Model-Agnostic and Fair Federated Learning Under Heterogeneous Preferences

Anonymous Authors[1]

Abstract

Federated learning (FL) provides an effective learning paradigm to jointly optimize a model based on datasets that are distributed across agents to protect privacy. Due to the heterogeneity in agent preferences and data distributions, ensuring the fairness of FL is critical. Our goal is to define a fairness notion for FL that is: (i) *fair to all* – does not "harm" agents with high-quality data due to other agents with low-quality data, (ii) *guaranteed to exist* – exists for all learning paradigms including DNNs, and (iii) *tunable to heterogeneity* – gives improved guarantees if agents have similar preferences or similar data distributions. The popular notions of *egalitarian fairness* and *weighted equity fairness* fail to satisfy (i), while the recently defined *CoreFed* fails to satisfy (ii).

In this paper, we achieve all three goals by introducing a new notion of *rankwise proportionality*. It ensures that the resulting model is "ranked" *proportionally highly* by all the agents, and therefore is *fair to all*. We then show that rankwise proportionality is *guaranteed to exist* for all ML models that are piecewise continuous (including DNNs). The idea of agents "ranking" the models naturally facilitates several measures of similarity between their preferences. We then establish guarantees for rankwise proportionality parameterized by agent heterogeneity with improved guarantees when heterogeneity is low (*tunable to heterogeneity*).

Finally, we design `RankFF`, a distributed fair FL algorithm, to train a rankwise proportional and Pareto optimal model. We show that `RankFF` is computationally efficient and outperforms baselines in terms of fairness and overall utility on different datasets.

1. Introduction

Federated Learning (FL) provides an effective distributed learning paradigm, where a group of agents, holding local data samples, can train a joint model without sharing their private data. Such a learning process can benefit significantly from rich heterogeneous data spread across agents, and it has been widely used for different applications such as autonomous vehicles (Elbir et al., 2020) and digital healthcare (Dayan et al., 2021; Xu et al., 2021).

Given the distributed nature of FL, fairness is a crucial desideratum for it. One of the most intriguing fairness aspects in FL is *representational parity*, which requires the final model to perform well on the data distributions of every agent. Some work (Huang et al., 2021; Yang et al., 2021) attempt to achieve this by establishing fairness at the agent selection phase, while others aim to prevent the global model from over-fitting to the data of particular agents (Li et al., 2020; Mohri et al., 2019).

An alternative approach to fairness in FL imports concepts from *social choice theory*. At a high level, fairness in FL resembles the classical *public decision making* setting in social choice theory, where the goal is to make a global decision (choose the final model from a set of feasible models) that is fair to the agents (similarly good performance on the local datasets of each agent). Relevant notions include: (1) *weighted equity fairness (Donahue & Kleinberg, 2021b)* – the final model balances the utility (or weighted utility) values of the agents[1]; (2) *egalitarian fairness (Donahue & Kleinberg, 2021b)* – the final model maximizes the smallest utility (or equivalently minimizes the largest loss); and (3) *core-stability based fairness (Chaudhury et al., 2022a)* – the final model ensures no set of agents can significantly increase their utility values by jointly training a model, using data which is local to these agents.

Chaudhury et al. show that the former two fairness notions are not robust to noisy data from agents, i.e., there are instances where the final trained model is tailored towards the data samples of agents with high noise. In a nutshell, if there is an agent i that incurs high loss (or low utility) for most

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1] This notion is called "Proportionality" in (Donahue & Kleinberg, 2021a). However, we call it weighted equity based fairness as achieving similar utility values for all agents is referred to as *equitable fairness* in social choice theory.

models that fit the data of the other agents well, then under both (1)and (2), the training algorithm will attempt to reduce the loss of agent i, thereby fitting the final model better with the data of agent i. This in turn will hurt the other agents (possibly with high-quality data). Core-Stability based fairness does not suffer from this weakness, as the fairness criterion does not compare the utility functions of the agents with one another; rather, it looks at how much utility the final model gives to an agent in comparison with the model that is trained solely on the data of this agent, and attempts to balance this ratio.[2] Despite its advantages, core-stability has shortcomings of its own:

- *Failure to exist in more general ML models like DNNs.* The existence of core-stable final models is only known when the utility functions of the agents are concave (or equivalently, the loss functions are convex) (Chaudhury et al., 2022a). One of the most important guarantees of a core-stable model c is *proportionality*, i.e., each agent's utility for c is at least $1/n$ times the agent's utility for any model, or equivalently $u_i(c) \geq u_i(c_i^*)/n$, where c_i^* is the model that maximizes agent i's utility. In the Apendix A, we show that even proportional final models (and consequently core-stable final models) fail to exist in smooth DNNs, thereby showing that the guarantees of Chaudhury et al. (2022a) do not hold for a broad spectrum of ML models.
- *Weak guarantees in practice.* Proportionality ensures that each agent receives a $1/n$ fraction of its best possible utility. In practice, this guarantee is weak for large n, and such guarantees are also achievable with high probability by a final model chosen uniformly at random. Better guarantees are achievable in practice since there is more similarity in the data distributions of agents compared to the worst-case heterogeneous distributions.

This motivates the main research questions of this paper:

> *Can we define a noise-robust fairness notion that is guaranteed to exist in a wide variety of ML models? Can we define and measure the heterogeneity of agent utilities? Can we show improved guarantees for our aforementioned fairness notion under limited heterogeneity?*

We answer all these questions in the affirmative.

Fairness Notion: Rankwise Proportionality. We define a new fairness notion, *rankwise proportionality*. Our notion is applicable to both *discrete* (set P is finite) and *continuous* settings (set P is measurable), and *ordinal* and *cardinal* preferences. For ease of presentation, we explain this concept in the discrete setting first, where an FL instance comprises of n agents, and a finite set of model vectors $P = \{\theta_1, \theta_2, \ldots, \theta_m\}$. Each agent a has a preference order over the model vectors, where $\theta_i \succ_a \theta_j$ denotes that she prefers θ_i over θ_j. Note that, it is easy to infer this order given her utility function u_a: $\theta_i \succ_a \theta_j$ if $u_a(\theta_i) > u_a(\theta_j)$, or $u_a(\theta_i) = u_a(\theta_j)$ and $i < j$. We define the rank of a model vector $\theta \in P$ for agent i, $\mathsf{rank}_i(\theta)$, as the fraction of model vectors that agent i prefers more than θ, i.e., $\mathsf{rank}_i(\theta) = |\{\alpha \in P \mid \alpha \succ_i \theta\}|/m$. Therefore, the lower the rank, the higher the preference for θ. We say that a model vector θ is rankwise proportional if and only if every agent prefers θ more than at least $1/n$ fraction of the model vectors, i.e., $\mathsf{rank}_i(\theta) \leq 1 - 1/n$ for all $i \in [n]$.

We remark that the notion of rankwise proportionality, though natural, has not been studied in social choice theory (to the best of our knowledge). A possible reason could be that, in standard public decision making settings such as voting, the number of alternatives m is significantly smaller than the number of agents (voters) n, i.e, $m \ll n$. Therefore, proportionality guarantees in terms of the number of agents n is not desirable in these settings. By contrast, in ML, the number of alternatives (feasible models) is significantly larger than the number of agents. Thus, Rankwise proportionality is a concept that can be very naturally studied at the intersection of ML and social choice theory, providing further evidence for the need for more interaction between the two fields.

We then extend this definition to the continuous FL setting, by defining $\mathsf{rank}_i(\theta) = \lambda(\{\alpha \mid u_i(\alpha) \geq u_i(\theta)\})/\lambda(P)$, where $\lambda(D)$ is the Lebesgue measure of D.[3] The only restriction on P is Lebesgue-measurability, in contrast to the Core-Fed algorithm of Chaudhury et al. (2022a), which crucially requires P to be convex.

Note that the fairness guarantee of rankwise proportionality is solely based on the *rank* of each model for each agent, and is therefore not sensitive to large perturbation or scaling of the utility values of an agent, as long as it does not change the rank of the model vector for the agent. Similar to Core-Fed, it is more robust to noise than egalitarian and weighted equity based fairness, as it does not compare the precise utility values of the agents with each other. However, in contrast to Core-Fed, we show that our guarantees are achievable for any ML model as long as the utility functions (or equivalently loss functions) are *piece-wise continuous*. We first establish the proof of existence in the discrete setting (Theorem 3.1) through a concise *pigeonhole argument*, and then we generalize the result to the continuous setting (Theorem 3.2).

Heterogeneity. Once we have the rank functions $\mathsf{rank}_i(\cdot)$ of

[2] As a result, even if there is an agent that incurs high loss for most models, the ratio of the utility (loss) derived from the final model and best model may not be relatively high for the noisy agent.

[3] We show in Proposition 1 that when the utility functions of the agents are piece-wise continuous, then the sets $\{c \mid u_i(c) \geq u_i(\theta)\}$ are Lebesgue measurable.

the agents, we can define the heterogeneity between two agents i and j as any well-studied distance between the rankings $\mathsf{rank}_i(\cdot)$ and $\mathsf{rank}_j(\cdot)$. To this end, we examine improved fairness guarantees (for rankwise proportionality) under standard distance measures d between rank functions like maximum displacement distance, footrule distance, and edit-distance (Caragiannis et al., 2016). Given a FL instance I, we define the heterogeneity $h(I)$ as $\max_{i,j\in[n]} d(\mathsf{rank}_i(\cdot), \mathsf{rank}_j(\cdot)))$. We say that a model θ is β-rankwise proportional if $\mathsf{rank}_i(\theta) \leq \beta$ for all $i \in [n]$.[4]

Concretely, given a FL instance I with heterogeneity $h(I)$, a $\beta(h(I), n)$ rankwise proportional (with $\beta(h(I), n) < 1 - 1/n$), Pareto-optimal model exists for any ML model, as long as the loss functions of the agents are piece-wise continuous (see Theorems 4.1, 4.2, 4.3 for the exact bounds on $\beta(h(I), n)$).

