argue that a model violating any of the above fairness/robustness definitions may lead to serious ethical problems in reality, which motivates us to achieve multi-level fairness and robustness simultaneously.

2.2.3. Multi-level Fairness and Robustness

Now we propose a novel and meaningful problem as below:

Problem 1 (**Multi-level Fairness and Robustness on FL**). Let the sensitive variable S be specified as existing client index c, the protected attribute(s) a, and newly added client index ad, respectively, then the dataset D can be split into groups $\{D_k^c | k = 1, 2, \dots, M^c\}$, $\{D_k^a | k = 1, 2, \dots, M^a\}$, and $\{D_k^{ad} | k = 1, 2, \dots, M^{ad}\}$, respectively. The task is to learn a federated model with small $Disparity$ and large $Robustness$ on $\{D_k^c\}$, $\{D_k^a\}$, and $\{D_k^{ad}\}$ simultaneously.

3. Unified Risk and Federated Optimization

In this section, we first propose a unified risk to guarantee multi-level fairness and robustness on FL. Then we develop an efficient federated optimization algorithm for it.

3.1. Unified Risk for Multi-level Fair and Robust FL

The essential issue of the existing single-level methods (including client level and attribute level) lies in the cases they consider are not bad enough, so that they cannot deal with complex distribution shifts or spurious correlation. From the view of distributionally robust optimization (DRO), we should construct a wide enough uncertainty set that not only contains the client level and attribute level, but also contains the worse cases to help to adapt to newly added clients.

As one possible way, we specify S as the combination of the client index and the given sensitive attribute(s). Specifically, we divide the local dataset D_i^c of client i into subgroups $D_i^c = \{D_{i,1}^u, D_{i,2}^u, \dots, D_{i,M_i}^u\}$ and consider the potential distribution shifts over them, where M_i is the number of subgroups on client i. Suppose that the samples of $D_{i,k}^u$ are drawn from the distribution $P_{i,k}^u$. Then we define:

$$\begin{aligned} &\mathcal{R}_{\text{unified}}(\theta) := \sup_{Q \in \mathcal{Q}^u} \{\mathbb{E}_{(x,y)\sim Q}[\ell(\theta,(x,y))]\}, \\ &\mathcal{Q}^u := \{\sum\nolimits_{i=1}^{N} \sum\nolimits_{k=1}^{M_i} \lambda_{i,k}^u P_{i,k}^u : \boldsymbol{\lambda}^u \in \Delta_{M-1}\}, \end{aligned} \tag{3}$$

where $M = \sum_{i=1}^N M_i$ is the total number of subgroups and $\lambda_{i,k}^u$ is the weight of k-th subgroup on client i.

3.2. Tractable and Efficient Federated Optimization

To optimize the proposed risk, we first introduce the empirical risk on k-th group of client i defined as $f_{i,k}(\theta) := \mathbb{E}_{(x,y)\sim \hat{P}_{i,k}^u}[\ell(\theta;(x,y))]$, where $\hat{P}_{i,k}^u$ is the empirical distribution over samples of group $D_{i,k}^u$. Then, with the techniques of DRO (Wiesemann et al., 2014; Duchi & Namkoong, 2017; Namkoong & Duchi, 2016), the problem of minimizing the risk in Eq. (3) can be rewritten as:

$$\min_{\theta} \max_{\boldsymbol{\lambda}^u \in \Delta_{M-1}} \{F(\theta, \boldsymbol{\lambda}^u) := \sum\nolimits_{i=1}^N \sum\nolimits_{k=1}^{M_i} \lambda_{i,k}^u f_{i,k}(\theta)\}, \tag{4}$$

To solve this minimax problem, we can alternately optimize model parameters θ and weights $\boldsymbol{\lambda}^u$. Instead of gradient ascent, we adopt mirror ascent method to update $\boldsymbol{\lambda}^u$. It is impractical to naively extend the mirror method to federated setting, since frequent communication is not allowed in FL. So we use an estimation technique to execute mirror update. Besides, we choose the negative entropy function $h(\boldsymbol{x}) = \sum_{i=1}^n (x_i \ln x_i - x_i)$ for simplified calculation. Then we can update $\boldsymbol{\lambda}^u$ as the following:

$$(\lambda_{i,k}^u)^{(r+1)} = \frac{(\lambda_{i,k}^u)^{(r)} e^{\gamma E v_{i,k}^{(r)}}}{\sum_{i=1}^N \sum_{k=1}^{M_i} (\lambda_{i,k}^u)^{(r)} e^{\gamma E v_{i,k}^{(r)}}}, \tag{5}$$

where r is global communication round, E is the number of local iterations, γ is stepsize, and $v_{i,k}$ is loss of subgroup $D_{i,k}^u$. The proposed update rule makes our algorithm more practical by significantly reducing computation complexity and communication cost.

Since communication costs are the principal constraint in FL (McMahan et al., 2017), we explore to further improve the convergence rate of the above federated algorithm by leveraging momentum acceleration techniques (Nesterov, 1983; Li et al., 2017; Ochs, 2018). Specifically, we additionally update model parameters as below:

$$\theta^{(r+1)E} = \widetilde{\theta}^{(r+1)E} + \beta_\theta(\widetilde{\theta}^{(r+1)E} - \widetilde{\theta}^{rE}), \tag{6}$$

and update group weights according to the following rule:

$$(\boldsymbol{\lambda}^u)^{(r+1)} = (\boldsymbol{\lambda}^u)^{(r)} + \beta_{\boldsymbol{\lambda}}((\widetilde{\boldsymbol{\lambda}}^u)^{(r+1)} - (\boldsymbol{\lambda}^u)^r), \tag{7}$$

where β_θ and $\beta_{\boldsymbol{\lambda}}$ are momentum coefficients. The second terms of step (6) and step (7) are momentum terms, which contains historical gradient information and helps speed up the convergence of our algorithm.

The details of Federated Mirror Descent Ascent with Momentum Acceleration (FMDA-M) are in Appendix.

4. Theoretical Analysis

In this section, we provide theoretical analysis for our proposed unified risk and convergence guarantee for FMDA-M.

Theorem 4.1. *Let $\hat{P}_i^c$, $\hat{P}_i^a$ and $\hat{P}_{i,k}^u$ be the empirical distributions over samples of local dataset D_i^c, D_i^a, and group $D_{i,k}^u$ respectively, $\hat{\mathcal{Q}}^c := \{\sum_{i=1}^N \lambda_i^c \hat{P}_i^c : \boldsymbol{\lambda}^c \in \Delta_{N-1}\}$ be the client-level uncertainty set, $\hat{\mathcal{Q}}^a := \{\sum_{i=1}^A \lambda_i^a \hat{P}_i^a : \boldsymbol{\lambda}^a \in \Delta_{A-1}\}$ be the attribute-level uncertainty set, and $\hat{\mathcal{Q}}^u := \{\sum_{i=1}^N \sum_{k=1}^{M_i} \lambda_{i,k}^u \hat{P}_{i,k}^u : \boldsymbol{\lambda}^u \in \Delta_{M-1}\}$ be the unified uncertainty set. We have*

$$\hat{\mathcal{Q}}^c \subseteq \hat{\mathcal{Q}}^u, \hat{\mathcal{Q}}^a \subseteq \hat{\mathcal{Q}}^u. \tag{8}$$

Moreover, let $\hat{\mathcal{R}}_{\text{client}}(\theta)$, $\hat{\mathcal{R}}_{\text{attribute}}(\theta)$ and $\hat{\mathcal{R}}_{\text{unified}}(\theta)$ be the empirical risks based on uncertainty sets $\hat{\mathcal{Q}}^c$, $\hat{\mathcal{Q}}^a$ and $\hat{\mathcal{Q}}^u$ respectively. We have

$$\hat{\mathcal{R}}_{\text{client}}(\theta) \le \hat{\mathcal{R}}_{\text{unified}}(\theta), \hat{\mathcal{R}}_{\text{attribute}}(\theta) \le \hat{\mathcal{R}}_{\text{unified}}(\theta). \tag{9}$$

Furthermore, assume that $\exists$ i and k, and j ($j \ne i$) s.t. the attribute of samples from $D_{i,k}^u$ is same as $D_{j,k}^u$ but $\hat{P}_{i,k}^u \ne \hat{P}_{j,k}^u$, and $\hat{P}_{i,k}^u \ne \hat{P}_l^c$ for $\forall$ l, then we have

$$(\hat{\mathcal{Q}}^c \cup \hat{\mathcal{Q}}^a) \subsetneq \hat{\mathcal{Q}}^u. \tag{10}$$

See Appendix for the proof. Theorem 4.1 shows that both client-level and attribute-level uncertainty sets are subsets of our proposed unified uncertainty set. Therefore, our proposed risk (3) provides an upper bound for both client-level risk and attribute-level risk, thereby optimizing it can guarantee client-level fairness and attribute-level fairness simultaneously. Actually, our proposed risk also considers the worse cases that the set union of client-level uncertainty set and attribute-level uncertainty set does not contain, which helps to deal with more complex distribution shifts and adapt to those newly added clients with agnostic distributions.

Theorem 4.2. *Let* $\mathcal{R}^u_{i,j}(\theta) := \mathbb{E}_{(x,y)\sim P^u_{i,j}}[\ell(\theta;(x,y))]$ *be a risk defined on* $D^u_{i,j}$, $\boldsymbol{\lambda}^u \in \Delta_{M-1}$ *be the group weights,* M *be the total number of groups,* $\bar{\mathcal{R}}^u(\theta)$ *be the average of group risks,* $d_{i,j} := (\mathcal{R}^u_{i,j}(\theta) - \bar{\mathcal{R}}^u(\theta))^2$ *and* $\mathrm{Var}(\mathcal{R}^u_{i,j}(\theta)) := \frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M_i} d_{i,j}$ *be the variance of group risks. If* $\|M\boldsymbol{\lambda}^u - \mathbf{1}\|_2^2 \leq \underset{i,j}{min}\{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} d_{i,j}}{d_{i,j}}\}$*, then there exists a constant* $C > 0$ *such that*

$$\mathcal{R}_{\text{unified}}(\theta) = \bar{\mathcal{R}}^u(\theta) + C\sqrt{\mathrm{Var}_{i\in[N],j\in[M_i]}(\mathcal{R}^u_{i,j}(\theta))}. \quad (11)$$

See Appendix for the proof. The theorem above shows that our proposed risk $\mathcal{R}_{\text{unified}}$ can be viewed as a combination of the average risk that helps improve the average performance and the variance term that encourages the model to have a uniform performance across different subgroups.

Theorem 4.3. *Suppose that each function* $f_{i,k}$ *is convex and L-smooth, global function* F *is linear in* $\boldsymbol{\lambda}$ *and L-smooth and the gradient w.r.t.* θ *and* λ*, model parameters* θ*, the variance of stochastic gradient method w.r.t.* θ *and* λ *are bounded. If we optimize (4) using FMDA-M algorithm with local iterations* $E = O(T^{\frac{1}{4}})$*, learning rate for model parameters* $\eta = O(T^{-\frac{1}{2}})$ *and stepsize for group weights* $\gamma = O(T^{-\frac{1}{2}})$*, and then it holds that*

$$\varepsilon_T \leq O(T^{-\frac{1}{2}}). \quad (12)$$

See Appendix for the proof. Here we give the convergence rate of the proposed FMDA-M algorithm in convex setting.

5. Experiments

In this section, we validate the effectiveness of our method on a real-world, large-scale, and challenging dataset. Additional experiment results and discussion are in Appendix.

5.1. Experimental Setup

Federated Dataset. We use ACS (Ding et al., 2021), which is collected from US Census surveys and consists of more than 1,500,000 samples. The goal is to predict whether an individual earns greater than 50,000 US dollars a year. We consider 50 states as clients in FL and choose gender as the sensitive attribute. We also evaluate different methods on Fashion-MNIST (Xiao et al., 2017), Digit-Five (Xu et al., 2018; Peng et al., 2019; Zhao et al., 2020), and Adult (Dua & Graff, 2017) datasets, and the details are in Appendix.

