Demystifying Local and Global Fairness Trade-offs in Federated Learning Using Information Theory

Faisal Hamman Sanghamitra Dutta

Abstract

In this paper, we present an information-theoretic perspective to group fairness trade-offs in federated learning (FL) with respect to sensitive attributes, such as, gender, race, etc. Existing works mostly focus on either *global fairness* (overall disparity of the model across all clients) or *local fairness* (disparity of the model at each individual client), without necessarily considering their trade-offs. There is a lack of understanding of the interplay between global and local fairness in FL, and if and when one implies the other. To address this gap, we leverage a body of work in information theory called partial information decomposition (PID) which first identifies three sources of unfairness in FL, namely, *Unique Disparity*, *Redundant Disparity*, and *Masked Disparity*. Using canonical examples, we demonstrate how these three disparities contribute to global and local fairness. This decomposition helps us derive fundamental limits and trade-offs between global or local fairness, particularly under data heterogeneity, as well as, derive conditions under which one implies the other. We also present experimental results on real-world datasets to support our theoretical findings. This work offers a more nuanced understanding of the sources of disparity in FL that can inform the use of local disparity mitigation techniques, and their convergence and effectiveness when deployed in practice.

1. Introduction

Federated learning (FL) is a framework where several parties *(clients)* collectively train machine learning models while retaining the confidentiality of their local data (Yang, 2020). With the growing use of FL in various high-stakes applications, such as finance, healthcare, recommendation systems, etc., it is crucial to ensure that these models do not discriminate against any demographic group based on sensitive features such as race, gender, age, nationality, etc. (Smith et al., 2016). While there are several methods to achieve group fairness in the centralized settings (Mehrabi et al., 2021a), these methods do not directly apply to a FL setting since each client only has access to their own local dataset, and hence, is restricted to only performing local disparity mitigation.

Some recent works (Du et al., 2021; Abay et al., 2020; Ezzeldin et al., 2021) focus on developing models that are fair when evaluated on the entire dataset across all clients, a concept known as *global fairness*. For example, several banks have decided to engage in a FL process to train a model that will determine loan qualifications without exchanging data among them. A globally fair model is one that does not discriminate against any protected group, when evaluated on the entire dataset across all the banks. On the other hand, *local fairness* considers the disparity of the model at each individual client, i.e., when evaluated on a client's local dataset. Local fairness is an important consideration, as the models will ultimately be deployed and used at the local client (Cui et al., 2021).

One might notice that global and local fairness evaluation can differ from each other when the local demographics at a client differ from the global demographics across the entire dataset (data heterogeneity, e.g., a bank with predominantly White customers). Previous research has mostly focused on achieving either global fairness (Du et al., 2021) or local fairness (Cui et al., 2021), without considering their trade-offs and interplay. There is a lack of understanding of the relationship between these two concepts, and if and when, one implies the other. It is generally understood that global and local fairness are the same when the data is i.i.d. across clients, but their interplay in other situations is not well-understood.

In this work, we aim to provide a fundamental understanding of group fairness trade-offs in the FL setting. We first formalize the notions of global and local disparity in FL using information theory. Next, we leverage a body of work within information theory called partial information decomposition (PID) to further identify three sources of disparity in FL that contribute to global and local disparity, namely, *Unique Disparity*, *Redundant Disparity*, and *Masked Disparity*. Using canonical examples, we provide a deeper understanding of these three sources of disparities. This decomposition is significant because of two reasons: (i) it helps us derive fundamental information-theoretic limits as well as trade-offs between global and local disparity, particularly under data heterogeneity; and (ii) it also helps us derive conditions under which either local or global fairness implies the other. We perform experiments to illustrate the presence of these disparities in practical scenarios. This work provides a more nuanced understanding of the interplay between these two fairness notions, that can better inform the use of disparity mitigation techniques, and their convergence and effectiveness when deployed in practice.

Our main contributions can be summarized as follows:

- **Partial information decomposition of global and local disparity into three sources of unfairness:** We formalize the notion of global and local fairness in federated learning using information theory. We first define global disparity as the mutual information between a model's prediction (denoted by $\hat{Y}$) and the sensitive attribute (denoted by Z), i.e., $\mathrm{I}(Z;\hat{Y})$ (Definition 1). Then, we show that local disparity can be represented as the conditional mutual information $\mathrm{I}(Z;\hat{Y}|S)$ where S denotes the client (Definition 2). We also demonstrate relationships between these information-theoretic terms and well-known fairness metrics such as statistical parity (see Lemma 1).

  Next, we propose a partial information decomposition (PID) that breaks down the global and local disparity into three components: *Unique Disparity*, *Redundant Disparity*, and *Masked Disparity*. We provide canonical examples to help understand these three sources of disparities in the context of FL (see Section 3.1).
- **Fundamental limits and trade-offs between local and global fairness:** With the use of the decomposed disparities, we have been able to uncover the fundamental information-theoretic limits and trade-offs between global and local disparities. We show the limitations of achieving global fairness using local fairness due to the redundant disparity (see Theorem 1) and the limitations of achieving local fairness using global fairness due to the masked disparity (see Theorem 2).
- **Understanding scenarios where local fairness implies global fairness and vice versa:** We also identify the conditions under which one form of fairness (local or global) implies the other. Specifically, we have established conditions under which local fairness can result in global fairness (Theorem 4) and conditions under which global fairness can result in local fairness (Theorem 5).
- **Experimental demonstrations:** We provide experimental evaluations on the Adult dataset to validate our theoretical findings. We demonstrate practical scenarios where unique, redundant, and masked disparities are prevalent, for two or more clients. We analyze the PID under various data heterogeneity scenarios with varying sensitive attribute distributions or varying synergy across clients.