Distributed Algorithm and Empirical Results. We develop a distributed learning algorithm `RankFF`, which computes the smallest β for which β-rankwise proportional and Pareto-optimal model exist, and outputs the desired model. The algorithm provably runs in polynomial time when the utility functions are concave.

Finally, we conduct comprehensive experiments on different datasets with around 100 agents and show that (i) `RankFF` chooses a fair final model and the utilitarian social welfare (sum of utilities of the agents) is larger than that of the final model chosen by baselines such as FedAvg and Core-Fed, and (ii) utilities generated from similar data distributions result in lower heterogeneity and tighter fairness guarantees, suggesting that our heterogeneity measures can indeed reflect the similarity (or dissimilarity) of local data distributions.

2. Preliminaries

Problem Setting. In any federated learning (FL) instance, there are n participating agents training a joint model $\theta \in P \subseteq \mathbb{R}^N$. In its full generality, P may be a discrete (finite) or a measurable set, and the preference of an agent i may be ordinal (ranking over θ's in P), or cardinal (defined by a utility function u_i). In the standard FL setting, an agent's preference over the models θ, depend on the losses they incur over her data distribution.

For agent i, let $\ell_i(\theta, \mathcal{D}_i)$ denote her *loss function value* for θ on samples drawn from her data distribution $\mathcal{D}_i$. One can define utility functions to capture preferences based on loss values next (this is to remain consistent with social choice theory literature). Similar to (Chaudhury et al., 2022a), for each agent i, we define a utility function $u_i \colon P \to \mathbb{R}_{\geq 0}$, as $u_i(\theta) = (\max_{\theta' \in P} \ell_i(\theta', \mathcal{D}_i)) - \ell_i(\theta, \mathcal{D}_i)$.

We define our novel fairness notions in Section 3. To measure the efficiency or non-wastefulness of our final solution, we use the standard notion of Pareto-optimality (PO).

Definition 1. (Pareto optimality) A model $\theta \in P$ is said to be Pareto optimal if there is no other model θ' that all agents prefer at least as much as θ, and at least one agent prefers θ' strictly more than θ. Formally,

$$\nexists \theta' \in P, \text{ s.t. } \forall i \in [n] : u_i(\theta') \geq u_i(\theta), \tag{1}$$

with at least one inequality strict.

3. Rankwise Proportionality and Its Existence

The fact that no approximation of core-stability, or even proportionality, is possible in the case of DNNs (See Appendix A) implores us to define a fairness notion that exists more broadly. In this section, we define rankwise proportionality in the most general setting, where agents' utilities may be ordinal or cardinal, and the set P of models is discrete or continuous. Then we show that rankwise proportional model always exists.

3.1. Rankwise Proportionality

To handle both ordinal and cardinal preferences, as well as discrete and continuous sets of P, we consider two settings for fair FL, discrete, and continuous.

Discrete FL setting. In the *discrete FL setting*, we assume set P is finite, and each agent has a ranking over models in P. Let $P = \{\theta_1, \theta_2, \ldots, \theta_m\}$. Let $\succ_i$ denote the ranking of agent i over P, i.e., $\theta \succ_i \theta'$ implies that i prefers θ over θ'. For each agent i, let σ_i denote agent i's *preference-ranked string* of models, i.e., $\sigma_i = \{\theta_{r_1}, \theta_{r_2}, \ldots, \theta_{r_m}\}$ implies that $\theta_{r_1} \succ_i \theta_{r_2} \succ_i \cdots \succ_i \theta_{r_m}$. Lastly, given $\sigma_i = \{\theta_{r_1}, \theta_{r_2}, \ldots, \theta_{r_m}\}$, we define $\mathsf{rank}_i(\theta_{r_k}) = (k-1)/m$. Intuitively, the rank of θ denotes the fraction of models that are preferred more than θ.

Continuous FL setting. In the *continuous FL setting*, we assume P is *Lebesgue-measurable*, and that each agent $i \in [n]$ has an associated utility function $u_i : P \to \mathbb{R}_{\geq 0}$. We only assume that the functions $u_i(\cdot)$ are Lebesgue-measurable. This is a very mild assumption, and is weaker than continuity; we show in Proposition 1 in Appendix C, that any *piece-wise continuous* function is Lebesgue-measurable.

We define the rank of a model c for an agent i analogously to the discrete setting. Let $P_i(c) = \{\theta \mid u_i(\theta) \geq u_i(c)\}$.[5] We define $\mathsf{rank}_i(c) = \frac{\lambda(P_i(c))}{\lambda(P)}$, where $\lambda(D)$ indicates the *Lebesgue-measure* of the set D, i.e., the high-dimensional

[4] A model is rankwise proportional when $\beta = 1 - 1/n$.

[5] Appendix C.1 elaborates on why the sets $P_i(c)$ are Lebesgue-measurable.

volume of D. Thus $\mathsf{rank}_i(c)$ intuitively measures the fraction of models that agent i prefers at least as much as the model c.

We now define our fairness metric. Given any FL instance (discrete or continuous), we say that θ is rankwise proportional if, for every agent θ is at least as good as a $1/n$ fraction of the entire set of models. Formally:

Definition 2. (Rankwise Proportionality) For the discrete setting we say that θ is rankwise proportional if for every agent i,

$$\mathsf{rank}_i(\theta) \leq \left\lceil m\left(1 - \frac{1}{n}\right)\right\rceil \cdot \frac{1}{m}.$$

For the continuous setting, we say that θ is rankwise proportional if $\mathsf{rank}_i(\theta) \leq \alpha$ for all agents i, where α is the smallest achievable rank for all agents larger than $1 - 1/n$, i.e., α is the minimum value such that for every agent $i \in [n]$, there exists a $\theta_i \in P$, such that $1 - 1/n < \mathsf{rank}_i(\theta_i) \leq \alpha$.

In the continuous setting, the rank function of an agent depends more subtly on the properties of the level sets of her utility function, and can change much more drastically. In particular, there may not be any θ for which $\mathsf{rank}_i(\theta) = (1 - 1/n)$, and that is the reason we need to define α. We note that if the level-sets of the utility functions are of zero measure then α will be arbitrarily close to $(1 - 1/n)$.

We also introduce the notion of β-rankwise proportionality in the discrete setting, which will be useful to prove guarantees parameterized by heterogeneity in Section 4.

Definition 3. (β-Rankwise Proportionality) We say that θ is β-rankwise proportional if for every agent i, $\mathsf{rank}_i(\theta) \leq \beta$. Note that when $\beta = \lceil m\left(1 - \frac{1}{n}\right)\rceil \cdot \frac{1}{m}$, and $\beta = \alpha$, β-rankwise proportionality coincides with rankwise proportionality in the discrete and the continuous case respectively.

Remark. We do not need any assumptions on P beyond measurability for our existence results on rankwise proportionality and heterogeneity (Sections 3 and 4). To prove polynomial running time for our algorithm in Section 5 we assume that P is closed and convex. In our experiments in Section 6, we do not need the convexity assumption on P.

Connection to Proportionality. *Proportionality* is a classical fairness notion within social choice theory (Kelly et al., 1998) that asks for existence of a $\theta \in P$ such that for each agent i, $u_i(\theta) \geq U_i/n$, where U_i is her maximum possible utility ($\max_{\theta' \in P} u_i(\theta')$). Equivalently, at θ, the *loss* in the utility compared to U_i is at most a multiplicative factor of $(1 - 1/n)$. Thus, while proportionality ensures a decrease of at most $(1 - 1/n)$ *in the percentage of the best possible*, rankwise proportionality ensures a decrease of at most $(1 - 1/n)$ *in the percentile of the best possible*. We note that these two notions are incomparable in general. However, in the discrete setting, if we define $u_i(\theta) = 1 - \mathsf{rank}_i(\theta)$ for all $i \in [n]$, then rankwise proportionality is proportionality.

Advantages of Rankwise Proportionality. Clearly, proportionality is inapplicable when agents' have only ordinal preferences. Second, even in the case of cardinal preferences (utilities), if the utility functions $u_i(\cdot)$ are not concave like in the case of DNNs, or if P is not convex and closed, a proportional model may not exist. In contrast, rankwise proportionality can be defined in all of these settings, and it always exists as we show next.

3.2. Existence in the Discrete FL Setting

Recall that in the discrete FL setting, agents rank the models belonging to a finite set P. We first show that in *any ML model*, a model satisfying rankwise proportionality exists.

Theorem 3.1. *Given a finite set of models* $P = \{\theta_1, \theta_2, \dots, \theta_m\}$, *and* any *set of agent preferences* $\langle \succ_1, \succ_2, \dots, \succ_n \rangle$, *there exists a rankwise proportional and Pareto-optimal* $\theta \in P$.

