Model-sharing Games: Analyzing Federated Learning Under Voluntary Participation

Kate Donahue, 1 Jon Kleinberg, 1, 2
1 Department of Information Science, Cornell University
2 Department of Computer Science, Cornell University
kdonahue@cs.cornell.edu, kleinber@cs.cornell.edu

Abstract
Federated learning is a setting where agents, each with access to their own data source, combine models learned from local data to create a global model. If agents are drawing their data from different distributions, though, federated learning might produce a biased global model that is not optimal for each agent. This means that agents face a fundamental question: should they join the global model or stay with their local model? In this work, we show how this situation can be naturally analyzed through the framework of coalitional game theory.

Motivated by these considerations, we propose the following game: there are heterogeneous players with different model parameters governing their data distribution and different amounts of data they have noisily drawn from their own distribution. Each player’s goal is to obtain a model with minimal expected mean squared error (MSE) on their own data. They have a choice of fitting a model based solely on their own data, or combining their learned parameters with those of some subset of the other players. Combining models reduces the variance component of their error through access to more data, but increases the bias because of the heterogeneity of distributions. In this work, we derive exact expected MSE values for problems in linear regression and mean estimation. We use these values to analyze the resulting game in the framework of hedonic game theory; we study how players might divide into coalitions, where each set of players within a coalition jointly constructs a single model. In a case with arbitrarily many players that each have either a “small” or “large” amount of data, we constructively show that there always exists a stable partition of players into coalitions.

1 Introduction
Imagine a situation as follows: a hospital is trying to evaluate the effectiveness of a certain procedure based on data it has collected from procedures done on patients in their facilities. It seems likely that certain attributes of the patient influence the effectiveness of the procedure, so the hospital analysts opt to fit a linear regression model with parameters $\theta$. However, because of the limited amount of data the hospital has access to, this model has relatively high error. Luckily, other hospitals also have data from implementations of this same procedure. However, for reasons of privacy, data incompatibility, data size, or other operational considerations, the hospitals don’t wish to share raw patient data. Instead, they opt to combine their models by taking a weighted average of the parameters learned by each hospital. If there are $M$ hospitals and hospital $i$ has $n_i$ samples, the combined model parameters would look like:

$$\hat{\theta}_f = \frac{1}{\sum_{i=1}^{M} n_i} \sum_{i=1}^{M} \hat{\theta}_i \cdot n_i$$

The situation described above could be viewed as a stylized model of federated learning. Federated learning is a distributed learning process that is currently experiencing rapid innovations and widespread implementation (Li et al. 2020; Kairouz et al. 2019). It is commonly used in cases where data is distributed across multiple agents and cannot be combined centrally for training. For example, federated learning is implemented in word prediction on cell phones, where transferring the raw text data would be infeasible given its large size (and sensitive content). The motivating factor for using federated learning is that access to more data will reduce the variance in a learned model, reducing its error.

However, there could be a downside to using federated learning. In the hospital example, it seems quite reasonable that certain hospitals might have different true generating models for their data, based on the differences in patient populations or variants of the procedure implementation, for example. Two dissimilar hospitals that are federating together will see a decrease in their model’s error due to model variance, but an increase in their error due to model bias. This raises some fundamental questions for each participating hospital - or, more generally, each agent $i$ considering federating. Which other agents should $i$ federate with in order to minimize its error? Will those other agents be interested in federating with $i$? Does there exist some stable arrangement of agents into federating clusters, and if so, what does that arrangement look like?

Numerous works have explored the issue of heterogeneous data in federated learning - we discuss specifically how they relate to ours in a later section. Often the goal in these lines of work is to achieve equality in error rates guaranteed to each agent, potentially by actively collecting more data or using transfer learning to ensure the model better fits local data. However, to our knowledge, there has not yet been work that systematically looks at the participation questions inherent in federated learning through the lens
of game theory — especially the theory of hedonic games, which studies the formation of self-sustaining coalitions.

In a hedonic game, players are grouped together into clusters or coalitions: the overall collection of coalitions is called a coalition structure. Each player’s utility depends solely on the identity of the other players in its coalition. A common question in hedonic games is the stability of a coalition structure. A coalition structure \( \Pi \) is core-stable (or “in the core”) if there does not exist a coalition \( C \) so that every player in \( C \) prefers \( C \) to its coalition in \( \Pi \). A coalition structure is individually stable if there does not exist a coalition \( C \subseteq \Pi \) so that a player \( i \notin C \) prefers \( C \cup \{i\} \) to its arrangement in \( \Pi \) and all players in \( C \) weakly prefer \( C \cup \{i\} \) to \( C \) (Bogomolnaia and Jackson 2002).

To explain the analogy of federated learning to hedonic games, we first consider that each agent in federated learning is a player in a hedonic game. A player is in coalition with other players if it is federating with them. Its cost is its expected error in a given federating cluster, which depends only on the identity of other players in its federating cluster. Players are assumed to be able to move between federating clusters only if doing so would benefit itself and not harm other players in the cluster it is moving to: notably, we allow players to freely leave a cluster, even if doing so would harm the players in the cluster it leaves behind.

**The present work: Analyzing federated learning through hedonic game theory** In this work, we use the framework of hedonic games to analyze the stability of coalitions in data-sharing applications that capture key issues in federated learning. By working through a sequence of deliberately stylized models, we obtain some general insights about participation and stability in these kinds of applications.

The first case we consider is where all players have the same number of data points. We show that in this case, when the number of data points \( n \) is fairly small, the only core-stable coalition structure is to have all players federating together. When \( n \) is large, the only core-stable coalition structure is to have all players separate. There exists a point case of intermediate \( n \) size where all coalition structures are core-stable. Next, in the case where all players have either one of two sizes (“small” or “large”), the analysis is more complicated, but we demonstrate constructively that there always exists a stable partition of players into clusters.

We are only able to produce these hedonic game theory results because of our derivations of exact error values for the underlying inference problems. We calculate these values for agents federating in two situation: 1) a mean estimation problem and 2) a linear regression problem. The error values depend on the number of samples each agent has access to, with the expectation taken over the values of samples each agent draws as well as the possible different true parameters of the data each player is trying to model. Our results are completely independent of the generating distributions used, relying only weakly on the ratio of two parameters.

Finally, we analyze two additional models, one of which ensures local learning is never in the core, and the other of which ensures the coalition structure where all players are federating together is the sole element in the core.

| Coalition structure | \( \text{err}_a(\cdot) \) | \( \text{err}_b(\cdot) \) | \( \text{err}_c(\cdot) \) |
|---------------------|----------------|----------------|----------------|
| \( \{a\}, \{b\}, \{c\} \) | 2 | 2 | 2 |
| \( \{a, b\}, \{c\} \) | 1.5 | 1.5 | 2 |
| \( \{a, b, c\} \) | 1.3 | 1.3 | 1.3 |

Table 1: The expected errors of players in each coalition when all three players have 5 samples each, with parameters \( \mu_c = 10, \sigma^2 = 1 \). Each row denotes a different coalition partition; for example, \( \{a, b\}, \{c\} \) indicates that players \( a \) and \( b \) are federating together while \( c \) is alone. Coalitions that are identical up to renaming of players are omitted.

| Coalition structure | \( \text{err}_a(\cdot) \) | \( \text{err}_b(\cdot) \) | \( \text{err}_c(\cdot) \) |
|---------------------|----------------|----------------|----------------|
| \( \{a\}, \{b\}, \{c\} \) | 2 | 2 | 0.4 |
| \( \{a, b\}, \{c\} \) | 1.5 | 1.5 | 0.4 |
| \( \{a\}, \{b, c\} \) | 2 | 1.72 | 0.39 |
| \( \{a, b, c\} \) | 1.55 | 1.55 | 0.41 |

Table 2: The expected errors of players in each coalition when players \( a \) and \( b \) have 5 samples each and player \( c \) has 25 samples, with parameters \( \mu_c = 10, \sigma^2 = 1 \).

2 Motivating example

To motivate our problem and clarify the types of analyses we will be exploring, we will first work through a simple motivating example. In this case, we will consider three players, \( a, b, \) and \( c \), who are each trying to solve a mean-estimation problem. Calculating the error each player can expect requires two parameters: \( \mu_c \), which reflects the average error each player experiences when sampling data from its own personal distribution, and \( \sigma^2 \), which reflects the average variance in the true parameters between players. We will discuss these more in future sections, including how to handle cases where they may be imperfectly known, but for now we will take them to be fixed.

We will first assume that each player has 5 samples from their local data distribution: Table 1 gives the error each player can expect in this situation. Note that because the players have the same number of samples, players in identical situations have identical errors. Every player sees its error minimized in the “grand coalition” \( \pi_g \) where all three players are federating together. This implies that the only arrangement that is stable (core-stable or individually stable) is \( \pi_g \). This is part of a broader pattern that we will more

| Coalition structure | \( \text{err}_a(\cdot) \) | \( \text{err}_b(\cdot) \) | \( \text{err}_c(\cdot) \) |
|---------------------|----------------|----------------|----------------|
| \( \{a\}, \{b\}, \{c\} \) | 0.4 | 0.4 | 0.4 |
| \( \{a, b\}, \{c\} \) | 0.7 | 0.7 | 0.4 |
| \( \{a, b, c\} \) | 0.8 | 0.8 | 0.8 |

Table 3: The expected errors of players in each coalition when players \( a, b, c \) each have 25 samples, with parameters \( \mu_c = 10, \sigma^2 = 1 \).
formally investigate in later sections.

Next, we will assume that player $c$ increases the amount of samples it has from 5 to 25: Table 3 demonstrates the error each player can expect in this situation. Here, the players have different preferences over which arrangement they would most prefer. The “small” players $a$ and $b$ would most prefer $\{a, b\} \{c\}$, whereas the “large” player $c$ would most prefer $\{a\} \{b, c\}$ or (identically) $\{b\} \{a, c\}$. However out of all of these coalition structures, only $\{a, b\}, \{c\}$ is stable (either core stable or individually stable). Note that $\{a\}, \{b, c\}$ is not core stable because the coalition $C = \{a, b\}$ is one where each player prefers $C$ to its current situation. It is also not individually stable because of the same reason: player $b$ could leave its coalition to join $\{a\}$, and the resulting set $\{a, b\}$ leads to a reduction in both of their errors.

Finally, we will assume that all three players have 25 samples: this example is shown in Table 4. As in Table 1 the players have identical preferences. However, in the case where they each had 5 samples, they minimized their error by being together. In the case where they each have 25 samples, the players minimize their error by being alone.

In later sections we will give theoretical results that explain this example more fully, but understanding the core-stable partitions here will help to build intuition for more general results.

3 Model and assumptions

Model and technical assumptions

This section introduces our model, which we will analyze in later sections. In our model, we assume that there is a fixed set of $[M]$ players. Player $j$ has a fixed number of samples, $n_j$. Though the number of samples is fixed, it is possible to analyze a varying number of samples by investigating all games involving the relevant number of samples. Each player draws their true parameters i.i.d. (independent and identically distributed) $(\theta_j, \varepsilon_j^2) \sim \Theta$. $\varepsilon_j^2$ represents the amount of noise present in the sampling process for a given player.

In the case of mean estimation, $\theta_j$ is a scalar representing the true mean of player $j$. Player $j$ draws samples i.i.d. from its true distribution: $Y \sim D_j(\theta_j, \varepsilon_j^2)$. Samples are drawn with variance $\varepsilon_j^2$ around the true mean of the distribution.

In the case of linear regression, $\theta_j$ is a $D$-dimensional vector representing the coefficients on the true classification function, which is also assumed to be linear. Each player draws $n_j$ input datapoints from their own input distribution $X_j \sim C_j$ such that $E_{x \sim C_j}[x^T x] = \Sigma_j$. They then noisy observations the outputs, drawing values i.i.d. $Y_j \sim D_j(X_j^T \theta_j, \varepsilon_j^2)$, where $\varepsilon_j^2$ again denotes the variance of how samples are drawn around the true mean.

We denote $\mu_e = E_{(\theta, \varepsilon) \sim \Theta}[\varepsilon^2]$; the expectation of the error parameter. In the mean estimation case, $\sigma^2 = \text{Var}(\theta_j)$ represents the variance around the mean. In the linear regression case, $\sigma^2 = \text{Var}(\theta_j^2)$. For simplicity, most of our analysis assumes each dimension in the coefficient is drawn independently from each other, but we relax this in the appendix.

We assume that each player knows how many samples it has access to. It may or may not have access to the data itself, but it does not know how its values (or its parameters) differ from the mean. For example, it does not know if the data it has is unusually noisy or if its true mean lies far from the true mean of other players.

All of the stability analysis results depend on the parameters $\mu_e$ and $\sigma^2$. However, the reliance is fairly weak: often the player only needs to know the ratio $\frac{\mu_e}{\sigma^2}$, and frequently only needs to know whether the number of samples a player has $n_j$ is larger or smaller than $\frac{\mu_e}{\sigma^2}$.

Much of this paper analyzes the stability of coalition structures. In particular, we focus on core stability and individual stability. Analyzing stability could be relevant because players can actually move between coalitions. However, even if players aren’t able to actually move, analyzing the stability of a coalition tells us something about its optimality for each set of players. We will refer to the coalition partition where all players are grouped together as the “grand coalition”.

Related works

Incentives and federated learning: [Blum et al. 2017] describes an approach to handling heterogeneous data where more samples are iteratively gathered from each agent in a way so that all agents are incentivized to participate in the grand coalition during federated learning. [Duan et al. 2021] builds a framework to schedule data augmentation and re-sampling. [Yu, Bagdasaryan, and Shmatikov 2020] demonstrates empirically that there can be cases where individuals get lower error with local training than federated and evaluates empirical solutions. [Wang et al. 2020] analyzes the question of when it makes sense to split or not to split datasets drawn from different distributions. Finally, [Blum et al. 2020] analyzes notions of envy and efficiency with respect to sampling allocations in federated learning.