Related works: There are various perspectives to fairness in FL (Shi et al., 2021). One such definition is *client-fairness* (Li et al., 2019), which aims to achieve equal performance across all client devices. In this work, we are instead interested in group fairness, i.e., fairness with respect to demographic groups based on gender, race, etc. Methods for achieving group fairness in a centralized machine learning setting (Hardt et al., 2016; Dwork et al., 2012; Kamishima et al., 2011; Pessach & Shmueli, 2022) may not directly apply in a FL setting since each client only has access to their local dataset. Existing works on group fairness in FL generally aim to develop models that achieve *global fairness*, without much consideration for the *local fairness* at each client (Ezzeldin et al., 2021). For instance, one approach to achieve global fairness in FL poses a constrained optimization problem to find the best model locally, while also ensuring that disparity at a client does not exceed a threshold and then aggregates those models (Chu et al., 2021; Rodríguez-Gálvez et al., 2021; Zhang et al., 2020). Other techniques involve bi-level optimization that aims to find the optimal global model (minimum loss) under the worst-case fairness violation (Papadaki et al., 2022; Hu et al., 2022; Zeng et al., 2021), or re-weighting mechanisms (Abay et al., 2020; Du et al., 2021), both of which often require sharing additional parameters with a server. More recently, (Cui et al., 2021) argue for local fairness, as the model will be deployed at the local client level, and propose constrained multi-objective optimization.

While previous works have made progress in attaining either global fairness (often with additional information sharing) or sometimes local fairness, their interplay has received less attention. In fact, the terms "global fairness" and "local fairness" have often been used somewhat loosely in the literature, without a clear understanding of their relationship, with the notable exceptions of (Ezzeldin et al., 2021; Cui et al., 2021) which claim that global and local fairness are equivalent when the dataset is i.i.d. across clients. Our work addresses this gap by formalizing both global and local fairness and deriving fundamental limits and trade-offs, particularly under data-heterogeneity, by leveraging a body of work in information theory called PID.

2. Preliminaries

Here, we introduce the FL setup and outline its key components. To begin, let K be the total number of federating clients. A client is represented as $S \in [K]$ where $[K] = \{1, 2, \ldots, K\}$. A client $S = k$ has a dataset $\hat{\mathcal{D}}_k = \{(x_i, y_i, z_i)\}_{i=1,\ldots n_k}$ where x_i denotes the input features, $y_i \in \{0, 1\}$ is the true label, $z_i \in \{0, 1\}$ is the sensitive attribute (1 for privileged group, 0 for unprivileged group), and n_k denotes the number of datapoints at client $S = k$. The entire dataset is given by $\hat{\mathcal{D}} = \cup_{k=1}^{K} \hat{\mathcal{D}}_k$. When denoting a random variable drawn from this dataset, we let X denote the input features, Z denote the sensitive attribute, and Y denote the true label. We also let $\hat{Y}$ represent the predictions of a model $f_\theta(X)$, where f_θ is a mapping from the input feature space to the output space, parameterized by its weights θ.

The standard FL framework aims to minimize the empirical risk $L(\theta)$ over each local dataset $\hat{\mathcal{D}}_k$:

$$\min_{\theta} L(\theta) = \min_{\theta} \frac{1}{K} \sum_{k=1}^{K} \alpha_k L_k(\theta). \tag{1}$$

Where, $L_k(\theta) = \frac{1}{n_k} \sum_{(x,y)\in\hat{\mathcal{D}}_k} l(f_\theta(x), y)$ is the local objective (or loss) at client k, and α_k is an importance coefficient (often equal across clients). $l(\cdot,\cdot)$ denotes a predefined loss function. In order to minimize the objective function (1), a decentralized approach is employed. Each client S trains on their own private dataset $\hat{\mathcal{D}}_k$ and provides their trained local model to a centralized server. The server aggregate the parameters of the local models to create a global model $f_\theta(x)$ using various techniques (Sah & Singh, 2022). E.g., the *FedAvg* algorithm (McMahan et al., 2017) is a popular approach that aggregates the parameters of local models by taking their average, which is then used to update the global model. This process is repeated until the global model achieves a satisfactory level of accuracy.

FL is a powerful technique for training high-performance global models under decentralized data. However, similar to the disparities that arise from centralized training (Mehrabi et al., 2021b), FL can also result in global models that discriminate against certain demographic groups. Next, we will explore the concept of group fairness in the context of FL and formally discuss two prevalent perspectives of fairness: global and local.

Definition 1 (Global Disparity). *The global disparity of a model f_θ with respect to Z is defined as $\mathrm{I}(Z, \hat{Y})$, the mutual information between Z and $\hat{Y}$ (where $\hat{Y} = f_\theta(X)$).*

This is related to a widely-used group fairness notion called statistical parity. Existing works (Hardt et al., 2016) define the global statistical parity of the model f_θ as: $\Pr(\hat{Y} = 1|Z = 1) = \Pr(\hat{Y} = 1|Z = 0)$. Global statistical parity is satisfied when Z is independent of $\hat{Y}$, which is equivalent to zero mutual information $\mathrm{I}(Z; \hat{Y}) = 0$. To further justify our choice of $\mathrm{I}(Z, \hat{Y})$ as a measure of global disparity, we provide a relationship between the absolute statistical parity gap and mutual information when they are non-zero in Lemma 1 (Proof in Appendix A).