Evaluation Metrics. We evaluate models' fairness and robustness on attribute level, client level, and agnostic distribution. In each kind of fairness, we use $Disparity$ in Eq. (1) to measure the degree of fairness, and use $Robustness$ in Eq. (2) to measure the robustness. Besides, we also report Equalized Odds (EO) at attribute level. The average accuracy of the models are similar/comparable for all algorithms, and reported in appendix.

Table 1. Experimental results on ASC dataset.

D: *Disparity* / R: *Robustness*	Client-level D	Client-level R	Attribute-level D	Attribute-level R	Attribute-level ΔEO	Agnostic D	Agnostic R
FedAvg [McMahan, *AISTATS'17*]	0.018	70.7	0.194	40.9	0.301	0.021	70.3
q-FFL [Li, *ICLR'20*]	0.017	71.4	0.188	41.7	0.291	0.022	70.9
TERM [Li, *ICLR'21*]	0.017	71.3	0.191	41.2	0.297	0.022	71.1
DRFA [Deng, *NeurIPS'20*]	0.017	71.8	0.194	40.8	0.303	0.020	71.8
FADE [Hong, *KDD'21*]	0.021	67.8	0.022	73.1	0.023	0.024	65.2
IndA (Individual-level)	0.018	70.4	0.172	43.5	0.272	0.021	71.6
FMDA-M (Ours)	**0.017**	**72.1**	**0.019**	**74.3**	**0.021**	**0.019**	**72.7**

Baselines. We compare the proposed FMDA-M algorithm with the following baselines: (i) FedAvg (McMahan et al., 2017): FedAvg is a commonly used algorithm in FL, which minimizes an average risk. (ii) q-FFL (Li et al., 2019a) (client level). (iii) TERM (Li et al., 2020) (client level). (iv) DRFA (Deng et al., 2020) (client level, an improvement on the AFL (Mohri et al., 2019)). (v) FADE (Hong et al., 2021) (attribute level). (vi) Individual-level Algorithm (denoted as IndA for convenience): We use the same algorithm as FMDA-M to solve the individual-level problem with objective (13) as a compared baseline.

Results and Discussion. As shown in Table 1, our proposed FMDA-M outperforms all baselines in terms of multi-level fairness and robustness on FL. For client level and attribute level, we find that FMDA-M also outperforms single-level baselines. We think the reason is that we adopts mirror method, which prevents the weights from being too hard to guarantee convergence stability. For agnostic distribution, FMDA-M also perform well due to the wider federated uncertainty set. We note that the performance gap between attribute-level method and FMDA-M on client-level fairness/robustness is not large, because the biased among clients may not strong enough. We also find FMDA-M achieves low EO, since EO can be viewed as a relaxed version of our fairness notion if S is specified as the combination of target label and sensitive attribute. Individual method is ideal but impractical since it is usually leads to the over pessimism problem in practice.

6. Conclusion

In this paper, we formulate the goal of multi-level fairness and robustness on FL, which is to achieve client-level, attribute-level and agnostic distribution fairness and robustness simultaneously. To achieve it, we propose a unified risk based on DRO and develop an efficient FMDA-M algorithm. Both theoretical analysis and experimental results demonstrate the effectiveness of our method.

7. Acknowledgements

This work was supported by the National Natural Science Foundation of China (U19B2043, U19B2042), Zhejiang Innovation Foundation(2019R52002) and Program of Zhejiang Province Science and Technology (2022C01044).

References

Awasthi, P., Cortes, C., Mansour, Y., and Mohri, M. Beyond individual and group fairness. *arXiv preprint arXiv:2008.09490*, 2020.

Biega, A. J., Gummadi, K. P., and Weikum, G. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, pp. 405–414, 2018.

Binns, R. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 514–524, 2020.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Cui, S., Pan, W., Liang, J., Zhang, C., and Wang, F. Addressing algorithmic disparity and performance inconsistency in federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Delage, E. and Ye, Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

Deng, Y., Kamani, M. M., and Mahdavi, M. Distributionally robust federated averaging. *NeurIPS*, 2020.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Du, W., Xu, D., Wu, X., and Tong, H. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 181–189. SIAM, 2021.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Duchi, J. and Namkoong, H. Variance-based regularization with convex objectives. *NeurIPS*, 2017.

Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pp. 119–133. PMLR, 2018.

Esfahani, P. M. and Kuhn, D. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *NeurIPS*, 29:3315–3323, 2016.

Hong, J., Zhu, Z., Yu, S., Wang, Z., Dodge, H. H., and Zhou, J. Federated adversarial debiasing for fair and transferable representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 617–627, 2021.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pp. 2029–2037. PMLR, 2018.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Kang, J., Xiong, Z., Niyato, D., Xie, S., and Zhang, J. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. *IEEE Internet of Things Journal*, 6(6): 10700–10714, 2019.

Konečnỳ, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016a.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016b.

Li, Q., Zhou, Y., Liang, Y., and Varshney, P. K. Convergence analysis of proximal gradient with momentum for nonconvex optimization. In *International Conference on Machine Learning*, pp. 2111–2119. PMLR, 2017.

Li, T., Sanjabi, M., Beirami, A., and Smith, V. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019a.

Li, T., Beirami, A., Sanjabi, M., and Smith, V. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020.

Li, T., Hu, S., Beirami, A., and Smith, V. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019b.

Liu, J., Shen, Z., Cui, P., Zhou, L., Kuang, K., Li, B., and Lin, Y. Stable adversarial learning under distributional shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8662–8670, 2021.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625. PMLR, 2019.

Mukherjee, D., Yurochkin, M., Banerjee, M., and Sun, Y. Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*, pp. 7097–7107. PMLR, 2020.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NIPS*, volume 29, pp. 2208–2216, 2016.

Nesterov, Y. E. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pp. 543–547, 1983.

Ochs, P. Local convergence of the heavy-ball method and ipiano for non-convex optimization. *Journal of Optimization Theory and Applications*, 177(1):153–180, 2018.

Oren, Y., Sagawa, S., Hashimoto, T. B., and Liang, P. Distributionally robust language modeling. *EMNLP*, 2019.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019.

Rawls, J. *Justice as fairness: A restatement*. Harvard University Press, 2001.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. *ICLR*, 2020.

Sharifi-Malvajerdi, S., Kearns, M., and Roth, A. Average individual fairness: Algorithms, generalization and experiments. *Advances in Neural Information Processing Systems*, 32:8242–8251, 2019.

Sinha, A., Namkoong, H., Volpi, R., and Duchi, J. Certifying some distributional robustness with principled adversarial training. *ICLR*, 2018.

Suresh, A. T., Felix, X. Y., Kumar, S., and McMahan, H. B. Distributed mean estimation with limited communication. In *International Conference on Machine Learning*, pp. 3329–3337. PMLR, 2017.

Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., and Zhou, Y. A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pp. 1–11, 2019.

Wang, Z., Fan, X., Qi, J., Wen, C., Wang, C., and Yu, R. Federated learning with fair averaging. *IJCAI*, 2021.

Wiesemann, W., Kuhn, D., and Sim, M. Distributionally robust convex optimization. *Operations Research*, 62(6): 1358–1376, 2014.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xu, R., Chen, Z., Zuo, W., Yan, J., and Lin, L. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3964–3973, 2018.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

Yu, F., Rawat, A. S., Menon, A., and Kumar, S. Federated learning with only positive labels. In *International Conference on Machine Learning*, pp. 10946–10956. PMLR, 2020.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Zhan, Y., Li, P., Qu, Z., Zeng, D., and Guo, S. A learning-based incentive mechanism for federated learning. *IEEE Internet of Things Journal*, 7(7):6360–6368, 2020.

Zhao, S., Wang, G., Zhang, S., Gu, Y., Li, Y., Song, Z., Xu, P., Hu, R., Chai, H., and Keutzer, K. Multi-source distilling domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12975–12983, 2020.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

The appendix is organized as the following: We discuss related work in Section A. We briefly state the relationship between our proposed fairness notion and other common fairness notions in section B. In section C, we discuss an ideal but impractical method of individual level. We present the details of FMDA-M algorithm in section D. In section E, we give the proof of our proposed theorems. At last, we provide additional experimental results and discussion in section F.

A. Related Work

Fairness in Machine Learning. Fairness in ML has attracted much attention, which can be divided into two branches: *individual fairness* and *group fairness* (Zemel et al., 2013; Awasthi et al., 2020; Binns, 2020). Individual fairness encourages the models to treat similar individuals similarly (Biega et al., 2018; Sharifi-Malvajerdi et al., 2019; Mukherjee et al., 2020), while group fairness requires that the model treats different groups equally (Dwork et al., 2012; 2018). Here we mainly focus the latter one, which is typically defined via some protected attribute(s) and a metric such as statistical parity, equalized odds (Hardt et al., 2016), and predictive parity (Chouldechova, 2017). In this paper, we consider how to learn a fair global model that achieves a uniform performance across groups in FL.

Federated Learning and Fairness. FL has received much attention as an important distributed learning paradigm (McMahan et al., 2017). The scope of federated learning studies is broad, which includes statistical challenges (Zhao et al., 2018; Yu et al., 2020), privacy protection (Truex et al., 2019), systematic challenges (Konečnỳ et al., 2016b;a; Suresh et al., 2017), fairness (Mohri et al., 2019; Li et al., 2019a), etc (Kairouz et al., 2019; Yang et al., 2019). The existing studies of fairness on FL can be divided into three categories: performance fairness across clients (Li et al., 2019a; Mohri et al., 2019; Deng et al., 2020; Li et al., 2021), model fairness defined on sensitive attributes (Du et al., 2021) and incentive mechanism (Kang et al., 2019; Zhan et al., 2020). The most relevant work is (Cui et al., 2021), which aims to address algorithm disparity and performance inconsistency in client level. However, they focus on the client level and train a federated model with similar performance and fairness metrics among different client, while the problem we study involves multiple levels. Specifically, the attribute-level fairness they define is a local fairness notion (in each client), while ours is a global fairness notion. Besides, we also pay attention to out-of-distribution fairness. In this paper, we focus on the performance fairness across clients (including existing clients and newly added clients) and sensitive attribute(s).

Distributionally Robust Optimization. There has been a surge of interest in distributionally robust optimization (DRO), which can deal with distribution shifts by considering a potential distribution set around the original distribution and optimizing the worst case (Delage & Ye, 2010; Wiesemann et al., 2014). There are mainly two definitions of distance between distributions in DRO: f*-divergences* and *Wasserstein distance*. The former method is effective when the support of the distribution is fixed (Duchi & Namkoong, 2017; Namkoong & Duchi, 2016; Duchi & Namkoong, 2021), while Wasserstein distance-based DRO considers the potential distributions with different supports and allows robustness to unseen data, but is difficult to optimize (Sinha et al., 2018; Esfahani & Kuhn, 2018; Liu et al., 2021). Recently, some studies about group DRO have emerged (Hu et al., 2018; Sagawa et al., 2020; Oren et al., 2019), which considers the distribution shifts over groups. In this paper, we extend DRO to FL setting for unified group fairness.

B. Discussion of Fairness Notions

Note that there are a lot of fairness notions. We argue that different notions correspond to different practical meanings, and there is no universally accepted fairness notion. Then we briefly discuss the relationship between our fairness notions and others. First, different from Equalized Odds (EO) and so on, our fairness notion naturally supports multi-attribute & multi-class settings, which is more **practical**. Then, for the binary setting, if we specify the sensitive variable S as sensitive attribute, our notion will degrade to Accuracy Parity. If we specify the S as the combination of target label and sensitive attribute, EO can be viewed as **a relaxed version** of our notion. Therefore, our notion is **general** and **flexible**. We recommend setting S based on expert knowledge in practice.