Transfer learning: [Mansour et al. 2020] and [Deng, Kamani, and Mahdavi 2020] both propose theoretical methods for using transfer learning to minimize error provided to agents with heterogeneous data. [Li et al. 2019] and [Martinez, Bertran, and Sapiro 2020] both provide methods to produce a more uniform level of error rates across agents participating in federated learning.

Clustering and federated learning: [Sattler, Muller, and Samek 2020] and [Shlezinger, Rini, and Eldar 2020] provide an algorithm to “cluster” together players with similar data distributions with the aim of providing them with lower error. They differ from our approach in that they consider the case where there is some knowledge of each player’s data distribution, where we only assume knowledge of the number of data points. Additionally, their approach doesn’t explicitly consider agents to be game-theoretic actors in the same way that this one does. Interestingly, [Guazzone, Anglano, and Sereno 2014] uses a game-theoretic framework to analyze feder-
ated learning, but with the aim of minimizing energy usage, not error rate.

Normative assumptions

This paper is primarily descriptive: it aims to model a phenomenon in the world, not to say whether that phenomenon is good or bad. For example, it could be that society as a whole values situations where many players federate together and might wish to require players to do so, regardless of whether this minimizes their error. It might be the case that society prefers all players, regardless of how many samples they have access to, have roughly similar error rates. Our use of the expected mean squared error is also worth reflecting on: it assumes that over- and under-estimates are equally costly and that larger mis-estimates are more costly. In a more subtle point, we are taking the expected MSE over parameter draws \( E(\theta, \epsilon^2) \sim \Theta \). A player with a true mean that happens to fall far from the mean might experience a much higher error than its expected MSE.

In the entirety of this paper, we are taking as fixed the requirement that data not be shared, either for privacy or technical capability reasons, and so are implicitly valuing that requirement more than the desire for lower error. We are also assuming that the problem at hand is completely encompassed by the machine learning task, which might omit the fact that non-machine learning solutions may be better suited. It also may be the fact that technical requirements other than error rate are more important: for example, the desire to balance the amount of computation done by each agent.

4 Expected error results

This paper’s first contribution is to derive exact expected values for the MSE of players under different situations. The fact that these values are exact allows us to precisely reason about each player’s incentives in later sections. We will state the main lemmas here and provide the proofs in Appendix B.

First, we provide results from mean estimation. Note that the local error decays exponentially with the number of samples the player has access to. The error player \( j \) gets while participating in federated error depends in a more complicated way on the number of samples player \( j \) has, as well as the number of samples each member in the coalition has.

**Lemma 4.1.** For mean estimation, the expected MSE of local estimation for a player with \( n_j \) samples is \( \frac{\mu_e}{n_j} \).

**Lemma 4.2.** For mean estimation, the expected MSE of federated estimation for a player with \( n_j \) samples is

\[
\frac{\mu_e}{N} + \frac{\sum_{i \neq j} n_i^2 + (N - n_j)^2}{N^2} \sigma^2
\]

where \( N = \sum_{i=1}^{M} n_i \).

Next, we provide results from linear regression. For each situation, we provide results that hold in general as well as simplified results that hold for the situation where \( X_j \) follows a certain distribution. Note that when the number of samples \( n_j \) is much larger than the dimension of the problem \( D \), the error values in the linear regression case take exactly the same form as those in mean estimation.

**Lemma 4.3.** For linear regression, the expected MSE of local estimation for a player with \( n_j \) samples is

\[
\mu_e \cdot \text{tr} \left[ \sum_j E_{X_j \sim X_j} \left[ (X_j^T X_j)^{-1} \right] \right]
\]

If the distribution of input values \( X_j \) is a \( D \)-dimensional multivariate normal distribution with 0 mean, then, the expected MSE of local estimation can be simplified to:

\[
\frac{\mu_e}{n_j - D - 2}
\]

**Lemma 4.4.** For linear regression, the expected MSE of federated estimation for a player with \( n_j \) samples is:

\[
L_j + \frac{\sum_{i \neq j} n_i^2 + (N - n_j)^2}{N^2} \frac{D}{n_j} \sum_{d=1}^{D} E_{x \sim X_j} [(x^d)^2] \cdot \sigma_d^2
\]

where \( L_j \) is equal to:

\[
\mu_e \sum_{i=1}^{M} \frac{n_i^2}{N^2} \frac{D}{n_i - D - 2}
\]

For linear regression, we define the hedonic game to have cost in local estimation equal to:

\[
\frac{\mu_e}{n_j} \cdot D
\]

In the federated case, we define the \( L_j \) term to be:

\[
\mu_e \sum_{i=1}^{M} \frac{n_i}{N} \cdot D
\]

In the limit of \( n_j \to \infty \), the cost is equal to the error rate the player experiences. For ease of analysis, we will operate with the cost rather than the error because cost closely approximates error, but has the added advantage of fitting exactly into the form of the error in the mean estimation game. This enables us to apply the coalition formation analysis in the next section to linear regression as well. In the next section we will focus on the mean estimation case, where we will use “cost” and “error” interchangeably.

5 Coalition formation: all players have same number of samples

In this section, we analyze the stability of coalition structures in two cases. We will use \( \pi_l \) to refer to the coalition partition where all players are alone and \( \pi_g \) to refer to the coalition where all players are together.

In the first case, we assume all players have the same number of samples: \( n_i = n \). The error for mean estimation simplifies greatly, as the lemma below shows.
Lemma 5.1. If all players have the same number of samples \( n \), then:

- If \( n < \frac{\mu_e}{\sigma^2} \), players minimize their error in \( \pi_g \).
- If \( n > \frac{\mu_e}{\sigma^2} \), players minimize their error in \( \pi_l \).
- If \( n = \frac{\mu_e}{\sigma^2} \), players are indifferent between any arrangement of players.

Proof. In the case that all players have the same number of samples, we can use \( n_l = n \) to simplify the error term:

\[
\frac{\mu_e}{M \cdot n} + \sigma^2 \left( \frac{M - 1}{M} \cdot \frac{n^2}{M^2} + \frac{(M - 1)^2 \cdot n^2}{M^2} \right)
\]

\[
= \frac{\mu_e}{M \cdot n} + \sigma^2 \frac{M - 1}{M}
\]

In order to see whether players would prefer a larger group (higher \( M \)) or a smaller group (smaller \( M \)), we take the derivative of the error with respect to \( M \):

\[
\frac{d}{dM} \left( \frac{\mu_e}{M \cdot n} + \sigma^2 \frac{M - 1}{M} \right) = \frac{-\mu_e}{M^2 \cdot n} - \sigma^2 \frac{1}{M^2} \cdot M
\]

This is positive when \( n > \frac{\mu_e}{\sigma^2} \): a player gets higher error the more players it is federating with. This is negative when \( n < \frac{\mu_e}{\sigma^2} \): a player gets lower error the more players it is federating with. This is \( 0 \) when \( n = \frac{\mu_e}{\sigma^2} \), which implies players should be indifferent between different arrangements. Plugging in for \( n = \frac{\mu_e}{\sigma^2} \) in the error equation gives:

\[
\frac{\mu_e}{M \cdot \mu_e} + \sigma^2 \frac{M - 1}{M} = \sigma^2
\]

which is equivalent to the error a player would get alone:

\[
\frac{\mu_e}{n} = \frac{\mu_e \cdot \sigma^2}{\mu_e} = \sigma^2
\]

This formulation allows us to classify the core of this problem in a very clean way. It turns out that, for all cases where \( n < \frac{\mu_e}{\sigma^2} \), \( \pi_g \) is the only element in the core: this continues until \( n = \frac{\mu_e}{\sigma^2} \), where all arrangements are in the core. For \( n > \frac{\mu_e}{\sigma^2} \), \( \pi_l \) becomes the only element in the core.

Lemma 5.2. If all players have the same number of samples \( n \), then:

- If \( n < \frac{\mu_e}{\sigma^2} \), \( \pi_g \) is the only partition that is core-stable.
- If \( n > \frac{\mu_e}{\sigma^2} \), \( \pi_l \) is the only partition that is core-stable.
- If \( n = \frac{\mu_e}{\sigma^2} \), any arrangement of players is core-stable.

Proof. If a partition \( \pi \) is optimal for every player, then it is core stable: there does not exist a coalition \( C \) where all players prefer \( C \) to \( \pi \), because there does not exist a coalition where any players prefer \( C \) to \( \pi \).

If a partition \( \pi \) is optimal for every player, then no other partition can be core stable: any set of players not in their optimal configuration could form a coalition \( C \) where all players would prefer \( C \).

In the case that players are indifferent between any arrangement, then for any partition \( \pi \) and any competing coalition \( C \), all players would be indifferent between \( \pi \) and \( C \), so \( \pi \) is core stable.

6 Coalition formation: Small & large player case

In this section, we add another layer of depth by allowing players to come in one of two “sizes”. “Small” players have \( n_s \) samples and “large” ones have \( n_L \) samples, with \( n_s < n_L \). We demonstrate that versions of the game in this pattern always have a stable partition by constructively producing an element that is stable. Note that this is not true in general of hedonic games. As discussed in Bogomolnaia and Jackson (2002), there are multiple instances where a game might have no stable partition.

To characterize this space, we divide it into cases depending on the relative size of \( n_s, n_L \). We will use the notation \( \pi(s, \ell) \) to denote a coalition with \( s \) small players and \( \ell \) large players, out of a total of \( S \) and \( L \) present. We will use \( \pi(s_1, \ell_1) \succ \pi(s_2, \ell_2) \) to mean that the small players prefer coalition \( \pi(s_1, \ell_1) \) to \( \pi(s_2, \ell_2) \) and \( \pi(s_1, \ell_1) \succ \pi(s_2, \ell_2) \) to mean the same preference, but for large players. All proofs from this section are present in Appendix C.

Case: \( n_s, n_L \geq \frac{\mu_e}{\sigma^2} \)

The first case is when \( n_s \) is fairly large: it turns out that each player minimizes their error by using local estimation, which implies that \( \pi_l \) is in the core. The intuitive explanation here is that the reduction in variance that federation would bring is more than offset by the increase in bias. Using the lemmas below tells us that when \( n_s > \frac{\mu_e}{\sigma^2} \), \( \pi_l \) is the only element in the core and when \( n_s = \frac{\mu_e}{\sigma^2} \) then any arrangement where the large players are alone are in the core.

Lemma 6.1. If \( n_i > \frac{\mu_e}{\sigma^2} \) for all \( i \in [M] \), then \( \pi_l \) is the unique element in the core.

Lemma 6.2. If \( n_i \geq \frac{\mu_e}{\sigma^2} \) for all \( i \in [M] \), with \( n_k > \frac{\mu_e}{\sigma^2} \) for at least one player \( k \), then any arrangement where the players with samples \( n_k > \frac{\mu_e}{\sigma^2} \) are alone is in the core.

Case: \( n_s, n_L \leq \frac{\mu_e}{\sigma^2} \)

Next, we explore the case where \( n_s \leq \frac{\mu_e}{\sigma^2} \). First, there are several building block lemmas we will find useful.

Lemma 6.3. If \( n_s \leq \frac{\mu_e}{\sigma^2} \), and \( n_s < n_L \), a small player in a coalition with \( s \) small players and \( \ell \) large players always prefers \( s \) as large as possible:

\[
s_2 > s_1 \quad \Rightarrow \quad \pi(s_2, \ell) \succ_0 \pi(s_1, \ell)
\]

Lemma 6.4. If \( n_e \geq \frac{\mu_e}{\sigma^2} \), and \( n_s < n_e \), a large player in a coalition with \( s \) small players and \( \ell \) large players always prefers \( \ell \) as small as possible:

\[
\ell_2 < \ell_1 \quad \Rightarrow \quad \pi(s, \ell_2) \succ_1 \pi(s, \ell_1)
\]

If \( n_k < \frac{\mu_e}{\sigma^2} \), then large players do not necessarily prefer \( \ell \) to be either as large as possible or as small as possible.

Lemma 6.5. Assume \( n_s \leq \frac{\mu_e}{\sigma^2} \), and \( n_s < n_L \). If \( n_L \leq \frac{\mu_e}{\sigma^2} \), a small player in a coalition with \( s \) small players and \( \ell \) large players always prefers \( \ell \) as large as possible. If \( n_L > \frac{\mu_e}{\sigma^2} \), as \( \ell \) is increased, the small player’s error first increases and then decreases: the player achieves a maximum error, and then overall error decreases with \( \ell \).
Lemma 6.6. Assume \( n_s \leq \frac{\mu_s}{\sigma_s^2} \), \( n_L \geq \frac{\mu_L}{\sigma_L^2} \), and \( n_s < n_L \). As \( s \) is increased, the large player’s error first decreases and then increases: the large player achieves a minimum level of error, and afterwards its error increases with \( s \).

If \( n_L < \frac{\mu_L}{\sigma_L^2} \), then large players do not necessarily prefer \( s \) to be either as large as possible or as small as possible.

These lemmas help us produce the theorem below. The proof of the theorem should not be immediately obvious from the above lemmas and is explained in detail in the appendix.

Theorem 6.7. If \( n_L \leq \frac{\mu_L}{\sigma_L^2} \) and \( n_s < n_L \), then the grand coalition \( \pi_g \) is core-stable.

Case: \( n_s < \frac{\mu_s}{\sigma_s^2} \), \( n_L > \frac{\mu_L}{\sigma_L^2} \)

Finally, we consider the case where \( n_s < \frac{\mu_s}{\sigma_s^2} \) and \( n_L > \frac{\mu_L}{\sigma_L^2} \).

We will prove the following theorem:

Theorem 6.8. Assume \( n_L > \frac{\mu_L}{\sigma_L^2} \). Then, there exists an arrangement of small and large players that is individually stable.

As discussed previously, individual stability is a notion of stability that is, roughly speaking, weaker than core stability. The theorem above raises the intriguing possibility that we could strengthen the result to show that there always exists a core-stable arrangement. Though we do not show that in this work, we believe it to be true and have included the additional lemma below that, while not used in our theorem, would be helpful in proving a core stability result.