Lemma 1. *[Relationship between Global Statistical Parity Gap and $\mathrm{I}(Z, \hat{Y})$] Let Z and $\hat{Y}$ be binary and $\Pr(Z = 0) = 1 - \Pr(Z = 1) = \alpha$. The global statistical parity gap $SP_{global} = |\Pr(\hat{Y} = 1|Z = 1) - \Pr(\hat{Y} = 1|Z = 0)|$ is bounded by $\frac{\sqrt{0.5\ \mathrm{I}(Z,\hat{Y})}}{2\alpha(1-\alpha)}$.*

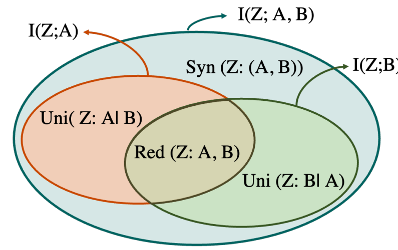

Figure 1. Venn diagram showing PID of $\mathrm{I}(Z; (A, B))$.

Similarly, existing literature (Ezzeldin et al., 2021) defines local statistical parity at a client k as: $\Pr(\hat{Y} = 1|Z = 1, S = k) = \Pr(\hat{Y} = 1|Z = 0, S = k)$. A critical observation that we make in this work is that: *local client unfairness can be quantified as the conditional mutual information* $\mathrm{I}(Z, \hat{Y}|S)$.

Definition 2 (Local Disparity). *The local disparity is defined as $\mathrm{I}(Z; \hat{Y}|S)$, the mutual information between Z and $\hat{Y}$ conditioned on S.*

Lemma 2. *$\mathrm{I}(Z, \hat{Y}|S) = 0$ if and only if $\Pr(\hat{Y} = 1|Z = 1, S = k) = \Pr(\hat{Y} = 1|Z = 0, S = k)$ at all clients.*

The proof (see Appendix A) uses the fact that

$$\mathrm{I}(Z, \hat{Y}|S) = \sum_{k=1}^{K} \Pr(S = k)\mathrm{I}(Z, \hat{Y}|S = k) \tag{2}$$

where $\mathrm{I}(Z, \hat{Y}|S = k)$ is the local mutual information at client k, and $\Pr(S = k) = \frac{n_k}{n}$, the proportion of data points at client k. Similar to Lemma 1, we can also get a relationship between SP_k and $\mathrm{I}(Z, \hat{Y}|S = k)$ when they are non-zero as expressed in Corollary 2 in Appendix A.

For the rest of this paper, we use $\mathrm{I}(Z, \hat{Y})$ to denote the global disparity and $\mathrm{I}(Z, \hat{Y}|S)$ to denote the local disparity.

2.1. Background on Partial Information Decomposition

The Partial Information Decomposition (PID) decomposes the mutual information $\mathrm{I}(Z; (A, B))$ about a random variable Z contained in the tuple (A, B) into four non-negative terms as follows :

$$\begin{aligned}\mathrm{I}(Z; (A, B)) = \mathrm{Uni}(Z, A|B) + \mathrm{Uni}(Z, B|A) \\ + \mathrm{Red}(Z; A, B) + \mathrm{Syn}(Z; (A, B))\end{aligned} \tag{3}$$

Here, $\mathrm{Uni}(Z, A|B)$ denotes the unique information about Z that is present only in A and not in B, $\mathrm{Red}(Z : (A, B))$ denotes the redundant information about Z that is present in both A and B, and $\mathrm{Syn}(Z; (A, B))$ denotes the synergistic information not present in either of A or B individually, but present jointly in (A, B).

Understanding PID: Let $Z = (Z_1, Z_2, Z_3)$ with $Z_1, Z_2, Z_3 \sim$ i.i.d. Bern(1/2). Let $A = (Z_1, Z_2, Z_3 \oplus N)$,

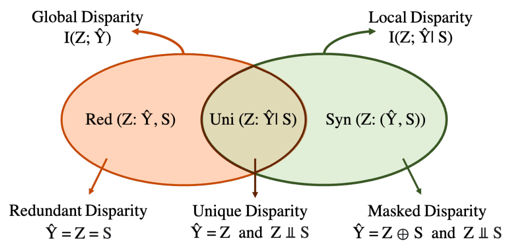

Figure 2. Venn diagram showing PID of Global and Local Disparity with canonical examples where each disparity is maximum.

$B = (Z_2, N)$, $N \sim$ Bern(½) is independent of Z. Here, $\mathrm{I}(Z; (A, B)) = 3$ bits.

The unique information about Z that is contained only in A and not in B is effectively contained in Z_1 and is given by $\mathrm{Uni}(Z{:}A|B) = \mathrm{I}(Z; Z_1) = 1$ bit. The redundant information about Z that is contained in both A and B is effectively contained in Z_2 and is given by $\mathrm{Red}(Z{:}(A, B)) = \mathrm{I}(Z; Z_2) = 1$ bit. Lastly, the synergistic information about Z that is not contained in either A or B alone, but is contained in both of them together is effectively contained in the tuple $(Z_3 \oplus N, N)$, and is given by $\mathrm{Syn}(Z{:}(A, B)) = \mathrm{I}(Z; (Z_3 \oplus N, N)) = 1$ bit. This accounts for the 3 bits in $\mathrm{I}(Z; (A, B))$. Here, B does not have any unique information about Z that is not contained in A, i.e., $\mathrm{Uni}(Z{:}B|A) = 0$.