C. Discussion of Individual-level Approach

Here we discuss an ideal but impractical way for unified group fairness by treating each sample as a group. Then, the framework will degenerate to an individual-level fairness method with risk:

$$\begin{aligned} \mathcal{R}_{\text{individual}}(\theta) &:= \sup_{Q \in \mathcal{Q}^{ind}} \left\{ \mathbb{E}_{(x,y)\sim Q}[\ell(\theta,(x,y))] \right\}, \\ \mathcal{Q}^{ind} &:= \{ Q | D_f\left(Q \| P\right) \leq r \}, \end{aligned} \tag{13}$$

where P is the data-generating distribution of full dataset D, $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex function with $f(1) = 0$, $D_f(Q\|P) = \int_{(\mathcal{X},\mathcal{Y})} f\left(\frac{dQ}{dP}\right) dP$ is *f-divergence* between distribution Q and P defined on $(\mathcal{X}, \mathcal{Y})$ and r is radius of uncertainty set $\mathcal{Q}^{ind}$. Note that the individual-level method is more capable of modeling the agnostic distribution by constructing a wide uncertainty set around distribution P. Intuitively, it may help to guarantee unified group fairness.

Unfortunately, the uncertainty set defined on individual level is usually overwhelmingly large leading to the over pessimism problem in practice (Hu et al., 2018; Sagawa et al., 2020; Liu et al., 2021). In fact, our proposed unified uncertainty set $\mathcal{Q}^u$ can be considered as a subset of $\mathcal{Q}^{ind}$ by imposing some structural constrains. Hence, By contrasting with risk (13), our proposed risk (3) provides a relatively tight upper bound for both client-level risk and attribute-level risk, and help to overcome this pessimism.

D. FMDA-M Algorithm

The details of our FMDA-M algorithm are summarized in Algorithm 1. FMDA-M consists of two main steps in each round: update of model parameters (lines 2 to 15 in Algorithm 1) and update of group weights (lines 16 to 23 in Algorithm 1).

Algorithm 1 FMDA-M algorithm

Input: Number of local iterations E, total number of iterations T, number of rounds $R = T/E$, stepsizes η, γ, momentum coefficients β_θ, $\beta_{\boldsymbol{\lambda}}$, sampling size of clients K, initialized model parameters $\theta^{(0)}$ and weight of k-th group of client i ${\lambda^u_{i,k}}^{(0)}$, $k = 1, 2, \dots, M_i$, $i = 1, 2, \dots, N$.

1: **for** $r = 0, 1, \dots, R-1$ **do**
2: Server samples a subset of clients $U^{(r)} \subset [N]$ with size of K according to probability $\lambda_i^{c(r)} = \sum_{j=1}^{M_i} {\lambda^u_{i,k}}^{(r)}$
3: Server samples t' from $rE+1$ to $(r+1)E$ uniformly
4: Server broadcasts $\theta^{(rE)}$ and ${\lambda^u_{i,k}}^{(r)}$ to corresponding client $i \in U^{(r)}$
5: **for** client $i \in U^{(r)}$ **do**
6: Set local model parameters $\theta_i^{(rE)} = \theta^{(rE)}$
7: **for** $t = rE, rE+1, ..., (r+1)E-1$ **do**
8: Select a group $D^u_{i,k}$ with probability ${\lambda^u_{i,k}}^{(r)} / \lambda_i^{c(r)}$
9: Sample data $\xi_i^{(t)}$ from $D^u_{i,k}$ uniformly
10: Update model: $\theta_i^{(t+1)} = \theta_i^{(t)} - \eta \nabla l\left(\theta_i^{(t)}; \xi_i^{(t)}\right)$
11: **end for**
12: **end for**
13: Client $i \in U^{(r)}$ sends $\theta_i^{((r+1)E)}$ and $\theta_i^{(t')}$ to the server
14: Server computes: $\widetilde{\theta}^{(r+1)E} = \frac{1}{K} \sum_{i \in U^{(r)}} \theta_i^{((r+1)E)}$
15: Server updates global model parameters with momentum: $\theta^{(r+1)E} = \widetilde{\theta}^{(r+1)E} + \beta_\theta(\widetilde{\theta}^{(r+1)E} - \widetilde{\theta}^{rE})$
16: Server computes: $\theta^{(t')} = \frac{1}{K} \sum_{i \in U^{(r)}} \theta_i^{(t')}$
17: Server samples a subset of clients $U_*^{(r)} \subset [N]$ with size of K uniformly
18: Server broadcasts $\theta^{(t')}$ to client $i \in U_*^{(r)}$
19: **for** client $i \in U_*^{(r)}$ do **do**
20: Compute loss $v_{i,k}^{(r)}$ of model $\theta^{(t')}$ on a minibatch of each group $D^u_{i,k}$
21: **end for**
22: Server computes: $(\widetilde{\lambda}^u_{i,k})^{(r+1)} = \frac{(\lambda^u_{i,k})^{(r)} e^{\gamma E v_{i,k}^{(r)}}}{\sum_{i=1}^N \sum_{k=1}^{M_i} (\lambda^u_{i,k})^{(r)} e^{\gamma E v_{i,k}^{(r)}}}$
23: Server updates global weights with momentum: $(\boldsymbol{\lambda}^u)^{(r+1)} = (\boldsymbol{\lambda}^u)^{(r)} + \beta_{\boldsymbol{\lambda}}((\widetilde{\boldsymbol{\lambda}}^u)^{(r+1)} - (\boldsymbol{\lambda}^u)^r)$
24: **end for**
25: **return** $\theta^{(T)}$, ${\lambda^u_{i,k}}^{(R)}$ $(k = 1, 2, \dots, M_i, i = 1, 2, \dots, N)$

E. Proof

E.1. Proof of Theorem 4.1

Proof. For $\forall Q \in \hat{\mathcal{Q}}^c$, it can be expressed as follows:

$$\begin{aligned} Q &= \sum_{i=1}^{N} \lambda_i^c \hat{P}_i^c \\ &= \sum_{i=1}^{N} \sum_{k=1}^{M_i} \lambda_i^c \frac{n_{i,k}^u}{n_i^c} \hat{P}_{i,k}^u, \end{aligned} \tag{14}$$

where $n_{i,k}^u$ is the sample size of group $D_{i,k}^u$, n_i^c is the sample size of local dataset D_i^c, $\sum_i^N \lambda_i^c = 1$ and $\lambda_i^c \geq 0$, $i = 1, 2, \dots, N$. If we let $\lambda_{i,k}^u := \lambda_i^c \frac{n_{i,k}^u}{n_i^c}$, it is obvious that $\sum_{i=1}^N \sum_{k=1}^{M_i} \lambda_i^c \frac{n_{i,k}^u}{n_i^c} = 1$ and $\lambda_i^c \frac{n_{i,k}^u}{n_i^c} \geq 0$, $i = 1, 2, \dots, N$, $k = 1, 2, \dots, M_i$, and thus $Q \in \hat{\mathcal{Q}}^u$. So we have

$$\hat{\mathcal{Q}}^c \subseteq \hat{\mathcal{Q}}^u. \tag{15}$$

Note that the feasible set (uncertainty set) of client-level empirical risk $\hat{\mathcal{R}}_{\text{client}}$ is a subset of that of group-level empirical risk $\hat{\mathcal{R}}_{\text{group}}$, so we have

$$\hat{\mathcal{R}}_{\text{client}}(\theta) \leq \hat{\mathcal{R}}_{\text{unified}}(\theta). \tag{16}$$

Similarly, we also have

$$\hat{\mathcal{Q}}^a \subseteq \hat{\mathcal{Q}}^u, \tag{17}$$

and

$$\hat{\mathcal{R}}_{\text{attribute}}(\theta) \leq \hat{\mathcal{R}}_{\text{unified}}(\theta). \tag{18}$$

Let the assumption in Theorem 4.2 hold: $\exists\, i$ and k, and j ($j \neq i$) s.t. the attribute of samples from $D_{i,k}^u$ is same as $D_{j,k}^u$ but $\hat{P}_{i,k}^u \neq \hat{P}_{j,k}^u$, and $\hat{P}_{i,k}^u \neq \hat{P}_l^c$ for $\forall\, l$. Note that $\hat{P}_{i,k}^u \notin \hat{\mathcal{Q}}^c$, $\hat{P}_{i,k}^u \notin \hat{\mathcal{Q}}^a$ and $\hat{P}_{i,k}^u \in \hat{\mathcal{Q}}^u$. Therefore, we have

$$(\hat{\mathcal{Q}}^c \cup \hat{\mathcal{Q}}^a) \subsetneq \hat{\mathcal{Q}}^u. \tag{19}$$

□

E.2. Relationship between Unified Uncertainty Set $\hat{\mathcal{Q}}^u$ and Individual-level Uncertainty Set $\hat{\mathcal{Q}}^{ind}$

Our proposed unified uncertainty set $\hat{\mathcal{Q}}^u$ can be considered as a subset of $\hat{\mathcal{Q}}^{ind}$ imposing some structural constrains. Compared with individual-level risk, our proposed group-based unified risk provides a relatively tight upper bound for both client-level risk and attribute-level risk. We give theoretical analysis as the following:

Proof. For $\forall Q \in \hat{\mathcal{Q}}^u$, it can be expressed as follows:

$$Q = \sum_{i=1}^{N} \sum_{k=1}^{M_i} \lambda_{i,k}^u \hat{P}_{i,k}^u. \tag{20}$$

Without loss of generality, if we define convex function $f(t)$ as $f(t) := t \cdot \log t$, then we have

$$\begin{aligned} D_f\left(Q \| \hat{P}\right) &= D_f\left(\sum_{i=1}^{N} \sum_{k=1}^{M_i} \lambda_{i,k}^u \hat{P}_{i,k}^u \| \sum_{i=1}^{N} \sum_{k=1}^{M_i} \frac{n_{i,k}^u}{n} \hat{P}_{i,k}^u\right) \\ &= \sum_{i}^{N} \sum_{k}^{M_i} \lambda_{i,k}^u \cdot \log \frac{\lambda_{i,k}^u}{n_{i,k}^u / n}. \end{aligned} \tag{21}$$

So if $\sum_{i=1}^N \sum_{k=1}^{M_i} \lambda_{i,k}^u \log(\frac{n}{n_{i,k}^u} \lambda_{i,k}^u) \leq r$, we have $Q \in \hat{\mathcal{Q}}^{ind}$ and thus

$$\hat{\mathcal{Q}}^u \subseteq \hat{\mathcal{Q}}^{ind}. \tag{22}$$

Note that the feasible set (uncertainty set) of group-level unified empirical risk $\hat{\mathcal{R}}_{\text{unified}}$ is a subset of that of individual-level empirical risk $\hat{\mathcal{R}}_{\text{ind}}$, so we have

$$\hat{\mathcal{R}}_{\text{unified}}(\theta) \leq \hat{\mathcal{R}}_{\text{ind}}(\theta). \tag{23}$$

□

E.3. Proof of Theorem 4.2

Proof. Recall that the group-level risk $\mathcal{R}_{\text{group}}(\theta)$ is defined as

$$\begin{aligned} \mathcal{R}_{\text{group}}(\theta) &:= \sup_{Q \in \mathcal{Q}^g} \left\{ \mathbb{E}_{(x,y)\sim Q}[\ell(\theta, (x,y))] \right\}, \\ \mathcal{Q}^g &:= \{ \sum_{i=1}^{M} \lambda_i^g P_i^g : \boldsymbol{\lambda}^g \in \Delta_{M-1} \}. \end{aligned} \tag{24}$$

With the techniques of distributional robustness optimization (Wiesemann et al., 2014; Duchi & Namkoong, 2017; Namkoong & Duchi, 2016), the problem of minimizing the risk in Eq. (24) can be rewritten as:

$$\begin{aligned} &\min_{\theta \in \Theta} \max_{\lambda_i} \sum_{i=1}^{M} \lambda_i^g \mathcal{R}_i^g(\theta), \\ &s.t. \sum_{i=1}^{M} \lambda_i^g = 1, \lambda_i^g \geq 0. \end{aligned} \tag{25}$$

