Fairness in model-sharing games

Kate Donahue
kdonahue@cs.cornell.edu
Department of Computer Science, Cornell University
USA

Jon Kleinberg
kleinberg@cornell.edu
Departments of Information and Computer Science,
Cornell University
USA

ABSTRACT
In many real-world situations, data is distributed across multiple self-interested agents. These agents can collaborate to build a machine learning model based on data from multiple agents, potentially reducing the error each experiences. However, sharing models in this way raises questions of fairness: to what extent can the error experienced by one agent be significantly lower than the error experienced by another agent in the same coalition? In this work, we consider two notions of fairness that each may be appropriate in different circumstances: egalitarian fairness (which aims to bound how dissimilar error rates can be) and proportional fairness (which aims to reward players for contributing more data). We similarly consider two common methods of model aggregation, one where a single model is created for all agents (uniform), and one where an individualized model is created for each agent. For egalitarian fairness, we obtain a tight multiplicative bound on how widely error rates can diverge between agents collaborating (which holds for both aggregation methods). For proportional fairness, we show that the individualized aggregation method always gives a small player error that is upper bounded by proportionality. For uniform aggregation, we show that this upper bound is guaranteed for any individually rational coalition (where no player wishes to leave to do local learning).

KEYWORDS
fairness, federated learning, game theory

ACM Reference Format:
Kate Donahue and Jon Kleinberg. 2023. Fairness in model-sharing games. In Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3543507.3583483

1 INTRODUCTION
In traditional machine learning, an agent (such as a hospital, school, or company) trains a model based on data it has locally. In many cases, the agent may only have access to a limited amount of local data, which may mean that any model created with it will have relatively high error. However, in many real-world scenarios, multiple agents all are trying to solve similar problems using related local data sets. For example, consider a hospital aiming to train a model predicting patient outcomes. Local learning would rely on the hospital’s own data from observations of its patients. However, it may be the case that multiple hospitals in the same area all similarly have local patient data related to this problem. This raises the question - could it be possible to leverage this related data to help build a better model?

Multiple research methods address this very problem. For example, data cooperatives and data trusts allow individuals to pool their data, often through the use of a trusted intermediary, in order to safety produce higher performance for those participating [13, 14, 25, 37]. Another method is federated learning, in which a central agent combines learned parameters (rather than the raw data itself) from multiple federating agents [17, 22]. More broadly, edge computing studies ways of leveraging data distributed among multiple nodes with communication limitations [43]. All of these methods have been the subject of multiple game theoretical analyses [3, 8, 9, 12, 15, 20, 26, 36, 45]. We will refer to situations where multiple agents cooperatively share data or model parameters as model-sharing games, using terminology from our prior work [8].

In general, model sharing can provide benefits to participating agents by allowing them to produce models with lower error. However, this may itself produce additional problems, especially around fairness. Often different agents using the same model will see different levels of error, potentially as a result of differences in how much data each agent contributed. While this disparity may be undesirable, a “fair” solution can have multiple definitions, reflecting different values and goals of the particular situation.

For example, consider a case with two hospitals A and B, where A has fewer relevant data points than B. It might be the case that A has less data because it is a smaller hospital with fewer resources. In this case, a reasonable fairness goal might be for the model sharing solution to rectify this inequity, producing errors for A and B that are approximately equal. The goal of having error rates that are the same or similar is reflected in egalitarian fairness. However, a different scenario might call for a different fairness goal. Consider the case where hospitals A and B have similar access to resources, but hospital A has less data because it has devoted less effort to collecting data. In this case, a reasonable fairness objective might be to reward player B for its higher contributions. One specific approach might be to have each agent’s error rate be inversely proportional to the amount of data it contributes: this motivates proportional fairness. While egalitarian and proportional fairness can’t be simultaneously achieved, each definition reflects goals that could be reasonable for different situations. Multiple papers separately discuss notions related to egalitarian or proportional fairness, in data cooperatives [32] or in federated learning [11, 21, 23, 27, 29, 42].
The present work:
In this work, we provide the first theoretical analysis and comparison of egalitarian and proportional fairness. Our paper is primarily analytical rather than prescriptive: we describe patterns of fairness that are and aren’t satisfied for commonly-used methods of model aggregation. In order for this analysis to be tractable theoretically, we build on the simplified model of model-sharing games we first proposed in Donahue and Kleinberg [8], which we summarize later in this work for convenience of the reader.

• For egalitarian fairness, we give a tight multiplicative bound on how dissimilar the error rates of collaborating agents can be. This bound varies based on the number of samples the largest agent contributes, as well as a property of the learning task related to the ratio of noise and bias. We show that this bound holds for a wide range of model-sharing methods satisfying a few straightforward properties.

• For proportional fairness, our results vary based on the model creation method. We identify proportionality as a natural baseline that emerges from the analysis: Proportionality would mean that if agent A has c times the number of samples as agent B, it would obtain 1/c times as much error; this is the property that would hold, for example, for the error rates if none of the agents shared any data. For fine-grained learning, where a personalized model is made for each agent, we show any coalition formation results in the sub-proportional error: that is, small players end up with error rates that are lower than proportionality would suggest. By contrast, for uniform model-sharing, where a single model is created for the entire coalition, we show that in general there exist coalitions that could display error rates that are arbitrarily disconnected from proportionality. That is, the error an agent experiences, compared to the error another agent experiences, could be either more or less than what the ratio of the size of their datasets suggests. However, we show that for uniform model-sharing, any coalition where small players experience error that is larger than proportionality fails to be individually rational: at least one agent could achieve lower error by leaving the coalition for local learning.

