Optimality and Stability in Federated Learning: A Game-theoretic Approach

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

Jon Kleinberg
Departments of Computer Science
and Information Science
Cornell University
kleinber@cs.cornell.edu

Abstract

Federated learning is a distributed learning paradigm where multiple agents, each only with access to local data, jointly learn a global model. There has recently been an explosion of research aiming not only to improve the accuracy rates of federated learning, but also provide certain guarantees around social good properties such as total error. One branch of this research has taken a game-theoretic approach, and in particular, prior work has viewed federated learning as a hedonic game, where error-minimizing players arrange themselves into federating coalitions. This past work proves the existence of stable coalition partitions, but leaves open a wide range of questions, including how far from optimal these stable solutions are. In this work, we motivate and define a notion of optimality given by the average error rates among federating agents (players). First, we provide and prove the correctness of an efficient algorithm to calculate an optimal (error minimizing) arrangement of players. Next, we analyze the relationship between the stability and optimality of an arrangement. First, we show that for some regions of parameter space, all stable arrangements are optimal (Price of Anarchy equal to 1). However, we show this is not true for all settings: there exist examples of stable arrangements with higher cost than optimal (Price of Anarchy greater than 1). Finally, we give the first constant-factor bound on the performance gap between stability and optimality, proving that the total error of the worst stable solution can be no higher than 9 times the total error of an optimal solution (Price of Anarchy bound of 9).

1 Introduction

Recent advances of machine learning techniques has made it possible to apply powerful prediction algorithms to a variety of domains. However, in real-world situations, data is often distributed across multiple locations and cannot be combined to a central repository for training. For example, consider patient medical data located at hospitals or student educational data at different schools. In each case, the individual agents (hospitals or schools) who hold the data wish to find a model that minimizes their error. However, the data at each location may be insufficient to train a robust model. Instead, the agents may prefer to build a model using data from multiple agents: multiple hospitals or schools. Collectively, the combined data may be able to produce a model with much higher accuracy, providing more powerful predictions to each agent and increasing overall welfare. However, it may be infeasible to transfer the data to some coordinating entity to build a global model: privacy, data size, and data format are all possible reasons that would make transferring data not a reasonable solution.

Federated learning is a novel distributed learning paradigm that aims to solve this problem (**?**). Data remains at separate local sites, which individual agents use to learn local model parameters or parameter updates. Then, only the parameters are transferred to the coordinating entity (for example,

Preprint. Under review.

a technology company), which averages together all of the parameters in order to form a single global model, which all of the agents use. Federated learning is a rapidly growing area of research (**???**).

However, research has also noted that federated learning, in its traditional form, may not be the best option for each agent (**????**). In the real world, agents may differ in their true distribution: the true model of patient outcomes at hospital A may differ from the true model at hospital B, for example. If these differences are large enough, federating agents may see their error increase under certain situations, potentially even beyond what they would have obtained with only local learning. For example, a player with relatively few samples may end up seeing its model "torqued" by the presence of a player with many samples. For this reason, agents may not wish to federate with every other potential agent.

Instead, each agent faces a choice: given the costs and benefits of federating with different players, it must determine which of the exponentially many combinations of players it would prefer to federate with. Simultaneously, every other agent is also attempting to identify and join a federating group that it prefers - and agents may have conflicting preferences. Prior work (**??**) has formulated this problem as a *hedonic game*, which each player derives some cost (error) from the coalition they join. The aim of such research has been to identify partitions of players that are *stable* against deviations, for varying definitions of stability. A hedonic game in general may not have any stable arrangements, so the area's contributions in the analysis of stability adds valuable insight into the incentives of federating agents.

However, this framework also leaves open multiple game theoretic questions. While the federating agents have individual incentives to reduce their error, society as a whole also has an interest in minimizing the overall error. In the school example, individual schools wish to find coalitions that work well on their own sub-populations, while the overall district or state may have an interest in finding an overall set of coalitions that minimizes the overall error. This analysis of a coalition partition's overall cost falls under the game theoretic notion of *optimality*.

One natural question relates to the tension between these two goals: the self-interested goal of the individual actors (stability) and the overall goal of reducing total cost (optimality). Given a that set of self-interested agents has found a stable solution, how far from optimal could it be? This is reflected by the *Price of Anarchy* of a game, the canonical approach to study optimality and stability jointly (**??**). The Price of Anarchy (PoA) is a ratio where the numerator is equal to the highest-cost stable arrangement and the denominator is equal to the lowest-cost arrangement (the optimal arrangement). It is lower bounded by 1, a bound that it achieves only if all stable arrangements are optimal. A higher Price of Anarchy value implies a greater trade off between stability and optimality, and bounding the Price of Anarchy for a particular game puts a limit on this trade-off. Federated learning is a situation where questions of stability have been analyzed, but to our knowledge there has been no systematic analysis of the Price of Anarchy in a model of federated learning.

The present work: A framework for optimality and stability in federated learning In this work, we make two main contributions to address this gap. First, we provide an efficient, constructive algorithm for calculating an optimal federating arrangement. Secondly, we prove the first-ever constant bound on the Price of Anarchy for this game, showing that the worst stable arrangement is no more than 9 times the cost of the best arrangement.

We begin Section 4 by defining optimality, drawing on a notion of weighted error derived from the standard objective in federated learning literature. The main contribution of this section is an efficient, constructive algorithm for calculating an optimal arrangement, along with a proof of its optimality. However, as demonstrated in Section 5, optimality and stability are not always simultaneously achieved. This section analyzes the Price of Anarchy, which measures how far from optimal the worst stable arrangement can be. First, we demonstrate that the optimal arrangement is not always stable. Next, we show that there exist sub-regions where the Price of Anarchy is equal to 1. Finally, this section proves an overall Price of Anarchy bound of 9, the first constant bound for this game.

It is worth emphasizing that, beyond the Price of Anarchy bound itself, part of the contribution of this work is the optimization and analysis to produce this bound. The proofs for this contribution are modular and illuminate multiple properties about the broader federated learning game under study. As such, these contributions could be useful for further investigating this model. For example, the modular structure of our proof is what enables us to establish stronger bounds for certain sub-cases.

2 Related work

Federated learning As we mentioned previously, federated learning has recently seen numerous advances. In this section, we highlight a few papers in federated learning that are especially related to our work.

The idea that agents might differ in their true models (that data might be generated non-i.i.d. across multiple agents) is commonly acknowledged in the federated learning literature. For example, **??** empirically demonstrate that federated learning, and especially privacy-related additions, can cause a wide disparity in error rates. Some techniques have been developed specifically towards this problem. For example, hierarchical federated learning adds an additional layer of hierarchical structure to federated learning, which could be used to reduce latency or to cluster together similar players (**??**). Many other works also relate to clustering, such as (**???????**). These works, which tend to be more applied than our work, may also differ in that they analyze situations where additional information is known, such as the data distribution at each location.

Other work aims to improve accuracy rates by selecting acquiring additional data (**??**). Some papers specifically analyze federated learning for high-stakes situations such as medical settings (**?????**). In general, all of these works have the goal of reducing the average error over all federating agents, which we will use to motivate our definition of optimality in later sections.

Game theory in federated learning The closest work to this current paper is **?**, which we discuss in greater detail in Section 3. Another paper using hedonic game theory to analyze federated learning games is **?**, which gives conditions for Nash stability in federated learning. Other works analyze the incentives of players to contribute resources towards federated learning: **?** analyzes fairness and efficiency in sampling additional points for federated learning and **?** analyzes incentives for agents to contribute computational resources in federated learning. Interestingly, multiple works take a game theoretic approach towards coalition formation in cloud computing, but with the aim of minimizing some cost besides error, such as electricity usage **??**.

3 Model and assumptions

We assume that there are M total agents (sometimes referred to as players). Each agent $i \in [M]$ has drawn n_i data points from their true local distribution $g(\theta_i)$, where θ_i are their true local parameters and $g(\cdot)$ is some true labeling function. A player's goal is to learn a model with low expected error on its own distribution. If a player opts for local learning, then it uses its local estimate of these parameters $g(\hat{\theta}_i)$ to predict future data points, obtaining error $err_i(\{i\})$. If a set of players C are federating together, we say that they are in a *coalition* or *cluster* together. They combine their local estimates of parameters into a single federated estimate, governed by the weighted average of their parameters:

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

A federating player i obtains error $err_i(C)$: note that this value may differ between players in the same coalition. For example, if player j has samples than player k, then $\hat{\theta}_C$ will be weighted more towards player j, meaning that player j will have lower expected error than k.

The weighted average method in Equation 1 is commonly used in federated learning (**?**). Because it is the most straightforward method, it is sometimes called "vanilla" federated learning. Alternative ways of federation might involve customizing the model for individuals, as in domain adaptation. For example, **?** models three methods of federation: vanilla (called "uniform"), as well as two models of domain adaptation.

There are multiple reasons why we opted to analyze the federation method in Equation 1 in this work. First of all, this federation method is the most straightforward method, and as such it is the natural candidate to begin analysis. Secondly, this federation method is the most interesting to analyze technically. Domain adaptation serves to increase the incentives of an individual agent to participate in federation: it reduces the tension between an individual's incentives and the optimal overall arrangement. Because of this, for Price of Anarchy it is more valuable and challenging to explore the case in Equation 1, where incentives are more opposed.

3.1 Theoretical model of federation from ?

Federated learning has been the subject of both applied and theoretical analysis; our focus here is on the theoretical side. In addition, for game theoretical reasoning to be feasible, we need a model that gives exact errors (costs) for each player, rather than bounds: these are needed in order to be able to argue that a certain arrangement is optimal, for example.

We opt to use the model developed in our prior work **?**, which produces the closed-form error value seen in Lemma 1 below. While we work within this model, we emphasize that **?** 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.

Lemma 1 (Lemma 4.2, from **?**)**.** *Consider a mean estimation task as follows: player j is trying to learn its true mean θ_j. It has access to n_j samples drawn i.i.d. $Y \sim \mathcal{D}_j(\theta_j, \epsilon_j^2)$, a distribution with mean θ_j and variance ϵ_j^2. Given a population of players, each has drawn parameters $(\theta_j, \epsilon_j^2) \sim \Theta$ from some common distribution Θ. A coalition C federating together 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 is:*

$$err_j(C) = \frac{\mu_e}{\sum_{i \in C} n_i} + \sigma^2 \cdot \frac{\sum_{i \in C, i \neq j} n_i^2 + \left(\sum_{i \in C, i \neq j} n_i\right)^2}{\left(\sum_{i \in C} n_i\right)^2} \tag{2}$$

where $\mu_e = \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta}[\epsilon_i^2]$ (the average noise in data sampling) and $\sigma^2 = Var(\theta_i)$ (the average distance between the true means of players).

Note that **?** also analyzes a linear regression game with a similar cost function, though in this work we will restrict our attention to the mean estimation game.

We use some of the same notion and modeling assumptions as **?**. For example, we use C to refer to a coalition of federating agents and Π 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: $N_C = \sum_{i \in C} n_i$. In a few lemmas we will re-use minor results proven in **?**, citing them for completeness.

For technical assumptions, we assume number of samples $\{n_i\}$ is fixed and known by all. We also assume that the parameters μ_e, σ^2 are approximately known: in particular, results will depend on whether the number of samples is larger or smaller than the critical threshold $\frac{\mu_e}{\sigma^2}$. We assume that a player does not know anything else about its own true parameters θ_i or the parameters of other players: for example, it does not know the true generating distribution Θ or if its true parameters are likely to lie far from the parameters of other players. We assume that each player has a goal of obtaining a model with low expected test error on its personal distribution, and that the federating coordinator is motivated to minimize some notion of total cost, but is otherwise impartial.

Finally, it is worth emphasizing key differences between this current work and **?**. The focus of **?** is defining a theoretical model of federated learning and analyzing the stability of such an arrangement. As such, it focuses solely on individual incentives, rather than overall societal welfare. On this other hand, this current work focuses on discussions of optimality (overall welfare) and Price of Anarchy. Finally, this paper work is in some ways more general: while some results in **?** only allow players to have two different numbers of samples ("small" or "large"), every result in our work holds for arbitrarily many players with arbitrarily many different numbers of samples. This distinction is a function of the questions analyzed in each paper: questions of stability (as in **?**) are much harder to analyze for players with arbitrarily many different sizes.

4 Optimality

We will begin with the question of optimality. As motivation, it is useful to consider the objective function of most federated learning papers **?**:

$$\min_{\theta} err_w(\theta) = \sum_{i=1}^{M} p_i \cdot err_i(\theta) =^* \frac{1}{\sum_{i=1}^{M} n_i} \sum_{i=1}^{M} n_i \cdot err_i(\theta)$$

While the weights can be any $p_i > 0, \sum_{i=1}^{M} p_i = 1$, the $*$ equality reflects the common setting where they are taken to be the empirical average. In this work, we will take the empirical average as our cost function:

Definition 1. *A coalition partition* Π *is optimal if it minimizes the weighted sum of errors across players, as defined below:*

$$f_w(\Pi) = \sum_{C \in \Pi} f_w(C) = \sum_{C \in \Pi} \sum_{i \in C} n_i \cdot err_i(C)$$

We will say that a coalition partition Π *is in* OPT *if it achieves minimal cost. Note that multiple partitions may achieve minimal cost, so* OPT *is a set of partitions.*

Because Π is a disjoint partition over the M players, $f_w(\Pi)$ is simply the error $err_w(\theta)$ scaled by a constant. Therefore, minimizing $f_w(\Pi)$ is equivalent to minimizing the weighted average of errors.

Some machine learning papers modify the empirical average objective to achieve other goals. For example, **??** consider variants where this goal is re-weighted in order to achieve certain fairness goals. Appendix A discusses other possible cost functions.

All of the above analysis holds for any model of federated learning. Lemma 2, below, gives the specific form of cost for federated learning using the model from **?**. The remaining analysis in this paper will assume this cost function. Proofs for results in this section are given in Appendix B.

Lemma 2. *Consider a partition* Π *made up of coalitions* $\{C_i\}$*. Then, using the error form given in Equation 2, the total cost of* Π *is given by*

$$f_w(\Pi) = \sum_{C \in \Pi} \left\{ \mu_e + \sigma^2 \cdot N_C - \sigma^2 \frac{\sum_{i \in C} n_i^2}{N_C} \right\}$$

The two most common arrangements in machine learning tasks are local learning (which we will denote by π_l) and the federation in the *grand coalition* (π_g), where all of the players are federating together in a single coalition. However, Lemmas 3 and 4 demonstrate that either of these could could perform arbitrarily poorly as compared the cost-minimizing (optimal) arrangement.

Lemma 3. $\forall \rho > 1$*, there exists a setting where local learning results in average error more than* ρ *times higher than optimal:* $\frac{f_w(\pi_l)}{f_w(OPT)} > \rho$*.*

Lemma 4. $\forall \rho > 1$*, there exists a setting where federating in the grand coalition results in average error more than* ρ *times higher than optimal:* $\frac{f_w(\pi_g)}{f_w(OPT)} > \rho$*.*

A priori, finding a partition of players that minimizes total cost seems extremely challenging. There are exponentially many options for partitions, and two lemmas above have shown that either of the most common choices could be arbitrarily far from optimal. However, next section will provide an efficient, constructive algorithm to calculate an optimal partition of players into federating coalitions.

4.1 Calculating an optimal arrangement

The main contribution of this section is Theorem 1 gives an algorithm for minimizing the total weighted error of the federating agents.