Inspired by (Duchi & Namkoong, 2017), we introduce an instrumental variable $\boldsymbol{u}$ defined as

$$\boldsymbol{u} := \boldsymbol{\lambda}^g - \frac{1}{K}\mathbf{1}, \tag{26}$$

where $\boldsymbol{\lambda}^g = (\lambda_1^g, \lambda_2^g, \ldots, \lambda_M)$ and $\boldsymbol{u} = (u_1, u_2, \ldots, u_M)$. Then the objective function of Eq. (25) can be rewritten as

$$\begin{aligned} &\sum_{i=1}^{M} \lambda_i^g \mathcal{R}_i^g(\theta) \\ =&\sum_{i=1}^{M} u_i \mathcal{R}_i^g(\theta) + \frac{1}{M}\sum_{i=1}^{M} \mathcal{R}_i^g(\theta) \\ =&\sum_{i=1}^{M} u_i \mathcal{R}_i^g(\theta) + \bar{\mathcal{R}}^g(\theta) \\ =&\sum_{i=1}^{M} u_i (\mathcal{R}_i^g(\theta) - \bar{\mathcal{R}}^g(\theta)) + \bar{\mathcal{R}}^g(\theta). \end{aligned} \tag{27}$$

With Cauchy–Schwarz inequality, we have

$$\begin{aligned} &\sum_{i=1}^{M} u_i (\mathcal{R}_i^g(\theta) - \bar{\mathcal{R}}^g(\theta)) + \bar{\mathcal{R}}^g(\theta) \\ \leq&\sqrt{\sum_{i=1}^{M} u_i^2}\sqrt{\sum_{i=1}^{M} (\mathcal{R}_i^g(\theta) - \bar{\mathcal{R}}^g(\theta))^2 + \bar{\mathcal{R}}^g(\theta)} \\ =&\bar{\mathcal{R}}^g(\theta) + \sqrt{\sum_{i=1}^{M} u_i^2}\sqrt{\mathrm{Var}(\mathcal{R}_i^g(\theta))}. \end{aligned} \tag{28}$$

The equality is attained if and only if

$$u_i = \sqrt{\frac{\|\boldsymbol{u}\|_2^2}{\sum_{i=1}^{M} d_i}} \cdot (\mathcal{R}_i^g(\theta) - \bar{\mathcal{R}}^g(\theta)). \tag{29}$$

Recall that $u_i = \lambda_i^g - \frac{1}{M}$, which requires that for $\forall i$,

$$\sqrt{\frac{\|\boldsymbol{u}\|_2^2}{\sum_{i=1}^M d_i}} \cdot (\mathcal{R}_i^g(\theta) - \bar{\mathcal{R}}^g(\theta)) \geq -\frac{1}{M}. \tag{30}$$

If $\|M\boldsymbol{\lambda}^g - \mathbf{1}\|_2^2 \leq \underset{i}{min}\{\frac{\sum_{i=1}^M d_i}{d_i}\}$, then for $\forall i$, we have

$$\sqrt{\frac{\|\boldsymbol{u}\|_2^2 \cdot d_i}{\sum_{i=1}^M d_i}} \leq \frac{1}{M}, \tag{31}$$

and thus Eq. (30) holds. □

E.4. Proof of Theorem 4.3

We analyze the convergence rate of Algorithm 1 by bounding the error ε_T defined as:

$$\varepsilon_T = \max_{\boldsymbol{\lambda}^u} \mathbb{E}[F(\bar{\theta}^{(T)}, \boldsymbol{\lambda}^u)] - \min_{\theta} \mathbb{E}[F(\theta, \bar{\boldsymbol{\lambda}^u}^{(T)})], \tag{32}$$

where T is the number of total iterations, $\bar{\theta}^{(T)}$ is the average of global model parameters of T iterations and $\bar{\boldsymbol{\lambda}^u}^{(T)}$ is the average of group weights of T iterations. Before it, we first introduce some technical lemmas.

Lemma E.1. *The stochastic gradient $\boldsymbol{u}^{(t)}$ defined as*

$$\begin{aligned}\boldsymbol{u}^{(t)} :=& \frac{1}{K} \sum_{i \in U^{(\lfloor \frac{t}{E} \rfloor)}} \nabla f_i(\theta_i^{(t)}; \xi_i^{(t)}) \\ =& \frac{1}{K} \sum_{i \in U^{(\lfloor \frac{t}{E} \rfloor)}} \sum_{j=1}^{M_i} \frac{{\lambda_{i,k}^u}^{(\lfloor \frac{t}{E} \rfloor)}}{\lambda_i^{c(\lfloor \frac{t}{E} \rfloor)}} \nabla f_{i,k}(\theta_i^{(t)}; \xi_i^{(t)})\end{aligned} \tag{33}$$

is unbiased, and its variance is bounded, which implies:

$$\begin{aligned}&\mathbb{E}_{\xi_i^{(t)}, U^{(\lfloor \frac{t}{E} \rfloor)}} \left[\frac{1}{K} \sum_{i \in U^{(\lfloor \frac{t}{E} \rfloor)}} \sum_{j=1}^{M_i} \frac{{\lambda_{i,k}^g}^{(\lfloor \frac{t}{E} \rfloor)}}{\lambda_i^{c(\lfloor \frac{t}{E} \rfloor)}} \nabla f_{i,k}(\theta_i^{(t)}; \xi_i^{(t)}) \right] \\ =& \mathbb{E}_{U^{(\lfloor \frac{t}{E} \rfloor)}} \left[\bar{\boldsymbol{u}}^{(t)} := \frac{1}{K} \sum_{i \in U^{(\lfloor \frac{t}{E} \rfloor)}} \sum_{j=1}^{M_i} \frac{{\lambda_{i,k}^u}^{(\lfloor \frac{t}{E} \rfloor)}}{\lambda_i^{c(\lfloor \frac{t}{E} \rfloor)}} \nabla f_{i,k}(\theta_i^{(t)}) \right] \\ =& \mathbb{E} \left[\sum_{i=1}^N \sum_{j=1}^{M_i} {\lambda_{i,k}^g}^{(\lfloor \frac{t}{E} \rfloor)} \nabla f_{i,k}(\theta_i^{(t)}) \right],\end{aligned} \tag{34}$$

$$\mathbb{E}\left[\|\boldsymbol{u}^{(t)} - \bar{\boldsymbol{u}}^{(t)}\|_2^2 \right] \leq \frac{B^2}{K}, \tag{35}$$

where $B > 0$ is a constant bound.

Proof. The stochastic gradient $\boldsymbol{u}^{(t)}$ is unbiased due to the fact that we sample the groups according to $\boldsymbol{\lambda}^{u(\lfloor \frac{t}{E} \rfloor)}$. The variance term is due to the assumption in Theorem 5.1. □

Inspired by (Li et al., 2019b), we introduce the gradient dissimilarity Γ defined as

$$\Gamma := \sup_{\theta, \boldsymbol{p} \in \Delta_{N-1}, i \in [N]} \sum_{j=1}^N p_j \|\nabla f_i(\theta) - \nabla f_j(\theta)\|_2^2, \tag{36}$$

where $f_i(\theta)$ is the local objective of client i.

Lemma E.2. *Define $\delta^{(t)} := \frac{1}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}} \|\theta_i^{(t)} - \theta^{(t)}\|_2^2$. For FMDA-M, the expected average squared norm distance of local models $\theta_i^{(t)}, i \in U^{(\lfloor \frac{t}{E}\rfloor)}$ and $\theta^{(t)}$ is bounded as follows:*

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\left[\delta^{(t)}\right] \leq 10\eta^2E^2\left(B^2 + \frac{B^2}{K} + \Gamma\right). \tag{37}$$

where expectation is taken over sampling of devices at each iteration.

Proof. Considering $rE \leq t \leq (r+1)E$, we have:

$$\begin{aligned}
&\mathbb{E}[\delta^{(t)}] \\
=&\mathbb{E}[\frac{1}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}} \|\theta_i^{(t)} - \theta^{(t)}\|_2^2] \\
\leq&\mathbb{E}[\frac{1}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}} \mathbb{E}\|\theta^{(rE)} - \sum_{s=rE}^{t-1}\eta\nabla f_i(\theta_i^{(s)};\xi_i^{(s)}) - \left(\theta^{(rE)} - \frac{1}{K}\sum_{i'\in U}\sum_{s=rE}^{t-1}\eta\nabla f_{i'}(\theta_{i'}^{(s)};\xi_{i'}^{(s)})\right)\|_2^2] \\
=&\mathbb{E}[\frac{1}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}} \|\sum_{s=rE}^{t-1}\eta\nabla f_i(\theta_i^{(s)};\xi_i^{(s)}) - \frac{1}{K}\sum_{i'\in U^{(\lfloor \frac{t}{E}\rfloor)}}\sum_{s=rE}^{t-1}\eta\nabla f_{i'}(\theta_{i'}^{(s)};\xi_{i'}^{(s)})\|_2^2] \\
\leq&\mathbb{E}[\frac{1}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}} \eta^2 E\sum_{s=rE}^{(r+1)E}\|\nabla f_i(\theta_i^{(s)};\xi_i^{(s)}) - \frac{1}{K}\sum_{i'\in U^{(\lfloor \frac{t}{E}\rfloor)}}\nabla f_{i'}(\theta_{i'}^{(s)};\xi_{i'}^{(s)})\|_2^2] \\
=&\eta^2E\mathbb{E}[\frac{1}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}}\sum_{s=rE}^{(r+1)E}\|\nabla f_i(\theta_i^{(s)};\xi_i^{(s)}) - \nabla f_i(\theta_i^{(s)}) + \nabla f_i(\theta_i^{(s)}) - \nabla f_i(\theta^{(s)}) + \nabla f_i(\theta^{(s)}) \\
&-\frac{1}{K}\sum_{i'\in U^{(\lfloor \frac{t}{E}\rfloor)}}\nabla f_{i'}(\theta^{(s)}) + \frac{1}{K}\sum_{i'\in U^{(\lfloor \frac{t}{E}\rfloor)}}\nabla f_{i'}(\theta^{(s)}) - \frac{1}{K}\sum_{i'\in U^{(\lfloor \frac{t}{E}\rfloor)}}\nabla f_{i'}(\theta_{i'}^{(s)}) \\
&+\frac{1}{K}\sum_{i'\in U^{(\lfloor \frac{t}{E}\rfloor)}}\nabla f_{i'}(\theta_{i'}^{(s)}) - \frac{1}{K}\sum_{i'\in U^{(\lfloor \frac{t}{E}\rfloor)}}\nabla f_{i'}(\theta_{i'}^{(s)};\xi_{i'}^{(s)})\|_2^2]
\end{aligned} \tag{38}$$

Using Jensen's inequality, we have:

$$\begin{aligned}
\mathbb{E}[\delta^{(t)}] \leq& 5\eta^2E\sum_{s=rE}^{(r+1)E}(B^2 + L^2\mathbb{E}[\frac{1}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}}\left\|\theta_i^{(s)} - \theta^{(s)}\right\|^2] + L^2\mathbb{E}[\frac{1}{K}\sum_{i'\in U^{(\lfloor \frac{t}{E}\rfloor)}}\left\|\theta_{i'}^{(s)} - \theta^{(s)}\right\|_2^2] \\
&+\mathbb{E}[\frac{1}{K}\sum_{i'\in U^{(\lfloor \frac{t}{E}\rfloor)}}\left\|\nabla f_i(\theta^{(s)}) - \nabla f_{i'}(\theta^{(s)})\right\|_2^2] + \frac{B^2}{K}) \\
\leq& 5\eta^2E\sum_{s=rE}^{(r+1)E}(B^2 + 2L^2\mathbb{E}[\delta^{(s)}] + \Gamma + \frac{B^2}{K}).
\end{aligned} \tag{39}$$