Taken together, these results indicate that smaller players (those who contribute fewer samples) in general obtain higher errors in model-sharing games. However, both fairness notions show that this disparity in performance is bounded: either by the multiplicative bound in egalitarian fairness, or by the model aggregation method in proportional fairness. Again, the focus of our work is descriptive: while these fairness results may or may not be desirable in different situations, our goal is to describe certain guarantees that arise from different model-sharing methods.

2 RELATED WORK AND FAIRNESS DEFINITIONS
Fairness in model-sharing is an especially rich topic: naturally, no one set of definitions is going to resolve the complex questions it raises. Much of the literature focuses on fairness in federated learning; for a survey of fairness issues here and methods to address them, refer to [34]. In this work, our goal is to theoretically evaluate two dominant notions of fairness in collaboratively building models1.

2.1 Egalitarian fairness
We adopt the terminology from Xu and Lyu [39] calling the first “egalitarian”. Egalitarian notions of fairness roughly revolve around ensuring agents have levels of error that are all roughly comparable. For example, Mohri et al. [29] works to minimize the error of the worst-off player, while Du et al. [10] considers a similar notion with an explicit fairness penalty. By contrast, Li et al. [21, 23] has a more flexible goal of minimizing the spread of error rates. Abay et al. [1] analyzes multiple causes of unfairness in federated learning and potential solutions. Egalitarian fairness has also been explored in data cooperatives [32]. Other papers that consider notions related to egalitarian fairness include [18, 31, 40, 41].

The motivation for this notion of fairness is that agents may differ in how much data they have for reasons beyond their control: persistent inequities in resources, for example. The goal of model-sharing, in this view, is to try and ameliorate these injustices by bringing error rates more close to each other. We formalize this idea in Definition 1.

DEFINITION 1. If a player $i$ in a coalition $C$ experiences error $err_i(C)$, then the coalition $C$ satisfies $\lambda$-egalitarian fairness if for some constant $\lambda \geq 1$, 
\[
\frac{err_i(C)}{err_j(C)} \leq \lambda \forall i, j \in C
\]

2.2 Proportional (collaborative) fairness
By contrast, proportional notions of fairness generally focus on ensuring that different agents are rewarded for contributing more to the overall model [11]. For example, Lyu et al. [27], Xu and Lyu [39] both use gradient-based methods to detect players contributing useful model updates (as opposed to noise) and reward them with more powerful variants of the model. Song et al. [35] takes a similar approach with the goal of allocating profit. Zhang et al. [42] proposes having a trusted party track the amount of effort each player is contributing, which will then be used to reward or penalize agents with access to different models.

Under the view of proportional fairness, the number of samples a player contributes to the shared model is under their control, so we wish to reward players who contribute more. In Definition 2, we formalize the idea of proportional fairness by describing how error rates relate to the number of samples that a player contributes.

DEFINITION 2. A coalition $C$ with satisfies $\alpha$-proportional fairness if for all pairs of players $i$ and $j$, with $n_i \leq n_j$, some $\alpha > 0$, 
\[
err_i(C) \leq \alpha \frac{n_j}{n_i}, \quad err_j(C) \forall i, j \in C, n_i \leq n_j
\]

Note that $\alpha < 1$ means that smaller players get lower error than proportionality would suggest, while $\alpha > 1$ means that smaller players get higher error than proportional. Either case may be desirable under different situations: we will restrict our attention to describing when either is possible to achieve.

1 Naturally, some fairness notions fall outside of these two notions, such as [5], which proposes and analyzes definitions fairness based on core stability of coalitions.
3 MODEL AND ASSUMPTIONS

3.1 Model sharing methods

Next, we will present the framework we will use to analyze model-sharing. We say that there are \( M \) total agents (sometimes referred to as players). Each agent \( i \in [M] \) has \( n_i \) data points, labeled by its true labeling distribution \( g(\theta_i) \), where \( \theta_i \) are their true local parameters and \( g(\cdot) \) is some labeling function (e.g., mean estimation, linear regression, deep learning). A player’s goal is to learn a model \( g(\hat{\theta}_i) \) with low expected error on its own distribution. If a player opts for local learning, then it uses its local estimate of these parameters \( \hat{\theta}_i \) to predict future data points. If a set of players \( C \) are building a model together, we say that they are in a coalition or cluster together. An aggregation method describes how a coalition aggregates information from participating agents to. For example, in federated learning, a federating method describes how players combine their learned parameters into federating models. The most common federating method is the weighted average method in Equation (1), often called “vanilla” or “uniform” federated learning ([28]):

\[
\hat{\theta}_C = \frac{1}{\sum_{i \in C} n_i} \sum_{i \in C} n_i \cdot \hat{\theta}_i
\]

(1)

Alternative ways of federation might involve customizing the model for the individual, as in domain adaptation. For example, [8] models three methods of federation: vanilla, as well as two models of domain adaptation, fine-grained and coarse-grained. In fine-grained federation, a personalized model is made for each player by weighting the contributions from every other player:

\[
\hat{\theta}_j = \sum_{i=1}^{M} v_{ji} \cdot \hat{\theta}_i
\]

(2)

with \( \sum_{i=1}^{M} v_{ji} = 1 \). “Optimal” fine-grained federation occurs when these \( v_{ji} \) weights are set so as to minimize error for each player and is the case we will consider throughout the rest of this paper.

A player \( i \) obtains error \( err_j(C) \), where \( m \) denotes the federating method being used. Note that in general \( err_j(C)m \neq err_j(C)m \); players in the same coalition may experience different error rates. For example, if player \( i \) has more samples than player \( j \), then \( \hat{\theta}_C \), as calculated in vanilla federation in Equation (1) will be weighted more towards player \( i \), meaning that player \( i \) will have lower expected error than \( j \). While these aggregation methods are motivated by federated learning, they could also be used to model combining methods in other settings, such as data coalitions.