Remark 1. *From Fig. 1, note that* $\mathrm{Uni}(Z{:}A|B)$ *can be viewed as the information-theoretic sub-volume of the intersection between* $\mathrm{I}(Z; A)$ *and* $\mathrm{I}(Z; A|B)$. *Similarly,* $\mathrm{Red}(Z{:}(A, B))$ *is the sub-volume between* $\mathrm{I}(Z; A)$ *and* $\mathrm{I}(Z; B)$. *Defining any one of the PID terms suffices to get the others. Here, we include a popular definition of* $\mathrm{Uni}(Z{:}A|B)$ *from (Bertschinger et al., 2014).*

Definition 3 (Unique Information). *Let* Δ *be the set of all joint distributions on* (Z, A, B) *and* Δ_p *be the set of joint distributions with the same marginals on* (Z, A) *and* (Z, B) *as the true distribution, i.e.,* $\Delta_p = \{Q \in \Delta : q(z, a) = \Pr(Z = z, A = a)$ *and* $q(z, b) = \Pr(Z = z, B = b)\}$. *Then,* $\mathrm{Uni}(Z{:}A|B) = \min_{Q \in \Delta_p} \mathrm{I}_Q(Z; A \mid B)$, *where* $\mathrm{I}_Q(Z; A \mid B)$ *is the conditional mutual information when* (Z, A, B) *have joint distribution* Q.

3. Main Results

In this section, we introduce a PID approach to understanding the fairness of FL models. First, we decompose global and local disparity into three sources of unfairness: unique disparity, redundant disparity, and masked disparity, and provide examples to illustrate and better understand these disparities in the context of FL.

3.1. Partial Information Decomposition of Disparity in Federated learning

Proposition 1. *The global and local disparity in a federated setting can be decomposed into PID terms as follows:*

$$\mathrm{I}(Z; \hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Red}(Z{:}\hat{Y}, S). \tag{4}$$

$$\mathrm{I}(Z; \hat{Y}|S) = \mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Syn}(Z{:}(\hat{Y}, S)). \tag{5}$$

We also refer to Fig. 2 for a pictorial illustration of this result. The proof follows directly from the relationship between different PID terms, as discussed in Remark 1.

The term $\mathrm{Uni}(Z{:}\hat{Y}|S)$ captures the information about sensitive attribute Z that is present only in the model prediction $\hat{Y}$ and not in the clients S. We refer to this as the **Unique Disparity**. The unique disparity represents the contribution of $\hat{Y}$ to our understanding of Z, separate from any information provided by S. It is important to note that this does not provide a complete picture of the model's disparity, as S may also contain redundant information with $\hat{Y}$ about Z.

The term $\mathrm{Red}(Z{:}\hat{Y}, S)$ denotes the redundant information about sensitive attribute Z that is present in both prediction $\hat{Y}$ and client S. We call this as the **Redundant Disparity**. The unique and redundant disparities make up the global disparity $\mathrm{I}(Z; \hat{Y})$.

The term $\mathrm{Syn}(Z{:}(\hat{Y}, S))$ represents the synergistic information about sensitive attribute Z that is not present in either $\hat{Y}$ or S individually, but is present jointly in $(\hat{Y}, S)$. We refer to this as the **Masked Disparity**, as it is only observed when $\hat{Y}$ and S are considered together. The unique and masked disparities make up the local disparity $\mathrm{I}(Z; \hat{Y}|S)$.

Canonical Examples. To better understand these concepts in a federated setting, we consider simple examples with binary model predictions, binary sensitive attribute, and two clients, i.e., $\hat{Y}, Z, S \in \{0, 1\}$. Note that $\mathrm{I}(Z; (\hat{Y}, S)) = H(Z) - H(Z|\hat{Y}, S) \leq H(Z) = 1$, i.e., the maximum disparity is 1 bit for this case.

Example 1 (Pure Uniqueness). *Let* $\hat{Y} = Z$ *and* $Z \perp\!\!\!\perp S$. *The model accepts only males from each client dataset. This model is both locally and globally unfair. Here,* $Red(Z; \hat{Y}, S) = 0$, $Uni(Z, \hat{Y}|S) = 1$, *and* $Syn(Z; (\hat{Y}, S)) = 0$

This is a scenario where the sensitive attributes are identically distributed across clients, such that $Z \perp\!\!\!\perp S$ (each client has the same proportion of privileged and unprivileged groups). If the model makes its predictions solely based on the sensitive attribute, i.e., $\hat{Y} = Z$, the unique disparity, $\mathrm{Uni}(Z, \hat{Y}|S) = 1$, since all information about Z is encoded in $\hat{Y}$ and none is present in S. The redundant disparity, $\mathrm{Red}(Z; \hat{Y}, S) = 0$, since $\hat{Y}$ and S do not share any information about Z. Similarly, $\text{Syn}(Z; (\hat{Y}, S)) = 0$, since jointly, $(\hat{Y}, S)$ do not contain any information about Z. As a result, both the global and local disparities, $\text{I}(Z; \hat{Y}) = \text{I}(Z; \hat{Y}|S) = 1$, indicating that the model is globally and locally unfair.