Lemma 6.9. Consider two coalitions \( \pi(s_1, \ell_1) \) and \( \pi(s_2, \ell_2) \) with \( s_2 < s_1 \). Then, it is not possible to pick \( \ell_2 \) so that \( \pi(s_1, \ell_1) \succ 0 \pi(s_2, \ell_2) \) and \( \pi(s_1, \ell_1) \succ 1 \pi(s_2, \ell_2) \).

Next, a proof of the theorem:

Proof. We will prove this directly by calculating an arrangement that is individually stable.

First, group all of the small players together. Then calculate \( \ell^* = \max \ell \) such that \( \pi(S, \ell) \succeq 1 \pi(0, 1) \): the largest number of large players that can be in the coalition such that the large players prefer this to being alone. Check if \( \pi(S, \ell) \prec 0 \pi(S, 0) \). If this is true, make the final arrangement \( \pi(S, 0) \), \( \pi(0, 1) \) – \( L \): by previous lemmas around how the small player’s error changes with \( \ell \), we know that if \( \pi(S, \ell) \prec 0 \pi(S, 0) \), then \( \pi(S, \ell') \prec 0 \pi(S, 0) \) for all \( \ell' < \ell \).

We will show that this is individually stable by showing no players wish to unilaterally deviate.

• No small player wishes to go to \( \pi(1, 0) \): reducing the number of small players in a group from \( S \) to 1 monotonically increases the error the small player faces.

• No small player wishes to go to \( \pi(1, 1) \): it is possible to reach this state by first going from \( \pi(S, 0) \) to \( \pi(S, 1) \) (which would increase error because \( 1 \leq \ell \) and \( \pi(S, \ell) \prec 0 \pi(S, 0) \)) and then from \( \pi(S, 1) \) to \( \pi(1, 1) \) (which would increase error because reducing the number of small players increases error).

• No large player can go to \( \pi(S, 1) \): this would increase the error of the small players.

• No large player wishes to go to \( \pi(0, 2) \): this would increase the error of both large players.

Next, we will consider the case where \( \pi(S, \ell') \geq 0 \pi(S, 0) \): in this case, we will show that \( \pi(S, \ell) \) is individually stable.

• No small player wishes to go from \( \pi(S, \ell') \) to \( \pi(1, 0) \). We can see that \( \pi(1, 0) \) has higher error because we know \( \pi(S, \ell') \geq 0 \pi(S, 0) \succ 0 \pi(1, 0) \).

• No small player wishes to go to \( \pi(1, 1) \). We can see \( \pi(1, 1) \) has higher error for the small player because \( \pi(S, \ell') \geq 0 \pi(S, 1) \succ 0 \pi(1, 1) \). The first inequality comes from the following reasoning: if \( \frac{d}{d \ell} \text{err}_0(\pi(S, \ell)) \) is negative at \( \ell = 1 \), then there is a monotonically increasing path of error from \( \ell' \) to 1. If \( \frac{d}{d \ell} \text{err}_0(\pi(S, \ell)) \) is positive at 1, then we know that \( \pi(S, 1) \prec 0 \pi(S, 0) \), whereas \( \pi(S, \ell') \geq 0 \pi(S, 0) \).

• No large player wishes to go to \( \pi(S, \ell' + 1) \): by definition of \( \ell' \), it would get greater error than in \( \pi(0, 1) \).

• No large player wishes to go to \( \pi(0, 2) \) for the same reason as above.

By this analysis, \( \pi(S, \ell') \) is individually stable.

Taken together, in this section we have shown that there always exists a stable partition of players into coalitions in the case where players come in two sizes. In the next section, we will consider other models of federation.

7 Other federation models

In the previous sections, we analyzed one type of federated learning: when the global model is produced by taking the weighted average of the parameters each player calculates on their own data:

\[
\hat{\theta}^f = \frac{1}{\sum_{i=1}^{M} n_i} \sum_{i=1}^{M} \hat{\theta}_i \cdot n_i
\]

With this federation method, we saw in previous sections that it is frequently the case that the situation where all players are federating together (\( \pi_g \)) is not stable. It might seem natural to investigate whether an alternate weighting mechanism might produce a stable \( \pi_g \) coalition structure. In this section, we investigate two alternate federating methods, both implemented in the mean-estimation case. These types of methods combine the global model with the local model: in this way, they can be seen as analogues to some kind of transfer learning.

All proofs are in Appendix D and Appendix A has a discussion of how this approach relates to other approaches related to combining data, such as Bayesian estimation and the James-Stein estimator.

w-weighting

In this case, each player has a parameter \( w_j \) that it uses to weight the global model with its own local model.

\[
\hat{\theta}_j^w = w_j \cdot \hat{\theta}_j + (1 - w_j) \cdot \frac{1}{N} \sum_{i=1}^{M} \hat{\theta}_i \cdot n_i
\]
Lemma 7.1. For $w$-weighting federated learning, the expected MSE of a player with $n_j$ samples is:

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

We are also able to reason about the optimal $w_j$ parameter. The lemma below tells us that each player would prefer federation, in some form, to being alone.

Lemma 7.2. The minimum error is always achieved when $w_j < 1$, implying that federation is always preferable to local learning.

However, there are two potential shortfalls of the $w_j$ method. The first is that, for a player $j$ in a coalition, increasing $n_i$ for $i \neq j$ can increase player $j$’s error. This is unfortunate because it means players are hurt when more samples are made available, even though having more samples improves overall knowledge of the parameters. For example, for $\mu_e = 10$, $\sigma^2 = 1$, consider two cases: one with a coalition of 4 players, each with $n_x = 30$, and one with a coalition of 3 players with $n_x = 30$ and one player with $n_x = 300$. In the first case, a small player has expected error 0.271, in the second case, 0.280.

The second shortfall is that, in some cases, $\pi_g$ is not core-stable. We can refer to the same example: the small player in a coalition with only the 3 small players would get error 0.278, which is less than the error it would get in $\pi_g$. In this situation, $\pi_g$ is not core-stable. The next section will explore a model that alleviates these drawbacks.

$v$-matrix

In this case, the federated estimate differs for each player and is:

$$\hat{\theta}_j = \sum_{i=1}^{M} v_{ji} \hat{\theta}_i$$

for $\sum_{i=1}^{M} v_{ji} = 1$. $\mathbf{v}_j$ is a length $M$ vector that denotes the weight player $j$ places on data taken from each of the players. Note that we can recover the $w$ weighting case with $v_{ji} = w + \frac{(1-w)n_i}{N}$ and $v_{ji} = (1-w) \cdot \frac{n_i}{N}$.

We can again derive exact error values:

Lemma 7.3. For mean estimation with the $v$-weighting federated learning method, the expected MSE of a player with $n_j$ samples is:

$$\mu_e \sum_{i=1}^{M} \frac{v_{ji}^2}{n_i} + \left( \sum_{i \neq j} v_{ji}^2 + \left( \sum_{i \neq j} v_{ji} \right)^2 \right) \cdot \sigma^2$$

Similarly, we can calculate the optimal $v$ weights for player $j$’s error rate.

Lemma 7.4. Define $V_i = \sigma^2 + \frac{\mu_e}{n_i}$. Then, the value of $v_{ji}$ that minimizes player $j$’s error is:

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

and

$$v_{jk} = \frac{1}{V_k} \cdot \frac{V_j - \sigma^2}{1 + V_j \sum_{i \neq j} v_{ji}}$$

From this analysis, a few properties become clear. To start with, $v_{jj}$ and $v_{jk}$ are always strictly between 0 and 1. This implies the following lemma:

Corollary 7.5. With $v$ weights set optimally, $\pi_g$ is optimal for each player.

Proof. Suppose by contradiction that some other coalition $\pi'$ gave player $j$ a lower error. WLOG, assume this coalition omitted player $k$. In this case, the $v$ weights for $\pi'$ can be represented as a length $M$ vector with 0 in the $k$th entry. However, set of weights is achievable in $\pi_g$; it is always an option to set a player’s coefficient $v_{jk}$ equal to 0. This contradicts the use of $v_j$ as an optimal weighting, so it cannot be the case that any player gets lower error in a different coalition.

Similarly, the fact that $\pi_g$ is optimal for every player implies that it is in the core, and that it is the only element in the core.

8 Conclusions and future directions

In this work, we have drawn a connection between a simple model of federated learning and the game theoretic tool of hedonic games. We used this tool to examine stable partitions of the space for two variants of the game. In service of this analysis, we computed exact error values for mean estimation and linear regression. Finally, we proposed and analyzed two other variants of federated learning that incentivize the formation of larger coalitions.

We believe that this framework is a simple and useful tool for analyzing the incentives of multiple self-interested agents in a learning environment. There are many fascinating extensions. For example, completely characterizing the core (including whether it is always non-empty) in the case of arbitrary number of samples $\{n_i\}$ is an obvious area of investigation. Besides this, it could be interesting to compute exact or approximate error values for cases beyond mean estimation and linear regression.

Acknowledgments

This work was supported in part by a Simons Investigator Award, a Vannevar Bush Faculty Fellowship, a MURI grant, AFOSR grant FA9550-19-1-0183; grants from the ARO and the MacArthur Foundation, and NSF grant DGE-1650441. We are grateful to A. F. Cooper, Thodoris Lykouris, Hakim Weatherspoon, and the AI in Policy and Practice working group at Cornell for invaluable discussions.
References

Abu-Mostafa, Y.; Lin, H.; and Magdon-Ismail, M. 2012. Learning from data: a short course: AMLbook. Google Scholar.

Anderson, T. W. 1962. An introduction to multivariate statistical analysis. Technical report, Wiley New York.

Blum, A.; Haghtalab, N.; Procaccia, A. D.; and Qiao, M. 2017. Collaborative PAC Learning. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., Advances in Neural Information Processing Systems 30, 2392–2401. Curran Associates, Inc. URL https://papers.nips.cc/paper/6833-collaborative-pac-learning.pdf.

Blum, A.; Haghtalab, N.; Shao, H.; and Phillips, R. L. 2020. Unpublished work, private correspondence.

Bogomolnaia, A.; and Jackson, M. O. 2002. The Stability of Hedonic Coalition Structures. Games and Economic Behavior 38(2): 201 – 230. ISSN 0899-8256. doi:https://doi.org/10.1006/game.2001.0877. URL http://www.sciencedirect.com/science/article/pii/S0899825601908772.

Casella, G. 1992. Illustrating empirical Bayes methods. Chemometrics and intelligent laboratory systems 16(2): 107–125.

Deng, Y.; Kamani, M. M.; and Mahdavi, M. 2020. Adaptive Personalized Federated Learning.

Duan, M.; Liu, D.; Chen, X.; Liu, R.; Tan, Y.; and Liang, L. 2021. Self-Balancing Federated Learning With Global Imbalanced Data in Mobile Systems. IEEE Transactions on Parallel and Distributed Systems 32(1): 59–71.

Efron, B.; and Morris, C. 1977. Stein’s paradox in statistics. Scientific American 236(5): 119–127.

Guazzone, M.; Anglano, C.; and Sereno, M. 2014. A Game-Theoretic Approach to Coalition Formation in Green Cloud Federations. 2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. doi:10.1109/ccgrid.2014.37. URL http://dx.doi.org/10.1109/CCGrid.2014.37.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; D’Oliveira, R. G. L.; Rouayheb, S. E.; Evans, D.; Gardner, J.; Garrett, Z.; Gascón, A.; Ghazi, B.; Gibbons, P. B.; Gruteser, M.; Harchaoui, Z.; He, C.; He, L.; Huo, Z.; Hutchinson, B.; Hu, S.; Jagg, M.; Javidi, T.; Joshi, G.; Khodak, M.; Konečný, J.; Korolova, A.; Koushanfar, F.; Koyejo, S.; Lepoint, T.; Liu, Y.; Mittal, P.; Mohri, M.; Nock, R.; Özgür, A.; Pagh, R.; Raykova, M.; Qi, H.; Ramage, D.; Raskar, R.; Song, D.; Song, W.; Stich, S. U.; Sun, Z.; Suresh, A. T.; Tranèr, F.; Vepakomma, P.; Wang, J.; Xiong, L.; Xu, Z.; Yang, Q.; Yu, F. X.; Yu, H.; and Zhao, S. 2019. Advances and Open Problems in Federated Learning.

Li, T.; Sanjabi, M.; Beirami, A.; and Smith, V. 2019. Fair Resource Allocation in Federated Learning.

Mansour, Y.; Mohri, M.; Ro, J.; and Suresh, A. T. 2020. Three Approaches for Personalization with Applications to Federated Learning.

Martinez, N.; Bertran, M.; and Sapiro, G. 2020. Minimax Pareto Fairness: A Multi Objective Perspective. In Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research. PMLR. URL https://proceedings.icml.cc/static/paper_files/icml/2020/1084-Paper.pdf.

Morris, C. N. 1986. Empirical Bayes: a frequency-Bayes compromise, volume Volume 8 of Lecture Notes–Monograph Series. 195–203. Hayward, CA: Institute of Mathematical Statistics. doi:10.1214/lms/11215540299. URL https://doi.org/10.1214/lms/11215540299.

Paquay, P. 2018. Learning-from-data-Solutions. URL https://github.com/ppaquay/Learning-from-Data-Solutions.

Paquet, F.; Muller, K.-R.; and Samek, W. 2020. Clustered Federated Learning: Model-Agnostic Distributed Multi-task Optimization Under Privacy Constraints. IEEE Transactions on Neural Networks and Learning Systems 1–13. ISSN 2162-2388. doi:10.1109/tnnls.2020.3015958. URL http://dx.doi.org/10.1109/TNNLS.2020.3015958.

Sellentin, E.; and Heavens, A. F. 2015. Parameter inference with estimated covariance matrices. Monthly Notices of the Royal Astronomical Society: Letters 456(1): L132–L136. ISSN 1745-3925. doi:10.1093/mnrasl/slv190. URL https://doi.org/10.1093/mnrasl/slv190.