3.2 Theoretical model from Donahue and Kleinberg [8]

All of the previous definitions are satisfied for any model-sharing scenario. However, in order for our theoretical fairness analysis to be feasible, we require a model that gives a functional form for \( err_j(C) \): the error each player experiences in a coalition. Moreover, we require a model that gives exact errors for each player, rather than bounds: these are needed in order to be able to analyze the ratio of the errors two players experience, for example.

We opt to use the model developed in our prior work Donahue and Kleinberg [8], which produces the closed-form error value seen in Lemma 1 below. While we work within this model, we emphasize that [8] asked different questions from this paper’s focus: our prior work focused on developing the federated learning model and analyzing the stability of federating coalitions, while our current work analyzes optimality and Price of Anarchy. (Separately, in [9] we similarly build on [8], but with the orthogonal goal of analyzing the optimality of federated learning: which arrangements minimize total error.)

**Lemma 1 (Lemma 4.2, from [8]).** Consider a mean estimation task as follows: player \( j \) is trying to learn its true mean \( \theta_j \). It has access to \( n_j \) samples drawn i.i.d. \( Y_i \sim \mathcal{D}_j(\theta_j, \sigma^2) \), a distribution with mean \( \theta_j \) and variance \( \sigma^2 \). Given a population of players, each has parameters \((\theta_j, \sigma^2_j) \sim \Theta \) taken from some distribution \( \Theta^2 \). A coalition \( C \) using uniform federation produces a single model based on the weighted average of local means (Eq. (1)). Then, the expected mean squared error player \( j \) experiences in coalition \( C \) using vanilla federation is:

\[
\text{err}_j(C) = \frac{\mu_e}{\sum_{i \in C} n_i} + \sigma^2 \left( \frac{\sum_{i \in C} n_i \sigma^2_i + \left( \sum_{i \in C, i \neq j} n_i \right)^2}{(\sum_{i \in C} n_i)^2} \right)^2
\]

where \( \mu_e = \mathbb{E}_{(\theta_i, \sigma^2_i) \sim \Theta} \mathbb{E}_{y_i \sim \mathcal{D}_i} \) (the average noise in data sampling) and \( \sigma^2 = \text{Var}(\theta) \) (the average dissimilarity or bias between the true means of players).

**Lemma 2 (Lemmas 4.1, 7.1, from [8]).** Consider the same setting as in Lemma 1, but where the coalition \( C \) was instead using personalized (fine-grained) federation. Then, the expected mean squared error player \( j \) experiences in the coalition with fine-grained federation is given by:

\[
\text{err}_j = \mu_e \left( \sum_{i \neq j} v_{ji}^2 \right)^{1/2} + \sigma^2 \left( \sum_{i \neq j} v_{ji}^2 \right)^{1/2} \cdot \sigma^2
\]

(4)

“Optimal” fine-grained federation occurs when each \( v_{ji} \) parameter is set so as to minimize error for player \( j \). Note that every player can simultaneously minimize their errors: player \( k \neq j \) can set their parameters \( v_{kj} \) separately from \( v_{ji} \). The values of \( v_{ji} \) that minimize error are given by:

\[
v_{ji} = \frac{1 + \sigma^2 \sum_{i \neq j} V_{ij}}{1 + \sum_{i \neq j} V_{ij}} \cdot \frac{1}{V_{ji}} \quad v_{jk} = \frac{1}{V_k} \cdot \frac{V_j - \sigma^2}{1 + \sigma^2 \sum_{i \neq j} V_{ij} V_j}
\]

for \( V_i = \sigma^2 + \frac{\mu_e}{n_i} \).

Note that [8] also analyzes a \( d \)-dimensional linear regression game and in particular show that the cost functions for this game are equivalent to the costs for mean estimation, so long as the number of samples \( n \) is much larger than the dimension of the problem \( d \). As a result, all of our findings directly extend from mean estimation to linear regression. Additionally, theoretical and empirical work ([16, 33]) indicates that power law rules for error curves are commonly used, indicating that functional forms similar to those proposed in [8] may be much more broadly applicable.

\(^{3}\text{Note that this does not imply that each player draws their parameters i.i.d, but rather that there exists some distribution of true model parameters among players. For example, } \Theta \text{ might be an (unknown) distribution of true coronavirus rates among towns, reflecting variability in true rates. By contrast, } \mathcal{D}_j(\theta_j, \sigma^2_j) \text{ would be the number of coronavirus cases in a particular town per day, which reflects noise in data collection.} \)
We use some of the same modeling assumptions as [8]. For example, we assume that each agent has a goal of obtaining a model with low expected test error on its personal distribution. For notation, we use $C$ to refer to a coalition of federating agents and $\Pi$ to refer to a collection of coalitions that partitions the $M$ agents. We will use $N_C$ to refer to the total number of samples present in coalition $C$: $N_C = \sum_{i \in C} n_i$. We will refer to $\frac{C}{\sum \mu C}$ as the noise/bias ratio, given that it reflects the level of noise in a learning problem, as compared to the average amount of bias of the true parameters of different players. We also assume that each agent is always "available" to participate in model-sharing, meaning that our setting is more close to cross-silo federated learning, as opposed to cross-device [38].

Finally, it is worth emphasizing key differences between this current work and [8]. The focus of [8] is defining a theoretical model of federated learning and taking a game-theoretic analysis of the stability of such an arrangement. While we use the theoretical model we developed in [8], in this work we analyze completely distinct questions: what are reasonable definitions of "fairness" in federated learning? When are certain fairness notions guaranteed to hold? Additionally, this paper is in some ways more general: while some key results in [8] only allow players to have two different numbers of samples ("small" or "large"), every result in this work holds for arbitrarily many players with arbitrarily many different numbers of samples.