Example 2 (Pure Redundancy)**.** *Let $\hat{Y} = Z = S$. Client $S = 0$ has all females with negative predicted outcomes and client $S = 1$ has all males with positive predicted outcomes. It is clear that this model achieves local fairness at each client however it is globally unfair. In terms of PID, Red$(Z; \hat{Y}, S) = 1$, Uni$(Z, \hat{Y}|S) = 0$, and Syn$(Z; (\hat{Y}, S)) = 0$.*

In this scenario, the sensitive attributes are skewed across clients, with client $S = 0$ containing only females ($Z = 0$) and client $S = 1$ containing only males ($Z = 1$). The model makes its predictions based on these sensitive attributes. In this case, the redundant disparity, $\text{Red}(Z; \hat{Y}, S) = 1$, since all information about Z is contained in both $\hat{Y}$ and S. The unique disparity, $\text{Uni}(Z, \hat{Y}|S) = 0$, since there is no information about Z in $\hat{Y}$ that is not present in S. Similarly, $\text{Syn}(Z; (\hat{Y}, S)) = 0$, since jointly, $(\hat{Y}, S)$ do not contain any information about Z that is not present in $\hat{Y}$ and S individually. As a result, the global disparity, $\text{I}(Z; \hat{Y}) = 1$, and the local disparity, $\text{I}(Z; \hat{Y}|S) = 0$, indicate that the model is globally unfair but locally fair. It is not surprising that the model is globally unfair, as the predictions are based on the sensitive attributes. However, since each client has only one protected group, the model exhibits local fairness. In general, pure redundant disparity is observed when $Z - S - \hat{Y}$ form a **Markov chain**, but Z and $\hat{Y}$ are correlated, i.e., $\hat{Y} = S$ and $S = g(Z)$ for some function g.

Example 3 (Pure Synergy)**.** *Let $\hat{Y} = Z \oplus S$ and $Z \perp\!\!\!\perp S$. The model accepts males from client $S = 0$ and females from client $S = 1$, while others are rejected. This model is not locally fair but globally fair. Here, Red$(Z; \hat{Y}, S) = 0$, Uni$(Z, \hat{Y}|S) = 0$, and Syn$(Z; (\hat{Y}, S)) = 1$.*

In this scenario, the sensitive attributes are identically distributed across clients, i.e., $Z \perp\!\!\!\perp S$. The model prediction is an XOR of the sensitive attribute Z and clients S, i.e., $\hat{Y} = Z \oplus S$. Thus, the model accepts males ($Z = 1$) from client $S = 0$ and females ($Z = 1$) from client $S = 1$, while rejecting all other individuals. The masked disparity $\text{Syn}(Z; (\hat{Y}, S)) = 1$, since $(\hat{Y}, S)$ specifies information about Z that is not specified either $\hat{Y}$ or S. With both $\hat{Y}$ and S having no information about Z, i.e., $\text{I}(Z; S) = \text{I}(Z; \hat{Y}) = 0$, it follows that there can not be any unique or redundant disparity, $\text{Red}(Z; \hat{Y}, S) = \text{Uni}(Z, \hat{Y}|S) = 0$. As a result, the model is locally unfair, with $\text{I}(Z; \hat{Y}|S) = 1$, but globally fair, with $\text{I}(Z; \hat{Y}) = 0$. The model achieves global fairness by balancing the local unfairness at each client.

These examples demonstrate pure uniqueness, pure redundancy, and pure synergy are achieved. In practice, it is usually a combination or mixture of these cases. i.e., non-zero Unique, Redundant, and Masked disparities. These also extend to multiple clients and sensitive attributes.

PID is an effective tool for identifying and quantifying the sources of disparity in FL, particularly when data is distributed non-uniformly across clients. By decomposing the global disparity into unique, redundant, and masked disparity, we have a more nuanced understanding of disparity that can inform the development of unfairness mitigation techniques in FL. In the next section, we will discuss the fundamental information-theoretic limits of the trade-off between local and global fairness.

3.2. Fundamental Limits and Tradeoffs Between Local and Global Fairness

In this section, we leverage the PID to show the fundamental limitations and trade-offs between local and global fairness. We illustrate that global fairness does not imply local fairness and vice versa.

Limitations in Achieving Global Fairness with Local Fairness. We investigate the use of local fairness to achieve global fairness, or scenarios where a model is trained to achieve local fairness and subsequently deployed at the global level. As clients only have access to their own datasets, applying local disparity mitigation methods at each client can be convenient. Works such as (Cui et al., 2021) argue that local fairness is important as models are deployed at the local client level. But what happens to global fairness? In Theorem 1, we formally demonstrate the limitations of this approach. Even if local clients are able to use some optimal local mitigation methods and model aggregation techniques to achieve local fairness, the global disparity may still be greater than zero.

Theorem 1 (Impossibility of Using Local Fairness to Attain Global Fairness)**.** *As long as redundant disparity $\text{Red}(Z{:}\hat{Y}, S) > 0$, the global disparity $\text{I}(Z, \hat{Y}) > 0$ even if local disparity goes to 0.*

In order to achieve local fairness, the unique and masked disparities must be reduced to zero. The proof leverages PID of global disparity into non-negative terms, namely, unique and redundant disparities as shown in (4). Recall, example 2 (Pure Redundancy), where local disparity was zero but the global disparity equals 1 as a result of the redundant disparity. In a practical setting, such as a FL scenario with two hospitals, there may be a dependence between the sensitive attribute and clients due to factors like location. E.g., a hospital may have predominantly White patients and another may have predominantly Black patients. If the model assigns $\hat{Y} = 1$ to individuals in the first and $\hat{Y} = 0$ to individuals in the second one, it would be locally fair but would fail to be globally fair due to a non-zero redundant disparity.

Limitation in Achieving Local Fairness with Global Fairness. We now consider the scenario where a model is trained to achieve global fairness and is subsequently deployed at the local client level. Theorem 2 formalizes the relationship between global and local fairness in FL.