Theorem 1. *Consider a set of players* $\{n_i\}$*. An optimal partition* Π *can be created as follows: first, start with every player doing local learning. Then, begin by grouping the players together in ascending order of size, stopping when the first player would increase its error by joining the coalition from local learning. Then, the resulting partition* Π *is optimal.*

Though the algorithm in Theorem 1 is straightforward, proving the optimality of the resulting partition Π requires several sub-lemmas. Each sub-lemma is a building-block that describes certain operations that either increase or decrease total cost. The proof of Theorem 1 largely consists of sequentially using these sub-lemmas in order to demonstrate the optimality of the calculated partition.

Statement and description of supporting lemmas First, Lemma 5 demonstrates a close relationship between movements of players that reduce total cost and movements of players that are in that

player's self-interest (recall that players always wish to minimize their expected error). Specifically, it shows that a player wishes to join a coalition from local learning if and only if that move would reduce total cost for the entire partition.

Lemma 5 (Equivalence of player preference and reducing cost)**.** *Take any coalition Q and any player j. Then, a player wishes to join that coalition (from local learning) if and only if doing so would reduce total cost. That is,*

$$f_w(\{n_j\}) + f_w(Q) \geq f_w(\{n_j\} \cup Q) \quad \Leftrightarrow \quad err_j(\{n_j\}) \geq err_j(\{n_j\} \cup Q)$$

Next, Lemma 6 shows that "swapping" the roles of two players (one doing local learning, one federating in a coalition) reduces total cost when the larger player is removed to local learning.

Lemma 6 (Swapping)**.** *Take any set Q including a player $n_j > n_k$, where the player n_k is doing local learning. Then, swapping the roles of players k and j always decreases total cost.*

$$f_w(Q \cup \{n_j\}) + f_w(\{n_k\}) > f_w(Q \cup \{n_k\}) + f_w(\{n_j\})$$

Lemmas 7 and 8 give results for when players are incentivized to leave or join a particular coalition: they show that such incentives are monotonic in the size of the player. By Lemma 5, these results also show the monotonicity of cost-reducing operations. Note that these lemmas are not equivalent: they differ in whether the reference player j is already in the coalition or not.

Lemma 7 (Monotonicity of joining)**.** *If a player of size n_j would prefer local learning to joining a coalition Q, then any player of size $n_k \geq n_j$ also prefers local learning to joining the same coalition. That is, for $n_k \geq n_j$,*

$$err_j(Q \cup \{n_j\}) \geq err_j(\{n_j\}) \quad \Rightarrow \quad err_k(Q \cup \{n_k\}) \geq err_k(\{n_k\})$$

Conversely, if a player j wishes to join Q, then any other player of size $n_k \leq n_j$ would have also wanted to join. That is, for $n_j \geq n_k$,

$$err_j(Q \cup \{n_j\}) \leq err_j(\{n_j\}) \quad \Rightarrow \quad err_k(Q \cup \{n_k\}) \leq err_k(\{n_k\})$$

Lemma 8 (Monotonicity of leaving)**.** *Take any coalition Q. Then, if any player $j \in Q$ of size n_j wishes to leave Q for local learning, then any player of size $n_k \geq n_j$ also wishes to leave for local learning. That is, for $n_k \geq n_j$*

$$err_j(Q) \geq err_j(\{n_j\}) \quad \Rightarrow \quad err_k(Q) \geq err_k(\{n_k\})$$

Conversely, if a player $j \in Q$ of size n_j does not *wish to leave Q for local learning, then any player $k \in Q$ of size $n_k \leq n_j$ also does not wish to leave. That is, for $n_k \leq n_j$*

$$err_j(Q) \leq err_j(\{n_j\}) \quad \Rightarrow \quad err_k(Q) \leq err_k(\{n_k\})$$

All of the above lemmas have analyzed situations where a single player is moving between coalitions. Lemma 8 analyzes cases where multiple players are rearranged simultaneously. Specifically, it provides an algorithm for combining together two separate groups (and then removing certain players) that is guaranteed to keep constant or reduce total cost.

Lemma 9 (Merging)**.** *Consider two groups of players, P, Q. First, merge together the two groups to form $P \cup Q$. Then, remove players from $P \cup Q$ to local learning, removing them in descending order of size. Stop removing players when the first player would prefer to stay (removing it would increase its error). Then, this overall process maintains or decreases total error. In other words,*

$$f_w(Q) + f_w(P) \geq f_w(\{Q \cup P\} \setminus L) + \sum_{i \in L} f_w(\{n_i\}) \tag{3}$$

where L is the set of large players removed in descending order of size. The inequality is strict so long as the final structure is not identical to the first, up to renaming of players, and it is not *the case that all the players have the exact same size.*

The proof of Theorem 1 is given simply by applying the lemmas sequentially to show that any other partition Π' can be converted to the described partition Π through a series of operations that decrease or hold constant total cost.

Coalition structure	$err_a(\cdot), n_a = 1$	$err_b(\cdot), n_b = 8$	$err_c(\cdot), n_c = 15$	$f_w(\Pi)$	$err_w(\Pi)$
$\{a,\}, \{b\}, \{c\}$	10	1.25	0.667	30	1.25
$\{a\}, \{b, c\}$	10	1.285	0.677	30.435	1.268
$\{a, c\}, \{b\}$	2.382	1.25	0.633	21.875	0.911
$\{a, b\}, \{c\}$	2.691	1.136	0.667	21.778	0.907
$\{a, b, c\}$	1.834	1.253	0.670	21.917	0.913

Table 1: Example with $\mu_e = 10, \sigma^2 = 1$ example with three players of size $n_a = 1, n_b = 8, n_c = 15$. Note that $\{a, b\}, \{c\}$ minimizes total cost, but is not individually stable: player a wishes to leave its coalition to join player c, which welcomes that player joining it. This produces $\{a, c\}, \{b\}$, which is the only individually stable arrangement, giving a Price of Anarchy value of $21.875/21.778 = 1.0045$.

5 Price of Anarchy

The previous section defined the "optimality" of a federating arrangement as its average error, and additionally provided an efficient algorithm to calculate a lowest-cost arrangement. Given that much of prior work (**??**) has studied the stability of cooperative games induced by federated learning, the next natural question is to study the relationship between stability and optimality. This section analyzes this relationship, using the canonical game theoretic tools of Price of Anarchy and Price of Stability. All proofs for this section are in Appendix C.

First, we will define the notions of stability under analysis, which are all drawn from standard cooperative game theory literature (**?**). A partition of players Π is *core stable* if there does not exist a set of players that all would prefer leave their location in Π and form a coalition together. A partition is *individually stable* (IS) if there does not exist a single player i that wishes to join some existing coalition C, where all members of C weakly prefer that i join. Our results will primarily use the notion of individual stability.

As a reminder, the Price of Anarchy (PoA) is the ratio between the worst (highest-cost) stable arrangement and the best (lowest-cost) arrangement. The Price of Stability is the ratio of the best stable arrangement and the best overall arrangement (regardless of if it is stable or not) (**?**). Note that the Price of Stability is 1 when there exists an optimal arrangement that is also stable.

First, we will show that for certain ranges of parameter space, the Price of Anarchy and/or Price of Stability are equal to 1. Specifically, Lemma 10 shows that when all players have relatively few samples (no more than $\frac{\mu_e}{\sigma^2}$ each), the grand coalition π_g is core stable, implying a Price of (Core) Stability of 1. Recall that μ_e and σ^2 are parameters of the federated learning model reflecting the average noise of the data and the average dissimilarity between federating agents, respectively.

Lemma 10. *For a set of players with $n_i \leq \frac{\mu_e}{\sigma^2}$ $\forall i$, the grand coalition π_g is always core stable.*

On the other hand, Lemma 11 shows that when all players have relatively many samples (at least $\frac{\mu_e}{\sigma^2}$ each), every core or individually stable arrangement is also optimal, which means that the Price of Anarchy for this situation is 1.

Lemma 11. *For a set of players with $n_i \geq \frac{\mu_e}{\sigma^2}$ $\forall i$, any arrangement that is core stable or individually stable is also optimal.*

However, it is *not* the case that either the Price of Stability or Price of Anarchy is always 1. Table 1 contains an example demonstrating this: there exists a simple three-player case where the optimal arrangement is not individually stable. However, the Price of Anarchy value here is quite small, which suggests the prospect that the Price of Anarchy in general could be bounded.

The main result of this section is Theorem 2, which proves a Price of Anarchy bound of 9 for this problem: the cost of the highest stable arrangement is no more than 9 times the cost of the optimal (lowest cost) arrangement.

Theorem 2 (Price of Anarchy)**.** *Denote Π_M to be a maximum-cost individually stable (IS) partition and Π_{opt} to be an optimal (lowest-cost) partition. Then,*

$$PoA = \frac{f_w(\Pi_M)}{f_w(\Pi_{opt})} \leq 9$$

In Theorem 2, the numerator is the cost of Π_M, a maximum-cost partition, and the denominator is Π_{opt}, an optimal (lowest-cost) partition. Recall that Definition 1 gives the cost of an arrangement as the weighted sum of the errors of the respective players. Therefore, to get an upper bound on the Price of Anarchy, we will upper bound the errors players experience in Π_M and lower bound on the error players experience in Π_{opt}.

Summary of proof technique Again, this section will show how the larger theorem is the result of several lemmas that act as building blocks. In particular, the lemmas will take two separate approaches towards creating the bound. Lemmas of the first type (12, 13, 14) all provide upper or lower bounds on the errors certain players can experience. These conditions depend on the size of the player (how many samples it has) and the size of the group it is federating with (how many samples in total the rest of the coalition has). For example, Lemmas 12 and 13 taken together show that a player with at least $\frac{\mu_e+\sigma^2}{2\sigma^2}$ samples has a worst-case error no more than 2 times its best-case error. The same pair of lemmas give a multiplicative bound of 9 for players with numbers of samples that falls between $\frac{\mu_e}{9\cdot\sigma^2}$ and $\frac{\mu_e+\sigma^2}{2\sigma^2}$. Finally, Lemmas 14 and 13 together give a factor of 7.5 for players with fewer than $\frac{\mu_e}{9\cdot\sigma^2}$ samples that are federating with other players of total size at least $\frac{\mu_e}{3\cdot\sigma^2}$. Taken together, these errors show that, for almost all cases, the highest error a player experiences is no more than 9 times higher than the lowest error it might experience.

The final case that needs to be addressed is when a player of size $\leq \frac{\mu_e}{9\cdot\sigma^2}$ is federating in a group with other players of total size $\leq \frac{\mu_e}{3\cdot\sigma^2}$. Lemma 15 handles this last case by an argument around stability. Specifically, it shows that any players in such an arrangement can only be stable if all of them are grouped together into a single federating coalition. In the proof of Theorem 2, this result ends up enabling an additive factor to the Price of Anarchy bound, which is absorbed into the other factors for a total Price of Anarchy value of 9.

Statement and description of supporting lemmas Next, we will walk through each lemma specifically. Lemma 12 gives an *upper* bound of $\frac{\mu_e}{n_i}$ on the error any player experiences in Π_M.

Lemma 12. *If Π_M is a maximum-cost IS partition, then $err_i(\Pi_M) \leq \frac{\mu_e}{n_i}$ for all players i.*

Proof. Because Π_M is individually stable, every player must get error no more than the error it would receive alone (doing local learning). By Lemma 1 with $C = n_i$, a player with samples n_i player gets error $\frac{\mu_e}{n_i}$ alone. □

Next, Lemma 13 provides *lower* bounds on the error a player can receive in Π_{opt}. It does this by bounding the minimum error a player could get in any arrangement. Again, because the cost of Π_{opt} is simply the weighted sum of errors of each individual player, this helps to upper bound the Price of Anarchy. First, Lemma 13 shows that for players with at least $\frac{\mu_e+\sigma^2}{2\sigma^2}$ samples, the lowest possible error it could experience is $\frac{1}{2}\cdot\frac{\mu_e}{n_j}$, which is a factor of 2 off from its worst-case error in Lemma 12. For players with fewer samples than $\frac{\mu_e+\sigma^2}{2\sigma^2}$, Lemma 13 says that the lowest error a player could experience is σ^2. This means that the ratio between the two errors is lower than 9 so long as $n_j \geq \frac{\mu_e}{9\cdot\sigma^2}$. Therefore, in order to get a factor of 9 bound for the overall Price of Anarchy, we need to handle the case of players with size $\leq \frac{\mu_e}{9\cdot\sigma^2}$, when players have very few samples.

Lemma 13. *Consider a player n_j and any set of players C. Then, we can lower bound the error player j recieves by federating with C:*

$$err_j(C \cup \{n_j\}) \geq \begin{cases} \frac{1}{2}\cdot\frac{\mu_e}{n_j} & n_j \geq \frac{\mu_e+\sigma^2}{2\sigma^2} \\ \sigma^2 & \textit{otherwise} \end{cases}$$

Lemma 14 is the first of two lemmas handling the case of players with very few samples. It shows that, if a player of size $\leq \frac{\mu_e}{3\cdot\sigma^2}$ is federating with a set of players of total size at least $\frac{\mu_e}{3\cdot\sigma^2}$, it is possible to *upper* bound on the error of players in Π_M by $7.5\cdot\sigma^2$. Given the lower bound of σ^2 in Lemma 13, these together show that there is a ratio of 7.5 at most between the error this player experiences in its best and worst-case arrangements.

Lemma 14. *Consider a player j federating with a coalition C. If the total number of samples N_C is at least $\frac{\mu_e}{3\sigma^2}$, then $err_j(C \cup \{n_j\}) \leq 7.25\cdot\sigma^2$.*

However, Lemma 14 does not handle one situation: what if a player of size $\leq \frac{\mu_e}{9 \cdot \sigma^2}$ is federating with a group of players of total size $\leq \frac{\mu_e}{3 \cdot \sigma^2}$? Lemma 15 addresses this last case: it shows that the only such arrangement that is stable is one where all such players are grouped together into a single arrangement. Note that this lemma is itself should not be obvious: it is composed of multiple sub-lemmas which are stated and proved in the appendix. The fact that there can be only one group of such players is used in the Theorem 2 to create an overall bound of 9.

Lemma 15. *Consider an arrangement of players, all of size $\leq \frac{\mu_e}{3\sigma^2}$, where at least one player is in a federating cluster where the total mass of its partners is no more than $\frac{\mu_e}{3\sigma^2}$. Then, the only stable arrangement of these players is to have all of them federating together.*

The full proof of Theorem 2 uses these lemmas collectively in order to get an overall Price of Anarchy bound of 9, showing that the worst individually stable arrangement has total cost no more than 9 times the optimal cost.

6 Conclusion

In this work, we have given the first Price of Anarchy bound for a game-theoretic model of federated learning. This bound quantifies a key tension between individual incentives and overall societal goals, answering a key question left open in prior literature. Beyond this bound, we also provide an efficient algorithm to calculate an optimal partition of players into federating coalitions, and have characterized conditions where the Price of Anarchy and/or Price of Stability is equal to 1.

There are multiple fascinating extensions to this work. To begin with, other definitions of societal cost (for example, weighting players' errors differently) could produce different Price of Anarchy bounds. Additionally, further work could model more sophisticated methods of federation, including models of domain adaptation. Finally, it would be interesting to explore other notions of societal interest. For example, one vein of research is fairness: how are error rates divided among federating players? Questions might revolve around the maximum gap in error rates between players and whether players that contribute more samples are always rewarded with lower error. Beyond these avenues, though, we believe that the broad topic of federated learning will continue to contain multiple useful and interesting research directions.