3.3 Research ethics and social impact
In Section 2 we began our discussion on fairness and ethics, which we will continue here. This work is primarily theoretical, but touches on questions with multiple possible societal implications. As we discussed previously, there are many reasonable definitions of fairness in model-sharing games (each of which could impact participants differently). In this work, we narrow our focus to two of the most commonly used and natural definitions, but it is important to note that this by no means is the last word on fairness in this area. For example, fairness could mean something besides error rates, such as equality of access to opportunities or access to recourse.

Besides fairness, model and data sharing also has privacy concerns. While federated learning involves sharing model parameters, rather than raw data, under some circumstances it still may be possible to learn information about the data distribution of specific players [4, 30]. Some players have explored the game-theoretic implications of privacy concerns in federated or data-sharing settings [2, 7, 24, 44], as well as [19], which builds on the model-sharing games framework from [8]. Finally, one core assumption of this paper is that the problem at hand will be improved by a machine learning model with lower error. In many situations, a different approach might be more suitable, such as an algorithm that is more explainable or a non-machine learning solution.

4 EGALITARIAN FAIRNESS
4.1 Motivation
In this section, we will analyze egalitarian fairness, where a typical goal is to ensure that all players have similar levels of error. This may be motivated by the fact that players may differ in their circumstances (differing amounts of data) for reasons out of their control. For example, imagine a low-resource hospital that is collaborating with a high-resource hospital. The low-resource hospital may have access to less local data for historical and/or continuing reasons of discrimination and unequal access to technology and other resources. This motivation would suggest that model-sharing algorithms should seek to close the gap the low resource and high resource hospital, making their error as close to equal as possible. In some cases, it may even be desirable for the low-resource (small) hospital to have lower error than the high-resource hospital, potentially as a way to begin to compensate for its lack of resources. In this section, we give theoretical bounds for what types of egalitarian fairness levels can be obtained.

4.2 Tight, widely-applicable upper bound
Theorem 1 shows that there exists a tight multiplicative bound for egalitarian fairness: that is, the error ratio between small and large players is upper bounded. This holds for any modular function, where definition 3, below, gives a formal definition of what this means. Later, we will directly show that both vanilla and fine-grained federation are modular. However, because the definition of “modular” encompasses multiple desirable properties of model-aggregation methods, we expect that a wide range of model-aggregation methods would satisfy them. Thus, Theorem 1 is best viewed as a relatively broad theorem potentially providing insight into multiple different aggregation scenarios.

Theorem 1. Any “modular” model-sharing method (as defined in Definition 3) satisfies the egalitarian error ratio bound:

$$\frac{\text{err}_s(C)}{\text{err}_l(C)} \leq 2 \cdot c + 1 \forall i, j \in C$$

so long as the largest player size is $\leq c \cdot \frac{n_s}{\sigma}$ (for $\frac{n_s}{\sigma}$ noise/bias ratio, for any $c > 0$). Moreover, this bound is tight (up to an additive factor of $\epsilon$).

Definition 3. Consider any coalition $C$ with players $s, l \in C$ such that $n_s \leq n_l$. Then, a model-sharing method is called modular if it satisfies the following five properties:

Property 1: The large player always has lower error than the small player:

$$\frac{\text{err}_s(C)}{\text{err}_l(C)} \geq 1$$

Property 2: The worst-case situation for the error ratio (the ratio of the small player’s error to the large player’s error) is always in the two-player case. That is,

$$\frac{\text{err}_s(C)}{\text{err}_l(C)} \leq \frac{\text{err}_s({n_s, n_l})}{\text{err}_l({n_s, n_l})}$$

Property 3: The error ratio increases as the large player gets more samples: the small player’s error either increases or else decreases more slowly than the large player’s error decreases. That is,

$$\frac{\partial}{\partial n_l} \text{err}_l({n_s, n_l}) \leq 0$$

Property 4: The error ratio decreases as the small player gets more samples: the large player’s error either increases or else decreases more slowly than the small player’s error decreases. That is,

$$\frac{\partial}{\partial n_s} \text{err}_s({n_s, n_l}) \leq 0$$
**Property 5:** As $\frac{n_s}{n_l} \to 0$ (the large player has many more samples than the small player), the error ratio converges to the following fraction:

$$\lim_{n_s/n_l \to 0} \frac{\text{err}_s(\{n_s, n_l\})}{\text{err}_l(\{n_s, n_l\})} = \frac{\mu_c}{n_l} + 2\sigma^2$$

This fraction reflects the fact that the large player’s error approaches that of local learning, while the small player’s error approaches that of simply using the model of the large player without contributing any data.

First, it may be useful to discuss the properties in Definition 3. Property 1 is simply stating that, for any pair of players both federating in the same coalition, the smaller of the two always has higher error. The remaining four properties contribute to the upper bound (and showing that this bound is tight). First, Property 2 states that the error ratio between any two players is highest when they are federating in a two-player coalition (without any other players). This property is especially helpful because the two-player coalition is much simpler to analyze. Properties 3 and 4 consider this two-player coalition, showing that the error ratio is highest when the large player has many more samples and the small player has very few. Finally, Property 5 describes the limit of the error ratio in the case where the large player has many more samples than the small player, stating that it must go to a fixed ratio depending on the noise-bias ratio, as well as the number of samples the large player has. Roughly speaking, the denominator of this term ($\frac{\mu_c}{n_l}$) approximates the error that the large player $\ell$ obtains while federating alone, while the numerator reflects the error that the small player obtains while using the model created almost entirely by the large player ($\mu_c$, plus a bias term of $2\sigma^2$).