Shlezinger, N.; Rini, S.; and Eldar, Y. C. 2020. The Communication-Aware Clustered Federated Learning Problem. In 2020 IEEE International Symposium on Information Theory (ISIT), 2610–2615.

Wang, H.; Hsu, H.; Diaz, M.; and Calmon, F. P. 2020. To Split or Not to Split: The Impact of Disparate Treatment in Classification.

Yu, T.; Bagdasaryan, E.; and Shmatikov, V. 2020. Salvaging Federated Learning by Local Adaptation.

Li, T.; Sanjabi, M.; Beirami, A.; and Smith, V. 2019. Fair Resource Allocation in Federated Learning.
A Relationship to other approaches

This section contains a high-level summary of similar approaches and how they relate to ours. Throughout we assume the goal is to estimate some unknown \( \theta_i \) given samples drawn \( Y_i \sim D(\theta_i) \).

A frequentist approach would take \( \theta_i \) to be a constant that would be estimated by the average of the given samples \( \frac{1}{n} \sum_{i=1}^{n} Y_i \).

A hierarchical Bayesian estimator assumes data is generated in the following way: data is drawn \( Y_i \sim D(Y|\theta_i) \). The parameter \( \theta_i \) is drawn \( \sim \Theta_i(\theta_i|\lambda_i) \), where hyperparameter \( \lambda_i \) is drawn from known distribution \( p(\lambda) \). Given some data, the parameter \( \theta_i \) can be estimated as follows

\[
p(\theta_i|Y_i) = \frac{p(Y_i|\theta_i)p(\theta_i)}{p(Y_i)} = \frac{p(Y_i|\theta_i)p(\lambda_i)p(\lambda_i)d\lambda}{\int p(\theta_i|\lambda_i)p(\lambda_i)d\lambda}.
\]

Parametric empirical Bayes (Morris 1986) is frequently described as an intermediate between these two viewpoints. Similar to the hierarchical Bayesian viewpoint, it assumes data is drawn \( Y_i \sim D(Y|\theta_i) \), with parameter \( \theta_i \) is drawn \( \sim \Theta_i(\theta_i|\lambda_i) \). However, it differs in that it estimates \( \lambda_i \) based on the data, producing \( \hat{\lambda}_i \). This estimate of the hyperparameter is used, along with the data, to estimate \( \theta_i \).

A related example is the James-Stein estimator (Efron and Morris 1977). The estimator assumes the following process: each of \( m \) players draws a single sample from a normal distribution with variance \( s^2 \),

\[
Y_i \sim \mathcal{N}(\theta_i, s^2)
\]

This is different from the empirical Bayes or Bayes case in that it is assumed that the means \( \theta_i \) are completely unrelated to each other. Nevertheless, it has been demonstrated that the James-Stein estimator:

\[
\hat{\theta}_{JS} = \left(1 - \frac{(m-2) \cdot s^2}{\|Y\|^2}\right) Y
\]

has lower expected MSE than simply using the drawn parameters \( Y_i \). In the case that the variance \( s^2 \) is not known perfectly, it can be estimated as \( \tilde{s}^2 \) using entire vector of data \( Y \).

Our method is similar at a high level to empirical Bayes: we assume each player draws data from a personal distribution governed by \( \theta_j \) and that the \( \theta_j \) terms are in turn drawn from some distribution \( \Theta \). However, one key difference is that all three methods discussed above assume knowledge of the distributions generating the data, or at least which family they are drawn from. For example, the James Stein estimator assumes a normal distribution: variants of it exist for different distributions, but not a version that works for all distributions. Similarly, a hierarchical Bayes or empirical Bayes viewpoint would require knowledge of the \( D, \Theta, p \) distributions. In our approach, we do not assume that we know the form of these generating distributions, only some summary values (mean and variance) of the distribution.

It is entirely possible that other approaches, especially those that assume knowledge of the generating distribution, will out-perform our approach in terms of the error guarantees they can provide. Our distribution-free approach allows it to be implemented in a broader range of situations.

Additionally, our approach is restricted to linear combinations of estimators such as \( \hat{\theta}^f = w \cdot \hat{\theta}_j + (1 - w) \sum_{i=1}^{M} \hat{\theta}_i \). It is possible that a method outside this situation, for example, something like \( \hat{\theta}^f = x \cdot \hat{\theta}_j + y \sum_{i=1}^{M} \hat{\theta}_i \), or something like the non-linear James Stein estimator, would produce better estimates.

B Expected error proofs

Mean estimation

Assume that that player \( j \in [M] \) draws parameters i.i.d \( (\theta_j, \epsilon_j^2) \sim \Theta \). Then, that player draws \( n_j \) i.i.d samples \( Y \sim D_j(\theta_j, \epsilon_j^2) \), where \( \theta_j \) is the true mean of their personal distribution \( \Theta_j \) and \( \epsilon_j^2 \) is the true variance of that distribution. The player can either choose to use local estimation, which corresponds to:

\[
\hat{\theta}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}
\]
or federation, which corresponds to:

\[
\hat{\theta}^f = \frac{1}{N} \sum_{i=1}^{M} \hat{\theta}_i \cdot n_i
\]

where \( N = \sum_{i=1}^{M} n_i \). We’re interested in the expected mean squared error of each different estimator. Note that there are two separate expectations taken over two separate quantities of randomness. One is randomness in draws, \( E_{Y \sim D(\theta, \epsilon^2)} \), and one is randomness over the parameter drawing, \( E_{(\theta_j, \epsilon_j^2) \sim \Theta} \).

**Lemma 4.1.** For mean estimation, the expected MSE of local estimation for a player with \( n_j \) samples is \( \frac{\mu_j}{n_j} \).

**Proof.** First, we apply expectation over data draws:

\[
E_{Y \sim D(\theta, \epsilon^2)} \left[ (\theta_j - \hat{\theta}_j)^2 \right] = Var(\hat{\theta}_j)
\]

\[
= Var \left( \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij} \right) = \frac{1}{n_j} \cdot n_j \cdot \epsilon_j^2 = \frac{\epsilon_j^2}{n_j}
\]

where we have used variance properties to simplify the term.

Next, we take the expectation over parameter draws:

\[
E_{(\theta_j, \epsilon_j^2) \sim \Theta} \left[ \frac{\epsilon_j^2}{n_j} \right] = \frac{\mu_j}{n_j}
\]

□

**Lemma 4.2.** For mean estimation, the expected MSE of federated estimation for a player with \( n_j \) samples is

\[
\frac{\mu_e}{N} + \frac{\sum_{i \neq j} n_i^2 + (N - n_j)^2}{N^2} \cdot \sigma^2
\]

where \( N = \sum_{i=1}^{M} n_i \).
Proof. Here, the quantity we are interested in is:

$$(\hat{\theta}^l - \theta_j)^2 = (\hat{\theta}^l - \theta_j + \theta_j - \theta_j)^2$$

First, we apply the expectation over data draws for each term. First, we note that the last term is 0:

$$\mathbb{E}_{Y \sim \mathcal{D}(\theta_j, \epsilon_j^2)} \left[ 2(\hat{\theta}^l - \theta_j) \cdot (\hat{\theta}^l - \theta_j) \right] = 0$$

Next, we look at the first term:

$$\mathbb{E}_{Y \sim \mathcal{D}(\theta_j, \epsilon_j^2)} \left[ (\hat{\theta}^l - \theta_j)^2 \right] = \text{Var}(\hat{\theta}^l)$$

$$= \frac{1}{N} \sum_{i=1}^{M} \theta_i \cdot n_j = \frac{1}{N^2} \left( \sum_{i=1}^{M} \text{Var}(\hat{\theta}_i^l) \cdot n_i^2 \right)$$

where we have used \( \text{Var}(\hat{\theta}_i^l) = \frac{n_i}{N} \) in the last step. Next, we take the expectation over the parameter draws:

$$\mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} \left[ \frac{1}{N^2} \left( \sum_{i=1}^{M} \epsilon_i^2 \cdot n_i \right) \right]$$

$$= \sum_{i=1}^{M} n_i \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\epsilon_i^2] = \frac{\mu_\epsilon}{N}$$

Finally, we look at the middle term. First, we rewrite:

$$(\theta_l - \theta_j)^2 = \left( \frac{1}{N} \sum_{i \neq j} n_i \cdot \theta_i - \theta_j \right)^2$$

$$= \left( \frac{1}{N} \sum_{i \neq j} n_i \theta_i - \frac{N - n_j}{N} \theta_j \right)^2$$

$$= \frac{1}{N^2} \left( \sum_{i \neq j} n_i \cdot \theta_i \right)^2 + \frac{(N - n_j)^2}{N^2} \theta_j^2 - 2 \sum_{i \neq j} (N - n_j) n_i \theta_i \theta_j$$

We can rewrite the first term as:

$$\frac{1}{N^2} \left( \sum_{i \neq j} n_i^2 \cdot \theta_i^2 \right) + \frac{1}{N^2} \sum_{i \neq j, i \neq k} n_i \cdot n_k \cdot \theta_i \cdot \theta_k$$

Note that we only have one source of randomness in this case: that of parameter draws from the distribution, \( \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} \). Taking the expectation with respect to this produces two distinct terms: \( \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2] \) and \( \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2] \).

First, we collect all terms involving \( \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2] \) to get:

$$\frac{\mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2]}{N^2} \left( \sum_{i \neq j} n_i^2 + (N - n_j)^2 \right)$$

Collecting terms with a \( \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2] \) coefficient gives:

$$\frac{\mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2]}{N^2} \left( \sum_{i \neq j, i \neq k} n_i \cdot n_k - 2 \sum_{i \neq j} (N - n_j) \cdot n_i \right)$$

Note that

$$(N - n_j)^2 = \sum_{i \neq j} n_i^2 + \sum_{i, k \neq j, k \neq i} n_i \cdot n_k$$

Substituting and rearranging the term inside the parentheses gives us:

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

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

So the overall coefficient becomes:

$$-\frac{\mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2]}{N^2} \left( \sum_{i \neq j} n_i^2 + (N - n_j)^2 \right)$$

We can combine the \( \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2] \) and \( \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2] \) terms because they share a common coefficient, giving us:

$$\sum_{i \neq j} n_i^2 + (N - n_j)^2 \left( \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2] - \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2] \right)$$

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

where we have used \( \sigma^2 = \text{Var}(\theta_j) = \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2] - \mathbb{E}_{(\theta_i, \epsilon_i^2) \sim \Theta} [\theta_i^2] \). The overall federated error becomes:

$$\frac{\mu_\epsilon}{N} + \sum_{i \neq j} n_i^2 + (N - n_j)^2 \sigma^2$$

\(\square\)

Linear regression

In this setting, we assume that each player \( j \in [M] \) draws parameters \( (\theta_j, \epsilon_j^2) \sim \Theta \), where \( \theta_j \) is a length \( D \) vector and \( \epsilon_j^2 \) is a scalar-valued variance parameter as before. The \( d \)th entry in the vector is \( \theta_j^2 \) and \( \text{Var}(\theta_j^2) = \sigma_j^2 \). We assume that each value \( \theta_j \) is drawn independently of the others. The main result of this section will assume that each dimension is drawn independently, for example that \( \theta_j^2 \) is independent of \( \theta_k^2 \), for \( k \neq l \), but will we will demonstrate how this can be relaxed. Each player draws \( n_j \) input data points from their own input distribution, \( X_j \sim \mathcal{X}_j \) such that \( \mathbb{E}_{x \sim \mathcal{X}_j} [x^2 | x] = \Sigma_j \); we assume that the input dimensions are continuous and may be correlated but not deterministic.
functions of each other. They then noisily observes the outputs, drawing $Y_j \sim D_j(X_j^T \theta_j, \epsilon_j^2)$. We use $\eta_j$ to denote the length $D$ vector of errors so that $Y_j = X_j^T \theta_j + \eta_j$. Each player uses ordinary least squares (OLS) to compute estimates of their parameters: We assume that $X_j^T X$ is invertible. This happens when the columns of $X$ are linearly independent: when $X$ is continuous and the columns aren’t deterministic functions of each other, this happens with probability 1.

$$\hat{\theta}_j = (X_j^T X_j)^{-1}Y_j = (X_j^T X_j)^{-1}(X_j \theta_j + \eta_j)$$

Using the above value results in local estimation, where federation would correspond to:

$$\hat{\theta}^j = \frac{1}{N} \sum_{i=1}^{M} \hat{\theta}_i \cdot n_i$$

The expected error that a set of estimates $\hat{\theta}$ is determined by the expectation of the following quantity.

$$(x^T \hat{\theta} - x^T \theta_j)^2$$

Here, the expectation is taken over four sources of randomness.

1. $E_{x \sim X_j}$: Drawing a new test point $x$ from the data distribution $X_j$.
2. $E_{Y_j \sim D_j(X_j^T \theta_j, \epsilon_j^2)}$: Drawing labels for the dataset $X_i$ from the distribution $D_j(X_j^T \theta_j, \epsilon_j^2)$.
3. $E_{X_i \sim X_j}$: Drawing the dataset $X_i$ from data distribution $X_j$.
4. $E_{(\theta_j, \epsilon_j^2) \sim \Theta}$: Drawing parameters $\theta_j, \epsilon_j^2$ for player $j$’s distribution.

$(x^T \hat{\theta} - x^T \theta_j)^2$ measures the expected error of a set of parameters at a particular point $x$: when the expectation is taken over all $x \sim X_j$, it represents the average error everywhere on the distribution. It might be not immediately clear, though, why $(x^T \hat{\theta} - x^T \theta_j)^2$ is the correct term to be considering. Other potential candidates might include:

1. $\|\hat{\theta}_j - \theta_j\|^2$
2. $(x^T \hat{\theta}_j - y)^2$ for $y = x^T \theta_j + \eta_j$.