Theorem 2 (Global Fairness Does Not Imply Local Fairness)**.** *As long as masked disparity* $\mathrm{Syn}(Z{:}(\hat{Y}, S)) > 0$, *there exist scenarios where global fairness is attained but local fairness is not.*

In order to achieve global fairness, the unique and redundant disparities must be reduced to zero. The proof leverages decomposition of local disparity into unique and masked disparities as shown in (5). Recall example 3 (Pure Synergy), where the model accepts males from client $S = 0$ and females from client $S = 1$, while rejecting all others. While this model is globally fair, it is not locally fair. This demonstrates that while it is possible to train a model to achieve global fairness, it may still exhibit disparity when deployed at the local level due to the canceling of disparities between clients. This effect is captured by the masked disparity.

Definition 4 (Difference Between Local and Global Disparity)**.** *The difference between global and local disparity is:* $\mathrm{I}(Z; \hat{Y}) - \mathrm{I}(Z; \hat{Y}|S) = \mathrm{I}(Z; \hat{Y}; S)$. *This term is the "interaction information," which, unlike other mutual-information-based measures, can be positive or negative.*

Interaction information quantifies the redundancy and synergy present in a system. In FL, positive interaction information indicates a system with high levels of redundancy and global disparity, while negative interaction information indicates a system with high levels of synergy and local disparity. Interaction information can inform the trade-off between local and global disparity.

3.3. Towards Understanding Scenarios where Local Fairness Can Imply Global Fairness and Vice Versa.

In this section, we identify the conditions for achieving global fairness through local fairness and vice versa.

Towards Achieving Global Fairness using Local Fairness. We have previously shown that even if local fairness is achieved, global fairness may still be non-zero due to the presence of redundant disparity. We now discuss the necessary and sufficient condition to achieve global fairness using local fairness.

Theorem 3 (Necessary and Sufficient Condition to Achieve Global Fairness Using Local Fairness)**.** *If local disparity* $\mathrm{I}(Z, \hat{Y}|S)$ *goes to zero, then the global disparity* $\mathrm{I}(Z, \hat{Y})$ *also goes to zero, if and only if the redundant disparity* $\mathrm{Red}(Z{:}\hat{Y}, S) = 0$.

Lemma 3. $Z \perp\!\!\!\perp S \implies \mathrm{Red}(Z{:}\hat{Y}, S) = 0$. *When* Z *and* S *are independent, the redundant disparity is zero.*

The results of Theorem 3 and Lemma 3 suggest that when the proportion of each protected group is equal across all clients, the redundant disparity will decrease to zero. Hence, when the local disparity goes to zero, the global disparity will also decrease to zero. However, in practice, this proportion is fixed since the dataset at each client cannot be changed, i.e., $\mathrm{I}(Z; S)$ is fixed. Therefore, we explore another more controllable condition to eliminate redundant disparity even when $\mathrm{I}(Z; S) > 0$.

Lemma 4. *If synergistic disparity* $\mathrm{Syn}(Z{:}(\hat{Y}, S)) = 0$, *the redundant disparity* $\mathrm{Red}(Z{:}\hat{Y}, S) = 0$ *if* $\hat{Y}$ *and* S *are independent* $\hat{Y} \perp\!\!\!\perp S$ *or* $\mathrm{I}(\hat{Y}; S) = 0$, *even if* $\mathrm{I}(Z; S) > 0$.

Theorem 4. *If local disparity goes to zero, then the global disparity also goes to zero, if the model prediction* $\hat{Y}$ *is independent of* S, *i.e.,* $\mathrm{I}(\hat{Y}; S) = 0$.

Theorem 4 demonstrates that, in order to reduce redundant disparity and achieve global fairness when there is a strong correlation between Z and S, one potential solution is to enforce independence between $\hat{Y}$ and S. This means that the model should make predictions at the same rate across all clients. The proofs are provided in Appendix B.

Remark 2. *It is worth noting that the independence between* $\hat{Y}$ *and* S *can be approximately achieved if the true* Y *and* S *are independent, as* $\hat{Y}$ *is an estimation of* Y. *However, it is often the case that* $Y \perp\!\!\!\perp S$ *is fixed due to the fixed nature of datasets at each client. The mutual information* $\mathrm{I}(Y; S)$ *can provide insight into the expected value of* $\mathrm{I}(\hat{Y}; S)$, *as the federated model typically aims to also achieve a reasonable level of accuracy. It may even be possible to enforce* $\hat{Y} \perp\!\!\!\perp S$ *at the cost of accuracy.*

Towards Achieving Local Fairness using Global Fairness. It was previously demonstrated that global fairness does not necessarily imply local fairness due to the presence of masked disparity. In Theorem 5, we explore a condition to eliminate masked disparity and achieve local fairness through global fairness.

Theorem 5. *Local disparity will always be less than global disparity if masked disparity* $\mathrm{Syn}(Z{:}(\hat{Y}, S)) = 0$.

Corollary 1. *The local disparity will always be less than global disparity if* $Z, \hat{Y}, S$ *form a Markov chain* $Z - \hat{Y} - S$.

4. Experimental Demonstrations

In this section, we provide experimental evaluations on real-world datasets to validate our theoretical findings. We investigate the PID of global and local fairness under various conditions and scenarios.

Dataset. UCL *Adult* dataset (Dua & Graff, 2017), which comprises over 40,000 data points. The objective is to predict whether the annual earnings of an individual is more than $50K$ per year. We select *gender* as a sensitive attribute, with Male as $Z = 1$ and Female as $Z = 0$.