Proof sketch. Our proof follows a pigeonhole argument. Consider agent 1. Assume that there is no rankwise proportional model, i.e., for each θ with rank at most $1 - 1/n$ for agent 1, there is an agent in $[n] \setminus \{1\}$ that ranks it higher than $1 - 1/n$. Since there are $m(1 - 1/n)$ many models that agent 1 ranks at most $1 - 1/n$ and there are $n - 1$ other agents, there exists an agent $i' \neq 1$ that ranks at least $m(1 - 1/n)/(n - 1) = m/n$ many models strictly larger than $1 - 1/n$. This is a contradiction, as there can be at most $m/n - 1$ models with rank strictly larger than $1 - 1/n$. Pareto-optimality follows immediately from the fact that models that Pareto-dominate a rankwise proportional model θ, are also rankwise proportional. The full proof can be found in Appendix B. □

Implications. Theorem 3.1 makes no assumptions about the agents' preferences or model properties, and therefore is essentially model agnostic. We note that the guarantees of this theorem are tight in general (See Example 1 in Appendix B). In addition, if agents' preferences are more *aligned*, we will see in Section 4 how these guarantees can be improved. Within ML, the fairness and efficiency guarantees of Theorem 3.1 is applicable in the following setting: Given the large amount of available pretrained models currently, such as the ones from PyTorch Hub (Paszke et al., 2017), TensorFlow Hub (Abadi et al., 2016), and Hugging Face (Wolf et al., 2020), it is common and practical for agents to evaluate their losses on the discrete set of pretrained models and provide only the ordering among them. This would help protect data privacy as well.

3.3. Existence in the Continuous FL Setting

Note that Theorem 3.1 is model-agnostic, i.e., the guarantees hold for any finite set P and any set of agent preferences. This strongly suggests that we should be able to achieve analogous guarantees for any continuous ML setting under mild assumptions: we can arbitrarily discretize the measurable set P, and look at the corresponding discrete instance which will admit a rankwise proportional and Pareto-optimal model. We validate this intuition.

For an agent i, we define the set A_i of *achievable* ranks as $A_i = \{\beta : \exists\ \theta \in P \text{ s.t. } \mathsf{rank}_i(\theta) = \beta\}$. We let $\alpha_i = \min A_i \cap (1 - 1/n, 1]$, and let $\alpha = \max_{i\in[n]} \alpha_i$. Thus, α is the smallest achievable rank larger than $1 - 1/n$, and is the parameter defining rankwise proportionality as per Definition 2. The main result of this section is the following.

Theorem 3.2. *In any FL instance with n agents and a measurable set of models P, where each agent i has a Lebesgue-measurable utility function u_i, there always exists a Pareto optimal model θ that is rankwise proportional.*

Proof Sketch. The core idea is still the pigeonhole argument used in the proof of Theorem 3.1. However, there are *significant technical details involved in proving Lebesgue-measurability of relevant sets (in particular the sets Z_i that we define ahead)*. Let $Z_i = \{\theta \in P \mid \mathsf{rank}_i(\theta) \le \alpha_i\}$. We show that $\lambda(Z_i) \ge \alpha_i \lambda(P) > (1 - 1/n) \cdot \lambda(P)$ as $\alpha_i > 1 - 1/n$.

Now, consider agent 1, and assume there exists no rankwise proportional model. Thus, for every $\theta \in Z_1$, there exists an agent i s.t. $\theta \in P \setminus Z_i$. Thus, $Z_1 \subseteq \cup_{i\in[n]\setminus\{1\}} P \setminus Z_i$, implying that

$$\lambda(Z_1) \le \lambda(\cup_{i\in[n]\setminus\{1\}} P \setminus Z_i) = \sum_{i\in[n]\setminus\{1\}} (1 - \alpha_i)\lambda(P) < (1 - 1/n)\lambda(P),$$

which is a contradiction. The detailed full proof is in the Appendix C.

Implications. Note that the guarantees in Theorem 3.2 holds for any measurable set P. This suggests that for FL instances, we can remove any set of models which have low utility values for all agents and only focus on those which are valued highly by some agents. We also remark that this is in contrast to the results in (Chaudhury et al., 2022a) that crucially require the convexity of P, assumption, even for proving the existence of a core-stable model. Finally, we remark that the guarantees of Theorem 3.2 are also tight, as Example 2 in Appendix C shows.

4. Heterogeneity and Improved Fairness Guarantees

In this section, we address our second main question: Can we improve the guarantees of Theorems 3.1 and 3.2 when the preferences of the individual agents are not drastically different?

To this end, given a FL instance I, we need to define a measure of heterogeneity $h(I)$ between the agents' preferences. Throughout this section, we define the heterogeneity in the discrete FL setting. We adapt the heterogeneity definitions to the continuous FL setting as follows: Given a continuous FL instance I, with the set of models P, discretize the space P of models into a grid with grid-length ε for a sufficiently small ε. This creates a discrete FL instance I_ε, where the finite set of models P_ε is the vectors corresponding to the grid points. We then define the heterogeneity in the continuous setting I as $h(I_\varepsilon)$.

Our heterogeneity measures are based on distance functions $d(\cdot)$ between rankings. Following the computational social choice literature (Caragiannis et al., 2016), we consider three distance measures between rankings: (i) maximum displacement distance, (ii) footrule distance, and (iii) edit distance. Once we fix the distance function d, the heterogeneity $h(I)$ of a FL instance I can be defined as $\max_{i,j\in[n]} d(\mathsf{rank}_i(\cdot), \mathsf{rank}_j(\cdot))$. We then improve the guarantees of Theorems 3.1 and 3.2 using $h(I)$ as a parameter. We defer the proofs to Appendix D.

Recall that a model $\theta \in P$ is β-rankwise proportional if $\mathsf{rank}_i(\theta) \le \beta$ for all $i \in [n]$. In what follows, we show upper bounds on β parameterized by $h(I)$.

4.1. Maximum Displacement Distance

The first distance measure we consider is the maximum displacement distance, which is the maximum difference in the ranks achieved by any feasible model in P.

Definition 4. The maximum displacement distance $\mathsf{md}(\mathsf{rank}_1(\cdot), \mathsf{rank}_2(\cdot))$ between two rankings $\mathsf{rank}_1(\cdot)$ and $\mathsf{rank}_2(\cdot)$ defined on P, is $\max_{c\in P} |\mathsf{rank}_1(c) - \mathsf{rank}_2(c)|$.

Theorem 4.1. *Given an instance I with $h(I) = \max_{i,j\in[n]} \mathsf{md}(\mathsf{rank}_i(\cdot), \mathsf{rank}_j(\cdot))$, where $\mathsf{md}(\cdot,\cdot)$ is the maximum displacement distance, there exists a model $\theta \in P$ which is $h(I)$-rankwise proportional.*

4.2. Footrule Distance

In contrast to the maximum displacement distance, the footrule distance looks at the "total displacement" (or average), i.e., the sum (average) of distances between the ranks of all feasible models in P. Formally,

Definition 5. The footrule distance $\mathsf{fd}(\mathsf{rank}_1(\cdot), \mathsf{rank}_2(\cdot))$ between two rankings $\mathsf{rank}_1(\cdot)$ and $\mathsf{rank}_2(\cdot)$, defined on P, is $\sum_{c\in P} |\mathsf{rank}_1(c) - \mathsf{rank}_2(c)|$.

Theorem 4.2. *Given an instance I with $h(I) = \max_{i,j\in[n]} \mathsf{fd}(\mathsf{rank}_i, \mathsf{rank}_j)$, where $\mathsf{fd}(\cdot,\cdot)$ is the footrule distance between two rank functions, there exists a model*

$\theta \in P$ which is $(\sqrt{2(n-1)h(I)/m})$-rankwise proportional.

4.3. Edit-Distance

The last similarity measure between rank functions that we consider is the *edit distance*, which intuitively measures the complexity of transforming one rank function into other. Recall that the *preference ranked string* σ_i of an agent is the m-length string defined by $\sigma_i[j] = \theta_j$, where $\mathsf{rank}_i(\theta_j) = (j-1)/m$. Now, given two agents 1 and 2, we define $\mathsf{ed}(\mathsf{rank}_1(\cdot), \mathsf{rank}_2(\cdot)) = \delta_D(\sigma_1, \sigma_2)$, where $\delta_D(\sigma_1, \sigma_2)$ denotes the edit distance (using insertions and deletions only) between two strings.

Theorem 4.3. *Given an instance I, with $h(I) = \max_{i,j\in[n]} \mathsf{ed}(\mathsf{rank}_i(\cdot), \mathsf{rank}_j(\cdot))$, where $\mathsf{ed}(\cdot,\cdot)$ denotes the edit distance between two rank functions, there exists a model $\theta \in P$, which is $\frac{1}{m}\cdot\big((n-1)h(I)\big)$-rankwise proportional.*

Discussion of Theorems 4.1, 4.2, 4.3. We remark that the best approximation achievable for rankwise proportionality is the minimum of the three guarantees in Theorems 4.1, 4.2, 4.3. We give empirical evidence in Section 6 that our measures of heterogeneity indeed reflect heterogeneity in the underlying data distributions of the agents. Finally, we note that, in this paper, we define the heterogeneity of an instance $h(I)$ as the maximum over all pairs of agents, the distance between the rank functions of the pair of agents. We believe that a finer heterogeneity measure could be the sum or average of distances over all pairs or a generalization of the distance functions $(\mathsf{md}(\cdot), \mathsf{ed}(\cdot), \mathsf{fd}(\cdot))$ to all n agents. We leave this as an interesting avenue for future research.

5. `RankFF`: Fair Federated Learning based on Rankwise Proportionality

In this section, we put forward our distributed learning algorithm `RankFF` that trains a rankwise proportional and Pareto optimal model. We assume that agent utility functions are concave; under this assumption `RankFF` runs in provably polynomial time.