Then we sum the above equation over $t = rE$ to $(r+1)E$ to get:

$$\begin{aligned}
\sum_{t=rE}^{(r+1)E}\mathbb{E}[\delta^{(t)}] \leq& 5\eta^2E\sum_{t=rE}^{(r+1)E}\sum_{s=rE}^{(r+1)E}\left(B^2 + 2L^2\mathbb{E}[\delta^{(s)}] + \Gamma + \frac{B^2}{K}\right) \\
=& 5\eta^2E^2\sum_{s=rE}^{(r+1)E}\left(B^2 + 2\mathbb{E}[\delta^{(s)}] + \Gamma + \frac{B^2}{K}\right) \\
\leq& 10\eta^2E^2\sum_{s=rE}^{(r+1)E}\left(B^2 + \Gamma + \frac{B^2}{K}\right).
\end{aligned} \tag{40}$$

Now we sum the above equation over $r = 0$ to $R-1$, and we have:

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\left[\delta^{(t)}\right] \leq 10\eta^2 E^2\left(B^2 + \frac{B^2}{K} + \Gamma\right). \tag{41}$$

□

Lemma E.3. *For FMDA-M, under the same conditions as in Theorem 5.1, for all θ, we have:*

$$\begin{aligned}\mathbb{E}\|\theta^{(t+1)} - \theta\|_2^2 \leq& \mathbb{E}\|\theta^{(t)} - \theta\|_2^2 - 2\eta\mathbb{E}\left[F(\theta^{(t)}, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E}\rfloor)}) - F(\theta, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E}\rfloor)})\right]\\ &+ L\eta\mathbb{E}\left[\delta^{(t)}\right] + \eta^2\mathbb{E}\|\bar{\boldsymbol{u}}^{(t)} - \boldsymbol{u}^{(t)}\|_2^2 + \eta^2 B^2 + \mathbb{E}\|\beta_\theta(\widetilde{\theta}^{(t+1)} - \widetilde{\theta}^t)\|_2^2.\end{aligned} \tag{42}$$

Proof. According to the stochastic gradient method, we have

$$\begin{aligned}\mathbb{E}\|\theta^{(t+1)} - \theta\|_2^2 =& \mathbb{E}\left\|\theta^{(t)} + \beta_\theta(\widetilde{\theta}^{(t+1)} - \widetilde{\theta}^t) - \eta\boldsymbol{u}^{(t)} - \theta\right\|_2^2\\ \leq& \mathbb{E}\|\theta^{(t)} - \eta\bar{\boldsymbol{u}}^{(t)} - \theta\|_2^2 + \eta^2\mathbb{E}\|\bar{\boldsymbol{u}}^{(t)} - \boldsymbol{u}^{(t)}\|_2^2 + \mathbb{E}\|\beta_\theta(\widetilde{\theta}^{(t+1)} - \widetilde{\theta}^t)\|_2^2\\ \leq& \mathbb{E}\|\theta^{(t)} - \theta^*\|_2^2 + \mathbb{E}[-2\eta\langle\bar{\boldsymbol{u}}^{(t)}, \theta^{(t)} - \theta^*\rangle] + \eta^2\mathbb{E}\|\bar{\boldsymbol{u}}^{(t)}\|_2^2\\ &+ \mathbb{E}\|\bar{\boldsymbol{u}}^{(t)} - \boldsymbol{u}^{(t)}\|_2^2 + \mathbb{E}\|\beta_\theta(\widetilde{\theta}^{(t+1)} - \widetilde{\theta}^t)\|_2^2.\end{aligned} \tag{43}$$

We first bound the second term in Eq. (43) by the properties of smoothness and convexity:

$$\begin{aligned}\mathbb{E}[-2\eta\langle\bar{\boldsymbol{u}}^{(t)}, \theta^{(t)} - \theta^*\rangle] =& \mathbb{E}_{U^{(\lfloor \frac{t}{E}\rfloor)}}\left[\frac{1}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}}(-2\eta\left\langle\nabla f_i(\theta_i^{(t)}), \theta^{(t)} - \theta_i^{(t)}\right\rangle)\right]\\ &+ \mathbb{E}_{U^{(\lfloor \frac{t}{E}\rfloor)}}\left[\frac{1}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}}(-2\eta\left\langle\nabla f_i(\theta_i^{(t)}), \theta_i^{(t)} - \theta^*\right\rangle)\right]\\ \leq& \mathbb{E}_{U^{(\lfloor \frac{t}{E}\rfloor)}}\left[\frac{2\eta}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}}\left[f_i(\theta_i^{(t)}) - f_i(\theta^{(t)})\right]\right]\\ &+ \mathbb{E}_{U^{(\lfloor \frac{t}{E}\rfloor)}}\left[\frac{2\eta}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}}\left[\frac{L}{2}\|\theta^{(t)} - \theta_i^{(t)}\|_2^2\right]\right]\\ &+ \mathbb{E}_{U^{(\lfloor \frac{t}{E}\rfloor)}}\left[\frac{2\eta}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}}\left[f_i(\theta) - f_i(\theta_i^{(t)})\right]\right]\\ =& \mathbb{E}_{U^{(\lfloor \frac{t}{E}\rfloor)}}\left[\frac{2\eta}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}}\left[f_i(\theta) - f_i(\theta^{(t)})\right]\right] + \mathbb{E}_{U^{(\lfloor \frac{t}{E}\rfloor)}}\left[\frac{2\eta}{K}\sum_{i\in U^{(\lfloor \frac{t}{E}\rfloor)}}\left[\frac{L}{2}\|\theta^{(t)} - \theta_i^{(t)}\|_2^2\right]\right]\\ =& -2\eta\mathbb{E}\left[\sum_{i=1}^{N}\lambda_i^{(\lfloor \frac{t}{E}\rfloor)}f_i(\theta^{(t)}) - \lambda_i^{(\lfloor \frac{t}{E}\rfloor)}f_i(\theta)\right] + L\eta\mathbb{E}\left[\delta^{(t)}\right]\\ =& -2\eta\mathbb{E}\left[F(\theta^{(t)}, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E}\rfloor)}) - F(\theta, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E}\rfloor)})\right] + L\eta\mathbb{E}\left[\delta^{(t)}\right].\end{aligned} \tag{44}$$

Then we bound the third term in Eq. (43) as follows:

$$
\begin{aligned}
&\eta^2\mathbb{E}\|\bar{\boldsymbol{u}}^{(t)}\|_2^2 \\
=&\eta^2\mathbb{E}\|\frac{1}{K}\sum_{i\in U^{(\lfloor\frac{t}{E}\rfloor)}}\sum_{j=1}^{M_i}\frac{\lambda_{i,k}^{u}{}^{(\lfloor\frac{t}{E}\rfloor)}}{\lambda_i^{c(\lfloor\frac{t}{E}\rfloor)}}\nabla f_{i,k}(\theta_i^{(t)})\|_2^2 \\
=&\eta^2\mathbb{E}\left\|\frac{1}{K}\sum_{i\in U^{(\lfloor\frac{t}{E}\rfloor)}}\nabla f_i(\theta_i^{(t)})\right\|_2^2 \\
\leq&\eta^2\frac{1}{K}\sum_{i\in U^{(\lfloor\frac{t}{E}\rfloor)}}\mathbb{E}\left\|\nabla f_i(\theta_i^{(t)})\right\|_2^2 \\
\leq&\eta^2B^2.
\end{aligned}
\tag{45}
$$

By plugging Eq. (44), Eq. (45) and Eq. (51) back to Eq. (43), we have:

$$
\begin{aligned}
\mathbb{E}\|\theta^{(t+1)}-\theta\|_2^2 \leq&\mathbb{E}\|\theta^{(t)}-\theta\|_2^2-2\eta\mathbb{E}\left[F(\theta^{(t)},\boldsymbol{\lambda}^{u(\lfloor\frac{t}{E}\rfloor)})-F(\theta,\boldsymbol{\lambda}^{u(\lfloor\frac{t}{E}\rfloor)})\right] \\
&+L\eta\mathbb{E}\left[\delta^{(t)}\right]+\eta^2\mathbb{E}\|\bar{\boldsymbol{u}}^{(t)}-\boldsymbol{u}^{(t)}\|_2^2+\eta^2B^2+\mathbb{E}\|\beta_\theta(\widetilde{\theta}^{(t+1)}-\widetilde{\theta}^{t})\|_2^2.
\end{aligned}
\tag{46}
$$

□

Lemma E.4. *The stochastic gradient at $\boldsymbol{\lambda}^u$ generated by Algorithm 1 is unbiased, and its variance is bounded, which implies:*

$$
\mathbb{E}[E\gamma\boldsymbol{v}]=\sum_{t=rE+1}^{(r+1)E}\gamma\nabla_{\boldsymbol{\lambda}^u}F(\theta^{(t)},\boldsymbol{\lambda}^u),
\tag{47}
$$

$$
\mathbb{E}[\|E\gamma\boldsymbol{v}-\sum_{t=rE+1}^{(r+1)E}\gamma\nabla_{\boldsymbol{\lambda}^u}F(\theta^{(t)},\boldsymbol{\lambda}^u)\|_2^2]\leq\gamma^2E^2\frac{B^2}{K},
\tag{48}
$$

where $B>0$ is a constant bound.

Proof. The stochastic gradient at $\boldsymbol{\lambda}^u$ is unbiased due to we sample the groups uniformly. The variance term is due to the assumption in Theorem 5.1. □

Lemma E.5. *For FMDA-M, under the assumption of Theorem 5.1, assuming the function h is an α-strongly convex function, then the following holds true for any $\boldsymbol{\lambda}^u\in\Delta_{m-1}$:*

$$
\begin{aligned}
\mathbb{E}[D_h(\boldsymbol{\lambda}^u||(\boldsymbol{\lambda}^u)^{(r+1)})]\leq&\mathbb{E}[D_h(\boldsymbol{\lambda}^u||(\boldsymbol{\lambda}^u)^{(r)})]-\sum_{t=rE+1}^{(r+1)E}\mathbb{E}[2\gamma(F(\theta^{(t)},\boldsymbol{\lambda}^{u(\lfloor\frac{t}{E}\rfloor)})-F(\theta^{(t)},\boldsymbol{\lambda}^g))] \\
&+\frac{\gamma}{2\alpha}\mathbb{E}\|\sum_{t=rE+1}^{(r+1)E}\nabla_{\boldsymbol{\lambda}^u}F(\theta^{(t)},\boldsymbol{\lambda}^u)\|_2^2+\mathbb{E}\|E\gamma\boldsymbol{v}^{(r)}-\sum_{t=rE+1}^{(r+1)E}\gamma\nabla_{\boldsymbol{\lambda}^u}F(\theta^{(t)},\boldsymbol{\lambda}^u)\|_2^2.
\end{aligned}
\tag{49}
$$

Proof. The proof is based on the update rule of mirror ascent and others are similar to Lemma E.1. □

Note that we can pick the right function h to make the assumption hold.

Now we can give the proof of Theorem 5.1.