7 Ethics and societal impact

Given this work's focus defining notions of optimality, there are important ethical considerations. In particular, "optimality" can be defined in multiple different ways: Section 4 motivates the definition we use and Appendix A discusses the merits of other definitions. In particular, it is worth emphasizing that "optimality" is a technical term in optimization and game theory which is always with respect to a given objective function and does not imply a more holistic notion of how desirable a certain solution is. For example, an arrangement could be "optimal" and still be unfair in how errors are distributed among players.

Although our methodology is application-agnostic, federated learning is a machine learning tool that could be applied towards positive goals (e.g. predicting patient outcomes at hospitals) or negative goals (e.g. used to surveil and control populations). It is also worth considering, for each application, whether there could be some other approach that would better address the need. For example, it may be worth considering whether approaches aiming at increasing the number of samples available for low-resource agents would do a better job of increasing the benefit of a federated learning solution. It may even be the case that a solution beyond machine learning would be preferable, such as interventions to reduce the need for a predictive model.

References

A Alternate definitions of optimality

The definition of cost used in this paper is given in Definition 1, which says that an arrangement is optimal if it minimizes the weighted sum of errors over players. As discussed previously, this definition is well-motivated by existing federated learning literature. Additionally, it matches the societal good perspective when the unit society cares about is at the level of the data point. For example, consider the example when the federating agents are hospitals and data points represent individual patients. Then, society as a whole likely cares about minimizing the overall error patients experience, which corresponds to the per-data-point notion of error.

However, other cost functions are worth discussing. For example, Definition 2 gives an unweighted notion of error:

Definition 2 (Unweighted cost)**.** *The unweighted cost function is given by summing the error over each of the players, without any weighting with respect to size:*

$$f_u(\Pi) = \sum_{C \in \Pi} f_u(C) = \sum_{C \in \Pi} \sum_{i \in C} err_i(C)$$

This definition might be better in a model where the unit society cares about is at the level of the agent. For example, consider a situation where the individual federating agent is a cell phone owned by a single person and data points are word predictions. Then, society as a whole might care about minimizing the sum of errors that individual cell phone users experience, which is given by the unweighted error function.

Finally, we may wish to consider some completely different weight function, given by the definition below:

Definition 3 (Arbitrary weights)**.** *The arbitrary cost metric is given by summing the weight over each of the players according to some weight* $\sum_{i \in [M]} p_i = 1$

$$f_a(\Pi) = \sum_{C \in \Pi} f_a(C) = \sum_{C \in \Pi} \sum_{i \in C} p_i \cdot err_i(C)$$

Definitions like this have been analyzed in **????**. For example, the set of weights $\{p_i\}$ could have fairness goals, attempting to up-weight players with higher error. Alternatively, it could represent some notion of the data quality players are contributing, with players producing more or lower-error players being weighted more.

In this work, we selected Definition 1 (weighted error) based on its standard use in the federated learning literature. Analysis of the same type (calculating an optimal arrangement and analyzing the Price of Anarchy) could be completed for any other definition of cost, but would require new proofs for calculation of optimal arrangements and for any Price of Anarchy bound.

B Optimality calculation

Lemma 2. *Consider a partition* Π *made up of coalitions* $\{C_i\}$*. Then, using the error form given in Equation 2, the total cost of* Π *is given by*

$$f_w(\Pi) = \sum_{C \in \Pi} \left\{ \mu_e + \sigma^2 \cdot N_C - \sigma^2 \frac{\sum_{i \in C} n_i^2}{N_C} \right\}$$

Proof.

$$f_w(C) = \sum_{j \in C} err_j(C) \cdot n_i = \sum_{j \in C} \left(\frac{\mu_e}{\sum_{i \in C} n_i} + \sigma^2 \frac{\sum_{i \neq j} n_i^2 + \left(\sum_{i \neq j} n_i\right)^2}{\left(\sum_{i \in C} n_i\right)^2} \right) \cdot n_j$$

$$= \sum_{j \in C} \frac{\mu_e}{\sum_{i \in C} n_i} \cdot n_j + \sigma^2 \sum_{j \in C} \frac{\sum_{i \neq j} n_i^2 + \left(\sum_{i \neq j} n_i\right)^2}{\left(\sum_{i \in C} n_i\right)^2} \cdot n_j$$

$$= \mu_e + \sigma^2 \sum_{j \in C} \frac{n_j \cdot \sum_{i \neq j} n_i^2 + n_j \cdot (N_C - n_j)^2}{N_C^2}$$

where we have used $N_C = \sum_{i \in C} n_i$. Focusing solely on the numerator of the second term, we simplify:

$$\sum_{j \in C} \left\{ n_j \cdot \sum_{i \neq j} n_i^2 + n_j \cdot N_C^2 + n_j^3 - 2N_C \cdot n_j^2 \right\} = \sum_{j \in C} n_j \cdot \sum_{i \in C} n_i^2 + N_C^2 \sum_{j \in C} n_j - 2N_C \sum_{j \in C} n_j^2$$

$$= N_C \cdot \sum_{i \in C} n_i^2 + N_C^3 - 2N_C \sum_{i \in C} n_i^2 = N_C^3 - N_C \cdot \sum_{i \in C} n_i^2$$

Combining this with the rest of the term gives:

$$\mu_e + \sigma^2 \cdot \frac{N_C^3 - N_C \cdot \sum_{i \in C} n_i^2}{N_C^2} = \mu_e + \sigma^2 \cdot N_C - \sigma^2 \frac{\sum_{i \in C} n_i^2}{N_C}$$

□

Lemma 3. *$\forall \rho > 1$, there exists a setting where local learning results in average error more than ρ times higher than optimal: $\frac{f_w(\pi_l)}{f_w(OPT)} > \rho$.*

Proof. We will prove this result by the setting where M players each have n samples, for $M > \rho$ and any $\mu_e, \sigma^2, n \in \mathbb{N}_{\geq 1}$ such that $n < \left(\frac{M}{\rho} - 1\right) \frac{\mu_e}{(M-1) \cdot \sigma^2}$.

In this simplified setting where all of the players have the same number of samples, the cost of a coalition C involving M players is given by:

$$\mu_e + \sigma^2 \cdot n \cdot M - \sigma^2 \cdot \frac{M \cdot n^2}{M \cdot n} = \mu_e + \sigma^2 \cdot n \cdot (M - 1)$$

For our given example, $n < \frac{\mu_e}{\sigma^2}$, which implies that "merging" any two groups A and B will reduce total cost:

$$\begin{aligned} f_w(A) + f_w(B) &= \mu_e + \sigma^2 \cdot n \cdot (M_A - 1) + \mu_e + \sigma^2 \cdot n \cdot (M_B - 1) \\ &> \mu_e + \sigma^2 \cdot n \cdot (M_A + M_B - 1) \\ &= f_w(A \cup B) \end{aligned}$$

This implies that the optimal cost is achieved by π_g, given by $\mu_e + \sigma^2 \cdot (M - 1)$. Conversely, the cost of having M players doing local learning is:

$$f_w(\pi_l) = \sum_{i=1}^{M} \left\{ \mu_e + \sigma^2 \cdot n - \sigma^2 \cdot \frac{n^2}{n} \right\} = \sum_{i=1}^{M} \mu_e = \mu_e \cdot M$$

Combining these facts gives:

$$\begin{aligned} \frac{f_w(\pi_l)}{f_w(OPT)} &= \frac{\mu_e \cdot M}{\mu_e + \sigma^2 \cdot (M-1) \cdot n} = \frac{M}{1 + \frac{\sigma^2}{\mu_e} \cdot (M-1) \cdot n} \\ &> \frac{M}{1 + \frac{\sigma^2}{\mu_e} \cdot (M-1) \cdot \frac{\mu_e}{(M-1) \cdot \sigma^2} \cdot \left(\frac{M}{\rho} - 1\right)} \\ &= \frac{M}{1 + \frac{M}{\rho} - 1} = \rho \end{aligned}$$

as desired. □

Lemma 4. *$\forall \rho > 1$, there exists a setting where federating in the grand coalition results in average error more than ρ times higher than optimal: $\frac{f_w(\pi_g)}{f_w(OPT)} > \rho$.*

Proof. We will prove this result by the setting where M players each have n samples, with $M > \frac{1}{\rho}$ and any $\mu_e, \sigma^2, n \in \mathbb{N}_{\geq 1}$ such that $n > \max\left[\frac{\mu_e}{\sigma^2 \cdot (M-1)} \cdot (\rho \cdot M - 1), \frac{\mu_e}{\sigma^2}\right]$.

The initial construction follows similarly to Lemma 3. For our given example, $n > \frac{\mu_e}{\sigma^2}$, which implies that "merging" any two groups A and B will *increase* total cost:

$$\begin{aligned} f_w(A) + f_w(B) &= \mu_e + \sigma^2 \cdot n \cdot (M_A - 1) + \mu_e + \sigma^2 \cdot n \cdot (M_B - 1) \\ &< \mu_e + \sigma^2 \cdot n \cdot (M_A + M_B - 1) \\ &= f_w(A \cup B) \end{aligned}$$

This implies that the optimal cost is achieved π_l. Using the value derived in the proof of Lemma 3, we have:

$$\begin{aligned} \frac{f_w(\pi_g)}{f_w(OPT)} &= \frac{\mu_e + \sigma^2 \cdot (M-1) \cdot n}{\mu_e \cdot M} \\ &= \frac{1 + \frac{\sigma^2}{\mu_e} \cdot (M-1) \cdot n}{M} \\ &> \frac{1 + \frac{\sigma^2}{\mu_e} \cdot (M-1) \cdot \max\left[\frac{\mu_e}{\sigma^2 \cdot (M-1)} \cdot (\rho \cdot M - 1), \frac{\mu_e}{\sigma^2}\right]}{M} \\ &\geq \frac{1 + \frac{\sigma^2}{\mu_e} \cdot (M-1) \cdot \frac{\mu_e}{\sigma^2 \cdot (M-1)} \cdot (\rho \cdot M - 1)}{M} \\ &= \frac{1 + \rho \cdot M - 1}{M} = \rho \end{aligned}$$

as desired. □

The proof of Theorem 1, below, relies on multiple sub-lemmas which are stated and proved immediately afterwards.

Theorem 1. *Consider a set of players $\{n_i\}$. An optimal partition Π can be created as follows: first, start with every player doing local learning. Then, begin by grouping the players together in ascending order of size, stopping when the first player would increase its error by joining the coalition from local learning. Then, the resulting partition Π is optimal.*

Proof. First, we note two special cases. If $\{n_i\} \leq \frac{\mu_e}{\sigma^2}$, then by Lemma 10 (stated and proved later in this appendix) the grand coalition π_g is core stable. For the grand coalition, core stability implies individual stability, so we know that every player prefers π_g to local learning. This implies that, following the steps given in this theorem, every player will prefer to join the growing coalition as opposed to doing local learning, and so the optimal arrangement is π_g.

Next, if $\{n_i\} > \frac{\mu_e}{\sigma^2}$, then by Lemma 5.3 in **?** every player minimizes their error in π_l (local learning). As a result, using the algorithm given in the statement of this theorem, every player will increase their error by combining with another player, so π_l is optimal. If $\{n_i\} \geq \frac{\mu_e}{\sigma^2}$ (some players have exactly $\frac{\mu_e}{\sigma^2}$ samples), then all players with $n_i = \frac{\mu_e}{\sigma^2}$ will be indifferent towards being merged with any other player also of size $\frac{\mu_e}{\sigma^2}$, but no player of size strictly greater than $\frac{\mu_e}{\sigma^2}$ will be able to be merged. The resulting optimal arrangement will have all of the players of size exactly $\frac{\mu_e}{\sigma^2}$ together, with all other players doing local learning, and will have cost identical to π_l.

Finally, we will consider the case where some players have size strictly less than $\frac{\mu_e}{\sigma^2}$ and some have strictly more. Call the partition calculated by following the steps of this theorem Π, and consider any other coalition partition Π'. We will convert Π' into Π using only cost reducing or maintaining steps, which will show that Π is optimal. We will refer to players with size $\leq \frac{\mu_e}{\sigma^2}$ as small, and players of size $> \frac{\mu_e}{\sigma^2}$ as large.

- If there are any coalitions where players would prefer to leave the coalition, remove them in order of descending size. Note: a coalition made up of only players of size smaller than $\frac{\mu_e}{\sigma^2}$ will never have players leave. A coalition made up of only players of size larger than $\frac{\mu_e}{\sigma^2}$ will always wish to have players leave. This reduces total cost by Lemma 5.
- Every coalition of size 2 or larger will have at least one small player in it. Begin merging all such coalitions (as well as any small players doing local learning), removing large players as necessary (in descending size, if they would prefer local learning). Note that the merging operation will never remove a small player, so it always strictly reduces the number of coalitions involving small players. This reduces cost by Lemma 9.
- When all of the small players are in one coalition, if there are large players in the coalition as well, check if they are the smallest possible large player. If not, swap them for smaller large players iteratively (ones that are doing local learning) until the players in the coalition are doing local learning. By Lemma 6, this reduces cost.
- Add large players in increasing order of size (if any wish to join). From Lemma 7 we know that if player n_i doesn't wish to join a coalition, then neither will any player of size $n_j \geq n_i$. From Lemma 5, adding any player that wishes to join reduces total cost.
- If no players wish to join, then remove large players in descending order of size if they would prefer local learning, which again from Lemma 5 would reduce cost. From Lemma 8, if a player of size n_i doesn't wish to leave, then all other players of size $n_j \leq n_i$ also do not wish to leave.