At a high level, the server iteratively performs a binary search to find the minimum rank r such that there exists a model c for which every agent i has $\mathsf{rank}_i(c) \le r$; among the models satisfying this constraint, the algorithm then finds one that maximizes the sum of utilities. In each iteration of the binary search, given a rank r, each agent computes a utility constraint $\alpha'(r)$ to satisfy the given rank, i.e., all the models in $C_i = \{c \in P \mid u_i(c) \ge \alpha'(r)\}$ has $\mathsf{rank}_i(c) \le r$. The server maximizes the sum of utility while favoring satisfaction of the constraints of all agents. Formally, we aim to solve the following convex program.

$$\begin{aligned} \text{maximize} \quad & \sum_{i\in[n]} u_i(\theta) \\ \text{subject to} \quad & u_i(\theta) \ge \alpha_i(r) \ \ \forall i \in [n] \\ & \theta \in P \end{aligned} \tag{2}$$

Note that the above program is convex because of the concavity of the utility functions of the agents, and the convexity of P. We show that any feasible solution θ to the above program has rank at most r for all agents, and an optimal solution will be Pareto-optimal as well.

Claim 1. *Given $\alpha_i(r)$ for all $i \in [n]$, an optimal solution θ to the program (2) has rank at most r for all agents $i \in [n]$ and is Pareto optimal.*

Proof. Let θ be a feasible solution to 2. Since $u_i(\theta) \ge \alpha_i(r)$, we have that $\{c \in P \mid u_i(c) \ge u_i(\theta)\} \subseteq \{c \in P \mid u_i(c) \ge \alpha_i(r)\}$, implying that $|\{c \in P \mid u_i(c) \ge u_i(\theta)\}| \le |\{c \in P \mid u_i(c) \ge \alpha_i(r)\}|$ (in the discrete setting) and $\lambda(\{c \in P \mid u_i(c) \ge u_i(\theta)\}) \le \lambda(\{c \in P \mid u_i(c) \ge \alpha_i(r)\})$ (in the continuous setting). Therefore, we have $\mathsf{rank}_i(c) \le r$ for all $i \in [n]$.

To show that θ is Pareto optimal, assume otherwise. Let $\theta' \in P$ be a model such that $u_i(\theta') \ge u_i(\theta)$ for all $i \in [n]$ with at least one strict inequality. Then, note that $u_i(\theta') \ge u_i(\theta) \ge \alpha_i(r)$ for all $i \in [n]$, and $\sum_{i\in[n]} u_i(\theta') > \sum_{i\in[n]} u_i(\theta)$ which is a contradiction as θ is an optimum solution to (2). □

5.1. Discrete FL Setting

To provide intuition, we first describe the algorithm in the discrete setting. Since the hypothesis space is discrete, each agent can evaluate all models on their local data, and compute the $\alpha_i(r)$. Concretely, $\alpha_i(r)$ is the $(rm)^{\text{th}}$ highest utility among the utilities of all m models in P. Agent i then sends back C_i and the utilities of the models in C_i to the server. After hearing from all agents, the server examines the intersection $\bigcap_i C_i$ and returns the model with the highest sum of utility. If $\bigcap_i C_i$ is empty, r is not a valid rank and the server increases r. Otherwise, the server decreases r. We present the full description of `RankFF` in the discrete FL setting as Algorithm 1 in Appendix E.

5.2. Continuous FL Setting

Now we describe the algorithm in the continuous setting. At the beginning of each iteration of the binary search, the server sends the current global model and target rank r to all agents. Each agent approximates the scalar $\alpha_i(r)$ given r as follows. First, observe that for a fixed α', agent i can determine an arbitrarily good approximation (say, $1 \pm \varepsilon$ approximation) of the volume/ Lebesgue measure of the convex set $\{c \in P \mid u_i(c) \geq \alpha'\}$ in randomized polynomial time[6] by the algorithm of Dyer et al. (1989). Since the volume of the set $\{c \in P \mid u_i(c) \geq \alpha'\}$ is decreasing with α', agent i can use $\log(1/\varepsilon)$ rounds of binary search on α' to determine the value at which $(1-\varepsilon)r\lambda(P) \leq \lambda(\{c \in P \mid u_i(c) \geq \alpha'\}) \leq (1+\varepsilon)r\lambda(P)$.

After every agent i has identified $\alpha_i(r)$, we leverage the penalty method to convert (2) to an unconstrained optimization program:

$$\text{minimize} \quad \Phi_k(\theta) \tag{3}$$

$$\Phi_k(\theta) = -\sum_{i\in[n]} u_i(\theta) + \sigma_k \sum_{i\in[n]} \left[\max(0, \alpha_i(r) - u_i(\theta))\right]^2, \tag{4}$$

where σ_k is the penalty coefficient.

Then, we optimize program (3) through gradient descent for T iterations in a distributed manner. Let $\Phi_{k,i}(\theta)$ be $-u_i(\theta) + \sigma_k \left[\max(0, \alpha_i(r) - u_i(\theta))\right]^2$. Since $\nabla_\theta \Phi_k(\theta) = \sum_{i\in[n]} \nabla_\theta \Phi_{k,i}(\theta)$, each agent $i \in [n]$ can compute the gradient of $\Phi_{k,i}(\theta)$ locally and send the gradient back to the server. The server takes the sum of all received gradients, updates the global model, and starts the next training round. The server increases the penalty coefficient σ_k by a factor of λ every q rounds. As $T \to \infty$, the solutions to (3) asymptotically converge to the solution of program (2). In practice, however, T is set to a finite number. Therefore, the server should check with the agents to see whether the optimization is successful, i.e., whether the constraints of program (2) are satisfied. The server increases r if the optimization fails and decreases r otherwise.

We note that in practice (e.g., a real-world FL system), it could be time-consuming to compute the Lebesgue measure via the algorithm of Dyer et al. (1989) for large neural networks. Thus, we introduce an alternative approximation to alleviate the computational burden. Concretely, each agent computes the gradient of the utility function with respect to the current global model $\nabla_\theta u_i(\theta)$ and samples J models along the gradient direction: $\hat{\theta}_{i,j} = \theta + \xi \cdot \frac{\nabla_\theta u_i(\theta)}{\|\nabla_\theta u_i(\theta)\|}, \xi \sim U(0,p), j = 1, 2, \cdots, J$, as we assume the $L2$-norm of gradients is bounded by p. Clearly, the sampled models construct a subset $P_i \subseteq P$, which is discrete. Now we can compute the rank for agent i as in the discrete setting on P_i. We present the full description of the approximation algorithm and RankFF in the continuous setting as Algorithm 2 and Algorithm 3 in Appendix E, respectively.

6. Empirical Evaluation

We evaluate our heterogeneity measures and algorithm RankFF in both discrete and continuous FL settings on rotated MNIST (LeCun et al., 2010) and CIFAR-10 (Krizhevsky, 2009) datasets. We first demonstrate the soundness of the distance measures in Section 4. Then, we compare our approach with two baseline algorithms, FedAvg (McMahan et al., 2017) and CoreFed (Chaudhury et al., 2022b). We show that models trained with RankFF achieve better proportionality guarantees for fairness than the baselines and comparable utilitarian social welfare.

6.1. Experiment Setup

FL Setting. We consider FL with 10 agents for MNIST and 100 agents for CIFAR-10. We introduce heterogeneity among agents by rotating images at different degrees, following the literature (Ghosh et al., 2020). We introduce 3 FL settings with different heterogeneity levels ($I_{\text{high}}, I_{\text{mid}}, I_{\text{low}}$) by controlling the number of clusters, where the agents with the same rotation degree belong to the same cluster. We defer the details to Table 3 in Appendix F.

Models. For MNIST, we use a CNN, which has two 5×5 convolution layers followed by two fully connected layers with ReLU activation. For CIFAR-10, we evaluate with a more complex network, VGG11 (Simonyan & Zisserman, 2014). In the discrete FL setting, the hypothesis set $P = \{\theta_1, \theta_2, \cdots, \theta_m\}$ is composed of m pre-trained models, which are trained on the same set of images with different rotations: $i \times 4$ degrees (MNIST) or $i \times 2$ degrees (CIFAR-10) for θ_i. We set m to be 100 and 200 for MNIST and CIFAR-10, respectively. We note that pre-training data and local data from agents are non-overlapping. In all our experiments, we define agent utility as $M - \ell_{\text{ce}}$, where ℓ_{ce} refers to the cross entropy loss on the agent's local test data. We set M to be 1.0 in our experiments.

Baselines. We compare our algorithm RankFF with FedAvg and CoreFed. Since both baselines only work in the continuous setting, we adapt them to the discrete setting as follows: each agent evaluates all the models in P and returns the model with the highest local utility. Then the server aggregates the selected models with the aggregation protocols of FedAvg and CoreFed. Finally, the server searches P for the model whose parameter has the smallest Euclidean distance to the aggregated model, since the aggregated model does not necessarily belong to P.

6.2. Experimental Results

Fairness evaluation. We demonstrate that our distributed algorithm RankFF achieves better fairness and utilitarian social welfare compared with baselines, where fairness refers to β-rankwise proportionality. We show the results in the discrete and continuous FL setting in Table 1 and Table 2, respectively. We can see that RankFF achieves a lower β in all settings, which bounds the maximum fraction of models in P that are preferred to the final global model

[6] Polynomial in the dimension of the model-vectors and $1/\varepsilon$.

Table 1. Comparison of fairness (β-rankwise proportionality) and utilitiarian social welfare ($\sum_{i\in[n]} u_i(\theta)$) with RankFF and baselines in the *discrete setting*.

FL Instance	Method	MNIST		CIFAR-10	
		β	$\sum_{i\in[n]} u_i(\theta)$	β	$\sum_{i\in[n]} u_i(\theta)$
I_{high}	FedAvg	0.97	6.68	1.000	57.089
	CoreFed	0.97	6.68	1.000	57.089
	RankFF	**0.39**	**6.99**	**0.770**	**64.226**
I_{mid}	FedAvg	0.60	9.09	0.875	68.560
	CoreFed	0.60	9.09	0.875	68.560
	RankFF	**0.19**	**9.32**	**0.325**	**68.914**
I_{low}	FedAvg	0.17	9.51	0.235	70.949
	CoreFed	0.17	9.51	0.235	70.949
	RankFF	**0.07**	**9.71**	0.235	70.949

Table 2. Comparison of fairness (β-rankwise proportionality) and utilitiarian social welfare ($\sum_{i\in[n]} u_i(\theta)$) with RankFF and baselines in the *continuous setting*.