Given these properties, it becomes straightforward to prove Theorem 1:

**Proof of Theorem 1.** Proving the $2c + 1$ bound is given by applying the four properties:

$$\frac{\text{err}_s(\{n_s, n_l\})}{\text{err}_l(\{n_s, n_l\})} \leq \frac{\mu_c}{n_l} + 2\sigma^2 \leq \frac{c^2}{n_l} + 2\sigma^2 = 2c + 1$$

Next, we will show that the bound is tight. Our goal is to show that, for all $\epsilon > 0$, there exists some set of parameters $n_s > 1, n_l \leq c \cdot \frac{\mu_c}{\sigma^2}, \mu_c, \sigma^2 \geq 0$ such that

$$\frac{\text{err}_s(\{n_s, n_l\})}{\text{err}_l(\{n_s, n_l\})} \geq 2c + 1 - \epsilon$$

The $\epsilon, \delta$ version of Property 4 can be written as: for all $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\frac{n_s}{n_l} < \delta \Rightarrow \frac{\mu_c}{n_l} + 2\sigma^2 - \frac{\text{err}_s(\{n_s, n_l\})}{\text{err}_l(\{n_s, n_l\})} < \epsilon$$

Setting $n_s = 1, n_l = c \cdot \frac{\mu_c}{\sigma^2}$ gives that Property 4 becomes:

$$\frac{\sigma^2}{c \cdot \mu_c} < \delta \Rightarrow 2c + 1 - \frac{\text{err}_s(\{n_s, n_l\})}{\text{err}_l(\{n_s, n_l\})} < \epsilon$$

$$\frac{\sigma^2}{c \cdot \mu_c} < \delta \Rightarrow \frac{\text{err}_s(\{n_s, n_l\})}{\text{err}_l(\{n_s, n_l\})} - \epsilon > 2c + 1 - \epsilon$$

Because $\sigma^2, \mu_c$ are free parameters, we can set $\sigma^2 = 1, \mu_c > \frac{1}{c} \delta$ in order to satisfy the precondition.

### 4.3 Proving aggregation methods are modular

The properties in Definition 3 are relatively logical, desirable properties we would expect many aggregation methods would satisfy. In this section, we will theoretically prove that there exist multiple federating methods that are modular. Specifically, Theorem 2 shows that uniform federation is modular and Theorem 3 shows that fine-grained federation is modular.

**Theorem 2.** Uniform federation is modular and therefore satisfies the $2 \cdot c + 1$ egalitarian fairness bound from Theorem 1.

**Theorem 3.** Fine-grained federation is modular and therefore satisfies the $2 \cdot c + 1$ egalitarian fairness bound from Theorem 1.

The proofs of these theorems are relatively long and involved, so they are deferred to the supplementary material (or the full version, available at https://arxiv.org/abs/2112.00818). However, we include a proof sketch for a portion of Theorem 2 to give intuition for the structure of these proofs.

**Proof sketch of Theorem 2.** In this proof, we will denote the uniform federation error by $err^u$. Because Definition 3 has five components, this proof will have five sections, but in this proof sketch, we will only present Property 2.

**Property 2:**

For the second property, we wish to show that the worst-case ratio of errors occurs in the two-player case. That is,

$$\frac{\text{err}_s(C)}{\text{err}_l(C)} \leq \frac{\text{err}_s(\{n_s, n_l\})}{\text{err}_l(\{n_s, n_l\})}$$

In order to prove this, we’ll show something stronger. Take any player $k \in C$ (with $k \neq s, l$). Then, we will show that the derivative of the ratio with respect to the size of player $k$ ($n_k$) is always negative:

$$\frac{\partial}{\partial n_k} \frac{\text{err}_s(C)}{\text{err}_l(C)} < 0$$

For uniform federation, we know that in the limit of a player contributing very few samples, its effect on the shared model approaches that of if it didn’t participate. Put another way,

$$\lim_{n_k \to 0} \frac{\text{err}_s(C)}{\text{err}_l(C)} = \text{err}_s(C \setminus \{n_k\})$$

and similarly for the large player. This tells us that we can compare any coalition $C$ into the two-player coalition $\{n_s, n_l\}$ by sending the size of every other player to 0. If we can show that $\frac{\partial}{\partial n_k} \frac{\text{err}_s(C)}{\text{err}_l(C)} < 0$, that means this process will only increase the ratio of the errors $\text{err}_s(C)/\text{err}_l(C)$. This would show that the worst-case ratio of errors occurs in the two-player coalition.

Next, we will start the proof. The derivative of the error ratio is given by:

$$\frac{\partial}{\partial n_k} \frac{\text{err}_s(C)}{\text{err}_l(C)} = \frac{\text{err}_s(C \setminus n_k)}{\text{err}_l(C \setminus n_k)} \cdot \frac{\text{err}_l(C)}{\text{err}_s(C)} - \frac{\text{err}_s(C)}{\text{err}_l(C)} \cdot \frac{\text{err}_s(C \setminus n_k)}{\text{err}_l(C \setminus n_k)}$$

This is negative whenever:

$$\frac{\text{err}_s(C \setminus n_k)}{\text{err}_l(C \setminus n_k)} < \frac{\text{err}_s(C \setminus n_k)}{\text{err}_s(C)} \quad \text{and} \quad \frac{\text{err}_l(C \setminus n_k)}{\text{err}_l(C)} < \frac{\text{err}_s(C \setminus n_k)}{\text{err}_s(C)}$$
To show this result, we will calculate the lefthand side (ratio relating to the small player’s error) and then show that it is less than the equivalent ratio for the large player.