The final arrangement exactly matches Π. □

Lemma 5 (Equivalence of player preference and reducing cost)**.** *Take any coalition Q and any player j. Then, a player wishes to join that coalition (from local learning) if and only if doing so would reduce total cost. That is,*

$$f_w(\{n_j\}) + f_w(Q) \geq f_w(\{n_j\} \cup Q) \quad \Leftrightarrow \quad err_j(\{n_j\}) \geq err_j(\{n_j\} \cup Q)$$

Proof. This proof will work by showing the forms of the inequalities are identical. We will start with the cost inequality:

$$\begin{aligned} f_w(\{n_j\}) + f_w(Q) &\geq f_w(\{n_j\} \cup Q) \\ \mu_e + \mu_e + \sigma^2 \cdot N_Q - \sigma^2 \cdot \frac{\sum_{i \in Q} n_i^2}{N_Q} &\geq \mu_e + \sigma^2 \cdot N_Q + \sigma^2 \cdot n_j - \sigma^2 \frac{\sum_{i \in Q} n_i^2 + n_j^2}{N_Q + n_j} \\ \mu_e &\geq \sigma^2 \cdot n_j - \sigma^2 \frac{\sum_{i \in Q} n_i^2 + n_j^2}{N_Q + n_j} + \sigma^2 \cdot \frac{\sum_{i \in Q} n_i^2}{N_Q} \end{aligned}$$

Bringing all terms over common denominator on the righthand side:

$$\begin{aligned} \mu_e &\geq \sigma^2 \frac{n_j \cdot (N_Q + n_j) \cdot N_Q - N_Q \sum_{i \in Q} n_i^2 - n_j^2 \cdot N_Q + N_Q \sum_{i \in Q} n_i^2 + n_j \cdot \sum_{i \in Q} n_i^2}{(N_Q + n_j) \cdot N_Q} \\ \mu_e &\geq \sigma^2 \frac{n_j \cdot (N_Q + n_j) \cdot N_Q - n_j^2 \cdot N_Q + n_j \cdot \sum_{i \in Q} n_i^2}{(N_Q + n_j) \cdot N_Q} \\ \mu_e &\geq \sigma^2 \frac{n_j \cdot N_Q^2 + n_j^2 \cdot N_Q - n_j^2 \cdot N_Q + n_j \cdot \sum_{i \in Q} n_i^2}{(N_Q + n_j) \cdot N_Q} \\ \mu_e &\geq \sigma^2 \frac{n_j \cdot N_Q^2 + n_j \cdot \sum_{i \in Q} n_i^2}{(N_Q + n_j) \cdot N_Q} \\ \mu_e &\geq \sigma^2 \cdot \frac{n_j}{N_Q + n_j} \cdot \frac{N_Q^2 + \sum_{i \in Q} n_i^2}{N_Q} \end{aligned}$$

Next, we will reduce the error inequality to the same form:

$$\begin{aligned}
err_j(\{n_j\}) &\geq err_j(\{n_j\} \cup Q) \\
\frac{\mu_e}{n_j} &\geq \frac{\mu_e}{N_Q + n_j} + \sigma^2 \cdot \frac{\sum_{i \in Q} n_i^2 + N_Q^2}{(N_Q + n_j)^2} \\
\frac{\mu_e}{n_j} - \frac{\mu_e}{N_Q + n_j} &\geq \sigma^2 \cdot \frac{\sum_{i \in Q} n_i^2 + N_Q^2}{(N_Q + n_j)^2} \\
\mu_e \cdot \frac{N_Q}{n_j \cdot (N_Q + n_j)} &\geq \sigma^2 \cdot \frac{\sum_{i \in Q} n_i^2 + N_Q^2}{(N_Q + n_j)^2} \\
\mu_e &\geq \sigma^2 \cdot \frac{(N_Q + n_j) \cdot n_j}{N_Q} \frac{\sum_{i \in Q} n_i^2 + N_Q^2}{(N_Q + n_j)^2} \\
\mu_e &\geq \sigma^2 \cdot \frac{n_j}{N_Q + n_j} \cdot \frac{N_Q^2 + \sum_{i \in Q} n_i^2}{N_Q}
\end{aligned}$$

as desired. □

Lemma 6 (Swapping)**.** *Take any set Q including a player $n_j > n_k$, where the player n_k is doing local learning. Then, swapping the roles of players k and j always decreases total cost.*

$$f_w(Q \cup \{n_j\}) + f_w(\{n_k\}) > f_w(Q \cup \{n_k\}) + f_w(\{n_j\})$$

Proof. We write out each side:

$$\mu_e + \sigma^2 \cdot N_Q + \sigma^2 \cdot n_j - \sigma^2 \frac{\sum_{i \in Q} n_i^2 + n_j^2}{N_Q + n_j} + \mu_e > \mu_e + \sigma^2 \cdot N_Q + \sigma^2 n_k - \sigma^2 \frac{\sum_{i \in Q} n_i^2 + n_k^2}{N_Q + n_k} + \mu_e$$

$$\sigma^2 \cdot n_j - \sigma^2 \cdot \frac{\sum_{i \in Q} n_i^2 + n_j^2}{N_Q + n_j} > \sigma^2 \cdot n_k - \sigma^2 \cdot \frac{\sum_{i \in Q} n_i^2 + n_k^2}{N_Q + n_k}$$

Dropping the common σ^2 term for clarity:

$$n_j - \frac{\sum_{i \in Q} n_i^2 + n_j^2}{N_Q + n_j} > n_k - \frac{\sum_{i \in Q} n_i^2 + n_k^2}{N_Q + n_k}$$

In order to prove the above inequality, we will consider the following fraction:

$$x - \frac{\sum_{i \in Q} n_i^2 + x^2}{N_Q + x}$$

and want to show this is increasing with respect to x. The derivative of the function gives:

$$\begin{aligned}
&1 - \frac{2x \cdot (N_Q + x) - (\sum_{i \in Q} n_i^2 + x^2) \cdot 1}{(N_Q + x)^2} \\
&= \frac{1}{(N_Q + x)^2} \cdot \left(N_Q^2 + x^2 + 2x \cdot N_Q - \left(2x \cdot (N_Q + x) - \left(\sum_{i \in Q} n_i^2 + x^2 \right) \right) \right) \\
&= \frac{1}{(N_Q + x)^2} \cdot \left(N_Q^2 + x^2 + 2x \cdot N_Q - 2x \cdot N_Q - 2x^2 + \left(\sum_{i \in Q} n_i^2 + x^2 \right) \right) \\
&= \frac{1}{(N_Q + x)^2} \cdot \left(N_Q^2 + \sum_{i \in Q} n_i^2 \right)
\end{aligned}$$

which is positive, as desired. This implies that the original inequality is satisfied, meaning that the swapping of the roles of players j, k decreases total cost. □

Lemma 7 (Monotonicity of joining)**.** *If a player of size n_j would prefer local learning to joining a coalition Q, then any player of size $n_k \geq n_j$ also prefers local learning to joining the same coalition. That is, for $n_k \geq n_j$,*

$$err_j(Q \cup \{n_j\}) \geq err_j(\{n_j\}) \quad \Rightarrow \quad err_k(Q \cup \{n_k\}) \geq err_k(\{n_k\})$$

Conversely, if a player j wishes to join Q, then any other player of size $n_k \leq n_j$ would have also wanted to join. That is, for $n_j \geq n_k$,

$$err_j(Q \cup \{n_j\}) \leq err_j(\{n_j\}) \quad \Rightarrow \quad err_k(Q \cup \{n_k\}) \leq err_k(\{n_k\})$$

Proof. The initial premise depends on whether or not the below inequality is satisfied:

$$err_j(Q \cup \{n_j\}) \geq err_j(\{n_j\})$$

$$\frac{\mu_e}{N_Q + n_j} + \sigma^2 \frac{\sum_{i \in Q} n_i^2 + \left(\sum_{i \in Q} n_i\right)^2}{(N_Q + n_j)^2} \geq \frac{\mu_e}{n_j}$$

Rearranging:

$$\sigma^2 \frac{\sum_{i \in Q} n_i^2 + \left(\sum_{i \in Q} n_i\right)^2}{(N_Q + n_j)^2} \geq \frac{\mu_e}{n_j} - \frac{\mu_e}{N_Q + n_j}$$

$$\sigma^2 \left(\sum_{i \in Q} n_i^2 + \left(\sum_{i \in Q} n_i \right)^2 \right) \geq (N_Q + n_j)^2 \cdot \left(\frac{\mu_e}{n_j} - \frac{\mu_e}{N_Q + n_j} \right)$$

$$\sigma^2 \left(\sum_{i \in Q} n_i^2 + \left(\sum_{i \in Q} n_i \right)^2 \right) \geq (N_Q + n_j)^2 \cdot \frac{\mu_e \cdot N_Q}{n_j \cdot (N_Q + n_j)}$$

$$\sigma^2 \left(\sum_{i \in Q} n_i^2 + \left(\sum_{i \in Q} n_i \right)^2 \right) \geq (N_Q + n_j) \cdot \frac{\mu_e \cdot N_Q}{n_j}$$

$$\sigma^2 \left(\sum_{i \in Q} n_i^2 + \left(\sum_{i \in Q} n_i \right)^2 \right) \geq \mu_e \cdot \frac{N_Q^2}{n_j} + \mu_e \cdot N_Q$$

The lefthand side is a constant independent of n_k and the righthand side is a constant plus a term that is decreasing in n_j. If the original inequality ($err_j(Q \cup \{n_j\}) \geq err_j(\{n_j\})$) is satisfied, then it will also be satisfied for any $n_k \geq n_j$ (implying $err_k(Q \cup \{n_k\}) \geq err_k(\{n_k\})$). Conversely, if the original inequality is not satisfied (so $err_j(Q \cup \{n_j\}) \leq err_j(\{n_j\})$), then it will also not be satisfied for any $n_k \leq n_j$ (implying $err_k(Q \cup \{n_k\}) \leq err_k(\{n_k\})$). □

Lemma 8 (Monotonicity of leaving)**.** *Take any coalition Q. Then, if any player $j \in Q$ of size n_j wishes to leave Q for local learning, then any player of size $n_k \geq n_j$ also wishes to leave for local learning. That is, for $n_k \geq n_j$*

$$err_j(Q) \geq err_j(\{n_j\}) \quad \Rightarrow \quad err_k(Q) \geq err_k(\{n_k\})$$

Conversely, if a player $j \in Q$ of size n_j does not *wish to leave Q for local learning, then any player $k \in Q$ of size $n_k \leq n_j$ also does not wish to leave. That is, for $n_k \leq n_j$*

$$err_j(Q) \leq err_j(\{n_j\}) \quad \Rightarrow \quad err_k(Q) \leq err_k(\{n_k\})$$

Proof. First, we will prove the first statement. Suppose by contradiction that some n_j wishes to leave, but another player of size $n_k \geq n_j$ does not wish to. First, we remove n_j for local learning, which by Lemma 5 reduces total cost. Next, we swap the role of players j and k, which by Lemma 6 again reduces or keeps constant total cost. We have constructed a series of operations that either reduce or

keep constant total cost, and results in an arrangement equivalent to simply removing player j. By Lemma 5, this means that player j originally would have wished to leave.

Next, we will prove the second statement. Suppose by contradiction that some player k wishes to leave, even though another player $n_j > n_k$ does not wish to leave. First, we remove player k to local learning: if it wishes to leave, then by Lemma 5 removing it reduces or keeps constant total cost. Then, by Lemma 6, we can reduce total cost by swapping it with the $n_j > n_k$ player. We have constructed a series of operations that either reduce or keep constant total cost, and results in an arrangement equivalent to simply removing player j. But this is exactly equivalent to just removing the n_j player, which we know from Lemma 5 must not reduce total cost (or else player j would wish to leave). □

Lemma 9 (Merging)**.** *Consider two groups of players, P, Q. First, merge together the two groups to form $P \cup Q$. Then, remove players from $P \cup Q$ to local learning, removing them in descending order of size. Stop removing players when the first player would prefer to stay (removing it would increase its error). Then, this overall process maintains or decreases total error. In other words,*

$$f_w(Q) + f_w(P) \geq f_w(\{Q \cup P\} \setminus L) + \sum_{i \in L} f_w(\{n_i\}) \tag{3}$$

where L is the set of large players removed in descending order of size. The inequality is strict so long as the final structure is not identical to the first, up to renaming of players, and it is not *the case that all the players have the exact same size.*

Proof. First, we have to reason about what L could be. We will say that player j with n_j samples is a largest element in $P \cup Q$, and WLOG $j \in P$. (If multiple players have n_j samples, it suffices to select one at random.). We will show that, in order to show Equation 3, it suffices to show that:

$$f_w(Q) + f_w(P) \geq f_w(Q \cup P \setminus n_j) + f_w(\{n_j\}) \tag{4}$$

First, assume L is empty. Then, every player wishes to stay in the final group. Then, Equation 3 becomes:

$$f_w(Q) + f_w(P) > f_w(Q \cup P)$$

From Lemma 5, we know that because player n_j doesn't wish to leave $Q \cup P$, removing it must increase total cost:

$$f_w(Q \cup P \setminus \{n_j\}) + f_w(\{n_j\}) > f_w(Q \cup P)$$

So, if we show that Equation 4 is satisfied, then this implies that Equation 3 is satisfied.
Next, we will assume $L = \{n_j\}$. Then, the statement we are trying to show is exactly Equation 4.
Finally, let's assume that $|L| \geq 2$: n_j is removed, but so are some others. Again, by Lemma 5, because these players prefer local learning to federation, adding them back in to the coalition increases cost, so

$$f_w(Q) + f_w(P \cup \{n_j\}) > f_w(Q \cup P \setminus n_j) + f_w(\{n_j\})$$

So, it suffices to consider Equation 4: if we prove that this is satisfied, it always implies that Equation 3 is satisfied.

Next, we will prove this statement:

$$f_w(Q) + f_w(P) \geq f_w(Q \cup P \setminus n_j) + f_w(\{n_j\})$$

Plugging in for the form of $f_w(\cdot)$ gives:

$$\mu_e + \sigma^2 \cdot N_Q - \sigma^2 \cdot \frac{\sum_{i \in Q} n_i^2}{N_Q} + \mu_e + \sigma^2 \cdot N_P - \sigma^2 \cdot \frac{\sum_{i \in P} n_i^2}{N_P}$$

$$\geq \mu_e + \sigma^2 \cdot N_Q + \sigma^2 \cdot (N_Q - n_j) - \sigma^2 \frac{\sum_{i \in Q} n_i^2 + \sum_{i \in P} n_i^2 - n_j^2}{N_Q + N_P - n_j} + \mu_e$$

Simplifying gives:

$$\sigma^2 \cdot n_j - \sigma^2 \cdot \frac{\sum_{i \in Q} n_i^2}{N_Q} - \sigma^2 \cdot \frac{\sum_{i \in P} n_i^2}{N_P} \geq -\sigma^2 \frac{\sum_{i \in Q} n_i^2 + \sum_{i \in P} n_i^2 - n_j^2}{N_Q + N_P - n_j}$$

For convenience, we'll drop the common σ^2 coefficient as we continue simplifying:

$$n_j - \frac{\sum_{i\in Q} n_i^2}{N_Q} - \frac{\sum_{i\in P} n_i^2}{N_P} \geq -\frac{\sum_{i\in Q} n_i^2 + \sum_{i\in P} n_i^2 - n_j^2}{N_Q + N_P - n_j}$$

$$n_j \geq \frac{\sum_{i\in Q} n_i^2}{N_Q} + \frac{\sum_{i\in P} n_i^2}{N_P} - \frac{\sum_{i\in Q} n_i^2 + \sum_{i\in P} n_i^2}{N_Q + N_P - n_j} + \frac{n_j^2}{N_Q + N_P - n_j}$$

$$n_j - \frac{n_j^2}{N_Q + N_P - n_j} \geq \frac{\sum_{i\in Q} n_i^2}{N_Q} + \frac{\sum_{i\in P} n_i^2}{N_P} - \frac{\sum_{i\in Q} n_i^2 + \sum_{i\in P} n_i^2}{N_Q + N_P - n_j}$$

$$n_j \cdot \frac{N_Q + N_P - n_j - n_j}{N_Q + N_P - n_j} \geq \left(\sum_{i\in Q} n_i^2\right) \cdot \left(\frac{1}{N_Q} - \frac{1}{N_Q + N_P - n_j}\right) + \left(\sum_{i\in P} n_i^2\right) \cdot \left(\frac{1}{N_P} - \frac{1}{N_Q + N_P - n_j}\right)$$

$$n_j \cdot \frac{N_Q + N_P - 2n_j}{N_Q + N_P - n_j} \geq \left(\sum_{i\in Q} n_i^2\right) \cdot \frac{N_P - n_j}{N_Q \cdot (N_Q + N_P - n_j)} + \left(\sum_{i\in P} n_i^2\right) \cdot \frac{N_Q - n_j}{N_P \cdot (N_Q + N_P - n_j)}$$

$$n_j \cdot (N_Q + N_P - 2n_j) \geq \left(\sum_{i\in Q} n_i^2\right) \cdot \frac{N_P - n_j}{N_Q} + \left(\sum_{i\in P} n_i^2\right) \cdot \frac{N_Q - n_j}{N_P}$$

$$N_Q + N_P - 2n_j \geq \left(\sum_{i\in Q} \frac{n_i^2}{n_j}\right) \cdot \frac{N_P - n_j}{N_Q} + \left(\sum_{i\in P} \frac{n_i^2}{n_j}\right) \cdot \frac{N_Q - n_j}{N_P}$$

Because n_j is the largest element, we can upper bound each term $\frac{n_i^2}{n_j}$ with n_i:

$$N_Q + N_P - 2n_j \geq (N_Q) \cdot \frac{N_P - n_j}{N_Q} + (N_P) \cdot \frac{N_Q - n_j}{N_P}$$

$$N_Q + N_P - 2n_j \geq N_P - N_j + N_Q - n_j$$

This gives an equality, and a strict inequality if $n_i < n_j$ for at least one player. Finally, we note that if the final structure is identical to the original structure, the cost is identical, so the inequality is similarly an equality. □

Lemma 10. *For a set of players with $n_i \leq \frac{\mu_e}{\sigma^2}$ $\forall i$, the grand coalition π_g is always core stable.*

Proof. For reference, **?** analyzes a restricted example of $n_i \leq \frac{\mu_e}{\sigma^2}$ case, where players come in two types, n_s, n_ℓ, both $\leq \frac{\mu_e}{\sigma^2}$. Theorem 6.7 in that work shows that the grand coalition π_g is core stable for the two-type case. This lemma extends that result to show that π_g is core stable for the broader case of $n_i \leq \frac{\mu_e}{\sigma^2}$, where players may come in more than two sizes.