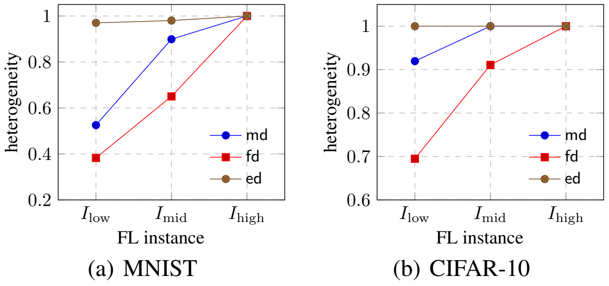
FL Instance	Method	MNIST		CIFAR-10	
		β	$\sum_{i\in[n]} u_i(\theta)$	β	$\sum_{i\in[n]} u_i(\theta)$
I_{high}	FedAvg	0.0792	8.73	0.1287	37.229
	CoreFed	0.0594	8.71	0.1287	37.225
	RankFF	**0.0495**	**9.15**	**0.0792**	**37.594**
I_{mid}	FedAvg	0.0594	9.23	0.1782	37.230
	CoreFed	0.0594	9.28	0.1485	37.230
	RankFF	**0.0495**	**9.58**	**0.0594**	**37.416**
I_{low}	FedAvg	0.0693	9.41	0.1485	37.230
	CoreFed	0.0594	9.38	0.1287	37.231
	RankFF	**0.0396**	**9.72**	**0.0396**	**37.540**

for every agent. In some cases, the differences are striking; for example, for CIFAR-10 and I_{low}, RankFF finds a model that each and every agent prefers to all but at most 4% of possible models (roughly speaking), whereas for FedAvg and CoreFed this value is 15% and 13%, respectively.

Moreover, RankFF achieves higher social welfare, especially when the heterogeneity level is high. We note that although FedAvg maximizes social welfare in the convex setting, RankFF can find a better local optimum for DNNs in our empirical evaluation.

Heterogeneity Measurement. We evaluate the effectiveness of the distance measures we described in Section 4 (md, fd and ed) in reflecting heterogeneity. In Figure 1, we show the heterogeneity $h(I)$ for each FL setting in the discrete setting. We normalize the values to $[0, 1]$ for the purpose of better visualization.

We observe that fd is the most effective in heterogeneity measurement on both datasets. Compared to md, the sum operation is more stable in the presence of extreme cases than the $\max$ operation. However, heterogeneity computed with md can still effectively distinguish different agent groups on MNIST. In addition, when m is large, edit distance changes when the ranks of two models are swapped in two agents with the same data distribution, even if their utilities are similar, thus leading to less significant similarity measurements compared to md and fd. In practice, we could combine different heterogeneity measures for better measurement, as we discuss in Section 4.



(a) MNIST (b) CIFAR-10

Figure 1. Heterogeneity computed with different distance functions for each FL instance. All the distance functions can effectively reflect heterogeneity in agent data distributions. The raw values are normalized.

7. Conclusion

In this paper, we have proposed a new notion of fairness in FL, namely rankwise proportionality, and have established its existence in a wide range of settings: any discrete or measurable set of models and ordinal or cardinal preferences of the agents. We note that these diverse settings arise naturally in many (practical) machine learning problems, e.g. DNNs, where our notion circumvents the *non-existence* issues that plague existing notions such as *proportionality* or *core-stability* from social choice theory. We have also provided empirical evidence that rankwise proportionality achieves strong fairness guarantees in practice. Taken together, our approach and results support the viewpoint that new concepts and techniques thrive at the intersection of ML and social choice theory.

In addition, we have initiated the study of fairness guarantees for FL under limited heterogeneity. While our results are encouraging and some measures (such as footrule distance) appear especially effective, there are compelling open questions regarding how to best define and utilize heterogeneity, including whether to consider the average distance between agents instead of the maximum.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pp. 265–283, 2016.

Caragiannis, I., Procaccia, A. D., and Shah, N. When do noisy votes reveal the truth? *ACM Trans. Economics and Comput.*, 4(3):15:1–15:30, 2016.

Chaudhury, B. R., Li, L., Kang, M., Li, B., and Mehta, R. Fairness in federated learning via core-stability. In *Thirty-sixth Conference on Neural Information Processing Systems*, 2022a.

Chaudhury, B. R., Li, L., Kang, M., Li, B., and Mehta, R. Fairness in federated learning via core-stability. *arXiv preprint arXiv:2211.02091*, 2022b.

Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., Liu, A., Costa, A. B., Wood, B. J., Tsai, C.-S., et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27 (10):1735–1743, 2021.

Donahue, K. and Kleinberg, J. M. Optimality and stability in federated learning: A game-theoretic approach. In *NeurIPS*, pp. 1287–1298, 2021a.

Donahue, K. and Kleinberg, J. M. Models of fairness in federated learning. *CoRR*, abs/2112.00818, 2021b.

Dyer, M. E., Frieze, A. M., and Kannan, R. A random polynomial time algorithm for approximating the volume of convex bodies. In *STOC*, pp. 375–381. ACM, 1989.

Elbir, A. M., Soner, B., and Coleri, S. Federated learning in vehicular networks. *arXiv preprint arXiv:2006.01412*, 2020.

Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33: 19586–19597, 2020.

Huang, T., Lin, W., Wu, W., He, L., Li, K., and Zomaya, A. Y. An efficiency-boosting client selection scheme for federated learning with fairness guarantee. *IEEE Trans. Parallel Distributed Syst.*, 32(7):1552–1564, 2021.

Kelly, F. P., Maulloo, A. K., and Tan, D. K. H. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49(3):237–252, 1998.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

LeCun, Y., Cortes, C., and Burges, C. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Li, T., Sanjabi, M., Beirami, A., and Smith, V. Fair resource allocation in federated learning. In *ICLR*. OpenReview.net, 2020.

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.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Tao, T. *An Introduction to Measure Theory*. Graduate Studies in Mathematics. American Mathematical Society, 2021. ISBN 9781470466404. URL https://books.google.com/books?id=k0lDEAAAQBAJ.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J., and Wang, F. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1):1–19, 2021.

Yang, M., Wang, X., Zhu, H., Wang, H., and Qian, H. Federated learning with class imbalance reduction. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 2174–2178. IEEE, 2021.

A. Non-Existence of Approximate Core-Stability for smooth DNNs

We briefly sketch the argument that no multiplicative approximation of core-stability is achievable in smooth DNNs.

Let $P = [0, 1]$ and $n = 2$. We define the utility functions of the two agents as follows:

$$u_1(\theta) = \begin{cases} L \cdot (\theta - \frac{1}{\sqrt{L}})^2 & \theta \leq \frac{1}{\sqrt{L}} \\ 0 & \theta > \frac{1}{\sqrt{L}} \end{cases}$$

$$u_2(\theta) = u_1(1 - \theta)$$

We first note that both utility functions are $2L$-smooth: Consider $u_1(\cdot)$ and two points θ_1 and $\theta_2 \in [0, 1]$. If both $\theta_1, \theta_2 \in [0, 1/\sqrt{L}]$, then

$$||\nabla u_1(\theta) - \nabla u_1(\theta_2)||_2 = 2L||\theta_2 - \theta_1||_2.$$

If $\theta_1, \theta_2 > 1/\sqrt{L}$, then

$$||\nabla u_1(\theta) - \nabla u_1(\theta_2)||_2 = 0.$$

Lastly, if $\theta_1 \leq 1/\sqrt{L}$, and $\theta_2 > 1/\sqrt{L}$, then

$$\nabla u_1(\theta_2) = \nabla u_1(1/\sqrt{L}) = 0,$$

and therefore

$$\begin{aligned} ||\nabla u_1(\theta_1) - \nabla u_1(\theta_2)||_2 &= ||\nabla u_1(\theta_1) - \nabla u_1(1/\sqrt{L})||_2 \\ &= 2L||\theta_1 - 1/\sqrt{L}||_2 \\ &\leq 2L||\theta_1 - \theta_2||_2. \end{aligned}$$

The highest possible utility for both agents is 1 (for agent 1, it is realized when $\theta = 0$, and for agent 2, it is realized when $\theta = 1$). However, for any θ, one of the agents will have a utility of 0: if $\theta \geq 1/\sqrt{L}$, then agent 1's utility will be 0, while if $\theta < 1/\sqrt{L}$, then agent 2's utility will be at most 0. Thus, at least one of the agents will have a utility of zero and therefore cannot be guaranteed any approximation of the proportionality guarantee (and also consequently core-stability guarantee).

B. Proof of Theorem 3.1

Theorem 3.1. *Given a finite set of models* $P = \{\theta_1, \theta_2, \ldots, \theta_m\}$*, and* any *set of agent preferences* $\langle \succ_1, \succ_2, \ldots, \succ_n \rangle$*, there exists a rankwise proportional and Pareto-optimal* $\theta \in P$.