Proof. With above lemmas, we can prove the Theorem 5.1. By the convexity of global function w.r.t. θ and its linearity in terms of $\boldsymbol{\lambda}^u$, we have:

$$
\begin{aligned}
&\mathbb{E}[F(\bar{\theta}, \boldsymbol{\lambda}^u) - \mathbb{E}[F(\theta, \bar{\boldsymbol{\lambda}}^u)] \\
\leq& \frac{1}{T}\sum_{t=1}^{T}\left\{\mathbb{E}\left[F(\theta^{(t)}, \boldsymbol{\lambda}^u)\right] - \mathbb{E}\left[F(\theta, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E} \rfloor)})\right]\right\} \\
\leq& \frac{1}{T}\sum_{t=1}^{T}\{\mathbb{E}\left[F(\theta^{(t)}, \boldsymbol{\lambda}^u)\right] - \mathbb{E}\left[F(\theta^{(t)}, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E} \rfloor)})\right] + \mathbb{E}\left[F(\theta^{(t)}, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E} \rfloor)})\right] - \mathbb{E}\left[F(\theta, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E} \rfloor)})\right]\} \\
\leq& \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{F(\theta^{(t)}, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E} \rfloor)}) - F(\theta, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E} \rfloor)})\} + \frac{1}{T}\sum_{r=0}^{R-1}\sum_{t=rE+1}^{(r+1)E}\mathbb{E}\{F(\theta^{(t)}, \boldsymbol{\lambda}^u) - F(\theta^{(t)}, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E} \rfloor)})\}.
\end{aligned}
\tag{50}
$$

We first bound the first term in Eq. (50). Sum the last term in Eq. (43) over $t = 0$ to $T - 1$ to get:

$$
\begin{aligned}
\sum_{t=0}^{T-1}\mathbb{E}\|\beta_\theta(\widetilde{\theta}^{(t+1)} - \widetilde{\theta}^{(t)})\|_2^2 =& \sum_{t=0}^{T-1}\mathbb{E}\|\beta_\theta^{t+1}(\widetilde{\theta}^{(1)} - \widetilde{\theta}^{(0)})\|_2^2 \\
\leq& \frac{\beta_\theta^2(1-\beta_\theta^{2T})}{1-\beta_\theta^2}\mathbb{E}\|\widetilde{\theta}^{(1)} - \widetilde{\theta}^{(0)}\|_2^2 \\
\leq& B
\end{aligned}
\tag{51}
$$

Then we plug Lemma E.1 and E.2 into Lemma E.3 and sum over $t = 1$ to T to get:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(F(\theta^{(t)}, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E} \rfloor)}) - F(\theta, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E} \rfloor)})) \leq& \frac{1}{2T\eta}\mathbb{E}\|\,\theta^{(0)} - \theta\|^2 + 5L\eta^2E^2\left(B^2 + \frac{B^2}{K} + \Gamma\right) + \frac{\eta B^2}{2} + \frac{\eta B^2}{2K} + \frac{B}{T} \\
\leq& \frac{D_{\mathcal{W}}^2}{2T\eta} + 5L\eta^2E^2\left(B^2 + \frac{B^2}{K} + \Gamma\right) + \frac{\eta B^2}{2} + \frac{\eta B^2}{2K} + \frac{B}{T}.
\end{aligned}
\tag{52}
$$

To bound the second term in Eq. (50), plugging Lemma E.4 into Lemma E.5, we have:

$$
\begin{aligned}
\frac{1}{T}\sum_{r=0}^{R-1}\sum_{t=rE+1}^{(r+1)E}\mathbb{E}(F(\theta^{(t)}, \boldsymbol{\lambda}^u) - F(\theta^{(t)}, \boldsymbol{\lambda}^{u(\lfloor \frac{t}{E} \rfloor)})) \leq& \frac{1}{\gamma T}D_h(\boldsymbol{\lambda}^u||(\boldsymbol{\lambda}^u)^{(0)}) + \frac{\gamma E}{2}B^2 + \frac{\gamma E B^2}{2K} \\
\leq& \frac{B^2}{\gamma T} + \frac{\gamma E B^2}{2} + \frac{\gamma E B^2}{2K}.
\end{aligned}
\tag{53}
$$

Taking max over $\boldsymbol{\lambda}^u$, min over θ, we have

$$
\begin{aligned}
\min_{\theta}\max_{\boldsymbol{\lambda}^u \in \Delta_{m-1}}\mathbb{E}[F(\bar{\theta}, \boldsymbol{\lambda}^u) - \mathbb{E}[F(\theta, \bar{\boldsymbol{\lambda}^u})] \leq& \frac{B^2}{2T\eta} + 5L\eta^2E^2\left(B^2 + \frac{B^2}{K} + \Gamma\right) + \frac{\eta B^2}{2} \\
&+ \frac{\eta B^2}{2K} + \frac{B}{T} + \frac{B^2}{\gamma T} + \frac{\gamma E B^2}{2} + \frac{\gamma E B^2}{2K}.
\end{aligned}
\tag{54}
$$

Plugging in $E = O(T^{\frac{1}{4}})$, $\eta = O(T^{-\frac{1}{2}})$, and $\gamma = O(T^{-\frac{1}{2}})$, we complete the proof.

$$
\max_{\boldsymbol{\lambda}^u \in \Delta_{m-1}}\mathbb{E}[F(\bar{\boldsymbol{w}}, \boldsymbol{\lambda}^u)] - \min_{\theta}\mathbb{E}[F(\boldsymbol{w}, \bar{\boldsymbol{\lambda}^u})] \leq O(T^{-\frac{1}{2}}),
\tag{55}
$$

$\square$

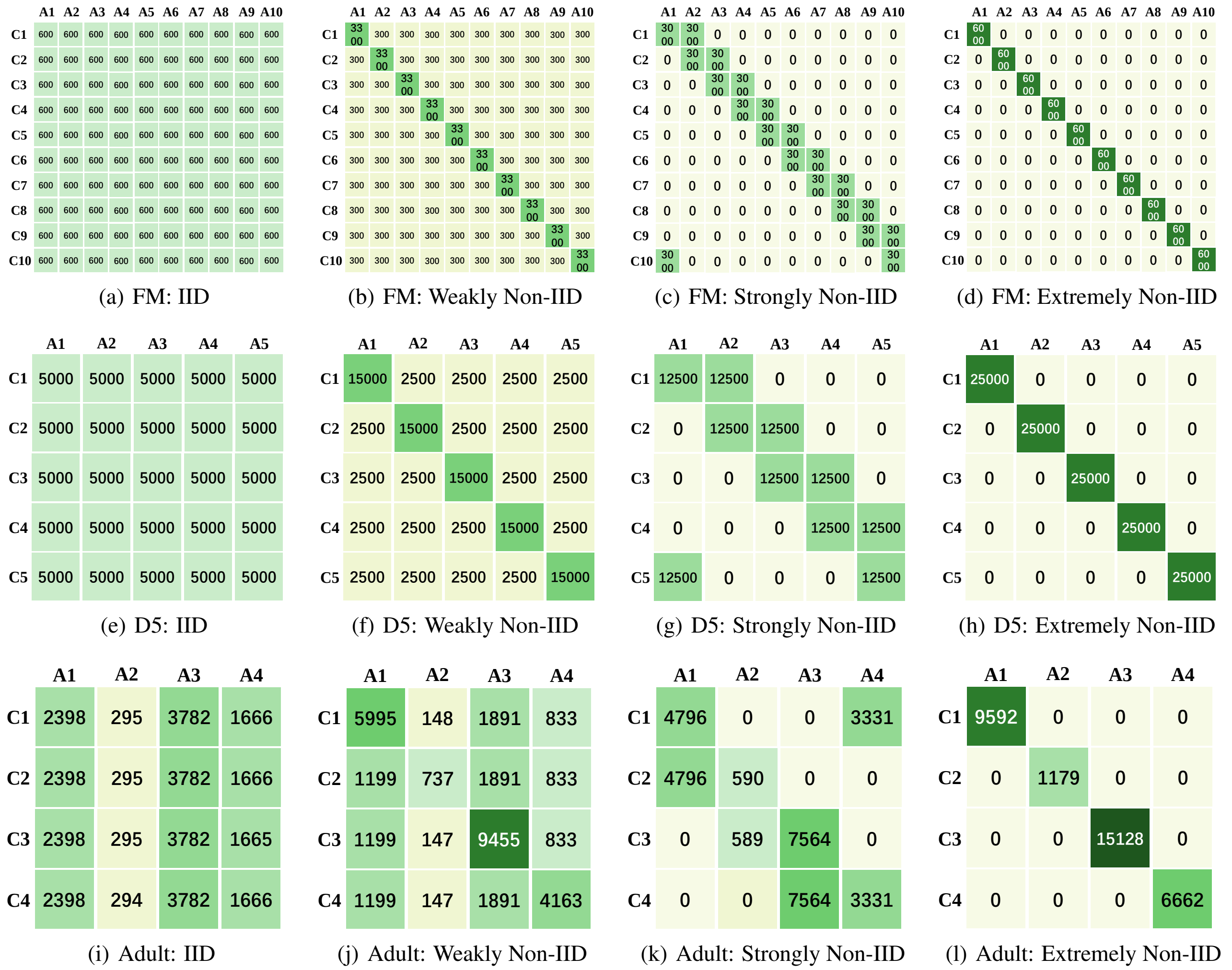


Figure 2. Various training distribution settings on Fashion-MNIST (FM), Digit-Five (D5) and Adult datasets. The numbers are sample sizes of each subgroup.

F. Additional Experimental Results

F.1. Additional Federated Datasets.

(1) Fashion-MNIST (FM) dataset (Xiao et al., 2017): FM is a classical image classification dataset containing 60,000 training examples with 10 categories. For FM, we set the target label as the sensitive attribute and consider ***label distribution shift*** across clients. As shown in Figure 2, we consider 4 different degrees of Non-IID settings: (a) IID, (b) weakly Non-IID, (c) strongly Non-IID, (d) extremely Non-IID. We run our algorithm and compared baselines on FM dataset with logistic regression model. (2) Digit-Five (D5) dataset (Xu et al., 2018; Peng et al., 2019; Zhao et al., 2020): D5 includes digit images with 10 categories sampled from 5 domains. For D5, we set the domain (i.e. data collection source) as the sensitive attribute and consider 4 different degrees of ***feature distribution shift*** across clients. We use a 2-layer CNN with a linear classifier. (3) UCI Adult dataset(Dua & Graff, 2017): Adult is a census dataset with 32,561 examples, and each sample has 14 features (including race, gender and so on) and a target label indicating whether the income is greater or less than \$50K. For Adult, we set gender and income as the protected attributes and consider 4 different degrees of ***unbalance*** setting (i.e. the amount of data varies greatly across both clients and attributes), which is hard to avoid in real FL scenario. We use a logistic regression model to predict the income.

Table 2. Experimental results of attribute-level fairness.

	Metrics	Robustness: $Robustness$ (%)				Fairness: $Disparity$			
Dataset	Method	IID	Weakly Non-IID	Strongly Non-IID	Extremely Non-IID	IID	Weakly Non-IID	Strongly Non-IID	Extremely Non-IID
Fashion MNIST	Centralized ERM	56.67±0.30				0.122±0.001			
	Centralized DRO	**68.98±0.35**				**0.081±0.002**			
	FedAvg	55.27±0.33	53.41±0.44	41.63±0.28	40.60±0.20	0.121±0.001	0.123±0.003	0.151±0.003	0.152±0.003
	DRFA	61.90±0.53	61.50±0.87	56.11±0.44	65.63±0.47	0.107±0.001	0.109±0.002	0.119±0.003	0.101±0.001
	IndA	55.04±0.85	58.46±0.99	45.91±0.81	51.35±0.61	0.139±0.006	0.136±0.006	0.145±0.006	0.130±0.007
	FMDA-M (Ours)	**68.06±0.42**	**66.60±0.68**	**68.31±0.38**	**67.72±0.23**	**0.082±0.002**	**0.086±0.002**	**0.085±0.002**	**0.087±0.002**
Digit Five	Centralized ERM	83.07±0.48				0.059±0.001			
	Centralized DRO	**84.48±0.47**				**0.046±0.002**			
	FedAvg	81.72±0.22	81.13±0.35	81.60±0.33	78.51±0.64	0.063±0.002	0.065±0.002	0.053±0.002	0.073±0.003
	DRFA	81.22±0.34	82.02±0.33	82.05±0.50	81.08±0.32	0.064±0.001	0.062±0.001	0.058±0.001	0.060±0.001
	IndA	81.14±0.28	81.03±0.38	81.14±0.21	78.99±0.51	0.067±0.003	0.067±0.002	0.065±0.002	0.070±0.004
	FMDA-M (Ours)	**85.10±0.23**	**83.91±0.24**	**84.20±0.22**	**81.98±0.36**	**0.043±0.001**	**0.044±0.002**	**0.054±0.001**	**0.051±0.001**
Adult	Centralized ERM	20.85±0.47				0.318±0.003			
	Centralized DRO	**70.42±0.56**				**0.031±0.003**			
	FedAvg	20.26±0.29	25.82±0.49	20.70±0.42	66.04±0.36	0.322±0.002	0.290±0.002	0.323±0.003	0.063±0.001
	DRFA	20.19±0.32	31.15±0.55	20.62±0.55	67.85±0.61	0.322±0.003	0.264±0.003	0.324±0.003	0.046±0.002
	IndA	21.60±0.58	25.69±0.61	21.90±0.63	65.21±0.68	0.315±0.004	0.292±0.003	0.324±0.005	0.079±0.004
	FMDA-M (Ours)	**70.20±0.35**	**71.16±0.41**	**70.74±0.44**	**70.57±0.48**	**0.031±0.002**	**0.030±0.002**	**0.031±0.002**	**0.032±0.001**

F.2. Additional Training Distribution Setting

Fashion-MNIST. For Fashion-MNIST, we set the target label as the sensitive attribute, which can take on 10 values. As shown in the first row of Figure 2, we consider *label distribution shift* across clients and split the training dataset into 10 clients in 4 manners: (a) IID, (b) weakly Non-IID, (c) strongly Non-IID, (d) extremely Non-IID.