First, we will assume by contradiction that there exists a set $A \subset C$, where C is the grand coalition, and where we assume that $err_j(A) < err_j(C)$ for every $j \in A$. We will then show that this violates the requirement that $n_i \leq \frac{\mu_e}{\sigma^2}$ for all $i \in C$, indicating that it is impossible for such a coalition A to exist.

By assumption,

$$err_j(C) > err_j(A)$$

Using $N_A = \sum_{i\in A} n_i$ and $N = \sum_{i\in C} n_i$ we have:

$$\frac{\mu_e}{N} + \sigma^2 \cdot \frac{\sum_{i\neq j} n_i^2 + (N - n_j)^2}{N^2} > \frac{\mu_e}{N_A} + \sigma^2 \cdot \frac{\sum_{i\in A, i\neq j} n_i^2 + (N_A - n_j)^2}{N_A^2}$$

Multiplying each side by n_j preserves the inequality:

$$\frac{\mu_e}{N} \cdot n_j + \sigma^2 \cdot \frac{\sum_{i\neq j} n_i^2 + (N - n_j)^2}{N^2} \cdot n_j > \frac{\mu_e}{N_A} \cdot n_j + \sigma^2 \cdot \frac{\sum_{i\in A, i\neq j} n_i^2 + (N_A - n_j)^2}{N_A^2} \cdot n_j$$

Next, we sum each side over all $j \in A$:

$$\sum_{j\in A} \left\{ \frac{\mu_e}{N} \cdot n_j + \sigma^2 \cdot \frac{\sum_{i\neq j} n_i^2 + (N - n_j)^2}{N^2} \cdot n_j \right\} > \sum_{j\in A} \left\{ \frac{\mu_e}{N_A} \cdot n_j + \sigma^2 \cdot \frac{\sum_{i\in A, i\neq j} n_i^2 + (N_A - n_j)^2}{N_A^2} \cdot n_j \right\}$$

We will evaluate this sum term by term. The μ_e terms are simplest:

$$\sum_{j \in A} \frac{\mu_e}{N} \cdot n_j = \frac{\mu_e}{N} \cdot N_A$$
$$\sum_{j \in A} \frac{\mu_e}{N_A} \cdot n_j = \mu_e$$

For evaluating the sum of the σ^2 coefficient, we will first note that we can rewrite the numerator:

$$\sum_{i \neq j} n_i^2 + (N - n_j)^2 = \sum_{i \neq j} n_i^2 + N^2 + n_j^2 - 2N \cdot n_j = \sum_{i \in C} n_i^2 + N^2 - 2N \cdot n_j$$

This means that the entire coefficient on the lefthand side can be rewritten as:

$$\begin{aligned}
\sum_{j \in A} \left\{ \sigma^2 \cdot \frac{\sum_{i \neq j} n_i^2 + (N - n_j)^2}{N^2} \cdot n_j \right\} &= \sum_{j \in A} \left\{ \sigma^2 \cdot \frac{N^2 + \sum_{i \in C} n_i^2 - 2N \cdot n_j}{N^2} \cdot n_j \right\} \\
&= \sum_{j \in A} \left\{ \sigma^2 \cdot \left(1 + \frac{\sum_{i \in C} n_i^2}{N^2} - 2\frac{n_j}{N} \right) \cdot n_j \right\} \\
&= \sigma^2 \cdot N_A + \sigma^2 \cdot N_A \frac{\sum_{i \in C} n_i^2}{N^2} - 2\frac{\sum_{i \in A} n_i^2}{N}
\end{aligned}$$

Similarly, we can rewrite the numerator of the σ^2 coefficient on the righthand side:

$$\sum_{i \neq j, i \in A} n_i^2 + (N_A - n_j)^2 = \sum_{i \neq j, i \in A} n_i^2 + N_A^2 + n_j^2 - 2N_A \cdot n_j = \sum_{i \in A} n_i^2 + N_A^2 - 2N_A \cdot n_j$$

Remember that $A \subset C$. Similarly, we can rewrite the entire coefficient as:

$$\begin{aligned}
\sum_{j \in A} \left\{ \sigma^2 \cdot \frac{\sum_{i \neq j, i \in A} n_i^2 + (N_A - n_j)^2}{N_A^2} \cdot n_j \right\} &= \sum_{j \in A} \left\{ \sigma^2 \cdot \frac{N_A^2 + \sum_{i \in A} n_i^2 - 2N_A \cdot n_j}{N_A^2} \cdot n_j \right\} \\
&= \sum_{j \in A} \left\{ \sigma^2 \cdot \left(1 + \frac{\sum_{i \in A} n_i^2}{N_A^2} - 2\frac{n_j}{N_A} \right) \cdot n_j \right\} \\
&= \sigma^2 \cdot N_A + \sigma^2 \cdot N_A \frac{\sum_{i \in A} n_i^2}{N_A^2} - 2 \cdot \sigma^2 \cdot \frac{\sum_{i \in A} n_i^2}{N_A} \\
&= \sigma^2 \cdot N_A + \sigma^2 \cdot \frac{\sum_{i \in A} n_i^2}{N_A} - 2 \cdot \sigma^2 \cdot \frac{\sum_{i \in A} n_i^2}{N_A} \\
&= \sigma^2 \cdot N_A - \sigma^2 \cdot \frac{\sum_{i \in A} n_i^2}{N_A}
\end{aligned}$$

Combining these terms back into the inequality gives:

$$\mu_e \cdot \frac{N_A}{N} + \sigma^2 \cdot N_A + \sigma^2 \cdot \frac{N_A}{N} \cdot \frac{\sum_{i \in C} n_i^2}{N} - 2\frac{\sum_{i \in A} n_i^2}{N} > \mu_e + \sigma^2 \cdot N_A - \sigma^2 \frac{\sum_{i \in A} n_i^2}{N_A}$$

Simplification:

$$\begin{aligned}
\mu_e \cdot \frac{N_A}{N} + \sigma^2 \cdot \frac{N_A}{N} \cdot \frac{\sum_{i \in C} n_i^2}{N} - 2\frac{\sum_{i \in A} n_i^2}{N} &> \mu_e - \sigma^2 \frac{\sum_{i \in A} n_i^2}{N_A} \\
\mu_e \cdot \frac{N_A}{N} + \sigma^2 \cdot \frac{N_A}{N} \cdot \frac{\sum_{i \in C} n_i^2}{N} - \sigma^2 \frac{\sum_{i \in A} n_i^2}{N} &> \mu_e - \sigma^2 \frac{\sum_{i \in A} n_i^2}{N_A} + \sigma^2 \frac{\sum_{i \in A} n_i^2}{N} \\
\frac{N_A}{N} \cdot \left(\mu_e + \sigma^2 \cdot \frac{\sum_{i \in C} n_i^2}{N} - \sigma^2 \frac{\sum_{i \in A} n_i^2}{N_A} \right) &> \mu_e + \sigma^2 \frac{\sum_{i \in A} n_i^2}{N} - \sigma^2 \frac{\sum_{i \in A} n_i^2}{N_A}
\end{aligned}$$

Note that the terms on the left and the right look very similar. We will strategically add and subtract a term on the left:

$$\frac{N_A}{N} \cdot \left(\mu_e + \sigma^2 \cdot \frac{\sum_{i \in C} n_i^2 - \sum_{i \in A} n_i^2 + \sum_{i \in A} n_i^2}{N} - \sigma^2 \frac{\sum_{i \in A} n_i^2}{N_A} \right) > \mu_e + \sigma^2 \frac{\sum_{i \in A} n_i^2}{N} - \sigma^2 \frac{\sum_{i \in A} n_i^2}{N_A}$$

Multiplying on the left side:

$$\frac{N_A}{N}\cdot\sigma^2\cdot\frac{\sum_{i\in C\setminus A} n_i^2}{N}+\frac{N_A}{N}\cdot\left(\mu_e+\sigma^2\cdot\frac{\sum_{i\in A} n_i^2}{N}-\sigma^2\frac{\sum_{i\in A} n_i^2}{N_A}\right)>\mu_e+\sigma^2\frac{\sum_{i\in A} n_i^2}{N}-\sigma^2\frac{\sum_{i\in A} n_i^2}{N_A}$$

Collecting terms:

$$\frac{N_A}{N}\cdot\sigma^2\cdot\frac{\sum_{i\in C\setminus A} n_i^2}{N}+\frac{N_A}{N}\cdot\left(\mu_e+\sigma^2\cdot\left(\sum_{i\in A} n_i^2\right)\cdot\left(\frac{1}{N}-\frac{1}{N_A}\right)\right)>\mu_e+\sigma^2\cdot\left(\sum_{i\in A} n_i^2\right)\cdot\left(\frac{1}{N}-\frac{1}{N_A}\right)$$

Changing the sign:

$$\frac{N_A}{N}\cdot\sigma^2\cdot\frac{\sum_{i\in C\setminus A} n_i^2}{N}+\frac{N_A}{N}\cdot\left(\mu_e-\sigma^2\cdot\left(\sum_{i\in A} n_i^2\right)\cdot\left(\frac{1}{N_A}-\frac{1}{N}\right)\right)>\mu_e-\sigma^2\cdot\left(\sum_{i\in A} n_i^2\right)\cdot\left(\frac{1}{N_A}-\frac{1}{N}\right)$$

Bringing across terms to the righthand side:

$$\frac{N_A}{N}\cdot\sigma^2\cdot\frac{\sum_{i\in C\setminus A} n_i^2}{N}>\left(1-\frac{N_A}{N}\right)\cdot\left(\mu_e-\sigma^2\cdot\left(\sum_{i\in A} n_i^2\right)\cdot\left(\frac{1}{N_A}-\frac{1}{N}\right)\right)$$

Bringing all coefficients of σ^2 to the lefthand side:

$$\frac{N_A}{N}\cdot\sigma^2\cdot\frac{\sum_{i\in C\setminus A} n_i^2}{N}+\left(1-\frac{N_A}{N}\right)\cdot\sigma^2\cdot\left(\sum_{i\in A} n_i^2\right)\cdot\left(\frac{1}{N_A}-\frac{1}{N}\right)>\left(1-\frac{N_A}{N}\right)\cdot\mu_e$$

Rewriting:

$$\frac{N_A}{N}\cdot\sigma^2\cdot\left(\sum_{i\in C\setminus A} n_i^2\right)+(N-N_A)\cdot\sigma^2\cdot\left(\sum_{i\in A} n_i^2\right)\cdot\left(\frac{1}{N_A}-\frac{1}{N}\right)>(N-N_A)\cdot\mu_e$$

We strategically rewrite the righthand side:

$$\frac{N_A}{N}\cdot\sigma^2\cdot\left(\sum_{i\in C\setminus A} n_i^2\right)+(N-N_A)\cdot\sigma^2\cdot\left(\sum_{i\in A} n_i^2\right)\cdot\left(\frac{1}{N_A}-\frac{1}{N}\right)>(N-N_A)\cdot\mu_e\cdot\frac{N_A}{N}+(N-N_A)\cdot\left(1-\frac{N_A}{N}\right)\cdot\mu_e$$

$$\frac{N_A}{N}\cdot\sigma^2\cdot\left(\sum_{i\in C\setminus A} n_i^2\right)+(N-N_A)\cdot\sigma^2\cdot\left(\sum_{i\in A} n_i^2\right)\cdot\left(\frac{1}{N_A}-\frac{1}{N}\right)>(N-N_A)\cdot\mu_e\cdot\frac{N_A}{N}+(N-N_A)\cdot N_A\cdot\left(\frac{1}{N_A}-\frac{1}{N}\right)\cdot\mu_e$$

We pull all of the terms over to the lefthand side:

$$\frac{N_A}{N}\cdot\left(\sum_{i\in C\setminus A} n_i\cdot\left(\sigma^2\cdot n_i-\mu_e\right)\right)+(N-N_A)\cdot\left(\frac{1}{N_A}-\frac{1}{N}\right)\cdot\left(\sum_{i\in A} n_i\cdot\left(n_i\cdot\sigma^2-\mu_e\right)\right)>0$$

Finally, we will show that the above inequality cannot hold. By assumption, $n_i\leq\frac{\mu_e}{\sigma^2}$ for all $i\in C$. This means that $\sigma^2\cdot n_i-\mu_e$ is negative for all $i\in C$. Because every other term on the lefthand side is positive (note that $\frac{1}{N_A}>\frac{1}{N}$), we know that the lefthand term is negative. However, the inequality is requiring that the term is positive. By this contradiction, we know that the initial assumption must have been wrong: so long as $n_i\leq\frac{\mu_e}{\sigma^2}$, there cannot be any set A such that each player strictly prefers A to C, so the grand coalition C is core stable. □

C Price of Anarchy

Lemma 11. *For a set of players with $n_i\geq\frac{\mu_e}{\sigma^2}\ \forall i$, any arrangement that is core stable or individually stable is also optimal.*

Type	Condition	Upper bound on $err_i(\Pi_M)$	Lower bound on $err_i(\Pi_{opt})$
T_0	$n_i \geq \frac{\mu_e+\sigma^2}{2\cdot\sigma^2}$	$\frac{\mu_e}{n_i}$, by Lemma 12.	$\frac{1}{2}\frac{\mu_e}{n_i}$, by Lemma 13
T_1	$\frac{\mu_e}{9\cdot\sigma^2} \leq n_i \leq \frac{\mu_e+\sigma^2}{2\sigma^2}$		σ^2, by Lemma 13
T_2	$n_i < \frac{\mu_e}{9\cdot\sigma^2}$ and is federating with other players of total mass at least $\frac{\mu_e}{3\sigma^2}$ in Π_M.	$7.25 \cdot \sigma^2$, by Lemma 14	
T_3	$n_i < \frac{\mu_e}{9\cdot\sigma^2}$ and is NOT federating with other players of total mass at least $\frac{\mu_e}{3\sigma^2}$ in Π_M.	Unbounded, but Lemma 15 gives a stability result.	

Table 2: Summary of relevant bounds for proof of Theorem 2.