In order to analyze this ratio, we will need the functional form of expected error each player receives while using uniform federation, which we take from Lemma 1 ([8]). The error of the small player can be written:

\[
\frac{\mu_e \cdot N + \sigma^2 \left( \sum_{i \in C} n_i^2 + N^2 - 2N \cdot n_s \right)}{N^2}
\]

where for notational convenience we have used \( N = \sum_{i \in C} n_i \) to denote the total number of samples among all players in coalition \( C \). The derivative of the error with respect to \( n_k \) is:

\[
\frac{\mu_e \cdot N + 2 \cdot \sigma^2 \left( (n_k + N - n_s) \cdot N - \left( \sum_{i \in C} n_i^2 + N^2 - 2N \cdot n_s \right) \right)}{N^3}
\]

\[
= \frac{-\mu_e \cdot N + 2 \cdot \sigma^2 \left( (n_k + N - n_s) \cdot N - \left( \sum_{i \in C} n_i^2 + N^2 - 2N \cdot n_s \right) \right)}{N^3}
\]

The term we’re interested in is the ratio of the derivative of the small player’s error (which we just calculated) to the error of the small player. Note that the error itself can be written as:

\[
\frac{\mu_e \cdot N + \sigma^2 \left( \sum_{i \in C} n_i^2 + N^2 - 2N \cdot n_s \right)}{N^2}
\]

So the ratio of the derivative to the overall error is:

\[
\frac{1}{N} \cdot \frac{-\mu_e \cdot N + 2 \cdot \sigma^2 \left( (n_k + N - n_s) \cdot N - \left( \sum_{i \in C} n_i^2 + N^2 - 2N \cdot n_s \right) \right)}{\mu_e \cdot N + \sigma^2 \left( \sum_{i \in C} n_i^2 + N^2 - 2N \cdot n_s \right)}
\]

What we would like to show is that the above term is less than the analogous term for the \( n \) variant, which can be symmetrically written as:

\[
\frac{1}{N} \cdot \frac{-\mu_e \cdot N + 2 \cdot \sigma^2 \left( (n_k + N - n) \cdot N - \left( \sum_{i \in C} n_i^2 + N^2 - 2N \cdot n \right) \right)}{\mu_e \cdot N + \sigma^2 \left( \sum_{i \in C} n_i^2 + N^2 - 2N \cdot n \right)}
\]

This is equivalent to proving:

\[
\frac{B + 2\sigma^2 \cdot N \cdot n_k}{A - 2 \cdot \sigma^2 \cdot N \cdot n} < \frac{B + 2\sigma^2 \cdot N \cdot n}{A - 2 \cdot \sigma^2 \cdot N \cdot n}
\]

for

\[
A = \mu_e \cdot N + \sigma^2 \cdot \left( \sum_{i \in C} n_i^2 + N^2 \right)
\]

\[
B = -\mu_e \cdot N + 2\sigma^2 \cdot \left( (n_k + N) \cdot N - \left( \sum_{i \in C} n_i^2 + N^2 \right) \right)
\]

We can cross multiply the inequality to get:

\[
(B + 2\sigma^2 \cdot N \cdot n_k) \cdot (A - 2 \cdot \sigma^2 \cdot N \cdot n) < (B + 2\sigma^2 \cdot N \cdot n) \cdot (A - 2 \cdot \sigma^2 \cdot N \cdot n_k)
\]

\[
0 < (B + 2\sigma^2 \cdot N \cdot n_k) \cdot (A - 2 \cdot \sigma^2 \cdot N \cdot n) - (B + 2\sigma^2 \cdot N \cdot n) \cdot (A - 2 \cdot \sigma^2 \cdot N \cdot n_k)
\]

Which simplifies to:

\[
0 < 2 \cdot (A + B) \cdot (n_k - n) \cdot N \cdot \sigma^2
\]

Because we have assumed that \( n_k > n \) (the larger player has more samples than the smaller player), this is true if \( A + B > 0 \). We can evaluate \( A + B \) as being:

\[
A + B = \mu_e N + \sigma^2 \left( \sum_{i \in C} n_i^2 + N^2 \right) - \mu_e N + 2\sigma^2 \left( (n_k + N) \cdot N - \left( \sum_{i \in C} n_i^2 + N^2 \right) \right)
\]

\[
= 2\sigma^2 \cdot (n_k + N) \cdot N - \sigma^2 \cdot \left( \sum_{i \in C} n_i^2 - N^2 \right)
\]

\[
= 2\sigma^2 \cdot n_k \cdot N + \sigma^2 \cdot \left( N^2 - \sum_{i \in C} n_i^2 \right) > 0
\]

as desired. This proves that

\[
\frac{\partial}{\partial n_k} \frac{\text{err}}{\text{err}_t} (C) < 0
\]

which by our prior reasoning indicates that for uniform federation, the worst-case error ratio occurs in the two-player coalition.

\[
\frac{\text{err}}{\text{err}_t} (C) \leq \frac{\text{err}}{\text{err}_t} (\{n_s, n_l\})
\]

\[\square\]

For a full proof of Theorem 2 (including Properties 1, 3-5) and a proof of Theorem 3 see the supplementary material.

Theorems 1, 2, 3, taken together, show that while small players must get higher error than larger players in the same federation, the error that small players experience is upper bounded. Specifically, this upper bound is a multiple of the error that the larger player experiences, and depends on the size of the larger player, as a function of the noise-bias ratio. Because this bound is tight, we know that no lower bound is possible.