The first candidate measures the difference in estimated parameters; however, we assume that the objective of learning is to have low error on predicting future points, rather than solely estimate the parameters. The second candidate represents the error of predicting an instance as opposed to a mean value: it ends up simply producing an additive increase in our overall error term. To see this, note that we can write

$$(x^T \hat{\theta}_j - y)^2 = (x^T \hat{\theta}_j - x^T \theta_j + x^T \theta_j - y)^2$$

$$= (x^T \hat{\theta}_j - x^T \theta_j + x^T \theta_j - y)^2$$

$$= (x^T \hat{\theta}_j - x^T \theta_j + x^T \theta_j - y)^2$$

The first term is the same error function we are considering. The middle term is 0 in expectation and the last term is $\mu_\epsilon$ in expectation, so this approach simply scales the error we were looking at by $\mu_\epsilon$.

**Lemma 4.3.** For linear regression, the expected MSE of local estimation for a player with $n_j$ samples is

$$\mu_\epsilon \cdot tr \left[ \sum_{X_j \sim X_j} \left[ (X_j^T X_j)^{-1} \right] \right]$$

If the distribution of input values $X_j$ is a $D$-dimensional multivariate normal distribution with 0 mean, then, the expected MSE of local estimation can be simplified to:

$$\frac{\mu_\epsilon}{n_j - D - 2} D$$

Note: portions of Abu-Mostafa, Lin, and Magdon-Ismail (2012) and Paquay (2018), especially problem 3.11, were helpful in formulating this approach. Sellentin and Heavens (2015) and Anderson (1962) were helpful in providing the connection to the Inverse Wishart.

**Proof.** First, note that:

$$x^T \theta_j - x^T \theta_j = x^T (\theta_j - (X_j^T X_j)^{-1}X_j^T Y_j)$$

$$= x^T (\theta_j - (X_j^T X_j)^{-1}X_j^T (X_j \theta_j + \eta_j))$$

$$= x^T (\theta_j - \theta_j - (X_j^T X_j)^{-1}X_j^T \eta_j)$$

$$= -x^T (X_j^T X_j)^{-1}X_j^T \eta_j$$

Then,

$$(x^T \theta_j - x^T \theta_j)^2 = \eta_j^T X_j (X_j^T X_j)^{-1} x x^T (X_j^T X_j)^{-1} X_j^T \eta_j$$

To simplify this, we note that the above quantity is a scalar. For a scalar, $a = tr(a)$, and for any matrix, $tr(AB) = tr(BA)$ through the cyclic property of the scalar.

$$= tr [\eta_j^T X_j (X_j^T X_j)^{-1} x x^T (X_j^T X_j)^{-1} X_j^T \eta_j]$$

$$= tr [xx^T (X_j^T X_j)^{-1} \eta_j \eta_j^T X_j (X_j^T X_j)^{-1}]$$

To evaluate this, we start by applying the various expectations, noting that expectation and trace commute. Applying $E_{\eta_j \sim D_j(0, \epsilon_j^2)}$ to the term above allows us to rewrite it as:

$$= tr [xx^T (X_j^T X_j)^{-1} X_j^T V X_j (X_j^T X_j)^{-1}]$$

where

$$V = E_{\eta_j \sim D_j(0, \epsilon_j^2)} [\eta_j \eta_j^T]$$

$\eta_j \eta_j^T$ is an $n_j \times n_j$ matrix. The $i$th diagonal is $(\eta_j^i)^2$, which has expectation $c_j^2$. Off diagonal entries have value $\eta_j^i \cdot \eta_j^k$ for $\ell \neq k$. Because the errors for each data point are drawn independently and with 0 mean, the expectation of this is 0. $E_{\eta_j \sim D_j(0, \epsilon_j^2)} [\eta_j \eta_j^T]$ is a diagonal matrix with $c_j^2$ along the diagonal: we can pull it out of the trace to obtain:

$$= c_j^2 tr [xx^T (X_j^T X_j)^{-1} X_j^T X_j (X_j^T X_j)^{-1}]$$

Taking the expectation over the drawn parameters gives:

$$= E_{\theta_j \sim \Theta} [c_j^2 tr [xx^T (X_j^T X_j)^{-1}]]$$

$$= \mu_\epsilon tr [xx^T (X_j^T X_j)^{-1}]$$
Taking the expectation over \( x \sim X_j \) gives:
\[
\mu_e \text{tr} \left[ \mathbb{E}_{x \sim X_j} [xx^T] (X_j^T X_j)^{-1} \right] = \mu_e \text{tr} \left[ \Sigma_j (X_j^T X_j)^{-1} \right]
\]
Finally, we take the expectation with respect to \( X_j \sim X_j \)
\[
\mu_e \text{tr} \left[ \Sigma_j E_{X_j \sim X_j} \left[ (X_j^T X_j)^{-1} \right] \right]
\]
Note that because the inverse and expectation do not commute, in general, we cannot simplify this without stronger assumptions.

There is one other situation where a particular case of linear regression gives us simpler results. As mentioned in the statement of the lemma, in this case we assume that the distribution of input values \( X_j \) is a 0-mean normal distribution with covariance matrix \( \Sigma_j \).

Note that, in general, \( \Sigma_j \neq E_{x \sim X_j} [xx^T] \). \( E_{x \sim X_j} [xx^T] \) has, along the diagonals, \( E_{x \sim X_j} [x_i x_j] \), and on the off-diagonals, \( E_{x \sim X_j} [x_i x_k] \). By contrast, the covariance matrix has the same term along the diagonals, but the off-diagonal term has \( E_{x \sim X_j} [x_i x_j] - E_{x \sim X_j} [x_i x_k] \). In the case we are looking at, the distribution is 0 mean, so the off-diagonal terms drop away, and \( Cov_{ij} = \Sigma_j \).

If this is the case, then \( (X_j^T X_j)^{-1} \) is distributed according to an Inverse Wishart distribution with parameters \( n_j - 1 \), \( D \), and covariance \( Cov_{ij} = \Sigma_j \). An unbiased estimator of \( Cov_{ij} \) is found by using \( \frac{1}{n_j} X_j^T X_j \), and the property of the inverse Wishart tells us that:

\[
\mathbb{E}_{X_j \sim X_j} \left[ \frac{1}{n_j - 1} X_j^T X_j \right] = \frac{n_j - 1}{n_j - D - 2} Cov_{ij}^{-1}
\]

Using these results, we can directly calculate the desired expectation:

\[
\mu_e \text{tr} \left[ \Sigma_j E_{X_j \sim X_j} \left[ (X_j^T X_j)^{-1} \right] \right] = \frac{\mu_e}{n_j - 1} \text{tr} \left[ \Sigma_j E_{X_j \sim X_j} \left[ \left( \frac{1}{n_j - 1} X_j^T X_j \right)^{-1} \right] \right] = \frac{\mu_e}{n_j - 1} \text{tr} \left[ \Sigma_j \frac{n_j - 1}{n_j - D - 2} Cov_{ij}^{-1} \right] = \frac{\mu_e}{n_j - 1} \text{tr} \left[ \Sigma_j \frac{n_j - 1}{n_j - D - 2} \Sigma_j^{-1} \right] = \frac{\mu_e}{n_j - 1} \text{tr} \left[ ID \right] = \frac{\mu_e}{n_j - D - 2} D \]

\[
\Sigma_j = \mathbb{E}_{X_j \sim X_j} [X_j^T X_j] + \text{Var}(X_j) = 1 + 0 = 1. \text{ Similarly, } X_j^T X_j = n_j \text{ deterministically, so the error term reduces to } \frac{\mu_e}{n_j} \text{ as before.}
\]

Note that, as expected, this does not simplify down to the mean estimation case for \( D = 1 \): that case would model a version of 1-dimensional linear regression, where it is necessary to estimate both \( \theta \) as well as \( \hat{x} \), the mean of the input distribution.

**Lemma 4.4.** For linear regression, the expected MSE of federated estimation for a player with \( n_j \) samples is:
\[
L_j = \frac{\mu_e + \sum_{i \neq j} n_i^2}{N^2} + \frac{(N - n_j)^2}{N^2} \sum_{d=1}^{D} E_{x \sim X_j} \left[ (x^d)^2 \right] \cdot \sigma_d^2
\]
where \( L_j \) is equal to:
\[
\mu_e \sum_{i = 1}^{M} \frac{n_i^2}{N^2} D
\]
\[
\frac{D}{N^2} n_i - D - 2
\]

**Proof.** Here, we will use \( \mathbb{E}_{Y \sim D(\theta, \epsilon^2)} \) to mean the expectation taken over all \( i \in [M] \) player’s data, given that all of the data influences the federated learning result.

\[
(x^T \theta_j - x^T \hat{\theta}^f)^2 = (x^T \theta_j - x^T \hat{\theta}^f + x^T \hat{\theta}^f - x^T \hat{\theta}^f)^2 = (x^T \theta_j - x^T \hat{\theta}^f)^2 + 2(x^T \theta_j - x^T \hat{\theta}^f) \cdot (x^T \hat{\theta}^f - x^T \hat{\theta}^f)
\]

Note that the expectation of the last term results in 0 because \( \mathbb{E}_{Y \sim D(\theta, \epsilon^2)} \left[ x^T \hat{\theta}^f - x^T \hat{\theta}^f \right] = 0 \). Next, we investigate the second equation in the sum.

\[
(x^T \theta_j - x^T \hat{\theta}^f)^2 = (x^T \sum_{i = 1}^{M} \frac{n_i}{N} \theta_i - x^T \sum_{i = 1}^{M} \frac{n_i}{N} \hat{\theta}_i)^2 = \left( \sum_{i = 1}^{M} \frac{n_i}{N} x^T (\theta_i - \hat{\theta}_i) \right)^2
\]

Expanding out the squared term gives us:

\[
\sum_{i = 1}^{M} \left( \frac{n_i}{N} x^T (\theta_i - \hat{\theta}_i) \right)^2 + \sum_{i = 1}^{M} \sum_{k \neq i} \left( \frac{n_i}{N} x^T (\theta_i - \hat{\theta}_i) \cdot \frac{n_k}{N} x^T (\theta_k - \hat{\theta}_k) \right)
\]

The second term ends up being relevant: because each set of parameters \( \theta_i \sim \Theta \) are drawn independently and because
each data set $X_i \sim \mathcal{X}_i$ are drawn independently, the $\theta_i - \hat{\theta}_i$ terms are independent of each other. Because each is 0 in expectation, the entire product has expectation 0. Applying the expectation to the first term and rewriting gives:

$$
\sum_{i=1}^{M} \frac{n_i}{N^2} (x^T \theta_i - x^T \hat{\theta}_i)^2
$$

The term inside the sum is exactly equivalent to the value we solved with the local estimation case: we can rewrite this as

$$
\mu e \sum_{i=1}^{M} \frac{n_i}{N^2} \text{tr} \left[ \Sigma_j E_{Y \sim D(\theta_i, \epsilon_l)} \left[ (X_i^T X_i)^{-1} \right] \right]
$$
or, if the necessary conditions are satisfied,

$$
\mu e \sum_{i=1}^{M} \frac{n_i}{N^2} D - D - 2
$$

Finally, we will explore the first term in the sum:

$$
(x^T \theta_j - x^T \theta^f_j)^2
$$

$$
= (x^T (\theta_j - \theta^f_j)^T x^T (\theta_j - \theta^f_j))
$$

Taking the expectation and using the cyclic property of the trace gives:

$$
= \text{tr} \left[ (\theta_j - \theta^f_j)^T x x^T (\theta_j - \theta^f_j) \right]
$$

Next, we focus on simplifying the term involving the $\theta$ values. Using the definition of $\theta^f_j$ gives:

$$
\left( \frac{N - n_j}{N} \theta_j - \sum_{i \neq j} \frac{n_i}{N \theta_i} \right) \left( \frac{N - n_j}{N} \theta_j - \sum_{i \neq j} \frac{n_i}{N \theta_i} \right)^T
$$

Expanding and taking the expectation gives us three terms. The first:

$$
\left( \frac{N - n_j}{N} \right)^2 \mathbb{E}_{\theta_j \sim \Theta} \left[ \theta_j \theta^T_j \right]
$$

$$
= \frac{1}{N^2} \left( \sum_{i \neq j} n_i^2 + \sum_{i, k \neq j, k \neq i} n_i \cdot n_k \right) \mathbb{E}_{\theta_j \sim \Theta} \left[ \theta^T_j \right]
$$

The second:

$$
-2 \frac{N - n_j}{N} \sum_{i \neq j} \mathbb{E}_{\theta_i, \theta_j \sim \Theta} \left[ \theta_j \theta^T_i \right]
$$

$$
= -2 \left( \frac{N - n_j}{N} \right)^2 \mathbb{E}_{\theta_i, \theta_j \sim \Theta} \left[ \theta_j \theta^T_i \right]
$$

$$
= -2 \frac{1}{N^2} \left( \sum_{i \neq j} n_i^2 + \sum_{i, k \neq j, k \neq i} n_i \cdot n_k \right) \mathbb{E}_{\theta_i, \theta_j \sim \Theta} \left[ \theta_j \theta^T_i \right]
$$

And the third:

$$
\mathbb{E}_{\theta_i \sim \Theta} \left[ \left( \sum_{i \neq j} \frac{n_i}{N} \theta_i \right) \left( \sum_{i \neq j} \frac{n_i}{N} \theta_i \right)^T \right]
$$

$$
= \sum_{i \neq j} \left( \frac{n_i}{N} \right)^2 \mathbb{E}_{\theta_i \sim \Theta} \left[ \theta_j \theta^T_j \right]
$$

$$
+ \sum_{i, k \neq j, k \neq i} \frac{n_i \cdot n_k}{N^2} \mathbb{E}_{\theta_i, \theta_j, \theta_k \sim \Theta} \left[ \theta_k \theta^T_i \right]
$$