Proof. By Lemma 5.3 in **?**, when all players have $\geq \frac{\mu_e}{\sigma^2}$ samples, each player with size $> \frac{\mu_e}{\sigma^2}$ minimizes its error by doing local learning. By the same lemma, each player of size exactly equal to $\frac{\mu_e}{\sigma^2}$ minimize their error in any arrangement with other players also of size $\frac{\mu_e}{\sigma^2}$. Taken together, this implies that the only stable arrangements are ones where all players of size $> \frac{\mu_e}{\sigma^2}$ are doing local learning and all players of size equal $\frac{\mu_e}{\sigma^2}$ are arranged in any grouping. Because all of these have equal error to the minimal error, the Price of Anarchy is equal to 1. □

Theorem 2 (Price of Anarchy)**.** *Denote Π_M to be a maximum-cost individually stable (IS) partition and Π_{opt} to be an optimal (lowest-cost) partition. Then,*

$$PoA = \frac{f_w(\Pi_M)}{f_w(\Pi_{opt})} \leq 9$$

Proof. This theorem is the result of multiple lemmas, each of which handle players of different sizes in different situations. Theorem 2 summarizes these contributions. Specifically, it divides players into four different types (T_0, T_1, T_2, T_3) based on their size and the group they are federating with in Π_M. These results are summarized in Table 2 and described below.

First, we note that by Lemma 12 the highest error any player can experience in Π_M is $\frac{\mu_e}{n_i}$, so the cost due to a particular player in Π_M is upper bounded by μ_e.

- Say that player $i \in T_0$ if $n_i \geq \frac{\mu_e+\sigma^2}{2\sigma^2}$. Lemma 13 shows that if $n_i \geq \frac{\mu_e+\sigma^2}{2\sigma^2}$, then $err_i(\Pi_{opt}) \geq \frac{1}{2}\frac{\mu_e}{n_i}$, so player i's contribution to the weighted cost is $\geq \frac{1}{2} \cdot \mu_e$.
- Say that player $i \in T_1$ if $\frac{\mu_e}{9\cdot\sigma^2} \leq n_i \leq \frac{\mu_e+\sigma^2}{2\sigma^2}$. Lemma 13 shows that $err_i(\Pi_{opt}) \geq \sigma^2$ for $n_i \leq \frac{\mu_e+\sigma^2}{2\sigma^2}$, so player i's contribution to the weighted cost is $\geq \sigma^2 \cdot n_i$.
- Say that player $i \in T_2$ if $n_i < \frac{\mu_e}{9\cdot\sigma^2}$ and if, in Π_M, it is federating with other players of total mass at least $\frac{\mu_e}{3\sigma^2}$. Then, by Lemma 14 $err_i(\Pi_M) \leq 7.25\sigma^2 \leq 7.5 \cdot \sigma^2$. Lemma 13 applies again and shows that $err_i(\Pi_{opt}) \geq \sigma^2$ for $n_i \leq \frac{\mu_e+\sigma^2}{2\sigma^2}$, so player i's contribution to the weighted cost is $\geq \sigma^2 \cdot n_i$.
- Say $i \in T_3$ if $n_i \leq \frac{\mu_e}{9\cdot\sigma^2}$ and if in Π_M it is *not* federating with other players of total mass at least $\frac{\mu_e}{3\cdot\sigma^2}$. Then, by Lemma 15 there is at most one group of such description in Π_M (or any IS arrangement) - call it A. What is this group's total contribution to the cost?

  $$\mu_e + \sigma^2 \cdot N_A - \sigma^2 \frac{\sum_{i\in A} n_i^2}{N_A} \leq \mu_e + \sigma^2 \cdot N_A - \sigma^2 \frac{N_A}{N_A} \leq^* \left(1 + \frac{1}{3} + \frac{1}{9}\right)\mu_e - \sigma^2 < 1.5\mu_e$$

  where in the step marked with $*$ we have upper bounded N_T by the knowledge that it contains a player of size $\leq \frac{\mu_e}{9\sigma^2}$ is federating with partners of total size no more than $\frac{\mu_e}{3\sigma^2}$. Note that N_T is the mass of the entire group containing T_3 players, and so may double-count the contributions of some players not in T_3.

Next, we bring these terms together to bound the overall result. Note that $f_w(\Pi)$ is a weighted cost that is obtained by multiplying player j's error by its number of samples n_j.

$$PoA = \frac{f_w(\Pi_M)}{f_w(\Pi_{opt})} \leq \frac{|T_0| \cdot \mu_e + |T_1| \cdot \mu_e + \sum_{i \in T_2} 7.5 \cdot \sigma^2 \cdot n_i + 1.5\mu_e}{|T_0| \cdot \frac{\mu_e}{2} + \sum_{i \in T_1} \sigma^2 \cdot n_i + \sum_{i \in T_2} \sigma^2 \cdot n_i + \sum_{i \in T_3} \sigma^2 \cdot n_i}$$

First, we note that if there do not exist any players in T_3, then we can write the bound as:

$$\frac{|T_0| \cdot \mu_e + |T_1| \cdot \mu_e + \sum_{i \in T_2} 7.5 \cdot \sigma^2 \cdot n_i}{|T_0| \cdot \frac{\mu_e}{2} + |T_1| \cdot \frac{\mu_e}{9} + \sum_{i \in T_2} \sigma^2 \cdot n_i} \leq 9$$

Suppose that $|T_3| \geq 1$. Then, the main goal is to absorb the additive $1.5 \cdot \mu_e$ term.

First, we consider the case where we have some player $n_j \geq \frac{\mu_e}{3\sigma^2}$, which we will show implies a PoA bound of 9. Any player of size $\geq \frac{\mu_e}{3 \cdot \sigma^2}$ must be in T_0 or T_1. First, we will assume that $j \in T_0$, so $|T_0| \geq 1$, meaning:

$$4.5 \cdot |T_0| \cdot \mu_e \geq |T_0| \cdot \mu_e + 1.5 \cdot \mu_e$$

This means the bound can be upper bounded by:

$$PoA \leq \frac{4.5|T_0| \cdot \mu_e + |T_1| \cdot \mu_e + \sum_{i \in T_2} 7.5 \cdot \sigma^2 \cdot n_i}{|T_0| \cdot \frac{\mu_e}{2} + |T_1| \cdot \frac{\mu_e}{9\sigma^2} + \sum_{i \in T_2} \sigma^2 \cdot n_i} \leq 9$$

Next, we consider the case where $j \in T_1$ and $|T_0| = 0$. Then, the upper bound becomes:

$$\begin{aligned}
PoA &< \frac{(|T_1| - 1) \cdot \mu_e + \mu_e + 7.5\sigma^2 \cdot \sum_{i \in T_2} n_i + 1.5\mu_e}{\sum_{i \neq j, i \in T_1} \sigma^2 \cdot n_i + \sigma^2 \cdot n_j + \sigma^2 \cdot \sum_{i \in T_2} n_i} \\
&< \frac{(|T_1| - 1) \cdot \mu_e + \mu_e + 7.5\sigma^2 \cdot \sum_{i \in T_2} n_i + 1.5\mu_e}{(|T_1| - 1) \cdot \frac{\mu_e}{9} + \frac{\mu_e}{3} + \sigma^2 \cdot \sum_{i \in T_2} n_i} \\
&< \frac{(|T_1| - 1) \cdot \mu_e + 2.5\mu_e + 9\sigma^2 \cdot \sum_{i \in T_2} n_i}{(|T_1| - 1) \cdot \frac{\mu_e}{9} + \frac{\mu_e}{3} + \sigma^2 \cdot \sum_{i \in T_2} n_i} \\
&< \frac{(|T_1| - 1) \cdot \mu_e + 3\mu_e + 9\sigma^2 \cdot \sum_{i \in T_2} n_i}{(|T_1| - 1) \cdot \frac{\mu_e}{9} + \frac{\mu_e}{3} + \sigma^2 \cdot \sum_{i \in T_2} n_i} \\
&= 9
\end{aligned}$$

Finally, we consider the case where all players have size $\leq \frac{\mu_e}{3\sigma^2}$. By Lemma 15, if there exist any players in T_3, then the entire arrangement is only stable if $\Pi_M = \pi_g = \Pi_{opt}$, giving a PoA of 1.

These proofs taken together show that the overall PoA is upper bounded by 9. □

Lemma 13. *Consider a player n_j and any set of players C. Then, we can lower bound the error player j recieves by federating with C:*

$$err_j(C \cup \{n_j\}) \geq \begin{cases} \frac{1}{2} \cdot \frac{\mu_e}{n_j} & n_j \geq \frac{\mu_e + \sigma^2}{2\sigma^2} \\ \sigma^2 & \textit{otherwise} \end{cases}$$

Proof. Player j's error when federating with the coalition C is:

$$err_j(C \cup \{n_j\}) = \frac{\mu_e}{N_C + n_j} + \sigma^2 \frac{\sum_{i \in C} n_i^2 + N_C^2}{(N_C + n_j)^2}$$

Given a fixed N_C, $\sum_{i \in C} n_i^2$ is minimized when all of the players besides j have size $n_i = \frac{N_C}{|C|}$, which means that $n_i^2 = \frac{N_C^2}{|C|^2}$. The error is thus lower bounded by:

$$err_j(C \cup \{n_j\}) \geq \frac{\mu_e}{N_C + n_j} + \sigma^2 \frac{\frac{N_C^2}{|C|} + N_C^2}{(N_C + n_j)^2}$$

This decreases with $|C|$, so we set $|C| = N_C$ to further lower bound the error:

$$\geq \frac{\mu_e}{N_C + n_j} + \sigma^2 \frac{N_C + N_C^2}{(N_C + n_j)^2}$$

Note that the "units" of this term might seem strange: the numerator of the σ^2 component involves a N_C and N_C^2. This is because we assumed that $\sum_{i \in C} n_i^2 \geq N_C$, which is correct in magnitude but which involves different units.

Next, we will lower bound this term by analyzing how it changes with N_C. First, we take the derivative with respect to N_C:

$$\frac{n_j \cdot (\sigma^2 - \mu_e + 2\sigma^2 \cdot N_C) - N_C \cdot (\mu_e + \sigma^2)}{(N_C + n_j)^3}$$

Case 1: Derivative always negative
In some situations, this derivative is always negative (the player j always prefers N_C as large as possible). When does this occur?

$$n_j \cdot (\sigma^2 - \mu_e + 2N_C \cdot \sigma^2) < (\mu_e + \sigma^2) \cdot N_C \quad \forall N_C$$

As $N_C \to \infty$, the $\sigma^2 - \mu_e$ additive term on the lefthand side becomes irrelevant, so what we require is

$$2\sigma^2 \cdot n_j \cdot N_C \leq (\mu_e + \sigma^2) \cdot N_C$$
$$n_j \leq \frac{\mu_e + \sigma^2}{2\sigma^2}$$

For players satisfying this premise, we can lower bound their error by sending $N_C \to \infty$ in the original error equation.

$$\lim_{N_C \to \infty} \left[\frac{\mu_e}{N_C + n_j} + \sigma^2 \frac{N_C + N_C^2}{(N_C + n_j)^2} \right] = \sigma^2$$

This implies that the player's error goes to σ^2 (from above), so is lower bounded by σ^2.
Case 2: Derivative sometimes negative, sometimes positive
Next, we'll consider the case where $n_j > \frac{\mu_e + \sigma^2}{2\sigma^2}$. The second derivative of the player's error with respect to N_C is:

$$2 \cdot \sigma^2 \cdot n_j - \mu_e - \sigma^2$$

which is greater than or equal to 0 in this case. In order to lower bound the overall error, we must bound the error when $N_C = 0$ (at its minimum value) and when the derivative with respect to N_C is 0 (local minimum). Note that when $N_C = 0$, player j's error is $\frac{\mu_e}{n_j}$, which is $> \frac{1}{2} \cdot \frac{\mu_e}{n_j}$, satisfying the premise. Next, we will consider the case where the derivative is equal to 0: In this case, the slope isn't always negative, so there must be some N_C such that the slope is equal to 0. This occurs when:

$$n_j \cdot (\sigma^2 - \mu_e) + N_C \cdot (2n_j \cdot \sigma^2 - \mu_e - \sigma^2) = 0$$
$$N_C = \frac{n_j \cdot (\mu_e - \sigma^2)}{2n_j \cdot \sigma^2 - \mu_e - \sigma^2}$$

Substituting in for this value of N_C into player j's error gives:

$$\frac{-\mu_e^2 - 2\mu_e \cdot \sigma^2 + 4n_j \cdot \mu_e \cdot \sigma^2 - (\sigma^2)^2}{-4n_j \cdot \sigma^2 + 4n_j^2 \cdot \sigma^2} = \frac{\mu_e}{n_j} \cdot \frac{-\mu_e - 2\sigma^2 + 4n_j \cdot \sigma^2 - \sigma^2 \cdot \frac{\sigma^2}{\mu_e}}{-4 \cdot \sigma^2 + 4n_j \cdot \sigma^2}$$

In order to prove that this whole term is lower bounded by $\frac{1}{2}\frac{\mu_e}{n_j}$, we will show that the coefficient on $\frac{\mu_e}{n_j}$ is lower bounded by $\frac{1}{2}$. Because $n_j \geq 1$, we know that the denominator is positive:

$$\frac{-\mu_e - 2\sigma^2 + 4n_j \cdot \sigma^2 - \sigma^2 \cdot \frac{\sigma^2}{\mu_e}}{-4 \cdot \sigma^2 + 4n_j \cdot \sigma^2} \geq \frac{1}{2}$$
$$-2\mu_e - 4\sigma^2 + 8n_j \cdot \sigma^2 - \frac{\sigma^4}{\mu_e} \geq -4\sigma^2 + 4n_j \cdot \sigma^2$$
$$-2\mu_e + 4n_j \cdot \sigma^2 - \frac{\sigma^4}{\mu_e} \geq 0$$
$$n_j \geq \frac{\mu_e}{2\sigma^2} + \frac{\sigma^2}{4\mu_e}$$

This is satisfied if the lower bound is smaller than or equal to $\frac{\mu_e+\sigma^2}{2\sigma^2}$. We can show this by noting that $\mu_e \geq \sigma^2$ for any avenue of interest (otherwise, $\frac{\mu_e}{\sigma^2} < 1$ and by Lemma 11 the only stable arrangement is to have all players doing local learning). This means that:

$$\frac{\mu_e}{2\sigma^2} + \frac{\sigma^2}{4\mu_e} \leq \frac{\mu_e}{2\sigma^2} + \frac{1}{4} = \frac{\mu_e + \frac{1}{2}\sigma^2}{2\sigma^2} < \frac{\mu_e + \sigma^2}{2\sigma^2}$$

as desired. This shows that:

$$err_j(C \cup \{n_j\}) \geq \frac{1}{2}\frac{\mu_e}{n_j}$$

□

Lemma 14. *Consider a player j federating with a coalition C. If the total number of samples N_C is at least $\frac{\mu_e}{3\sigma^2}$, then $err_j(C \cup \{n_j\}) \leq 7.25 \cdot \sigma^2$.*

Proof. The error a player n_j experiences is given by:

$$err_j(C \cup \{n_j\}) = \frac{\mu_e}{n_j + N_C} + \sigma^2 \frac{\sum_{i \in l} n_i^2 + N_C^2}{(N_C + n_j)^2}$$

Given a fixed total sum N_C, the $\sum_{i \in l} n_i^2$ term is maximized when all of the mass is on a single partner. So the overall cost can be upper bounded by:

$$< \frac{\mu_e}{N_C + n_j} + \sigma^2 \frac{2N_C^2}{(N_C + n_j)^2}$$

Taking the derivative with respect to N_C gives:

$$\begin{aligned} -\frac{\mu_e}{(N_C + n_j)^2} + \sigma^2 \frac{4N_C \cdot (N_C + n_j)^2 - 4N_C^2 \cdot (N_C + n_j)}{(N_C + n_j)^4} &= -\frac{\mu_e}{(N_C + n_j)^2} + \sigma^2 \frac{4N_C \cdot n_j}{(N_C + n_j)^3} \\ &= \frac{-\mu_e \cdot (N_C + n_j) + 4\sigma^2 N_C \cdot n_j}{(N_C + n_j)^3} \end{aligned}$$

Next, we will upper bound player j's error based on the sign of the derivative with respect to N_C.