5 PROPORTIONAL FAIRNESS

5.1 Motivation

Next, we consider proportional fairness, which revolves around the idea that players who contribute more data should be rewarded with lower error rates. This notion relies on the view that players do have control over how much data they obtain. For example, imagine again two hospitals, this time with comparable levels of resources, but one of them has devoted a substantially larger share of its annual budget towards collecting high-quality local data. As a result, it has substantially more local data it can use for training. A model-sharing coalition that gives both hospitals roughly similar experiences, and depends on the size of the larger player, as a function of the noise-bias ratio. Because this bound is tight, we know that no lower bound is possible.
5.2 Fine-grained (personalized) always satisfies $\alpha \leq 1$

Our first result is about fine-grained federation - when the aggregated model is personalized for each participating player. In this case, Theorem 4 gives us an extremely clean result: proportional error always satisfies $\alpha \leq 1$. That is, for any two players that are federating, $n_s \leq n_l$, the small player always obtains error that is upper bounded by the error of the large player, multiplied by $\frac{n_s}{n_l}$.

**Theorem 4.** Optimal fine-grained federation always has $\alpha \leq 1$: the smaller player experiences lower error than proportionality would suggest.

**Proof sketch.** In order to show that $\alpha \leq 1$, we will show that:

$$err_s(C) \leq err_l(C) \cdot \frac{n_l}{n_s} \quad n_s \leq n_l$$

or:

$$err_s(C) \cdot n_s \leq err_l(C) \cdot n_l$$

where $err_s(C)$ and $err_l(C)$ are the error that the small and large players respectively obtain while participating in coalition C and using fine-grained federation.

In order to analyze this, we use the functional form of error for fine-grained federation from Lemma 2:

$$\frac{\mu}{V_j} \sum_{i=1}^{M} v_{ij} \cdot \frac{1}{n_i} + \left( \frac{\sum v_{ij}}{\sum_{i \neq j} v_{ij}} \right)^2 \cdot \sigma^2$$

(5)

where the parameters $v_{ij}$ are set so as to minimize error:

$$v_{jj} = \frac{1 + \sigma^2 \sum_{i \neq j} v_{ij}}{1 + V_j \sum_{i \neq j} v_{ij}}$$

for $V_j = \sigma^2 + \frac{\mu}{V_j}$. We’re interested in the term given by $err_s(C) \cdot n_s$.

Plugging in the $v_{jk}$ parameters into Eq. (5) gives:

$$\mu \cdot \frac{1 + \sigma^2 \sum_{i \neq j} \frac{1}{V_i}}{(1 + V_j \sum_{i \neq j} \frac{1}{V_i})^2} \cdot n_j \cdot \sigma^2 \cdot \left( \frac{1 + V_j \sum_{i \neq j} \frac{1}{V_i} - 1 - \sigma^2 \sum_{i \neq j} \frac{1}{V_i}}{1 + V_j \sum_{i \neq j} \frac{1}{V_i}} \right)^2$$

and:

$$+ n_j \cdot \left( \frac{\mu}{\sum_{i \neq j} \left( \frac{1}{V_j} + \frac{1}{V_i} \right)} \right) \cdot V_i$$

Strategic simulation (details in full proof) shows that the above equation can be dramatically simplified to this form:

$$\frac{\mu}{V_j} \sum_{i=1}^{M} \frac{1}{V_i} \left( 1 + \sigma^2 \sum_{i \neq j} \frac{1}{V_i} \right)$$

Next, in order to show that $\alpha \leq 1$, we need to show that:

$$\frac{\mu}{V_j} \sum_{i=1}^{M} \frac{1}{V_i} \left( 1 + \sigma^2 \sum_{i \neq j} \frac{1}{V_i} \right) \leq \frac{\mu}{V_l} \sum_{i=1}^{M} \frac{1}{V_i} \left( 1 + \sigma^2 \sum_{i \neq j} \frac{1}{V_i} \right)$$

Further analysis shows that this holds exactly whenever:

$$V_l \leq V_s$$

Note that because $V_l = \sigma^2 + \frac{\mu}{V_l}$,

$$n_s \leq n_l \implies V_l \geq V_s$$

This means that the overall proof is concluded: using optimal fine-grained federation, sub-proportional error is always satisfied. □

5.3 Vanilla federation satisfies $\alpha \leq 1$ if individually rational

While Theorem 4 gives a very clean result for fine-grained federation, the situation for vanilla federation is much less clear. Lemma 3 shows empirically that there exist situations where a federating coalition using vanilla federation could have $\alpha$ values on either side of $\alpha = 1$.

**Lemma 3.** There exist cases where a coalition using vanilla federation satisfies $\alpha < 1$, $\alpha = 1$, or $\alpha > 1$: that is, the smaller player could experience error that is lower than, equal to, or greater than proportionality would suggest.

**Proof.** This proof is constructive: we give an explicit example of such situations. Throughout, we use the error form given in Lemma 1 with $\mu = 10, \sigma^2 = 1$. We will consider the case where there are exactly two players: $n_s = 6$, and $n_l$ with a varying number of samples.

The first row gives the error that the small and large players experience with $n_l = 20$. Note that the implied $\alpha$ value (sixth column) is less than 1, which is because the ratio of their errors (given in the third column) is strictly less than the (inverse) ratio of their sizes (given in the fifth column). This means that the small player gets lower error than proportional scaling with respect to their sizes would suggest. The second row gives the error that the players experience with $n_l = 30$. Here, the ratio of their errors exactly matches the ratio of their sizes: this is a case of exact proportionality. Finally, the third row gives the error the players experience when $n_l = 40$. Here the ratio of their errors is strictly greater than the ratio of their sizes: this means $\alpha > 1$.