If we combine all three terms and collect coefficients, we get:

$$
\frac{1}{N^2} \sum_{i \neq j} n_i^2 \left( 2 \mathbb{E}_{\theta_i \sim \Theta} \left[ \theta_j \theta^T_j \right] - 2 \mathbb{E}_{\theta_i, \theta_j \sim \Theta} \left[ \theta_j \theta^T_i \right] \right)
$$

$$
+ \frac{1}{N^2} \sum_{i, k \neq j, k \neq i} n_i \cdot n_k \left( \mathbb{E}_{\theta_i, \theta_j \sim \Theta} \left[ \theta_j \theta^T_j \right] - \mathbb{E}_{\theta_i, \theta_j, \theta_k \sim \Theta} \left[ \theta_j \theta^T_i \right] \right)
$$

$$
= \left( \mathbb{E}_{\theta_i \sim \Theta} \left[ \theta_j \theta^T_j \right] - \mathbb{E}_{\theta_i, \theta_j \sim \Theta} \left[ \theta_j \theta^T_i \right] \right)
$$

$$
\cdot \frac{1}{N^2} \left( 2 \sum_{i \neq j} n_i^2 + \frac{1}{N^2} \sum_{i, k \neq j, k \neq i} n_i \cdot n_k \right)
$$

$$
= \left( \mathbb{E}_{\theta_i \sim \Theta} \left[ \theta_j \theta^T_j \right] - \mathbb{E}_{\theta_i, \theta_j \sim \Theta} \left[ \theta_j \theta^T_i \right] \right)
$$

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

We can recombine this term with the component involving the trace to rewrite it as:

$$
\text{tr} \left[ \Sigma_j \left( \mathbb{E}_{\theta_j \sim \Theta} \left[ \theta_j \theta^T_j \right] - \mathbb{E}_{\theta_i, \theta_j \sim \Theta} \left[ \theta_j \theta^T_i \right] \right) \right]
$$

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

Next, we need to reason about the difference in the expected terms. In this setting, we are assuming that each coefficient is drawn separately from the other coefficients. $\mathbb{E}_{\theta_i \sim \Theta} \left[ \theta_i \theta^T_i \right]$ has, on the $d$th element of the diagonal, $\mathbb{E}_{\theta_i \sim \Theta} \left[ \theta^2_i \right]$ and on the off-diagonal terms in the $l$, $k$th entry, $\mathbb{E}_{\theta_i \sim \Theta} \left[ \theta_l \theta_k^T \right]$. Here, we are assuming that $\theta_l$ and $\theta_k$ are independent, so this equals $\mathbb{E}_{\theta_l \sim \Theta} \left[ \theta_l \right] \cdot \mathbb{E}_{\theta_k \sim \Theta} \left[ \theta_k \right]$. We relax this assumption below. $\mathbb{E}_{\theta_i, \theta_j \sim \Theta} \left[ \theta_i \theta^T_j \right]$ has, on the $d$th element of the diagonal, $\mathbb{E}_{\theta_i \sim \Theta} \left[ \theta^2_i \right]$, and on the $l$, $k$th off-diagonal term has the same value as the other matrix: $\mathbb{E}_{\theta_i \sim \Theta} \left[ \theta_l \theta_k^T \right]$ and $\mathbb{E}_{\theta_i \sim \Theta} \left[ \theta_l \theta_k \right]$. The difference between these two matrices is a diagonal matrix with $\mathbb{E}_{\theta_i \sim \Theta} \left[ \theta^2_i \right]$ and $\mathbb{E}_{\theta_i \sim \Theta} \left[ \theta_l \theta_k \right]$. The difference between these two matrices is a diagonal matrix with $\mathbb{E}_{\theta_i \sim \Theta} \left[ \theta^2_i \right] - \mathbb{E}_{\theta_i \sim \Theta} \left[ \theta^2_i \right] = \sigma^2_d$ on the diagonal, where $\sigma^2_d$ represents the variance of the $d$th coefficient. That turns the term involving the trace into a simple sum:

$$
\sum_{i \neq j} \frac{n_i^2}{N^2} \cdot (N - n_j)^2 \sum_{d=1}^{D} \frac{1}{N^2} \sum_{i \neq j} \left( x^d \right)^2 \cdot \sigma^2_d
$$
In the proof above, we assumed that the draw of parameter value $\theta^i_l$ is independent of $\theta^j_k$, for $l \neq k$. A case where this might not occur is when these values are correlated: say, the value drawn for $\theta^i_l$ is anti-correlated with the parameter drawn for $\theta^j_k$. (Note that we still assume draws are independent across players: $\theta^i_l$ is independent of $\theta^i_j$ and $\theta^j_k$.) Relaxing this assumption is not hard and would change the results in the following way: the off-diagonal terms of the difference would no longer be 0. Instead, the off-diagonal $l, k$th entry becomes

$$E_{\theta^i_l \sim e} [\theta^i_l \cdot \theta^j_k] - E_{\theta^i_j \sim e} [\theta^i_j \cdot \theta^j_k]$$

Performing the matrix multiplication with $\Sigma_j$ turns this into:

$$\sum_{d=1}^{D} E_{x \sim X_j} [(x^d)^2] \cdot \sigma_d^2 + \sum_{i \neq d} E_{x \sim X_j} [x^d \cdot x^i]$$

$$\cdot (E_{\theta^i_l \sim e} [\theta^i_l \cdot \theta^j_k] - E_{\theta^i_j \sim e} [\theta^i_j \cdot \theta^j_k])$$

Our final value for this component of the error would be the same form, but with a slightly different coefficient. □

C Supporting lemmas for strict federation case

Lemma 6.1. If $n_i > \frac{\mu_e}{\sigma^2}$ for all $i \in [M]$, then $\pi_i$ is the unique element in the core.

Proof. We will show this by showing that every player minimizes their error by being alone in $\pi_i$. We will use $N_Q$ to be the sum of elements within a coalition $Q$. $Q$ could be the coalition equal to all players ($\pi_\sigma$) or some strict subset, but we will assume it contains at least 2 elements. We will show that every player gets higher error in $Q$ than it would get alone. We wish to show:

$$\frac{\mu_e}{N_Q} + \sigma^2 \sum_{i \neq j, i \in Q} n_i^2 + (N_Q - n_j)^2 > \frac{\mu_e}{n_j}$$

Cross multiplying gives:

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

Rewriting:

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

The righthand side can be rewritten as:

$$\mu_e \cdot N_Q^2 - \mu_e \cdot N_Q \cdot n_j = \mu_e \cdot (N_Q - n_j)^2 + \mu_e \cdot n_j \cdot (N_Q - n_j)$$

Then, we can prove the inequality by splitting it up into two terms. The first:

$$\sigma^2 \cdot n_j \cdot (N_Q - n_j)^2 > \mu_e \cdot (N_Q - n_j)^2$$

which is true because $n_j \cdot \sigma^2 > \mu_e$. The second:

$$\sigma^2 \cdot n_j \cdot \sum_{i \neq j, i \in Q} n_i^2 > \mu_e \cdot n_j \cdot (N_Q - n_j)$$

which is satisfied because, for each player,

$$\sigma^2 \cdot n_i^2 > \mu_e \cdot n_i$$

because $\sigma^2 \cdot n_i > \mu_e$.

□

Lemma 6.2. If $n_i \geq \frac{\mu_e}{\sigma^2}$ for all $i \in [M]$, with $n_k > \frac{\mu_e}{\sigma^2}$ for at least one player $k$, then any arrangement where the players with samples $n_k > \frac{\mu_e}{\sigma^2}$ are alone in the core.

Proof. To prove this, we can note that, in the proof above, any coalition $Q$ with at least one player $n_i > \frac{\mu_e}{\sigma^2}$ would satisfy the desired inequality: all players participating would get higher error than they could alone. This shows that any coalition involving a player with more samples than $\frac{\mu_e}{\sigma^2}$ is infeasible. We have previously shown that all players with $n_i = \frac{\mu_e}{\sigma^2}$ get equal error no matter their arrangement. □

Lemma 6.3. If $n_s \leq \frac{\mu_e}{\sigma^2}$ and $n_s < n_\ell$, a small player in a coalition with $s$ small players and $\ell$ large players always prefers $s$ as large as possible:

$$s_2 > s_1 \Rightarrow \pi(s_2, \ell) > \pi(s_1, \ell)$$

Proof. To prove this, we will show that the derivative of the small player’s error with respect to $s$ is always negative. The error is:

$$\frac{\mu_e}{s \cdot n_s + \ell \cdot n_\ell} + \sigma^2 \left[ (s-1) \cdot n_s^2 + \ell \cdot n_\ell^2 + (s-1) \cdot n_s + \ell \cdot n_\ell \right]^2$$

The derivative with respect to $s$ is:

$$\frac{n_s \cdot (s \cdot n_s \cdot (n_s \cdot \sigma^2 - \mu_e))}{(s \cdot n_s + \ell \cdot n_\ell)^3} - \frac{n_s \cdot (\ell \cdot n_\ell \cdot (\mu_e + 2n_\ell \cdot \sigma^2 - 3n_s \cdot \sigma^2))}{(s \cdot n_s + \ell \cdot n_\ell)^3}$$

Showing that the derivative is negative is equivalent to showing that the term below is negative:

$$s \cdot n_s \cdot (n_s \cdot \sigma^2 - \mu_e) - \ell \cdot n_\ell \cdot (\mu_e + 2n_\ell \cdot \sigma^2 - 3n_s \cdot \sigma^2)$$

We can break this term into multiple components:

$$s \cdot n_s \cdot (n_s \cdot \sigma^2 - \mu_e) \leq 0$$

because $n_s \leq \frac{\mu_e}{\sigma^2}$. We can rewrite a second term as:

$$\mu_e - n_s \cdot \sigma^2 + 2\sigma^2(n_\ell - n_s)$$

We know that $\mu_e - n_s \cdot \sigma^2 \geq 0$, and because $n_\ell > n_s$,

$$2n_\ell \cdot \sigma^2 - 2n_s \cdot \sigma^2 > 0$$

These facts, taken together, show that the derivative is always negative. □

Lemma 6.4. If $n_\ell \geq \frac{\mu_e}{\sigma^2}$ and $n_s < n_\ell$, a large player in a coalition with $s$ small players and $\ell$ large players always prefers $\ell$ as small as possible:

$$\ell_2 < \ell_1 \Rightarrow \pi(s, \ell_2) > \pi(s, \ell_1)$$

If $n_\ell < \frac{\mu_e}{\sigma^2}$, then large players do not necessarily prefer $\ell$ to be either as large as possible or as small as possible.
Proof. To prove this, we will show that the derivative of the large player’s error with respect to $\ell$ is always positive when $n_\ell \geq \frac{\mu_s}{2\sigma^2}$. The error is:

$$\frac{\mu_e}{s \cdot n_s + \ell \cdot n_\ell} + \sigma^2 \cdot \left( (s-1) \cdot n_s^2 +\ell \cdot n_\ell^2 + ((s-1) \cdot n_s + \ell \cdot n_\ell) \cdot (s \cdot n_s + \ell \cdot n_\ell) \right)^2$$

The derivative with respect to $\ell$ is:

$$n_\ell \cdot (\ell \cdot n_\ell \cdot (\mu_e - \sigma^2 \cdot n_s + \sigma^2 (n_\ell - n_s))) \cdot (\ell \cdot n_\ell + s \cdot n_s)^2$$

We wish to show that the numerator is positive. We can break it into multiple components:

$$n_\ell \cdot \sigma^2 - \mu_e \geq 0$$

because $n_\ell \geq \frac{\mu_s}{2\sigma^2}$. We can rewrite the second term as

$$\frac{\mu_e}{s \cdot n_s + \ell \cdot n_\ell} \cdot n_\ell \cdot (\mu_e - \sigma^2 \cdot n_s + \sigma^2 (n_\ell - n_s))$$

which is negative because $n_\ell \geq \frac{\mu_s}{2\sigma^2}$ and $n_\ell > n_s$.

Next, we consider the case where $n_\ell < \frac{\mu_s}{2\sigma^2}$. The first term is now negative and the second term is the sum of two terms: one is positive and one is negative, so the overall derivative could be either positive or negative. \(\square\)

Lemma 6.5. Assume $n_\ell \leq \frac{\mu_s}{\sigma^2}$ and $n_s < n_\ell$. If $n_\ell \leq \frac{\mu_s}{2\sigma^2}$, a small player in a coalition with $s$ small players and $\ell$ large players always prefers $\ell$ as large as possible. If $n_\ell > \frac{\mu_s}{2\sigma^2}$, as $\ell$ is increased, the small player’s error first increases and then decreases: the player achieves a maximum error, and then overall error decreases with $\ell$.

Proof. The small player’s error is

$$\frac{\mu_e}{s \cdot n_s + \ell \cdot n_\ell} + \sigma^2 \cdot \left( (s-1) \cdot n_s^2 +\ell \cdot n_\ell^2 + ((s-1) \cdot n_s + \ell \cdot n_\ell) \cdot (s \cdot n_s + \ell \cdot n_\ell) \right)^2$$

The derivative with respect to $\ell$ is:

$$n_\ell \cdot (\ell \cdot n_\ell \cdot (\mu_e - \sigma^2 \cdot n_s + \sigma^2 (n_\ell - n_s))) \cdot (\ell \cdot n_\ell + s \cdot n_s)^2$$

The derivative is negative when the term below is positive:

$$\ell \cdot n_\ell \cdot (\mu_e - \sigma^2 \cdot n_s + \sigma^2 (n_\ell - n_s)) + s \cdot n_s \cdot (\mu_e - \ell \cdot \sigma^2)$$

The first term (multiplying $\ell \cdot n_\ell$) is always positive. For $n_\ell \leq \frac{\mu_s}{2\sigma^2}$ the second term is also positive or zero, so the derivative is always negative.