Case 1: Derivative with respect to N_C always positive:
This occurs when the numerator is positive for all $N_C \geq 0$, or

$$\begin{aligned} -\mu_e \cdot (N_C + n_j) + 4\sigma^2 N_C \cdot n_j &> 0 \\ N_C \cdot (4\sigma^2 \cdot n_j - \mu_e) &> \mu_e \cdot n_j \end{aligned}$$

To begin with, we must have that $4\sigma^2 \cdot n_j > \mu_e$ or else the lefthand side is negative, so $n_j > \frac{\mu_e}{4\sigma^2}$. Given that, the error is largest when N_C is set to its largest value of $\frac{\mu_e}{3\sigma^2}$.

$$\begin{aligned} \frac{\mu_e}{3 \cdot \sigma^2} \cdot (4\sigma^2 n_j - \mu_e) &> \mu_e \cdot n_j \\ 4\sigma^2 n_j - \mu_e &> 3\sigma^2 \cdot n_j \\ n_j &> \frac{\mu_e}{\sigma^2} \end{aligned}$$

If this is the case, what is the maximum amount of error that n_j receives? The error is in the form:

$$\frac{\mu_e}{N_C + n_j} + \sigma^2 \frac{2N_C^2}{(N_C + n_j)^2}$$

We know that this is maximized when $N_C \to \infty$. In this case, μ_e term goes to 0. The σ^2 term (by L'Hôpital's rule) goes to:

$$\sigma^2 \frac{4N_C}{2(N_C + n_j)} \to 2\sigma^2$$

Case 2: Derivative with respect to N_C is always negative
Next, we'll consider the inverse case where the derivative is always negative. This occurs when:

$$N_C \cdot (4\sigma^2 \cdot n_j - \mu_e) < \mu_e \cdot n_j \quad \forall N_C$$

This has to be true for all N_C, which implies that the $4\sigma^2 \cdot n_j - \mu_e$ term is negative, or $n_j \leq \frac{\mu_e}{4\sigma^2}$. If this is the case, the maximal error is achieved when the N_C term is smallest ($\frac{\mu_e}{3\cdot\sigma^2}$). Plugging into the error form gives us:

$$\begin{aligned}
\frac{\mu_e}{\frac{\mu_e}{3\cdot\sigma^2}+n_j} + \sigma^2 \frac{2\cdot\frac{\mu_e^2}{9\cdot\sigma^4}}{\left(\frac{\mu_e}{3\cdot\sigma^2}+n_j\right)^2} &= \frac{\mu_e\cdot\left(\frac{\mu_e}{3\cdot\sigma^2}+n_j\right)+\frac{2\mu_e^2}{9\cdot\sigma^2}}{\left(\frac{\mu_e}{3\cdot\sigma^2}+n_j\right)^2} \\
&= \frac{\mu_e\cdot\left(\frac{\mu_e}{3\cdot\sigma^2}+n_j\right)+\frac{2\mu_e^2}{9\cdot\sigma^2}}{\frac{1}{9\sigma^4}\cdot\left(\mu_e+3\sigma^2\cdot n_j\right)^2} \\
&< \frac{9\sigma^4\mu_e\cdot\left(\frac{\mu_e}{3\cdot\sigma^2}+n_j\right)+2\cdot\mu_e^2\cdot\sigma^2}{\mu_e^2} \\
&= 3\sigma^2 + 9\sigma^2\cdot\frac{\sigma^2}{\mu_e}\cdot n_j + 2\sigma^2 \\
&< 5\sigma^2 + 9\sigma^2\cdot\frac{\sigma^2}{\mu_e}\cdot\frac{\mu_e}{4\sigma^2} \\
&= 7.25
\end{aligned}$$

where in the last step we have used that $n_j \leq \frac{\mu_e}{4\sigma^2}$.

Case 3: when the derivative with respect to N_C is sometimes positive and sometimes negative Using the values above, we know this occurs when $\frac{\mu_e}{4\sigma^2} \leq n_j \leq \frac{\mu_e}{\sigma^2}$. First, we'll confirm that the error first decreases and then increases with N_C. The derivative is:

$$N_C\cdot\left(4\sigma^2\cdot n_j-\mu_e\right)-\mu_e\cdot n_j$$

Here, we are assuming that the coefficient on N_C is either 0 or positive, so the second derivative with respect to N_C is positive. Given that the derivative is negative at some point, it must be negative for small N_C. We know from Case 1 that as $N_C \to \infty$, the error goes to $2\sigma^2$, so in order to bound the entire space, we only need to bound the error at the smallest value of N_C, which is $\frac{\mu_e}{3\cdot\sigma^2}$. The first few steps are identical to Case 2:

$$\frac{\mu_e}{\frac{\mu_e}{3\cdot\sigma^2}+n_j} + \sigma^2 \frac{2\cdot\frac{\mu_e^2}{9\cdot\sigma^4}}{\left(\frac{\mu_e}{3\cdot\sigma^2}+n_j\right)^2} = \frac{\mu_e\cdot\left(\frac{\mu_e}{3\cdot\sigma^2}+n_j\right)+\frac{2\mu_e^2}{9\cdot\sigma^2}}{\left(\frac{\mu_e}{3\cdot\sigma^2}+n_j\right)^2} = \frac{\mu_e\cdot\left(\frac{\mu_e}{3\cdot\sigma^2}+n_j\right)+\frac{2\mu_e^2}{9\cdot\sigma^2}}{\frac{1}{9\sigma^4}\cdot\left(\mu_e+3\sigma^2\cdot n_j\right)^2}$$

In the next step, though, we use that $\frac{\mu_e}{4\sigma^2} \leq n_j \leq \frac{\mu_e}{\sigma^2}$.

$$\begin{aligned}
&< \frac{9\sigma^4\mu_e\cdot\left(\frac{\mu_e}{3\cdot\sigma^2}+n_j\right)+2\cdot\mu_e^2\cdot\sigma^2}{(\mu_e+\frac{3}{4}\mu_e)^2} \\
&= \frac{3\sigma^2+9\sigma^2\cdot\frac{\sigma^2}{\mu_e}\cdot n_j+2\sigma^2}{\frac{49}{16}} \\
&< \frac{16}{49}\cdot\left(5\sigma^2+9\sigma^2\cdot\frac{\sigma^2}{\mu_e}\cdot\frac{\mu_e}{\sigma^2}\right) \\
&= \frac{16}{49}\cdot 14\cdot\sigma^2 \\
&< 5\sigma^2
\end{aligned}$$

Of the three cases, the highest bound is $7.25\cdot\sigma^2$. □

Lemma 15, below, relies on Lemmas 16, 17, and 18, which are stated and proved immediately after the proof of Lemma 15.

Lemma 15. *Consider an arrangement of players, all of size $\leq \frac{\mu_e}{3\sigma^2}$, where at least one player is in a federating cluster where the total mass of its partners is no more than $\frac{\mu_e}{3\sigma^2}$. Then, the only stable arrangement of these players is to have all of them federating together.*

Proof. By Lemma 16, we know that every player in every group welcomes the addition of any other player. Therefore, in order to prove that this arrangement isn't individually stable, we simply have to prove that a player would wish to move.

We will consider a cluster A with elements $i \in T_3$ present. We know that there exists at least one element in A s.t. the mass of its partners $(N_A - n_i)$ is less than $\frac{\mu_e}{3\sigma^2}$. This implies also that $N - n_a < \frac{\mu_e}{3\sigma^2}$ for n_a the largest element in A. We also know that $n_a < \frac{\mu_e}{3\sigma^2}$ because we know that there exists some other element in the cluster with $N_A - n_i < \frac{\mu_e}{3\sigma^2}$.

Next, let's suppose there exists some other cluster B, such that all elements are $\leq \frac{\mu_e}{3\sigma^2}$ in size. We will consider some n_b largest player in B. There are four possible cases:

1. $n_a \geq n_b, N_A - n_a \geq N_B - n_b$: Unstable by Lemma 17 (player b wishes to move to A).
2. (Symmetric to above) $n_a \leq n_b$, $N_A - n_a \leq N_B - n_b$: Unstable by Lemma 17 (player a wishes to move to B).
3. $n_a > n_b$, $N_A - n_a < N_B - n_b$. Note that in this case, we know that $N_A - n_a \leq \frac{\mu_e}{3\sigma^2}$, so we satisfy the conditions of Lemma 18, and thus player a would prefer to join B.
4. $n_a < n_b$, $N_A - n_a > N_B - n_b$. In this case, we know that $\frac{\mu_e}{3\sigma^2} > N_A - n_a > N_B - n_b$, so we again satisfy the conditions of Lemma 18, and thus player b would prefer to join A.

□

Lemma 16. *A group of players where each has size $n_i \leq \frac{\mu_e}{3\sigma^2}$ always welcomes the addition of another player of size $n_k \leq \frac{\mu_e}{3\sigma^2}$.*

Proof. For this section, we will rewrite the form of the error that a player experiences while federating with a coalition C. Specifically, we will write the error in the form below, where a_i refers to the number of players with number of samples n_i.

$$\frac{\mu_e}{\sum_{i=1}^{M} a_i \cdot n_i} + \sigma^2 \frac{\sum_{i \neq j} a_i \cdot n_i^2 + (a_j - 1) \cdot n_j^2 + (\sum_{i \neq j} a_i \cdot n_i + (a_j - 1) \cdot n_j)^2}{(\sum_{i=1}^{M} a_i \cdot n_i)^2}$$

Setting $N = \sum_{i=1}^{M} a_i \cdot n_i$ gives:

$$\frac{\mu_e}{N} + \sigma^2 \cdot \frac{\sum_{i \neq j} a_i \cdot n_i^2 + (a_j - 1) \cdot n_j^2 + (N - n_j)^2}{N^2}$$

In order to prove that any player j welcomes the addition of any other player k, we will show that the derivative with respect to a_k is always negative. This means that player j always sees its error decrease with the addition of another player of size n_k. As we take the derivative, the coefficient on the μ_e term in the error value becomes:

$$-\frac{\mu_e \cdot n_k}{N^2} = -\frac{\mu_e \cdot n_k \cdot N^2}{N^4}$$

The derivative of the coefficient on the σ^2 term becomes:

$$\frac{\sigma^2}{N^4} \cdot \left(\left(n_k^2 + 2(N - n_j) \cdot n_k\right) \cdot N^2 - \left(\sum_{i \neq j} a_i \cdot n_i^2 + (a_j - 1) \cdot n_j^2 + (N - n_j)^2 \right) \cdot 2 \cdot N \cdot n_k \right)$$

So, the overall derivative is negative if:

$$\mu_e \cdot n_k \cdot N^2 > \sigma^2 \cdot \left(\left(n_k^2 + 2(N - n_j) \cdot n_k\right) \cdot N^2 - \left(\sum_{i \neq j} a_i \cdot n_i^2 + (a_j - 1) \cdot n_j^2 + (N - n_j)^2 \right) \cdot 2 \cdot N \cdot n_k \right)$$

We pull out and cancel common terms:

$$\mu_e \cdot n_k \cdot N^2 > \sigma^2 \cdot n_k \cdot N \cdot \left((n_k + 2N - 2n_j) \cdot N - 2 \left(\sum_{i \neq j} a_i \cdot n_i^2 + (a_j - 1) \cdot n_j^2 + (N - n_j)^2 \right) \right)$$

$$\mu_e \cdot N > \sigma^2 \cdot \left((n_k + 2(N - n_j)) \cdot (N - n_j + n_j) - 2 \left(\sum_{i \neq j} a_i \cdot n_i^2 + (a_j - 1) \cdot n_j^2 + (N - n_j)^2 \right) \right)$$

Strategically expanding:

$$\mu_e \cdot N > \sigma^2 \cdot \left(n_k \cdot N + 2(N - n_j)^2 + 2n_j \cdot N - 2n_j^2 - 2\left(\sum_{i \neq j} a_i \cdot n_i^2 + (a_j - 1) \cdot n_j^2 \right) - 2(N - n_j)^2 \right)$$

Collecting:

$$\mu_e \cdot N > \sigma^2 \cdot \left(N \cdot (n_k + 2n_j) - 2 \sum_{i=1}^{M} a_i \cdot n_i^2 \right)$$

Substituting in for N:

$$\mu_e \cdot \sum_{i=1}^{M} a_i \cdot n_i > \sigma^2 \cdot \left(\sum_{i=1}^{M} a_i \cdot n_i \cdot (n_k + 2n_j) - 2 \sum_{i=1}^{M} a_i \cdot n_i^2 \right)$$

$$0 > \sum_{i=1}^{M} a_i \cdot n_i \cdot (\sigma^2 \cdot n_k + 2\sigma^2 \cdot n_j - 2\sigma^2 \cdot n_i - \mu_e)$$

Our goal is to show that this is negative if $n_i \leq \frac{\mu_e}{3\sigma^2}$ for all i.