Evaluation. We define global and local disparities as mutual information measures. To estimate these values, we use the python `dit` package (James et al., 2018), which includes PID functions that allows us to decompose the global and local disparities into unique, redundant, and masked disparities. We implement the definition of unique information from (Bertschinger et al., 2014).

Demonstrating the Disparities. First, we demonstrate scenarios with unique, redundant, and masked disparities for the Adult dataset using two clients, $S = 0, 1$. We do this by strategically splitting the dataset between the clients and training our federated model using *FedAvg* (McMahan et al., 2017). For context, a model in the centralized case, achieved an accuracy of 84.67% and disparity $\mathrm{I}(Z; \hat{Y}) = 0.03537$.

Scenario 1 *Unique Disparity on Adult dataset.* Unique disparity is observed when the sensitive attribute is evenly distributed across clients ($Z \perp\!\!\!\perp S$). To achieve this, we randomly distribute the dataset among the two clients to ensure $\mathrm{I}(Z; S) = 0$. The global and local disparity is 0.0359 bits. Decomposing these we get the unique disparity as 0.0359 and zero redundant and masked disparity, indicating that the source of the disparity is solely from the dependence between the model predictions and sensitive attributes, and not from S. This matches the centralized case.

Scenario 2 *Redundant Disparity on Adult dataset.* Redundant disparity occurs when there is high heterogeneity of sensitive attributes across clients ($Z \approx S$). To achieve this, we distribute the dataset so as client $S = 0$ contains mainly females $Z = 0$ and client $S = 1$ contains mainly males, $Z = 1$, resulting in $\mathrm{I}(Z; S) = 0.8486$. The model trained had a global disparity of 0.0431 and a local disparity of 0.0014. By decomposing these, we find that the redundant disparity is 0.0431, masked disparity is 0.0014, and zero unique disparity.

Scenario 3 *Masked Disparity on Adult dataset.* The masked disparity is observed when the model predictions $\hat{Y} = Z \oplus S$. To attain this, we distribute the dataset such that the first client dataset contains males ($Z = 1$) with true labels $Y = 1$ and females ($Z = 0$) with true labels $Y = 0$. The second client dataset contains the remaining (males with $Y = 0$ and females with $Y = 1$). The trained model had a local disparity of 0.1761 and a global disparity of 0.0317. With a masked disparity of 0.1761, redundant disparity of 0.0317, and zero unique disparity. The non-zero redundant disparity is due to the way we split the data, which resulted in $\mathrm{I}(Z; S) = 0.2409$.

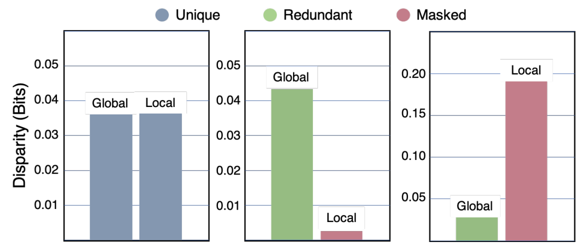

Figure 3. Plot demonstrating scenarios with unique, redundant, and masked disparities for the Adult dataset two client case. *(left)* Unique disparity when sensitive attributes are equally distributed across clients. *(center)* Redundant disparity when there is a dependency between clients and sensitive attributes. *(right)* Masked disparity when model predictions $\hat{Y} \approx Z \oplus S$.

We summarize the three scenarios in Fig. 3. Additionally, we evaluate the effects of using local fairness mitigation technique on the various disparities present. This is achieved by incorporating a statistical parity regularizer at each individual client. The results are presented in Table 1.

Table 1. Table illustrates the effects of using a naive local disparity mitigation technique on the various scenarios. It proved efficacious only when Unique disparity is present (scenario 1). However, with high redundancy or synergy (scenarios 2 &3), the utilization of the disparity mitigation technique exacerbated disparities.

	Loc.	Glob.	Uniq.	Red.	Mas.
Scenario 1	0.0359	0.0359	0.0359	0.0000	0.0000
+ fairness	0.0062	0.0062	0.0062	0.0000	0.0000
Scenario 2	0.0014	0.0431	0.0000	0.0431	0.0014
+ fairness	0.0110	0.0626	0.0000	0.0626	0.0110
Scenario 3	0.1761	0.0317	0.0000	0.0317	0.1761
+ fairness	0.0935	0.0418	0.0053	0.0365	0.0882

PID of Disparity under Heterogeneous Sensitive Attribute Distribution. We analyze the PID of local and global disparities under different sensitive attribute distributions across clients. We train the model with two clients, each having equal-sized datasets. We use $\alpha = \Pr(Z = 0|S = 0)$ to represent sensitive attribute heterogeneity. Note that for a fixed α, the proportions of sensitive attributes at the other client are fixed. For example since $\Pr(Z = 0) = 0.33$ for the Adult dataset, $\alpha = 0.33$ results in even distribution of sensitive attribute across the two clients. Our results are summarized in Fig. 4 and Table 2.

Observing Levels of Masked Disparity. Here, we aim to gain a deeper understanding of the circumstances leading to masked disparities. Through scenario 3, we showed how high masked disparities can occur. However, the level of synergy portrayed in the example may not always be present in reality. We attempt to quantify this using a metric *synergy level*. The synergy level (λ) measures how closely the model prediciton $\hat{Y}$ aligns with the XOR of Z and S.