If $n_\ell > \frac{\mu_s}{2\sigma^2}$, then the term starts out as positive and (as $\ell$ increases) becomes more smaller, eventually becoming negative and staying negative. This means that the derivative starts out as negative and (as $\ell$ increases) becomes larger and larger, eventually becoming and staying positive. \(\square\)

Lemma 6.6. Assume $n_\ell \leq \frac{\mu_s}{\sigma^2}$, $n_s \geq \frac{\mu_s}{\sigma^2}$, and $n_s < n_\ell$. As $s$ is increased, the large player’s error first decreases and then increases: the large player achieves a minimum level of error, and afterwards its error increases with $s$.

If $n_\ell < \frac{\mu_s}{\sigma^2}$, then large players do not necessarily prefer $s$ to be either as large as possible or as small as possible.

Proof. The large player’s error is:

$$\frac{\mu_e}{s \cdot n_s + \ell \cdot n_\ell} + \sigma^2 \cdot \left( (s-1) \cdot n_s^2 +\ell \cdot n_\ell^2 + ((s-1) \cdot n_s + \ell \cdot n_\ell) \cdot (s \cdot n_s + \ell \cdot n_\ell) \right)^2$$

The derivative with respect to $s$ is:

$$n_s \cdot (\ell \cdot n_\ell (\mu_e - n_s \cdot \sigma^2)) \cdot (\ell \cdot n_\ell + s \cdot n_s)^2$$

The derivative is negative when the term below is positive:

$$\ell \cdot n_\ell (\mu_e - n_s \cdot \sigma^2) + n_s \cdot s \cdot (\mu_e - \ell \cdot \sigma^2 + \sigma^2 (n_s - n_\ell))$$

The first term is always positive or zero. The second term is always negative. As $s$ increases, the overall sum starts out as positive, and then becomes negative. This means that the derivative starts out as negative and then becomes positive: the large player achieves a minimum amount of error, and then overall error increases with $s$.

Next, we consider the case where $n_\ell < \frac{\mu_s}{\sigma^2}$. Again, the first term is positive or zero. The second term, though, is composed of a sum: one component is positive and one is negative, so it is not necessarily clear whether the overall sum is positive or negative. \(\square\)

Theorem 6.7. If $n_\ell \leq \frac{\mu_s}{2\sigma^2}$ and $n_s < n_\ell$, then the grand coalition $\pi_g$ is core-stable.

Proof. As a reminder, the small players always prefer $s$ as large as possible, and for $n_\ell \leq \frac{\mu_s}{2\sigma^2}$ they also prefer $\ell$ as large as possible, so $\pi(S, L) = \pi_g$, minimizes error for small players. For this reason, any defection coalition that has $\pi(s > 0, \ell)$ is infeasible because the small players would get higher error.

The only kind of defections we need to consider are in the form $\pi(0, \ell)$. We will consider $\pi(0, L)$ and show that the large players prefer $\pi(S, L)$ to $\pi(0, L)$: $\pi(S, L) > \pi(0, L)$. In the case that $n_\ell < \frac{\mu_s}{\sigma^2}$, $\pi(0, L) \succ \pi(0, L)$, so any other arrangement is also not a possible defection. In the case that $n_\ell = \frac{\mu_s}{\sigma^2}$, $\pi(0, L) = \pi(0, L)$, so similarly any other defection is not possible. What we’d like to show is:

$$\frac{\mu_e}{s \cdot n_s + \ell \cdot n_\ell} + \sigma^2 \cdot \left( (s-1) \cdot n_s^2 +\ell \cdot n_\ell^2 + ((s-1) \cdot n_s + \ell \cdot n_\ell) \cdot (s \cdot n_s + \ell \cdot n_\ell) \right)^2$$

$$< \frac{\mu_e}{\ell \cdot n_\ell} + \sigma^2 \cdot (s \cdot n_s + \ell \cdot n_\ell)^2$$
Cross multiplying turns the condition into:

\[
\mu_c \cdot (s \cdot n_s + \ell \cdot n_\ell) \cdot \ell^2 \cdot n_\ell^2 \\
+ \sigma^2 \cdot (s \cdot n_s^2 + (\ell - 1) \cdot n_\ell^2 + s \cdot n_s + (\ell - 1) \cdot n_\ell)^2 \cdot \ell^2 \cdot n_\ell^2 \\
< \mu_c \cdot \mu_c \cdot (s \cdot n_s + \ell \cdot n_\ell)^2 + \sigma^2 \cdot (s \cdot n_s + \ell \cdot n_\ell)^2 \\
\cdot (s \cdot n_s + \ell \cdot n_\ell)^2
\]

If we collect the \(\mu_c\) terms, we get:

\[
\mu_c \cdot \ell \cdot n_\ell \cdot (s \cdot n_s + \ell \cdot n_\ell) \cdot (s \cdot n_s + \ell \cdot n_\ell - \ell \cdot n_\ell) \\
= \mu_c \cdot \ell \cdot n_\ell \cdot (s \cdot n_s + \ell \cdot n_\ell) \cdot s \cdot n_s
\]

If we collect the \(\sigma^2\) terms, we get:

\[
\sigma^2 \cdot \ell^2 \cdot n_\ell^2 \cdot (s \cdot n_s + (\ell - 1) \cdot n_\ell^2 \\
+ (s \cdot n_s + (\ell - 1) \cdot n_\ell)^2 - (\ell - 1) \cdot (s \cdot n_s + \ell \cdot n_\ell^2)
\]

First, we expand the first squared term and combine it with another term:

\[
s \cdot n_s^2 + (\ell - 1) \cdot n_\ell^2 + s^2 \cdot n_s^2 + 2 \cdot s \cdot (\ell - 1) \cdot n_\ell + (\ell - 1)^2 \cdot n_\ell^2 \\
= n_s^2 \cdot s \cdot (s + 1) + 2 \cdot s \cdot (\ell - 1) \cdot n_\ell + n_\ell^2 \cdot (\ell - 1) \cdot \ell
\]

Multiplied by \(\ell\), it becomes:

\[
\ell \cdot (n_s^2 \cdot s \cdot (s + 1) + 2 \cdot s \cdot (\ell - 1) \cdot n_\ell + n_\ell^2 \cdot (\ell - 1) \cdot \ell)
\]

Expanding out the second squared term gives us:

\[
s^2 \cdot n_s^2 + 2 \cdot s \cdot \ell \cdot n_s \cdot n_\ell + \ell^2 \cdot n_\ell^2
\]

When we multiply this by \(- (\ell - 1)\), it becomes

\[-(\ell - 1) \cdot (s^2 \cdot n_s^2 + 2 \cdot s \cdot \ell \cdot n_s \cdot n_\ell + \ell^2 \cdot n_\ell^2)
\]

Next, we combine similar terms in both sums. First, we start with coefficients of \(n_s^2\):

\[
\ell \cdot n_s^2 \cdot s \cdot (s + 1) - (\ell - 1) \cdot n_s^2 \cdot s^2 \\
= n_s^2 \cdot s \cdot (s + 1) - (\ell - 1) \cdot s \\
= n_s^2 \cdot s \cdot (s + 1) - (\ell - 1) \cdot s + s \\
= n_s^2 \cdot s \cdot (\ell + s)
\]

Next, we do the next term, which involves coefficients of \(n_s \cdot n_\ell\):

\[
2 \cdot s \cdot (\ell - 1) \cdot n_s \cdot n_\ell - 2 \cdot (\ell - 1) \cdot s \cdot n_s \cdot n_\ell \\
= 0
\]

And the last one term, with coefficients of \(n_\ell^2\):

\[
\ell^2 \cdot (\ell - 1) \cdot n_\ell^2 - (\ell - 1) \cdot \ell^2 \cdot n_\ell^2 \\
= 0
\]

If we multiply the only nonzero term and multiply by the terms we pulled out, it becomes:

\[
n_s^2 \cdot s \cdot (\ell + s) \cdot \ell \cdot n_\ell^2 \cdot \sigma^2
\]

Next, we return this to our inequality. What we’re trying to show is:

\[
\ell \cdot n_\ell^2 \cdot n_s^2 \cdot (\ell + s) \cdot \sigma^2 < \mu_c \cdot \ell \cdot n_\ell \cdot (s \cdot n_s + \ell \cdot n_\ell) \cdot s \cdot n_s
\]

Cancelling some terms:

\[
n_\ell \cdot n_s \cdot (\ell + s) \cdot \sigma^2 < \mu_c \cdot (s \cdot n_s + \ell \cdot n_\ell)
\]

Expanding out terms:

\[
\sigma^2 \cdot (s \cdot n_s \cdot n_\ell + s \cdot n_\ell \cdot n_s) < \mu_c \cdot (s \cdot n_s + \ell \cdot n_\ell)
\]

We can prove this by splitting up piecewise:

\[
\sigma^2 \cdot n_s < \mu_c \cdot \ell \cdot n_\ell
\]

because \(\sigma^2 \cdot n_s < \mu_c\). Similarly,

\[
\sigma^2 \cdot n_\ell \cdot s \cdot n_s \leq \mu_c \cdot s \cdot n_s
\]

because \(\sigma^2 \cdot n_\ell \leq \mu_c\).
Secondly, we know that the error the small player experiences first increases and then decreases as \( \ell \) increases. Is it possible to pick an \( \ell_{2} < \ell_{2}' \) so that the small player gets lower error there than in \( \pi(s_{1}, \ell_{1}) \)?

Suppose that the derivative of \( err_{0}(s_{2}, \ell) \) with respect to \( \ell \) is positive at \( \ell = \ell_{2} \); then, reducing \( \ell \) from \( \ell_{2}' \) to \( \ell_{2} \) might reduce the small player’s error. However, for every point where the small player’s derivative is positive, \( err_{0}(s_{2}, \ell) > err_{0}(s_{2}, 0) > err_{0}(S, 0) \): the small player would not wish to move here because it would get strictly higher error than it would get in \( \pi(S, 0) \).

Suppose instead that the derivative of \( err_{0}(s_{2}, \ell) \) with respect to \( \ell \) is negative or zero at \( \ell = \ell_{2} \). Then, if \( \ell_{2} < \ell_{2}' \), reducing the number of large players in the coalition from \( \ell_{2}' \) to \( \ell_{2} \) would increase the error of the small players. This is also not an allocation that the small players would prefer.

Increasing \( \ell \), so that \( \ell_{2} > \ell_{2}' \) sufficiently large, would satisfy the small player, but we already showed that it would increase the error the large player experiences. As a result, it is not possible to pick an allocation that both the small and large players prefer to \( \pi(s_{1}, \ell_{1}) \).

\[ \square \]

**D Supporting lemmas for other models**

*(w-best, v-matrix)*

In this case, the federated estimate differs for each player and is

\[
\hat{\theta}_{j}^{w} = w \cdot \hat{\theta}_{j} + (1 - w) \cdot \frac{1}{N} \sum_{i=1}^{M} \hat{\theta}_{i} \cdot n_{i}
\]

for \( w \in [0, 1] \). Note that \( w \) can be \( w_{j} \): that is, each player can have a different \( w \) parameter. For readability we will use \( w_{j} \) in the lemma below.

**Lemma 7.1.** For \( w \)-weighting federated learning, the expected MSE of a player with \( n_{j} \) samples is:

\[
\mu_{e} \left( \frac{w^{2}}{n_{j}} + \frac{1 - w^{2}}{N} \right) + \sum_{i \neq j} n_{i}^{2} + \frac{(N - n_{j})^{2}}{N^{2}} \cdot (1 - w^{2}) \sigma^{2}
\]

**Proof.** We can use a similar process to the pure federated case.

\[
(\hat{\theta}_{j}^{w} - \theta_{j})^{2} = (\hat{\theta}_{j}^{w} - \theta_{j}^{w})^{2} + (\theta_{j}^{w} - \theta_{j})^{2} + 2 \cdot (\hat{\theta}_{j}^{w} - \theta_{j}^{w}) \cdot (\theta_{j}^{w} - \theta_{j})
\]

The expectation of the last term (taken with respect to the data draws) is 0. Next, we look at the first term:

\[
\mathbb{E}_{Y \sim D(\theta_{j}, \epsilon_{j})}[ (\hat{\theta}_{j}^{w} - \theta_{j}^{w})^{2} ] = Var(\hat{\theta}_{j}^{w})
\]

\[
= Var \left( w \cdot \hat{\theta}_{j} + (1 - w) \cdot \frac{1}{N} \sum_{i=1}^{M} \hat{\theta}_{i} \cdot n_{i} \right)
\]

\[
= w^{2}Var(\hat{\theta}_{j}) + \left( \frac{1 - w^{2}}{N^{2}} \right) Var \left( \frac{1}{N} \sum_{i=1}^{M} \hat{\theta}_{i} \cdot n_{i} \right)
\]

\[
+ 2 \cdot \frac{w \cdot (1 - w)}{N} \text{Cov} \left( \hat{\theta}_{j}, \frac{1}{N} \sum_{i=1}^{M} \hat{\theta}_{i} \cdot n_{i} \right)
\]

Simplifying some of the terms:

\[
Var \left( \sum_{i=1}^{M} \hat{\theta}_{i} \cdot n_{i} \right) = \sum_{i=1}^{M} n_{i}^{2} \cdot Var(\hat{\theta}_{i})
\]

and

\[
Cov \left( \hat{\theta}_{j}, \sum_{i=1}^{M} \hat{\theta}_{i} \cdot n_{i} \right) = \sum_{i=1}^{M} Cov(\hat{\theta}_{j}, \hat{\theta}_{i} \cdot n_{i}) = n_{j} \cdot Var(\hat{\theta}_{j})
\]

where we have used that \( \hat{\theta}_{j} \) is independent of \( \hat{\theta}_{i} \) for \( i \neq j \). Collecting like terms gives:

\[
Var(\hat{\theta}_{j}) \left( w^{2} + \frac{(1 - w)^2 \cdot n_{j}^{2}}{N^{2}} + \frac{2 \cdot w \cdot (1 - w) \cdot n_{j}}{N} \right)
\]

\[
+ \frac{(1 - w)^2}{N^{2}} \sum_{i \neq j} n_{i}^{2} \cdot Var(\hat{\theta}_{i})
\]