In later analyses, we will show that all arrangements with $\alpha > 1$ fail to be individually rational: at least one federating agent would prefer local learning. For the example with $n_l = 40$, the error form
from Lemma 1 gives that the large player would obtain error of 0.25 with local learning. This is strictly less than its federating error of 0.251, so it is incentivized to leave the federating coalition. 

Note that Table 1 also illustrates the bound derived in Theorem 1. Given that the error/bias ratio here is \( \frac{\alpha}{\mu} = \frac{10}{10} = 1 \), the bound scales with the size of the largest player, in multiples of 10. The bound in the fifth column is always strictly greater than the ratio in the third column, illustrating the result from Theorem 1.

Next, Theorem 5 gives conditions for when \( \alpha \leq 1 \) is guaranteed for vanilla federation: specifically, whenever the coalition is individually rational. A coalition \( C \) is individually rational if each player in the coalition prefers being in \( C \) to doing local learning. Individual rationality is always a key goal of federated learning, indicating that many model-sharing scenarios might already achieve it [6].

**Theorem 5.** Any individually rational coalition using vanilla federation always has \( \alpha \leq 1 \): the smaller player experiences lower error than proportionality would suggest.

As mentioned previously, \( \alpha \leq 1 \) means that smaller players (contributing fewer samples to the federating model) experience lower error than proportionality would suggest. This may or may not be desirable, depending on the motivation for model sharing (as described in Section 2). The full proof of Theorem 5 is given in the supplementary material, but we include a proof sketch below.

**Proof Sketch.** We will show this result by proving the contrapositive: if \( \alpha > 1 \) is violated, then at least one player wishes to leave that coalition for local learning (so individual rationality is violated).

In our analysis, we will consider two players \( n_1 < n_2 \), though the players could be arbitrarily situated with respect to \( \frac{n_1}{n_2} \). We will assume that all of the players are federating together in coalition \( A \). For notational convenience, we will write the federating coalition \( A \) of interest as \( A = C \cup \{n_1\} \), so \( C \) refers to all players except \( n_1 \).

The case where \( \alpha > 1 \) is:

\[
err_L(A) \cdot n_2 \geq err_L(A) \cdot n_1
\]

The full proof, in the supplementary material, shows that this inequality can be rewritten as:

\[
-2\sigma^2 \cdot N_C + \frac{\mu_C}{\mu} \cdot N_C + \frac{\sigma^2}{\sigma} \cdot \left( \sum_{i \in C} n_i^2 + N_C^2 \right) \leq n_1
\]

Equation (6) gives a lower bound on how many samples the large player has to have before \( \alpha \leq 1 \) is violated. Next, the proof obtains a similar lower bound for how many samples the large player would need before it would prefer to defect: we will show this lower bound is lower than Equation (6) above.

The case when the large player wishes to defect from \( A \) to local learning (or is ambivalent about defecting) is when:

\[
err_L(n_1) \leq err_L(A)
\]

The full version of the proof shows that this can be simplified to:

\[
-2\sigma^2 \cdot N_C + \frac{\mu_C}{\mu} \cdot N_C + \frac{\sigma^2}{\sigma} \cdot \left( \sum_{i \in C} n_i^2 + N_C^2 \right) \leq n_1
\]

The full proof concludes by showing that the bound in Equation (7) is lower than the bound in Equation (6): any situation where the \( n_1 \) term is large enough so that \( \alpha > 1 \), we know that the large player already wishes to leave. 

Overall, the bound in Theorem 5 is different from the results on egalitarian fairness in Section 4. For egalitarian fairness, we obtained a single bound that held for a wide range of model aggregation methods. In this section, by contrast, we show that the bounds we obtain are specific to the model aggregation method. In Theorem 4, we show that for fine-grained federation, \( \alpha \leq 1 \) is always satisfied. However, for vanilla federation, we obtain a guarantee that holds whenever a coalition is individually rational: whenever all collaborating players obtain lower error than they would in local learning. The result in Theorem 5 could be useful if participating is truly voluntary: then, any coalition where smaller players experience error greater than proportional could never exist, because it would unravel into local learning. However, this result could still be useful even if joining a coalition isn’t something individual players could opt out of. Theorem 5 shows that any coalition with \( \alpha > 1 \) is one that also has the problem that some players receive higher error than they would in local learning - which indicates a broader failure of model-sharing, which may prompt any organizer to change their model aggregation methods or coalition structure.

### 6 DISCUSSION

In this work, we have used a theoretical model to analyze the fairness properties of model-sharing games. We formalized two notions of fairness already under analysis in data cooperative and federated learning research and gave conditions for when those fairness properties are or are not satisfied. For egalitarian fairness, we gave a tight multiplicative bound on the gap in error rates between any two federating players. For proportional fairness, we showed that sub-proportional fairness \((\alpha \leq 1)\) is satisfied always for fine-grained federation, and for any individually rational coalition for vanilla federation.

Future research could consider questions in two directions. First: which other notions of fairness might be relevant? It might be the case that an additive version of egalitarian fairness would be more appropriate for certain situations, for example, or a version of proportional fairness based on some metric related to the quality of data points, rather than how their quantity. Secondly, which fairness properties might other methods of model aggregation display? While we have analyzed two federation methods (vanilla and fine-grained), it could be fruitful to analyze other methods to determine how well they achieve certain fairness results.

### ACKNOWLEDGMENTS

This work was supported in part by a Simons Investigator Award, a Vannevar Bush Faculty Fellowship, MURI grant W911NF-19-0217, AFOSR grant FA9550-19-1-0183, ARO grant W911NF19-1-0057, a Simons Collaboration grant, a grant from the MacArthur Foundation, and NSF grant DGE-1650441. We are grateful to the AI, Policy, and Practice working group at Cornell for invaluable discussions.