Proof. Consider a bipartite graph $G = ([n], P, E)$, where there is an edge $(i, \theta) \in E$ for $i \in [n]$ and $\theta \in P$ iff $\mathsf{rank}_i(\theta) \leq \alpha$ where $\alpha = \lceil m(1 - 1/n) \rceil \cdot \frac{1}{m}$. By the definition of rank, each $i \in [n]$ has $(m \cdot \alpha + 1)$ incident edges. Since $\alpha \geq 1 - 1/n$, we have:

$$|E| = n \cdot m \cdot \alpha > m(n - 1).$$

Now suppose for the sake of contradiction that there is no $\theta \in P$ such that for all i, $\mathsf{rank}_i(\theta) \leq \alpha$. Equivalently, each θ has $\leq n - 1$ incident edges in G. Thus:

$$|E| \leq m \cdot (n - 1),$$

which is a contradiction.

The above argument ensures the existence of a rankwise proportional model θ. Then the model $\theta^* = \arg\min_{\theta \in P} \max_i \mathsf{rank}_i(\theta)$ is both rankwise proportional and Pareto-optimal. □

Example 1. This example proves that the guarantee in Theorem 3.1 is tight. Concretely, we show an instance with n agents and m models $P = \{a_1, a_2, \ldots, a_m\}$ in which there is no model $\theta \in P$ s.t. $\mathsf{rank}_i(\theta) < 1 - 1/n$ for all $i \in [n]$. We let $m = n \cdot k$ for an integer k; the example can be generalized when $n \nmid m$ as well.

For $j \in [n]$, let s_j be the string of length k given by $s_j[\ell] = a_{(j-1)k+\ell}$, for $\ell \in [k]$. Then the preference-ranked string σ_i of agent $i \in [n]$ is given by:

$$\sigma_i = s_i \circ s_{i+1} \circ \cdots \circ s_n \circ s_1 \circ \cdots \circ s_{i-1},$$

where $\circ$ denotes concatenation. We illustrate this for $n = m = 3$:

$$\begin{aligned} \sigma_1 &: \theta_1\ \theta_2\ \theta_3 \\ \sigma_2 &: \theta_2\ \theta_3\ \theta_1 \\ \sigma_1 &: \theta_3\ \theta_1\ \theta_2 \end{aligned}$$

We now show that for each model θ_r for $r \in [m]$, there is an agent $i \in [n]$ s.t. $\mathsf{rank}_i(\theta_r) \geq 1 - 1/n$. Let $r = (j-1)k + \ell$, where $j \in [n]$ and $\ell \in [k]$. Then $\theta_r \in s_j$ by definition. From the construction of the preference strings, note that for the agent $i = (j \mod n + 1)$, the string s_j and hence the model θ_r appears among the least ranked k models for i, hence $\mathsf{rank}_i(\theta_r) \geq \frac{(n-1)k}{nk} = 1 - 1/n$. Thus, for the above instance, there is no model $\theta \in P$ s.t. $\mathsf{rank}_i(\theta) < 1 - 1/n$ for every $i \in [n]$, thus showing the guarantee of Theorem 3.1 is tight.

C. Measurabilty and Proof of Theorem 3.2

We refer the reader to standard texts on measure theory e.g. (Tao, 2021) for the definition of a Lebesgue-measurable set. We show below that all sets and functions defined in the current work are Lebesgue-measurable (measurable for short).

C.1. Preliminaries

First, we note that since the space P of models is a non-empty, closed, convex subset of $\mathbb{R}^N$, P is measurable and $\lambda(P) > 0$. We state the definition of a measurable function.

Definition 6. Let $f : D \to \mathbb{R}$ be a function defined on a measurable set D. Then f is measurable if and only if for every $r \in \mathbb{R}$ the set $\{x \in D : f(x) \geq r\}$ is measurable.

In our work, the only assumption we make is that the agent utility functions $\{u_i(\cdot)\}_{i\in[n]}$ are Lebesgue-measurable. This is a very mild assumption, and is weaker than continuity, since any piece-wise continuous function is measurable, as we show below in Proposition 1.

With Definition 6, it is easy to see that the upper-level sets $L_i(\ell) = \{\theta : u_i(\theta) \geq \ell\}$ are measurable when the utility function u_i is measurable. In particular, the sets $P_i(\theta) = \{\theta' : u_i(\theta') \geq u_i(\theta)\} = L_i(u_i(\theta))$ are also measurable.

Proposition 1. *Let $f : D \to \mathbb{R}$ be a piece-wise continuous function defined on a measurable set D. Then f is Lebesgue-measurable.*

Proof. Define $f^{-1}(U) = \{x \in D : f(x) \in U\}$ to be the inverse image of U under f. One can show using Definition 6 that f is measurable iff $f^{-1}(U)$ is measurable for every open set U of $\mathbb{R}$.

We first prove the theorem for a continuous function f. Consider any open set U of $\mathbb{R}$, and take any $x \in f^{-1}(U)$. Since U is open, there is a small-enough neighborhood of $f(x)$ contained in U, i.e., $\mathcal{B}(f(x), \varepsilon) \subseteq U$ for some $\varepsilon > 0$[7]. By continuity of f, there is a neighborhood $\mathcal{B}(x, \varepsilon')$ s.t. $f(x') \in \mathcal{B}(f(x), \varepsilon)$ for all $x' \in B(x, \varepsilon')$. implying that $\mathcal{B}(x, \varepsilon') \subseteq f^{-1}(U)$. Thus for every $x \in f^{-1}(U)$, there is a small neighborhood of x contained in $f^{-1}(U)$, showing that $f^{-1}(U)$ is an open set, and hence is measurable. Thus, f is a measurable function.

Now suppose that f is piece-wise continuous. Then there is a partition of $\mathbb{R}$ into countably many intervals $X_1, X_2, \ldots$ s.t. f is continuous on each A_i. Let $Y_i = f(X_i)$. Then for any open set U, we have

$$f^{-1}(U) = f^{-1}(\cup_i (U \cap Y_i)) = \cup_i f^{-1}(U \cap Y_i).$$

Since f is continuous, $f^{-1}(U \cap Y_i)$ is measurable. Since countable union of measurable sets is measurable, $f^{-1}(U)$ is also measurable for any open set U, thus showing f is measurable. □

C.2. Proof of Theorem 3.2

Theorem 3.2. *In any FL instance with n agents and a measurable set of models P, where each agent i has a Lebesgue-measurable utility function u_i, there always exists a Pareto optimal model θ that is rankwise proportional.*

We first state the following useful lemma, which effectively asserts that the fraction of models with rank at most β is at least β.

Lemma 1. *Let $Z_i = \{\theta \in P : \mathsf{rank}_i(\theta) \leq \alpha_i\}$. Then Z_i is Lebesgue-measurable and $\lambda(Z_i) \geq \alpha_i \cdot \lambda(P)$.*

Proof. We first show that set $Z_i = \{\theta : \lambda(P_i(\theta)) \leq \beta\}$ is measurable, for any $\beta \in \mathbb{R}$.

Let $L_i(\ell) = \{\theta : u_i(\theta) \geq \ell\}$ be the upper-level set of u_i at level ℓ. Note that for $\ell \geq \ell'$, we have $L_i(\ell) \subseteq L_i(\ell')$.

Claim 2. *Let $\ell^* = \inf\{\ell \in \mathbb{R} : \lambda(L_i(\ell)) < \beta\}$. Then $Z_i = \cup_{\ell \geq \ell^*} L_i(\ell)$.*

Proof. For any $\theta \in Z_i$, let $\ell_0 = u_i(\theta)$ and consider any non-decreasing sequence $\{\ell_k\}_{k\in\mathbb{N}}$ with $\lim_{n\to\infty} \ell_n = \ell_0$. Then we have $\{L_i(\ell_k)\}_{k\in\mathbb{N}}$ is inclusion-wise non-decreasing, i.e., $L_i(\ell_0) \supseteq L_i(\ell_1) \supseteq \ldots$. By the property of measure, this implies $\{\lambda(L_i(\ell_k))\}_{k\in\mathbb{N}}$ is non-increasing. Since $L_i(\ell_0) = P_i(\theta)$, we have $\lambda(L_i(\ell_0)) \leq \beta$. Since the sequence $\{\ell_k\}_{k\in\mathbb{N}}$ is non-increasing, $\lambda(L_i(\ell_k)) \leq \beta$. Thus, for $\ell^* = \inf\{\ell : \lambda(L_i(\ell)) < \beta\}$ we have $Z_i = \bigcup_{\ell \geq \ell^*} L_i(\ell)$, because our choices of $\theta \in Z_i$ and the sequence was arbitrary. □

With the above claim, observe that $Z = \cup_{\ell \geq \ell^*} L_i(\ell) = \{\theta : u_i(\theta) \geq \ell^*\} = L_i(\ell^*)$. Since u_i is a measurable function, Definition 6 implies that Z is a measurable set.

Since $\alpha_i \in A_i$, there exists a $\theta \in P$ such that $\mathsf{rank}_i(\theta) = \alpha_i$, i.e., $\lambda(P_i(\theta)) = \alpha_i \cdot \lambda(P)$. Now consider any $\theta' \in P_i(\theta)$, i.e., $u_i(\theta') \geq u_i(\theta)$. Thus by the definition of $P_i(\cdot)$, we have $P_i(\theta') \subseteq P_i(\theta)$. This implies that $\lambda(P_i(\theta')) \leq \lambda(P_i(\theta)) = \alpha_i \cdot \lambda(P)$. By definition of rank, we obtain $\mathsf{rank}_i(\theta') \leq \alpha_i$, i.e., $\theta' \in Z$.

This shows that $P_i(\theta) \subseteq Z$. Thus $\lambda(Z) \geq \lambda(P_i(\theta)) = \alpha_i \cdot \lambda(P)$, as claimed. □

We now turn to the proof of the theorem.