Digit-Five. For Digit-Five, we set the domain (i.e. data collection source) as the sensitive attribute, which can take on 5 values. As shown in the second row of Figure 2, we consider *feature distribution shift* across clients and split the training dataset into 5 clients in 4 manners: (e) IID, (f) weakly Non-IID, (g) strongly Non-IID, (h) extremely Non-IID.

Adult. For Adult, we set the combination of gender and income (target label) as the sensitive attribute, which can take on 4 values: high-income male, low-income male, high-income female, low-income female. As shown in the third row of Figure 2, we consider both *label distribution shift*, *feature distribution shift* and *unbalance* across clients and split the training dataset into 4 clients in 4 manners: (i) IID, (j) weakly Non-IID, (k) strongly Non-IID, (l) extremely Non-IID.

F.3. Hyper-parameter Setting

Fashion-MNIST. We use a logistic regression model to predict the target of images. We set the number of local iterations $E = 10$, the number of rounds $R = 500$, batchsize is 50, learning rate for model parameters η=1e-2, stepsize for group weights γ=1e-2 and momentum coefficients $\beta_\theta = \beta_{\boldsymbol{\lambda}} = 0.4$.

Digit-Five. We use a 2-layer CNN with a linear classifier to classify the images. We set the number of local iterations $E = 10$, the number of rounds $R = 500$, batchsize is 50, learning rate for model parameters η=2e-2 and stepsize for group weights γ=2e-2 and momentum coefficients $\beta_\theta = \beta_{\boldsymbol{\lambda}} = 0.4$.

Adult. We use a logistic regression model to predict the income. We set the number of local iterations $E = 10$, the number of rounds $R = 500$, batchsize is 50, learning rate for model parameters η=1e-2 and stepsize for group weights γ=1e-2 and momentum coefficients $\beta_\theta = \beta_{\boldsymbol{\lambda}} = 0.4$.

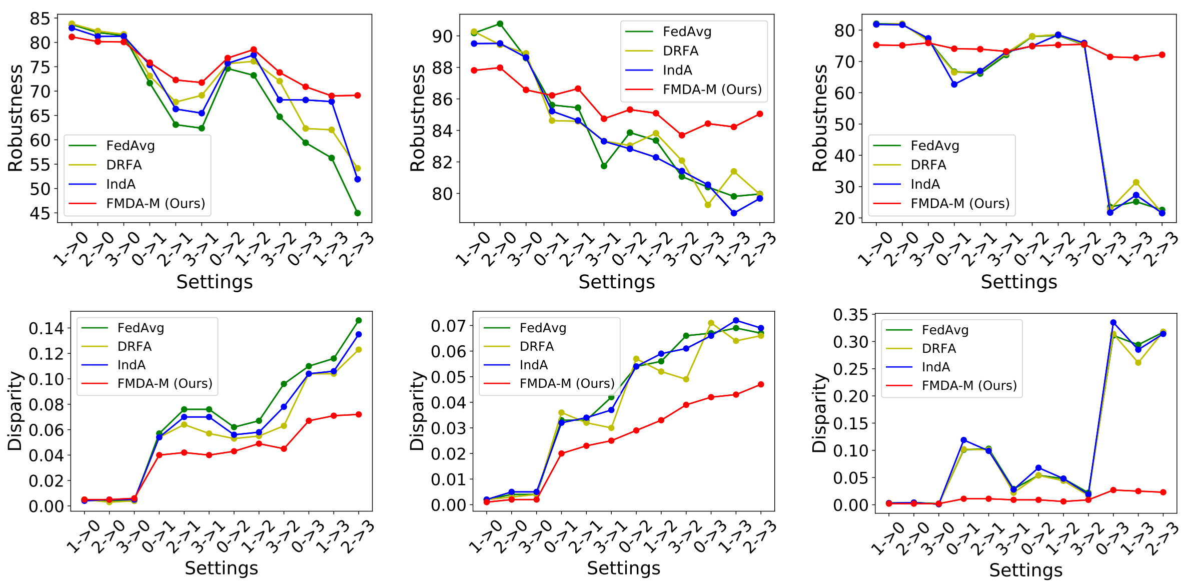
F.4. Results of Attribute-level Fairness

We evaluate the attribute-level fairness of models trained by FMDA-M and compared baselines. The results are reported in Table 2. From the results, we observe that FMDA-M outperforms baselines on three datasets, in terms of both the metric $Disparity$ and $Robustness$. As we analyzed in the previous section, the client-level method is not flexible enough to deal with distribution shifts over attributes, and the individual-level method is too conservative to perform well in practice. By constructing an appropriate uncertainty set, FMDA-M achieves good performance which is very similar to centralized DRO, even in Non-IID settings.

The results on Adult dataset demonstrate that the unbalance of dataset is a great challenge for training a fair model, especially

Table 3. Experimental results of client-level fairness.

	Metrics	Robustness: *Robustness* (%)				Fairness: *Disparity*			
Dataset	Method	IID	Weakly Non-IID	Strongly Non-IID	Extremely Non-IID	IID	Weakly Non-IID	Strongly Non-IID	Extremely Non-IID
Fashion MNIST	FedAvg	84.34±0.47	70.24±0.64	65.92±0.47	44.04±0.38	**0.004±0.001**	0.059±0.002	0.091±0.006	0.146±0.004
	DRFA	**84.39±0.45**	71.25±0.88	70.93±0.80	66.54±0.68	0.004±0.001	0.055±0.003	0.064±0.002	0.099±0.002
	IndA	83.82±0.71	72.75±1.03	69.69±1.09	53.00±1.06	0.004±0.001	0.057±0.003	0.081±0.008	0.125±0.006
	FMDA-M (Ours)	81.05±0.36	**72.78±0.63**	**73.82±0.59**	**68.45±0.28**	0.005±0.001	**0.039±0.002**	**0.042±0.002**	**0.069±0.002**
Digit Five	FedAvg	**90.58±0.30**	84.79±0.42	84.42±0.33	74.97±0.40	0.001±0.000	0.035±0.001	0.054±0.002	0.083±0.005
	DRFA	90.23±0.34	86.02±0.59	83.23±0.41	79.86±0.30	0.001±0.000	0.031±0.001	0.052±0.002	0.060±0.002
	IndA	89.61±0.58	84.13±0.77	82.97±0.71	77.64±0.99	0.002±0.001	0.036±0.002	0.056±0.003	0.074±0.004
	FMDA-M (Ours)	87.81±0.25	**86.18±0.36**	**85.45±0.37**	**81.34±0.34**	**0.001±0.000**	**0.023±0.001**	**0.037±0.002**	**0.048±0.003**
Adult	FedAvg	**82.11±0.43**	68.61±0.48	77.95±0.49	66.89±0.44	0.004±0.001	0.091±0.004	0.055±0.003	0.067±0.002
	DRFA	82.01±0.33	69.77±0.55	**78.29±0.57**	68.53±0.60	0.004±0.001	0.082±0.003	0.053±0.002	0.051±0.002
	IndA	81.31±0.61	68.53±0.61	77.98±0.74	64.72±0.90	0.003±0.001	0.090±0.005	0.052±0.004	0.074±0.004
	FMDA-M (Ours)	75.26±0.52	**74.03±0.47**	74.98±0.43	**71.84±0.48**	**0.003±0.000**	**0.012±0.001**	**0.010±0.001**	**0.029±0.002**



(a) Fashion-MNIST (b) Digit-Five (c) Adult

Figure 3. Experimental results of robustness (the top row) and fairness (the bottom row) on agnostic distribution. IID, weakly Non-IID, strongly Non-IID and extremely Non-IID are denoted by setting 0, setting 1, setting 2 and setting 3, respectively. The coordinate $i \rightarrow j$ of horizontal axis means that we train a federated learning model in setting i and test it in setting j.

in FL where the data distribution of each client is unknown to others. We find that the overfitting appears in FedAvg, DRFA and individual-level method due to the jumbo sample size of the low-income male group. By contrast, our proposed FMDA-M samples from each subgroup according to the group weights, thus overcoming this challenge.

Note that FMDA-M also outperforms DRFA in extremely Non-IID setting, where the unified group fairness optimized in our FMDA-M is exactly the client-level fairness optimized in DRFA. The reason for this is that DRFA uses naive gradient ascent method to update weights and the projection operator usually leads to the very hard weights, which may affect the stability of algorithm. By contrast, our algorithm adopts Eq. (5) to generate smoother weights and it helps the model to converge.

F.5. Results of Client-level Fairness

We evaluate the client-level fairness of models trained by different algorithm and report the results in Table 3. We observe that FMDA-M is able to guarantee the accuracy of the worst-performing client and decrease *Disparity* in most training distribution settings. We also find that, unlike other methods of which the performance decreases significantly with increasing degree of Non-IID, FMDA-M shows extremely stable performance under various settings, which is thanks to the weights update rule (5) we adopt.

We note that *Robustness* of FMDA-M is slightly lower than FedAvg and DRFA in IID setting, because the distributions of clients are very similar and the $Robustness$ will degrade to average accuracy, which is in line with the optimization objective of FedAvg and DRFA. Indeed, our FMDA-M significantly improves client-level fairness in Non-IID settings (more challenging and more common in reality), though occasionally with a small performance sacrifice in IID setting.

F.6. Results of Agnostic Distribution Fairness

To evaluate the agnostic distribution fairness, we simulate the newly added clients as follows: we train each federated model in one of the the training distribution settings (e.g., IID setting), but test under other three settings (e.g., weakly Non-IID, strongly Non-IID and extremely Non-IID settings) where the distributions of clients are different and agnostic from the existing clients.

The results of agnostic distribution fairness are shown in Figure 3. We find that our FMDA-M outperforms compared baselines in terms of both $Robustness$ and $Disparity$ in most cases, which illustrates that our FMDA-M is better adapted to new distributions. As we state before, the proposed FMDA-M considers a larger uncertainty set but with appropriate degrees of freedom, so the model trained by FMDA-M can deal with kinds of new distributions. However, the resulting model can be overly pessimistic when the radius of uncertainty set is too large. Besides, the individual-level method is hard to optimize, and that is why the individual-level method does not perform well.