Using \( \sigma^{2}(\hat{\theta}_{i}) = \frac{\epsilon_{i}^{2}}{n_{i}} \) simplifies this to:

\[
\frac{\epsilon_{j}^{2}}{n_{j}} \left( w^{2} + \frac{(1 - w)^2 \cdot n_{j}^{2}}{N^{2}} + \frac{2 \cdot w \cdot (1 - w) \cdot n_{j}}{N} \right)
\]

\[
+ \frac{(1 - w)^2}{N^{2}} \sum_{i \neq j} n_{i} \cdot \epsilon_{i}^{2}
\]

Taking the expectation \( \mathbb{E}_{(\theta_{j}, \epsilon_{j}) \sim \Theta} \) on each side produces:

\[
\frac{\mu_{e}}{n_{j}} \left( w^{2} + \frac{(1 - w)^2 \cdot n_{j}^{2}}{N^{2}} + \frac{2 \cdot w \cdot (1 - w) \cdot n_{j}}{N} \right)
\]

\[
+ \frac{(1 - w)^2}{N^{2}} \sum_{i \neq j} n_{i} \cdot \mu_{e}
\]

Simplifying the coefficient gives:

\[ = \frac{\mu_{e}}{n_{j}} \left( w^{2} + \frac{2 \cdot w \cdot (1 - w) \cdot n_{j}}{N} \right) + \frac{\mu_{e} \cdot (1 - w)^2}{N} \sum_{i=1}^{M} n_{i} \]

\[ = \frac{\mu_{e}}{n_{j}} \left( w^{2} + \frac{2 \cdot w \cdot (1 - w) \cdot n_{j}}{N} \right) + \frac{\mu_{e} \cdot (1 - w)^2}{N} \sum_{i=1}^{M} n_{i} \]

Finally, we look at the last term. First, we rewrite:

\[
(\theta_{j} - \theta_{j})^{2} = \left( w \cdot \theta_{j} + \frac{1 - w}{N} \sum_{i=1}^{M} \theta_{i} \cdot n_{i} - \theta_{j} \right)^{2}
\]

\[
= \left( \frac{1 - w}{N} \sum_{i \neq j} \theta_{i} \cdot n_{i} - \frac{(1 - w) \cdot (N - n_{j})}{N} \theta_{j} \right)^{2}
\]

\[
= \left( \frac{1 - w}{N} \sum_{i \neq j} \theta_{i} \cdot n_{i} \right)^{2} + \frac{(1 - w)^2 \cdot (N - n_{j})^{2}}{N^{2}} \theta_{j}^{2}
\]
\[-2(1-w)^2 \cdot \frac{(N-n_j)}{N^2} \cdot \theta_j \cdot \sum_{i \neq j} \theta_i \cdot n_i\]

The first term can be rewritten as:
\[
\frac{(1-w)^2}{N^2} \left( \sum_{i \neq j} \theta_i^2 \cdot n_i^2 \right) + \sum_{i,k \neq j, i \neq k} \theta_i \cdot \theta_k \cdot n_i \cdot n_k
\]

If we take the expectation of the entire sum with respect to drawing parameters for each player, we will obtain some terms involving \(E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i^2]\) and some involving \(E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i]^2\). Collecting terms involving \(E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i^2]\) gives:
\[
\frac{E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i^2]}{N^2} \cdot (1-w)^2 \left( \sum_{i \neq j} n_i^2 + (N-n_j)^2 \right)
\]

Collecting terms involving \(E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i]^2\) gives:
\[
\frac{E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i]^2}{N^2} \cdot (1-w)^2 \left( \sum_{i,k \neq j, i \neq k} n_i \cdot n_k \right)
\]

Note that
\[(N-n_j)^2 = \sum_{i \neq j} n_i^2 + \sum_{i,k \neq j, i \neq k} n_i \cdot n_k
\]

Substituting and rearranging the term inside the parentheses gives us:
\[(N-n_j)^2 - \sum_{i \neq j} n_i^2 - 2(N-n_j)^2 = - \sum_{i \neq j} n_i^2 - (N-n_j)^2
\]

So the overall coefficient becomes:
\[-\frac{\frac{E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i^2]}{N^2} \cdot (1-w)^2 \left( \sum_{i \neq j} n_i^2 + (N-n_j)^2 \right)}{N^2} \cdot (1-w)^2
\]

We can combine the \(E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i^2]\) and \(E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i]^2\) terms because they share a common coefficient, giving us:
\[
\frac{\sum_{i \neq j} n_i^2 + (N-n_j)^2}{N^2} \cdot (1-w)^2.
\]

\[
\left( E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i^2] - E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i]^2 \right)
\]

where we have used \(\sigma^2 = Var(\theta_j) = E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i^2] - E_{(\theta, \epsilon^2_\theta) \sim \Theta} [\theta_i]^2\). The overall federated error becomes:
\[
\mu_e \cdot \frac{\sum_{i \neq j} n_i^2 + (N-n_j)^2}{N^2} \cdot (1-w)^2 \cdot \sigma^2
\]

**Lemma 7.2.** The minimum error is always achieved when \(w_j < 1\), implying that federation is always preferable to local learning.

**Proof.** Taking the derivative of the error with respect to \(w\) produces:
\[
2\mu_e \left( \frac{w_j - w}{n_j} \right) - 2 \sum_{i \neq j} \frac{n_i^2 + (N-n_j)^2}{N^2} \cdot (1-w) \cdot \sigma^2
\]

Setting this equal to 0 and solving for \(w\) produces \(w_j\) equal to:
\[
\frac{1}{n_j} \cdot \left( n_j \cdot \sigma^2 \right)
\]

Note that this value \(w_j\) depends on the player \(j\) that it is in reference to. It is also always strictly between 0 and 1.

To confirm that this is a point of minimum error rather than maximum, we can take the second derivative of the error, which gives a result that is always positive:
\[
2\mu_e \left( \frac{1}{n_j} - \frac{1}{N} \right) + 2 \sum_{i \neq j} \frac{n_i^2 + (N-n_j)^2}{N^2} \cdot \sigma^2
\]

**v-matrix**

In this case, the federated estimate differs for each player and is:
\[
\hat{\theta}_j = \sum_{i=1}^{M} v_{ji} \theta_i
\]

for \(\sum_{i=1}^{M} v_{ji} = 1\). Note that we can recover the \(w\) weighting case with \(v_{ji} = w + \frac{1-w}{n_j} \cdot \theta_i\) and \(v_{ji} = (1-w) \cdot \theta_i\).

**Lemma 7.3.** For mean estimation with the \(w\)-weighting federated learning method, the expected MSE of a player with \(n_j\) samples is:
\[
\mu_e \sum_{i=1}^{M} \frac{v_{ji}^2}{n_j} + \left( \sum_{i \neq j} v_{ji}^2 + \left( \sum_{i \neq j} v_{ji} \right)^2 \right) \cdot \sigma^2
\]

**Proof.** We again use a similar approach to the \(w\) method and the pure federated case.
\[
(\hat{\theta}_j - \theta_j)^2 = (\hat{\theta}_j - \theta_j - \theta_j + \theta_j)^2
\]

\[
= (\hat{\theta}_j - \theta_j)^2 + (\theta_j - \theta_j)^2 + 2 \cdot (\hat{\theta}_j - \theta_j) \cdot (\theta_j - \theta_j)
\]

The expectation of the last term (taken with respect to the data draws) is 0. Next, we look at the first term and take the expectation with respect to the data draws:
\[
E_{Y \sim D(\theta, \epsilon^2_\theta)} [(\hat{\theta}_j - \theta_j)^2] = Var(\hat{\theta}_j) = Var \left( \sum_{i=1}^{M} v_{ji} \theta_i \right)
\]

\[
= \sum_{i=1}^{M} v_{ji}^2 Var(\theta_i) = \sum_{i=1}^{M} v_{ji}^2 \frac{\sigma^2}{n_i}
\]
Taking the expectation with respect to the parameters $E_{(\theta, \epsilon^2) \sim \Theta}$ gives:

$$
\mu_c \sum_{i=1}^{M} \frac{v_{ji}^2}{n_i}
$$

Finally, we look at the last term. First, we rewrite:

$$
(\theta_j^v - \theta_j)^2 = \left( (v_{jj} - 1) \cdot \theta_j + \sum_{i \neq j} v_{ji} \cdot \theta_i \right)^2
$$

We know that $1 = v_{jj} + \sum_{i \neq j} v_{ji}$, so this sum can be rewritten as:

$$
\left( \sum_{i \neq j} v_{ji} \cdot (\theta_i - \theta_j) \right)^2
$$

Taking the expectation with respect to the parameters draws produces terms that involve both $E_{(\theta, \epsilon^2) \sim \Theta}[\theta_i^2]$ and $E_{(\theta, \epsilon^2) \sim \Theta}[\theta_i^2]$. Collecting like terms gives:

$$
E_{(\theta, \epsilon^2) \sim \Theta}[\theta_i^2] \left( 2 \sum_{i \neq j} v_{ji}^2 + \sum_{i \neq j, i \neq k} v_{ji} \cdot v_{jk} \right)
$$

Taking the expectation with respect to the parameters draws produces terms that involve both $E_{(\theta, \epsilon^2) \sim \Theta}[\theta_i^2]$ and $E_{(\theta, \epsilon^2) \sim \Theta}[\theta_i^2]$. Collecting like terms gives:

$$
E_{(\theta, \epsilon^2) \sim \Theta}[\theta_i^2] \left( 2 \sum_{i \neq j} v_{ji}^2 + \sum_{i \neq j, i \neq k} v_{ji} \cdot v_{jk} \right)
$$

Taking the expectation with respect to the parameters draws produces terms that involve both $E_{(\theta, \epsilon^2) \sim \Theta}[\theta_i^2]$ and $E_{(\theta, \epsilon^2) \sim \Theta}[\theta_i^2]$. Collecting like terms gives:

$$
E_{(\theta, \epsilon^2) \sim \Theta}[\theta_i^2] \left( 2 \sum_{i \neq j} v_{ji}^2 + \sum_{i \neq j, i \neq k} v_{ji} \cdot v_{jk} \right)
$$

We can rewrite the coefficient using that $(\sum_{i \neq j} v_{ji})^2 = \sum_{i \neq j} v_{ji}^2 + \sum_{i, k \neq j, i \neq k} v_{ji} \cdot v_{jk}$:

$$
2 \sum_{i \neq j} v_{ji}^2 + \sum_{i, k \neq j, i \neq k} v_{ji} \cdot v_{jk}
$$

For an overall error value of:

$$
\mu_c \sum_{i=1}^{M} \frac{v_{ji}^2}{n_i} + \left( \sum_{i \neq j} v_{ji}^2 + \sum_{i \neq j} v_{ji} \right)^2 \cdot \sigma^2
$$

**Lemma 7.4.** Define $V_i = \sigma^2 + \frac{\mu_c}{n_i}$. Then, the value of $\{v_{ji}\}$ that minimizes player $j$’s error is:

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

and

$$
v_{jk} = \frac{1}{V_k} \cdot \frac{V_j - \sigma^2}{1 + \sum_{i \neq j} \frac{1}{v_i}}
$$

**Proof.** To minimize, we will take the derivative of player $j$’s error with respect to the $v_{jk}$ weight. Note that we only have $v_{jj} = 1 - \sum_{i \neq j} v_{ji} = 1 - v_{jk} - \sum_{i \neq j, i \neq j} v_{jj}$ so $v_{jk}$ appears twice in the component involving $\mu_c$. Rewriting the error gives:

$$
\mu_c \sum_{i \neq j} \frac{v_{ji}^2}{n_i} + \mu_c \left( 1 - \sum_{i \neq j} v_{jj} \right)^2
$$

Taking the derivative with respect to $v_{jk}$ gives:

$$
\mu_c 2v_{jk} - 2\mu_c \left( 1 - \sum_{i \neq j} v_{ji} \right) + \sigma^2 \left( 2 \sum_{i \neq j} v_{ji} + 2v_{jk} \right)
$$

To confirm that we are finding a minimum rather than a maximum, we note that the second derivative is always positive:

$$
\mu_c \frac{2}{n_k} + \mu_c \frac{2}{n_j} + \sigma^2 (2 + 2) > 0
$$

We first simplify the derivative by substituting in for $v_{jj}$:

$$
\mu_c 2v_{jk} - 2\mu_c \frac{v_{ji}}{n_j} + 2\sigma^2 (1 - v_{jj} + v_{jk}) = 0
$$

And then solve for $v_{jk}$ to obtain:

$$
v_{jk} = \frac{v_{jj} \cdot \frac{\sigma^2 + \frac{\mu_c}{n_j}}{\sigma^2 + \frac{\mu_c}{n_k}}} - \frac{\sigma^2}{\sigma^2 + \frac{\mu_c}{n_k}}
$$

To find $v_{jj}$, we use that all of the weights sum up to 1:

$$
v_{jj} + \sum_{i \neq j} v_{ji} \cdot \frac{\sigma^2 + \frac{\mu_c}{n_j}}{\sigma^2 + \frac{\mu_c}{n_k}} - \sum_{i \neq j} \frac{\sigma^2}{\sigma^2 + \frac{\mu_c}{n_k}} = 1
$$

$$
v_{jj} \left( 1 + \sum_{i \neq j} \frac{\sigma^2}{\sigma^2 + \frac{\mu_c}{n_k}} \right) - \sum_{i \neq j} \frac{\sigma^2}{\sigma^2 + \frac{\mu_c}{n_k}} = 1
$$

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

Next, we define $V_j = \sigma^2 + \frac{\mu_c}{n_i}$. This allows us to rewrite the term as:

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

Similarly, we can rewrite:

$$
v_{jk} = \frac{1}{V_k} \cdot \frac{V_j - \sigma^2}{1 + V_j \sum_{i \neq j} \frac{1}{V_j}}
$$

□