Proof of Theorem 3.2. For $i \in [n]$ and $\theta \in P$, define the function $f_i(\theta) = \mathbb{1}(\mathsf{rank}_i(\theta) \leq \alpha)$, and let $g(\theta) = \sum_i f_i(\theta)$. We show later that f_i and g are Lebesgue-measurable.

Suppose for the sake of contradiction there is no model θ that is rankwise proportional, i.e., for every $\theta \in P$ there is some $i \in [n]$ s.t. $\mathsf{rank}_i(\theta) > \alpha$. Then $\sum_i f_i(\theta) \leq n - 1$ for every $\theta \in P$. Thus:

$$\begin{aligned} \int_{\theta\in P} g(\theta) d\theta &= \int_{\theta\in P} \sum_i f_i(\theta) d\theta \\ &\leq \int_{\theta\in P} (n-1) d\theta \\ &= (n-1) \cdot \lambda(P). \end{aligned} \tag{5}$$

[7] $\mathcal{B}(x, \varepsilon)$ is the open ball of radius ε centered at x.

On the other hand, notice that by changing the order of summation and using $\alpha = \max_i \alpha_i$ we obtain:

$$\begin{aligned}\int_{\theta\in P} g(\theta)d\theta &= \sum_i \int_{\theta\in P} f_i(\theta)d\theta \\ &= \sum_i \int_{\theta\in P} \mathbb{1}(\mathsf{rank}_i(\theta)\le\alpha)d\theta \\ &\ge \sum_i \int_{\theta\in P} \mathbb{1}(\mathsf{rank}_i(\theta)\le\alpha_i)d\theta \\ &= \sum_i \lambda(Z_i),\end{aligned}\tag{6}$$

where $Z_i = \{\theta \in P : \mathsf{rank}_i(\theta) \le \alpha_i\}$. Using Lemma 1 together with $\alpha_i > 1 - 1/n$, we obtain:

$$\begin{aligned}\int_{\theta\in P} g(\theta)d\theta &\ge \sum_i \lambda(Z_i) \ge \sum_i \alpha_i \cdot \lambda(P) \\ &> \sum_i (1-1/n)\cdot\lambda(P) \\ &= (n-1)\cdot\lambda(P).\end{aligned}$$

Putting (5) and (6) together leads to the desired contradiction, showing the existence of a rankwise proportional model θ. Since a Pareto-improvement over θ is also rankwise proportional, we conclude that there exists a Pareto-optimal rankwise proportional model. □

Completing the proof of Theorem 3.2: We now show that the indicator function $f_i(\theta) = \mathbb{1}(\mathsf{rank}_i(\theta) \le \alpha)$ is Lebesgue-measurable for $\alpha \in \mathbb{R}$. This is immediate using Definition 6 and the previous proof. Since the sum of measurable functions is measurable, the function $g = \sum_i f_i$ is also measurable. Lastly, note that the integrals are well-defined since the Lebesgue integral is well-defined for every non-negative Lebesgue-measurable function.

Example 2. Let $P = [0,1]$. Consider two agents with $u_1(\theta) = \theta$, and $u_2(\theta) = 1-\theta$. Then $\mathsf{rank}_1(\theta) = \frac{\lambda([\theta,1])}{\lambda([0,1])} = 1-\theta$, and $\mathsf{rank}_2(\theta) = \frac{\lambda([0,\theta])}{\lambda([0,1])} = \theta$. As Theorem 3.2 indicates, $\theta^* = 1/2$ is a model s.t. $\mathsf{rank}_1(\theta^*) = \mathsf{rank}_2(\theta^*) = 1 - 1/n$, since $n = 2$. But there is no model $\theta \in P$ s.t. $\mathsf{rank}_1(\theta) < 1/2$ and $\mathsf{rank}_2(\theta) < 1/2$, showing Theorem 3.2 is tight. Moreover, for $n, m \ge 2$, Example 1 can be adapted for the continuous case.

D. Missing Proofs from Section 4

D.1. Proof of Theorem 4.1

Theorem 4.1. *Given an instance I with $h(I) = \max_{i,j\in[n]} \mathsf{md}(\mathsf{rank}_i(\cdot), \mathsf{rank}_j(\cdot))$, where $\mathsf{md}(\cdot,\cdot)$ is the maximum displacement distance, there exists a model $\theta \in P$ which is $h(I)$-rankwise proportional.*

Proof. Choose θ to be the best model preferred by agent 1, i.e., $\mathsf{rank}_1(\theta) = 0$. Then by the definition of md and $h(I)$, we have $\mathsf{md}(\mathsf{rank}_1, \mathsf{rank}_i) = |\mathsf{rank}_i(\theta) - \mathsf{rank}_1(\theta)| \le h(I)$, for any $i \ge 2$. Thus, $\mathsf{rank}_i(\theta) \le h(I)$ for all $i \ge 2$, as desired. □

D.2. Proof of Theorem 4.2

Theorem 4.2. *Given an instance I with $h(I) = \max_{i,j\in[n]} \mathsf{fd}(\mathsf{rank}_i, \mathsf{rank}_j)$, where $\mathsf{fd}(\cdot,\cdot)$ is the footrule distance between two rank functions, there exists a model $\theta \in P$ which is $(\sqrt{2(n-1)h(I)/m})$-rankwise proportional.*

Proof. Let $r = \sqrt{2(n-1)mh(I)}$. For sake of contradiction assume that there exists no model $\theta \in P$, such that $\mathsf{rank}_i(\theta) \le r/m$ for all $i \in [n]$. Consider agent 1. For $j \in [r]$, let θ_j denote the model such that $\mathsf{rank}_1(\theta_j) = (j-1)/m$. Note that by our assumption, for each $j \in [r]$, there exists an agent $a(j) \in [n] \setminus \{1\}$, such that $\mathsf{rank}_{a(j)}(\theta_j) > r/m$, implying that $|\mathsf{rank}_{a(j)}(\theta_j) - \mathsf{rank}_1(\theta_j)| > (r-j+1)/m$.

Then observe the following:

$$\begin{aligned}&\sum_{i\in[n]\setminus\{1\}} \mathsf{fd}(\mathsf{rank}_1, \mathsf{rank}_i) \\ \ge &\sum_{i\in[n]\setminus\{1\}} \sum_{j\in[r]} |\mathsf{rank}_i(\theta_j) - \mathsf{rank}_1(\theta_j)| \\ \ge &\sum_{j\in[r]} |\mathsf{rank}_{a(j)}(\theta_j) - \mathsf{rank}_1(\theta_j)| \\ > &\frac{1}{m}\cdot\sum_{j\in[r]}(r-j+1) = \frac{r(r+1)}{2m}\end{aligned}\tag{7}$$

Also, note that since $h(I)$ is an upper bound for the footrule distance between any two rank functions, we have $\sum_{i\in[n]\setminus\{1\}} \mathsf{fd}(\mathsf{rank}_i, \mathsf{rank}_1) \le (n-1)h(I)$. With (7), this implies $r(r+1) < 2(n-1)mh(I)$, which does not hold for $r = \sqrt{2(n-1)mh(I)}$.

Thus there must exist a $\theta \in P$ s.t. $\mathsf{rank}_i(\theta) \le r/m$ for all $i \in [n]$, for $r = \sqrt{2(n-1)mh(I)}$. □

D.3. Proof of Theorem 4.3

Theorem 4.3. *Given an instance I, with $h(I) = \max_{i,j\in[n]} \mathsf{ed}(\mathsf{rank}_i(\cdot), \mathsf{rank}_j(\cdot))$, where $\mathsf{ed}(\cdot,\cdot)$ denotes the edit distance between two rank functions, there exists a model $\theta \in P$, which is $\frac{1}{m}\cdot\big((n-1)h(I)\big)$-rankwise proportional.*

We make two intuitive observation about edit-distance which will be useful in the proof of Theorem 4.3.

Lemma 2. *Consider two strings x and y on Σ, such that $\delta_D(x, y) \leq \beta$. Let $x[i] = y[j] = a \in \Sigma$ be a character that is matched in x and y by the ED-optimal alignment. Then $|i - j| \leq \beta/2$.*

Proof. First note that the edit distance $\delta_D(x, y)$ of two strings x and y can be expressed in terms of the length of the longest common subsequence $\mathsf{lcs}(x, y)$ as follows:

$$\delta_D(x, y) = |x| + |y| - 2 \cdot \mathsf{lcs}(x, y)$$

Now assume indices $i, j \in [m]$ are such that $x[i] = y[j] = a$ is character matched in x and y be the ED-optimal alignment of x and y. Without loss of generality assume $i \geq j$. For sake of contradiction assume $i - j > \beta/2$, where $\delta_D(x, y) = \beta$. Let $x[a, b]$ denote the substring of x between the indices a and b (inclusive), for $a, b \in [m]$. Then observe that:

$$\begin{aligned} &\delta_D(x[1, i-1], y[1, j-1]) \\ &= (i-1) + (j-1) - 2 \cdot \mathsf{lcs}(x[1, i-1], y[1, j-1]) \\ &\geq (i-1) + (j-1) - 2 \cdot (j-1) = i - j, \end{aligned} \tag{8}$$

since $\mathsf{lcs}(x, y) \leq \min |x|, |y|$. Likewise, observe that:

$$\begin{aligned} &\delta_D(x[i+1, m], y[j+1, m]) \\ &= (m-i) + (m-j) - 2 \cdot \mathsf{lcs}(x[i+1, m], y[j+1, m]) \\ &\geq (m-i) + (m-j) - 2 \cdot (m-i) = i - j. \end{aligned} \tag{9}$$

Putting (8) and (9) together, we get:

$$\begin{aligned} &\delta_D(x, y) \\ &= \delta_D(x[1, i-1], y[1, j-1]) + \delta_D(x[i+1, m], y[j+1, m]) \\ &\geq 2(i - j) > \beta, \end{aligned} \tag{10}$$

which is a contradiction since $\delta_D(x, y) = \beta$. Thus it must be that $|i - j| \leq \beta/2$. □