F.7. Average Performance

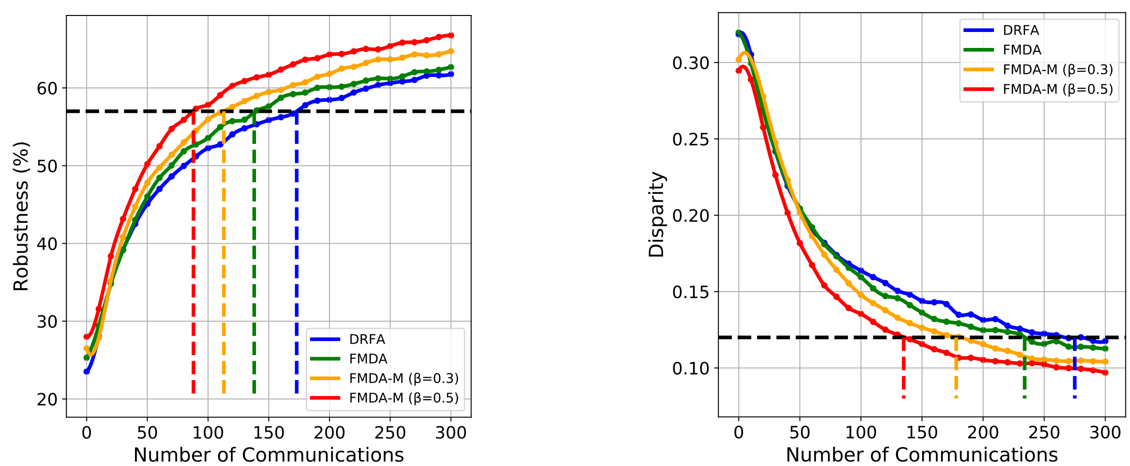
Table 4 shows that the average accuracy of model over groups and clients are similar for all algorithms on Fashion-MNIST and Digit-Five. On Adult, FMDA-M improves the average performance over groups, while the average accuracy over clients of the FMDA-M is slightly lower than the compared baselines. There is a serious imbalance in Adult dataset. Specifically, as shown in the third row of Figure 2, the sample size of low-income male group is much larger than high-income female group. Our FMDA-M focuses on the performance of minority group. As we see in the previous experiments, our FMDA-M improves the accuracy of the minority group by about 50% (Table 2). And it is inevitable that a bit of performance of majority group will be lost. There is a large amount of data of low-income male group on clients, and thus the average accuracy over clients will go down. Indeed, our FMDA-M significantly improves attribute-level fairness and shows extremely stable performance on agnostic distribution, though occasionally with a small average performance sacrifice.

F.8. Efficiency and Ablation Study

To evaluate the efficiency of our FMDA-M algorithm and demonstrate how it works, we run the following algorithms on Fashion-MNIST dataset in extremely Non-IID setting: (i) DRFA (Deng et al., 2020): DRFA algorithm is proposed to solve the min-max problem in federated setting. (ii) FMDA. (iii) FMDA-M ($\beta = \beta_\theta = \beta_{\boldsymbol{\lambda}} = 0.3$). (iv) FMDA-M ($\beta = \beta_\theta = \beta_{\boldsymbol{\lambda}} = 0.5$). Note that the above algorithms share the same optimization objective. We report the learning curves of models in terms of attribute-level robustness and fairness over 300 rounds of communications, as shown in Figure 4. Results in other settings are in the appendix.

The results show that the proposed FMDA outperforms DRFA in terms of convergence rate. The most likely reason is that FMDA adopt mirror ascent based on Bregman divergence, instead of projection operation based on Euclidean distance, to update the group weights, which prevents the weights from being too hard to guarantee convergence stability. We observe that FMDA-M is more efficient and can achieve the same level as others with fewer number of communication rounds, because the momentum term can modify the direction of the current gradients to accelerate convergence.

To demonstrate how the momentum term works, we do further experiments on Fashion-MNIST dataset. By denoting the parameters of convergent federated model as θ_{opt} and denoting the parameters of current federated model as $\theta_{current}$, we can define the optimal update direction as $d_{opt} = \theta_{current} - \theta_{opt}$. Then we can get the cosine similarity between d_{opt} and current gradient g, and the cosine similarity between d_{opt} and the gradient modified by momentum d_{mod}. The results are shown in Figure 6. We observe that the gradient modified by momentum d_{mod} is more similar to the optimal update direction as d_{opt}, which means that the momentum term can correct the gradients and accelerate the convergence.

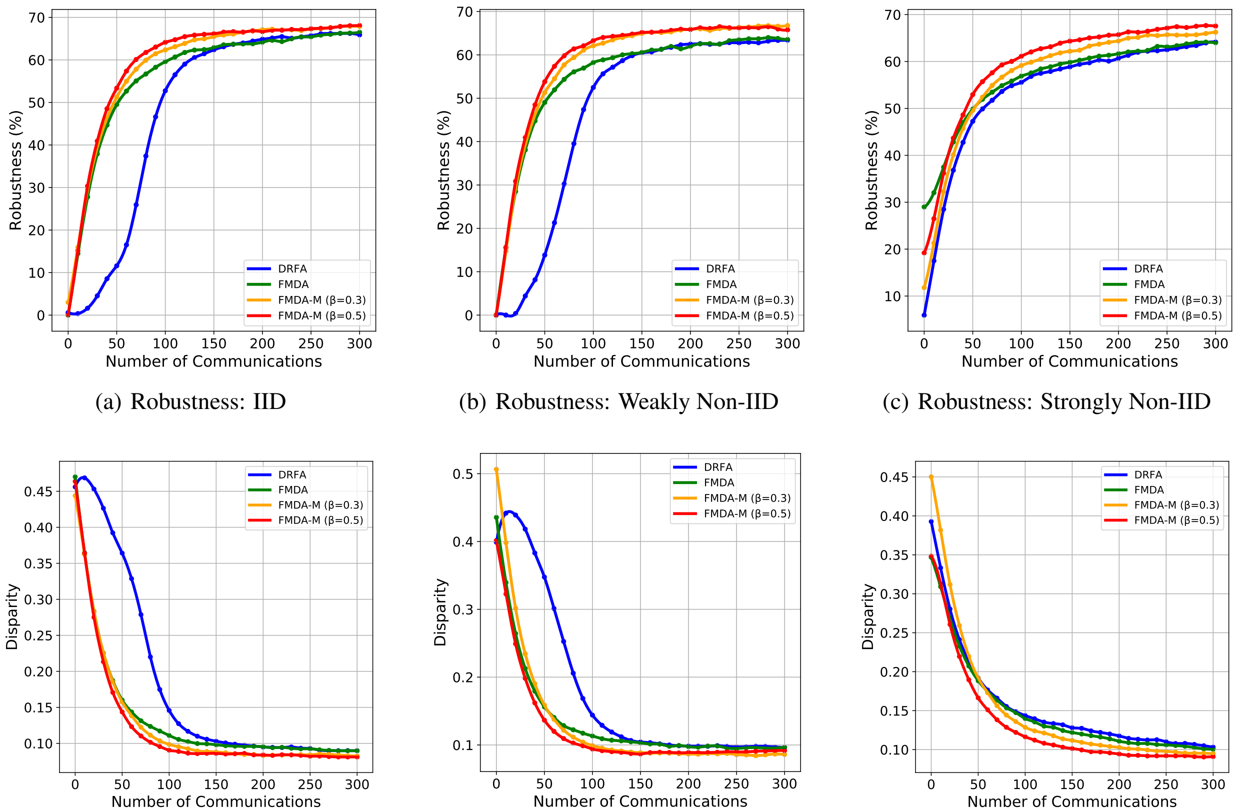


(a) Robustness (b) Fairness

Figure 4. Experimental results of efficiency of different algorithms on Fashion-MNIST in extremely Non-IID setting at attribute level.

Table 4. Average accuracy of the global model over attributes and clients.

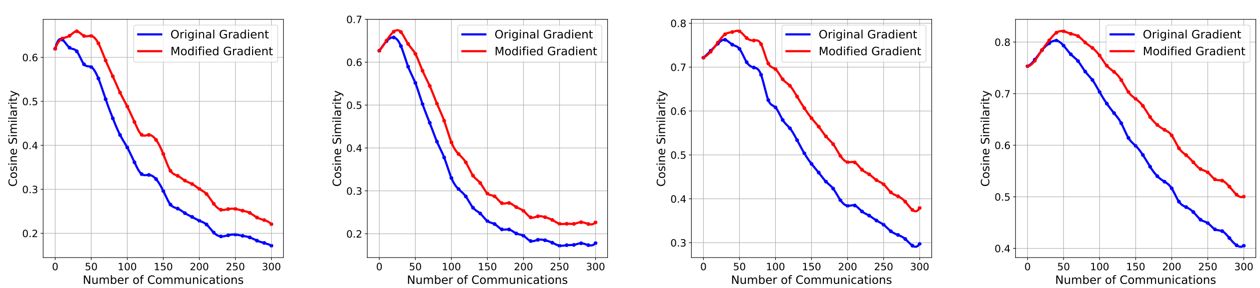
	Metrics	Average Accuracy over Attributes: Avg_Acc (%)				Average Accuracy over Clients: Avg_Acc (%)			
Dataset	Method	IID	Weakly Non-IID	Strongly Non-IID	Extremely Non-IID	IID	Weakly Non-IID	Strongly Non-IID	Extremely Non-IID
Fashion MNIST	FedAvg	83.47±0.11	83.20±0.10	81.75±0.11	80.96±0.10	85.28±0.07	84.80±0.06	83.05±0.08	82.23±0.09
	DRFA	83.46±0.16	83.13±0.11	81.83±0.12	81.04±0.23	85.31±0.10	84.79±0.12	83.19±0.13	82.44±0.15
	IndA	82.92±0.37	82.73±0.32	81.47±0.21	81.46±0.39	84.30±0.13	84.76±0.11	82.77±0.15	82.86±0.18
	FMDA-M (Ours)	82.31±0.25	82.63±0.20	81.24±0.18	80.56±0.22	81.95±0.18	84.03±0.12	82.73±0.18	81.83±0.13
Digit Five	FedAvg	89.36±0.26	88.89±0.24	89.46±0.30	88.11±0.25	90.80±0.10	90.37±0.09	90.53±0.13	89.03±0.14
	DRFA	89.17±0.20	89.03±0.24	89.06±0.19	87.93±0.24	90.31±0.12	90.58±0.08	89.69±0.16	89.09±0.16
	IndA	88.03±0.31	87.95±0.29	88.34±0.34	87.38±0.35	89.87±0.19	89.84±0.14	89.74±0.15	88.77±0.20
	FMDA-M (Ours)	88.94±0.21	88.01±0.28	88.18±0.20	86.64±0.24	88.78±0.12	89.31±0.16	89.34±0.18	88.96±0.13
Adult	FedAvg	64.49±0.19	66.27±0.15	64.52±0.16	73.79±0.22	82.38±0.15	80.52±0.16	83.36±0.12	73.81±0.23
	DRFA	64.76±0.23	67.64±0.25	64.59±0.18	74.13±0.24	82.41±0.16	80.50±0.17	83.44±0.11	73.94±0.24
	IndA	65.07±0.28	65.87±0.29	64.31±0.26	73.25±0.31	81.88±0.24	80.37±0.19	83.12±0.14	73.67±0.29
	FMDA-M (Ours)	74.63±0.18	74.67±0.19	74.74±0.16	74.60±0.21	75.61±0.27	75.81±0.21	75.66±0.32	74.13±0.29



(a) Robustness: IID (b) Robustness: Weakly Non-IID (c) Robustness: Strongly Non-IID

(d) Fairness: IID (e) Fairness: Weakly Non-IID (f) Fairness: Strongly Non-IID

Figure 5. Efficiency of different algorithms on Fashion-MNIST in IID, weakly Non-IID and strongly Non-IID settings at attribute level.



(a) IID (b) Weakly Non-IID (c) Strongly Non-IID (d) Extremely Non-IID

Figure 6. Cosine similarity between optimal update direction d_{opt} and original gradient g (and gradient modified by momentum d_{mod}).