First, we look over the portion of the sum equal to the k index. This term is equal to:

$$a_k \cdot n_k \cdot (2\sigma^2 \cdot n_j - \sigma^2 \cdot n_k - \mu_e)$$

which is negative, given our conditions. Next, we look at the j term in the sum:

$$a_j \cdot n_j \cdot (\sigma^2 \cdot n_k - \mu_e)$$

which is also negative. The remaining portions of the sum can be written as:

$$(N - a_j \cdot n_j - a_k \cdot n_k) \cdot (\sigma^2 \cdot n_k + 2\sigma^2 \cdot n_j - \mu_e) - 2\sigma^2 \sum_{i \neq j,k} a_i \cdot n_i^2$$

which we would like to show is negative. We can maximize this term by holding N constant and minimizing the negative portion by setting $n_i = 1$ for all other players besides j, k. This gives us an upper bound of:

$$\leq (N - a_j \cdot n_j - a_k \cdot n_k) \cdot (\sigma^2 \cdot n_k + 2\sigma^2 \cdot n_j - \mu_e) - 2\sigma^2 (N - a_j \cdot n_j - a_k \cdot n_k)$$
$$= (N - a_j \cdot n_j - a_k \cdot n_k) \cdot (\sigma^2 \cdot n_k + 2\sigma^2 \cdot n_j - \mu_e - 2\sigma^2)$$

Given the condition that $n_k, n_j \leq \frac{\mu_e}{3\sigma^2}$, we know that the coefficient is no more than

$$3\sigma^2 \frac{\mu_e}{3\sigma^2} - \mu_e - 2\sigma^2 < 0$$

Taken together, this shows that the derivative of player j's error with respect to a_k is negative, which means that player j always sees its error decrease with the addition of another player k. □

Lemma 17. *Assume we have two groups of players, A and B with all players of size $\leq \frac{\mu_e}{3\sigma^2}$. Then, if either of the two conditions below are satisfied, the arrangement is not individually stable.*

1. *There exists $a \in A, b \in B$ such that $n_a = n_b$.*
2. *There exists $a \in A, b \in B$ such that $n_a > n_b$ and $N_A - n_a \geq N_B - n_b$. (Note that this could be defined symmetrically with respect to B).*

Proof. First, we will assume that player a does not wish to move to B (if this is not true, then we already know that the arrangement is not IS). This tells us that:

$$err_a(A) \leq err_a(B \cup \{n_a\})$$

Next, we will derive sufficient conditions for player b to wish to move to A, or

$$err_b(A \cup \{n_b\}) < err_b(B)$$

We will use the shorthand of $N'_A = N_A - n_a$ and $N'_B = N_B - n_b$. From the form of each player's error as in Lemma 1, we can derive conditions for the difference in errors experienced by two players in the same coalition. Consider a coalition C and two players $j, k \in C$, with $n_k \geq n_j$ Then,

$$\begin{aligned}
err_j(C) - err_k(C) &= \sigma^2 \cdot \frac{\sum_{i \neq j} n_i^2 + (N_C - n_j)^2}{N_C^2} - \sigma^2 \cdot \frac{\sum_{i \neq k} n_i^2 + (N_C - n_k)^2}{N_C^2} \\
&= \sigma^2 \cdot \frac{n_k^2 - n_j^2 + (N_C - n_j)^2 - (N_C - n_k)^2}{N_C^2} \\
&= \sigma^2 \cdot \frac{n_k^2 - n_j^2 + (N_C^2 + n_j^2 - 2n_j \cdot N_C) - (N_C^2 + n_k^2 - 2n_k \cdot N_C)}{N_C^2} \\
&= \sigma^2 \cdot \frac{-2n_j \cdot N_C + 2n_k \cdot N_C}{N_C^2} \\
&= 2\sigma^2 \cdot \frac{N_C \cdot (n_k - n_j)}{N_C^2} \\
&= 2\sigma^2 \cdot \frac{n_k - n_j}{N_C}
\end{aligned}$$

We can apply this derivation to obtain two equalities:

$$err_b(A \cup b) = err_a(A \cup b) + 2\sigma^2 \frac{n_a - n_b}{N'_A + n_a + n_b}$$

$$err_a(B \cup a) = err_b(B \cup a) - 2\sigma^2 \frac{n_a - n_b}{N'_B + n_a + n_b}$$

So, rewriting the first inequality tells us that:

$$err_a(A) \leq err_b(B \cup a) - 2\sigma^2 \frac{n_a - n_b}{N'_B + n_a + n_b}$$

Pulling over:

$$err_a(A) + 2\sigma^2 \frac{n_a - n_b}{N'_B + n_a + n_b} \leq err_b(B \cup a)$$

Note that because all of the players are of size $\leq \frac{\mu_e}{3 \cdot \sigma^2}$, we know by Lemma 16 that every player welcomes the addition of every other player, so

$$err_b(B \cup a) < err_b(B)$$

In order to complete the proof, we need to show that $err_b(A \cup \{n_b\})$ is less than $err_a(A) + 2\sigma^2 \frac{n_a - n_b}{N'_B + n_a + n_b}$. Again, because all of the players are of size $\leq \frac{\mu_e}{3 \cdot \sigma^2}$, we know from Lemma 16 that every player welcomes the addition of every other player, so

$$err_a(A \cup \{n_b\}) + 2\sigma^2 \frac{n_a - n_b}{N'_B + n_a + n_b} < err_a(A) + 2\sigma^2 \frac{n_a - n_b}{N'_B + n_a + n_b}$$

From our prior relation, we know that

$$err_b(A \cup b) - 2\sigma^2 \frac{n_a - n_b}{N'_A + n_a + n_b} + 2\sigma^2 \frac{n_a - n_b}{N'_B + n_a + n_b} = err_a(A \cup \{n_b\}) + 2\sigma^2 \frac{n_a - n_b}{N'_B + n_a + n_b}$$

Rewriting the term on the left tells us that what we want to show is:

$$err_b(A \cup b) \leq err_b(A \cup b) + 2\sigma^2 \cdot (n_a - n_b) \cdot \left(\frac{1}{N'_B + n_a + n_b} - \frac{1}{N'_A + n_a + n_b} \right)$$

Now, we can apply our case analysis. If $n_a = n_b$, then the added coefficient is 0, so the final inequality holds. The inequality also holds if the fractional coefficient is positive or 0, or

$$\begin{aligned}
\frac{1}{N'_B + n_a + n_b} &\geq \frac{1}{N'_A + n_a + n_b} \\
N'_A + n_a + n_b &\geq N'_B + n_a + n_b \\
N'_A &\geq N'_B
\end{aligned}$$

which is exactly the second criteria. □

Lemma 18. *Assume we have two groups of players, A and B, with all players of size $\leq \frac{\mu_e}{3\sigma^2}$. Define n_a, n_b to be the largest players in A, B respectively. Assume that $n_a > n_b$ and $N_A - n_a < N_B - n_b$, with $N_A - n_a \leq \frac{\mu_e}{3\sigma^2}$. Then, player a would prefer to join B.*

Proof. We will show that the preconditions imply that player a would wish to move to group B, or else

$$err_a(B \cup \{n_a\}) < err_a(A)$$

Or, rewritten out,

$$\frac{\mu_e}{N_B + n_a} + \sigma^2 \frac{\sum_{i \in B} n_i^2 + N_B^2}{(N_B + n_a)^2} < \frac{\mu_e}{N_A} + \sigma^2 \frac{\sum_{i \in A, i \neq a} n_i^2 + (N_A - n_a)^2}{N_A^2}$$

We will upper and lower bound the costs on both sides by taking the worst and best case scenario for how the B and A players can be arranged, respectively. We have already showed that we can minimize the total arrangement of fixed total mass by dividing it into players of size exactly 1, so the player sizes equal to 1, or

$$\sum_{i \in A, i \neq a} n_i^2 \geq N_A - n_a$$

Conversely, let's try to upper bound the B sum. Previously, we did this by grouping all of the mass into a single player. In this case, we can't do this - we've assumed that the n_b term is the largest of them, so the most we can set them to be equal to is n_b exactly. However, the same reasoning still holds: if we keep the total $N_B - n_b$ constant but rearrange them into groups of maximum size b, we only increase total cost. To see why, consider that we have x, y with $x \geq y$, and some $x \leq b \leq x + y$. Then, we wish to show that:

$$x^2 + y^2 < b^2 + (x + y - b)^2$$

Expanding:

$$x^2 + y^2 < b^2 + (x + y - b)^2 = b^2 + b^2 + x^2 + y^2 - 2b \cdot x - 2b \cdot y + 2x \cdot y$$

Cancelling common terms means we want to show:

$$2b \cdot x + 2b \cdot y < 2b^2 + 2x \cdot y$$

$$x + y < b + \frac{x \cdot y}{b} < b + \frac{b \cdot y}{b} = b + y$$

which is satisfied.

This result tells us that this process (grouping them into players of exactly size n_b, plus at most one player of size $< n_b$) does maximize the total sum, subject to this constraint. We will again use the shorthand of $N'_A = N_A - n_a$ and $N'_B = N_B - n_b$. Excluding player n_b, the mass is N'_B, so the number of copies of n_b that we can make is $\frac{N'_B}{n_b} := c + \epsilon$, for integer c and $\epsilon \in [0, 1)$. If we know that $\epsilon = 0$ (which is always achievable), then we know that:

$$\sum_{i \in B, i \neq b} n_i^2 \leq c \cdot n_b^2 = \frac{N'_B}{n_b} \cdot n_b^2 = N'_B \cdot n_b$$

What if $\epsilon > 0$? Then,

$$\sum_{i \in B, i \neq b} n_i^2 \leq c \cdot n_b^2 + (\epsilon \cdot n_b)^2 < c \cdot n_b^2 + \epsilon \cdot n_b^2 = \frac{N'_B}{n_b} \cdot n_b^2 = N'_B \cdot n_b$$

So, in either way, the $N'_B \cdot n_b$ term is an upper bound. This means that the worst-case scenario for us to show that:

$$\frac{\mu_e}{N'_B + n_a + n_b} + \sigma^2 \frac{N'_B \cdot n_b + n_b^2 + (N'_B + n_b)^2}{(N'_B + n_a + n_b)^2} < \frac{\mu_e}{N'_A + n_a} + \sigma^2 \frac{N'_A + (N'_A)^2}{(N'_A + n_a)^2}$$

We'll work by upper bounding the lefthand side. First, we'll replace the N'_B. First, we'll also look at the derivative with respect to N'_B, which gives:

$$\frac{n_a(-\mu_e + 3n_b \cdot \sigma^2 + 2N_b \cdot \sigma^2) - (n_b + N'_B)(\mu_e + n_b \sigma^2)}{(n_a + n_b + N'_B)^3}$$

The numerator can be rewritten as:

$$-\mu_e \cdot (n_a + n_b + N'_B) - \sigma^2 \cdot n_b \cdot (N'_B + n_b) + 3\sigma^2 \cdot n_a \cdot n_b + 2\sigma^2 \cdot n_a \cdot N'_B$$

We can show that this is negative because:

$$N'_B \cdot (-\mu_e + 2\sigma^2 \cdot n_a) < 0$$

since $n_a \leq \frac{\mu_e}{3\sigma^2}$. Similarly,

$$n_b \cdot (-\mu_e + 3\sigma^2 \cdot n_a) \leq 0$$

Because the derivative with respect to N'_B is negative, we can over-bound it by setting it to its smallest value: $N'_A + 1$ (or N'_A, for simplicity). This means that we can upper bound the lefthand side by writing:

$$\frac{\mu_e}{N'_A + n_a + n_b} + \sigma^2 \frac{N'_A \cdot n_b + n_b^2 + (N'_A + n_b)^2}{(N'_A + n_a + n_b)^2} < \frac{\mu_e}{N'_A + n_a} + \sigma^2 \frac{N'_A + (N'_A)^2}{(N'_A + n_a)^2}$$

Next, we'll work on replacing the n_b term on the lefthand side. We start out by taking the derivative of the lefthand side with respect to n_b. This gives us:

$$\frac{-\mu_e \cdot (n_a + N'_A) + \sigma^2 \cdot N'_A \cdot (N'_A + 3n_a) + n_b \cdot (-\mu_e + \sigma^2 \cdot (N'_A + 4n_a))}{(N'_A + n_a + n_b)^3}$$

We will inspect the sign of the derivative, which is given by the numerator. Specifically, we will show that the derivative is always negative or 0 at $n_b = 0$, and is either negative forever, or else is negative and then positive. This implies that the lefthand side of the overall equation is either always decreasing in n_b (implying that we can upper bound it by setting $n_b = 0$) or else is decreasing and then increasing (in which case the upper bound is either at $n_b = 0$ or $n_b = n_a$).

First, we will prove our claim about the derivative. At $n_b = 0$, the derivative is:

$$-\mu_e \cdot (n_a + N'_A) + \sigma^2 \cdot N'_A \cdot (N'_A + 3n_a)$$

We want to show this is negative, or:

$$\sigma^2 \cdot N'_A \cdot (3n_a + N'_A) \leq \mu_e \cdot (n_a + N'_A)$$

Upper bounding the lefthand side:

$$3\sigma^2 \cdot N'_A \cdot (n_a + N'_A) \leq \mu_e \cdot (n_a + N'_A)$$

$$3\sigma^2 \cdot N'_A \leq \mu_e$$

which is satisfied by assumption. So, we know that the derivative starts out as 0 or negative. If the coefficient on n_b (equal to $\sigma^2 \cdot (4n_a + N'_A) - \mu_e$) is negative, then the lefthand side of the overall equation is always decreasing as n_b increases - so the upper bound at $n_b = 0$ suffices. Otherwise, the curve is decreasing, then increasing.

Upper bound at $n_b = 0$
This bound is fairly straightforward. What we want to show is:

$$\frac{\mu_e}{N'_A + n_a} + \sigma^2 \frac{N'^2_A}{(N'_A + n_a)^2} < \frac{\mu_e}{N'_A + n_a} + \sigma^2 \frac{N'_A + (N'_A)^2}{(N'_A + n_a)^2}$$

which is obviously true.

Upper bound at $n_b = n_a$
This bound is trickier. (Note that technically, the upper bound is at $n_a - 1$, but it is simpler to over-bound with n_a). What we'd like to show is:

$$\frac{\mu_e}{N'_A + 2n_a} + \sigma^2 \frac{N'_A \cdot n_a + n_a^2 + (N'_A + n_a)^2}{(N'_A + 2n_a)^2} \leq \frac{\mu_e}{N'_A + n_a} + \sigma^2 \frac{N'_A + N'^2_A}{(N'_A + n_a)^2}$$

We can write:

$$\mu_e \cdot \left(\frac{1}{N'_A + n_a} - \frac{1}{N'_A + 2n_a} \right) = \mu_e \cdot \frac{n_a}{(N'_A + n_a) \cdot (N'_A + 2n_a)}$$

Next, we move on to the σ^2 portion. Note that we can simplify the lefthand side, since:

$$N_A'^2 + n_a^2 + 2N_A' \cdot n_a + n_a^2 + N_A' \cdot n_a = N_A'^2 + 2n_a^2 + 3N_A' \cdot n_a = (N_A' + n_a) \cdot (N_A' + 2n_a)$$

So, the inequality we'd like to show becomes:

$$\sigma^2 \frac{N_A' + n_a}{N_A' + 2n_a} - \sigma^2 \frac{N_A' + N_A'^2}{(N_A' + n_a)^2} \leq \mu_e \cdot \frac{n_a}{(N_A' + n_a) \cdot (N_A' + 2n_a)}$$

Simplifying the lefthand side gives:

$$\sigma^2 \cdot \frac{(N_A' + n_a)^3 - (N_A' + N_A'^2) \cdot (N_A' + 2n_a)}{(N_A' + 2n_a) \cdot (N_A' + n_a)^2} \leq \mu_e \cdot \frac{n_a}{(N_A' + n_a) \cdot (N_A' + 2n_a)}$$
$$\sigma^2 \cdot \frac{(N_A' + n_a)^3 - (N_A' + N_A'^2) \cdot (N_A' + 2n_a)}{N_A' + n_a} \leq \mu_e \cdot n_a$$

We can make the lefthand side larger by making the negative part smaller - specifically, replacing the $(N_A' + 2n_a)$ with a $(N_A' + n_a)$. This gives us:

$$\sigma^2 \cdot \frac{(N_A' + n_a)^3 - (N_A' + N_A'^2) \cdot (N_A' + n_a)}{N_A' + n_a} \leq \mu_e \cdot n_a$$
$$\sigma^2 \cdot \left((N_A' + n_a)^2 - (N_A' + N_A'^2)\right) \leq \mu_e \cdot n_a$$

Expanding out the lefthand side gives us:

$$\sigma^2 \cdot (N_A'^2 + n_a^2 + 2n_a \cdot N_A' - N_A' - N_A'^2) \leq \mu_e \cdot n_a$$
$$\sigma^2 \cdot (n_a^2 + 2n_a \cdot N_A' - N_A') \leq \mu_e \cdot n_a$$

Again, we can make the lefthand side larger by dropping the negative portion:

$$\sigma^2 \cdot (n_a^2 + 2n_a \cdot N_A') \leq \mu_e \cdot n_a$$
$$\sigma^2 \cdot (n_a + 2N_A') \leq \mu_e$$

Which is satisfied because we require N_A', n_a both $\leq \frac{\mu_e}{3\sigma^2}$. Note that, while this is a $\leq$, because we know that $n_b < n_a$, the overall inequality is strict. □