Lemma 3. *Consider two strings x, y that are permutations of Σ with $\delta_D(x, y) \leq \beta$. Then for any set $A \subseteq \Sigma$ with $|A| = \beta/2 + 1$, there exists $c \in A$ s.t. c is a character that is matched in x and y be the ED-optimal alignment.*

Proof. Let $A \subseteq \Sigma$ with $|A| = \beta/2 + 1$. Suppose none of the characters in A are matched in the ED-optimal alignment of x and y. Since x and y are permutations of Σ, each character $c \in A$ contributes a cost of 2 to the edit distance: one each for the character c appearing in x and y, i.e., an insertion and a deletion. Then $\delta_D(x, y) \geq 2 \cdot (\beta/2 + 1) = \beta + 2$, which contradicts $\delta_D(x, y) \leq \beta$. □

We prove Theorem 4.3 by induction on n. For $n = 2$, we invoke Lemma 3 setting x and y to be the preference-ranked strings σ_1 and σ_2, and letting $A = \sigma_1[1, \ldots, (h(I)/2 + 1)]$, i.e., the $(h(I)/2 + 1)$ models ranked best by agent 1. Recall here that $h(I) = \mathsf{ed}(\sigma_1, \sigma_2)$. Then Lemma 3 shows that there exists a $\theta \in P$ s.t. θ is matched by the ED-optimal alignment of σ_1 and σ_2. Moreover since $\theta \in A$ we know $\mathsf{rank}_1(\theta) \leq (h(I)/2)/m$. Then by Lemma 2 we have $\mathsf{rank}_2(\theta) \leq \mathsf{rank}_1(\theta) + h(I)/(2m)$. Thus, $\mathsf{rank}_2(\theta) \leq \frac{h(I)}{m}$, showing the theorem statement holds for $n = 2$.

Now assume the statement holds for $n - 1$, for $n \geq 3$. Let $\ell_{n-1} = (n-2)h(I)$. Consider the set of preferences $\langle \succ_1, \ldots, \succ_{n-1} \rangle$. By the induction hypothesis, there exists a $\theta_1 \in P$, such that $\mathsf{rank}_i(\theta_1) \leq \ell_{n-1}/m$ for all $i \in [n-1]$. Equivalently there is a $\theta_1 \in P$ which is common to all the prefixes $\sigma_1[1 \ldots \ell_{n-1}], \ldots, \sigma_{n-1}[1 \ldots \ell_{n-1}]$. Now, construct the strings σ_i^2 for all $i \in [n-1]$, by deleting θ_1 from σ_i. By our induction hypothesis, there must be a θ_2, common in the prefixes $\sigma_1^2[1 \ldots \ell_{n-1}], \ldots, \sigma_{n-1}^2[1 \ldots \ell_{n-1}]$. Thereafter, we construct σ_i^3 for all $i \in [n-1]$, by deleting θ_2 from σ_i^2, and continue the same argument for $h(I)/2$ many iterations and determine $\theta_1, \theta_2, \ldots, \theta_{h(I)/2+1}$.

By construction, note that for each $i \in [n-1]$ and $j \in \{1, \ldots, (h(I)/2 + 1)\}$, we have $\mathsf{rank}_i(\theta_j) \leq (\ell_{n-1} + h(I)/2)/m$. Now consider the strings σ_1 and σ_n. Since $\delta_D(\sigma_1, \sigma_n) \leq h(I)$, Lemma 3 implies that there exists a model θ_r for some $r \in \{1, \ldots, (h(I)/2 + 1)\}$ which is matched by the ED-optimal alignment of σ_1 and σ_n. Then, by Lemma 2, $\mathsf{rank}_n(\theta_r) \leq \mathsf{rank}_1(\theta_r) + h(I)/m \leq (\ell_{n-1} + h(I)/2)/m + h(I)/(2m) = ((n-1)h(I))/m$. Therefore, $\theta_r \in P$ is a model such that for all $i \in [n]$, $\mathsf{rank}_i(\theta) \leq \frac{1}{m} \cdot ((n-1)h(I))$.

E. Distributed Algorithm

Here we present the full description of `RankFF` in Algorithm 1 (discrete setting) and Algorithm 3 (continuous setting).

F. Details in Experiment Setup

Here, we present how we create FL instances with different heterogeneity levels. Recall that we introduce heterogeneity by rotating the local data of each agent by a specific degree. Table 3 shows the rotation list of agents for different FL Instances.

Table 3. FL instances with different heterogeneity levels for MNIST (10 agents). For CIFAR-10 (100 agents), we repeat the rotation list of MNIST 10 times for each setting so that the number of clusters remains the same.

FL Instance	Degree of Rotation for 10 Agents
I_{high}	0, 0, 20, 20, 40, 60, 100, 120, 180, 200 (8 clusters)
I_{mid}	0, 0, 0, 0, 0, 20, 20, 20, 40, 60 (4 clusters)
I_{low}	0, 0, 0, 0, 0, 0, 20, 20, 20, 20 (2 clusters)

Algorithm 1 `RankFF` (Discrete)

```
Input: Server models $\theta_1, \theta_2, \cdots, \theta_m$
Output: Model weights $\theta$, rankwise proportionality $\beta$
$\beta \leftarrow 1, r \leftarrow 1, C(r) \leftarrow \emptyset$;
for $h = 0, 1, \cdots, \lfloor \log_2 m \rfloor$ do
    Server sends $\theta_1, \theta_2, \cdots, \theta_m$ to agents;
    for $i \in [n]$ in parallel do
        $i$ computes the utility for $\theta_1, \theta_2, \cdots, \theta_m$;
        $\alpha_i(r) \leftarrow$ the $\lfloor rm \rfloor^{\text{th}}$ highest utility;
        $C_i \leftarrow \{c \in P \mid u_i(c) \geq \alpha_i(r)\}$;
        $i$ sends $C_i$ and $u_i(c)$ for $c \in C_i$ to server;
    end for
    if $\bigcap_i C_i$ is not empty then
        $\beta \leftarrow r, C(r) \leftarrow \bigcap_i C_i, r \leftarrow r - 0.5^h$;
    else
        $r \leftarrow r + 0.5^h$;
    end if
end for
$\theta \leftarrow \arg\max_{\theta \in C(r)} \sum_{i \in [n]} u_i(\theta)$.
```

Algorithm 2 `LocalApprox`

```
Input: Model $\theta$, agent $i$, target rank $r$
Parameter: number of models to sample $J$, norm bound of sampled gradients $p$
Output: $\alpha_i(r)$
$c \leftarrow 1$
for $j = 1, 2, \cdots, J$ do
    $\xi_j \sim U(0, p)$; {sample $\xi$ uniformly}
    $\hat{\theta}_j \leftarrow \theta + \xi_j \cdot \frac{\nabla_\theta u_i(\theta)}{\|\nabla_\theta u_i(\theta)\|}$;
    compute the utility $u_i(\hat{\theta}_j)$;
end for
$\hat{\theta}_{\pi_1}, \hat{\theta}_{\pi_2}, \cdots, \hat{\theta}_{\pi_J} \leftarrow \textsc{Sort}\{\hat{\theta}\}_j$ such that $u_i(\hat{\theta}_{\pi_j}) \geq u_i(\hat{\theta}_{\pi_k}), \forall j < k$;
$j^* \leftarrow \lfloor r \times J \rfloor$
$\alpha_i(r) \leftarrow \hat{\theta}_{\pi_{j^*}}$
```

Algorithm 3 `RankFF` (Continuous)

```
Input: Number of iterations of binary search $H$, number of rounds for optimizating with penalty method $T$, initial penalty coefficient $\sigma_0$, penalty scaling factor $\lambda$, interval of penalty scaling $q$
Output: Model weights $\theta^T$, rankwise proportionality $\beta$
$r \leftarrow 1$; {initialize the rank as 1}
for $h = 0, 1, \cdots, H$ do
    Server sends $\theta^t$ to agents;
    for $i \in [n]$ in parallel do
        $\alpha_i(r) \leftarrow$ LocalApprox$(\theta^t, i, r)$;
        $i$ sends $\alpha_i(r)$ to server;
    end for
    for $t = 0, 1, \cdots, T-1$ do
        for $i \in [n]$ in parallel do
            $i$ computes $\nabla_{\theta^t} \Phi_{k,i}(\theta^t)$ on its local dataset $\mathcal{D}_s$, where $\Phi_{k,i}(\theta^t)$ is defined in (4);
            $i$ sends $\nabla_{\theta^t} \Phi_{k,i}(\theta^t)$ to server;
        end for
        Server aggregates the gradients following
            $$\nabla_{\theta^t} \Phi_k(\theta^t) \leftarrow \sum_{i \in [n]} \nabla_{\theta^t} \Phi_{k,i}(\theta^t);$$
        Server updates $\theta^{t+1}$ following
            $$\theta^{t+1} \leftarrow \theta^t - \nabla_{\theta^t} \Phi_k(\theta^t);$$
        Server scales the penalty coefficient by $\lambda$ if $T \mod q = 0$;
    end for
    for $i \in [n]$ in parallel do
        $i$ computes $u_i(\theta^{t+1})$;
        $i$ sends $u_i(\theta^{t+1})$ to server;
    end for
    if $u_i(\theta^{t+1}) \geq \alpha_i(r), \forall i \in [n]$ then
        $\beta \leftarrow r, r \leftarrow r - 0.5^h$;
    else
        $r \leftarrow r + 0.5^h$;
    end if
end for
```