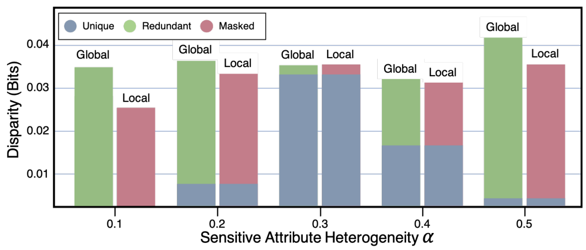

Figure 4. Plot illustrates PID of global and local fairness varying levels of sensitive attribute heterogeneity (α) and two clients case. When α is close to 0.3, the data is split evenly across clients, resulting in a higher level of unique disparity. As α deviates from 0.3, i.e., higher dependency between Z and S, the unique disparity decreases while redundant and masked disparity increase.

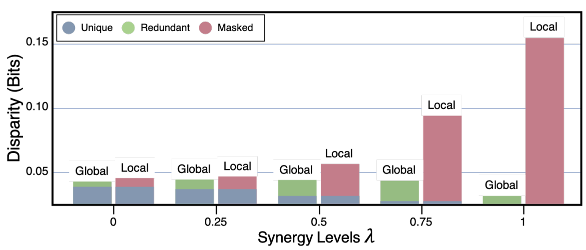

Figure 5. The plot demonstrates the relationship between the synergy level (λ) and global and local fairness. As the synergy level increases, the masked disparity and local disparity also increases as expected.

A value of 1 represents perfect alignment, while a value of 0 indicates independence. To achieve a high synergy level, we apply the method outlined in Scenario 3. To decrease the level, we randomly switch data points between clients until the synergy level reaches 0. The number of data points switched controls the level of synergy, ranging from 0 to 1. We conduct experiments with varying levels of synergy to observe the impact on masked disparity. The results are summarized in Fig. 5 and Table 4.

Experiments with Multiple Clients. Here, we examine scenarios involving multiple clients. Observations are similar to the two-client case we previously studied. We experiment with $K = 10$ clients. we study the disparities that occur when the data distribution is near i.i.d. distributed among the clients. We manipulate the proportion of sensitive attributes in the first half of the clients by using α. $\alpha = \Pr(Z = 0|S = 1, \ldots, 5)$. $\alpha = 0.33$ would correspond to the case where the sensitive attribute is distributed independently among the clients. We choose values of α that are close to 0.33. These distributions closely emulate the realistic scenarios that occur in the real world. In this scenario, we observe the presence of all types of disparities. Our results are summarized in Fig. 4 and Table 6.

Conclusions and Future Work: In this study, we explored

Table 2. The PID of global and local disparity for varying sensitive attribute heterogeneity α.

α	$\mathrm{I}(Z;S)$	Loc.	Glob.	Uniq.	Red.	Mas.
0.1	0.1877	0.0262	0.0342	0.000	0.0342	0.0262
0.2	0.0575	0.0336	0.0364	0.0064	0.0301	0.0273
0.3	0.0032	0.0363	0.0365	0.0332	0.0032	0.0031
0.4	0.0154	0.0311	0.0319	0.0186	0.0133	0.0125
0.5	0.0957	0.0368	0.0413	0.0023	0.0390	0.0345

Table 3. PID of global and local disparity under varying synergy levels λ.

λ	$\mathrm{I}(Z;S)$	Loc.	Glob.	Uniq.	Red.	Mas.
0	0.0035	0.0402	0.0373	0.0338	0.0035	0.0063
0.25	0.0113	0.0486	0.0419	0.0308	0.0111	0.0178
0.5	0.0299	0.0536	0.0335	0.0127	0.0208	0.0410
0.75	0.0846	0.0932	0.0366	0.0023	0.0343	0.0909
1	0.2409	0.1644	0.0149	0.0000	0.0150	0.1644

the intricacies of fairness in the federated setting using PID. We were able to uncover three distinct types of disparities: Unique, Redundant, and Masked Disparity. By utilizing canonical examples, we were able to illustrate how these disparities contribute to both global and local unfairness, and establish fundamental limits and trade-offs between them. Our experimental results further reinforced our findings and highlighted the various scenarios in which unique, redundant, and masked disparities occur in a practical setting. This work provides a more nuanced understanding of the sources of disparity in FL, which can inform the use of local disparity mitigation techniques, and their convergence and effectiveness when deployed in practice. Future studies could also investigate how this approach could be extended to other fairness metrics, such as equalized odds.

References

Abay, A., Zhou, Y., Baracaldo, N., Rajamoni, S., Chuba, E., and Ludwig, H. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

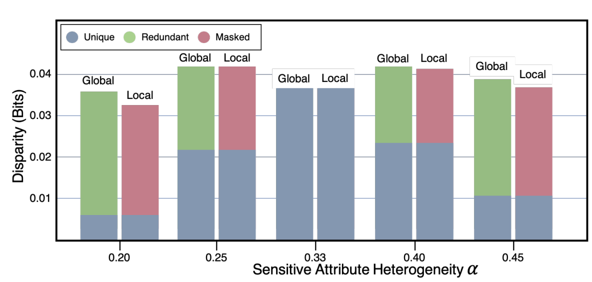

Figure 6. The plot shows the PID of disparities when the data is near i.i.d. among $K = 10$ clients. All types of disparities can be observed. The value $\alpha = 0.33$ represents the case where the data is i.i.d. and only unique disparity is observed.

Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., and Ay, N. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

Chu, L., Wang, L., Dong, Y., Pei, J., Zhou, Z., and Zhang, Y. Fedfair: Training fair models in cross-silo federated learning. *arXiv preprint arXiv:2109.05662*, 2021.

Cui, S., Pan, W., Liang, J., Zhang, C., and Wang, F. Addressing algorithmic disparity and performance inconsistency in federated learning. *Advances in Neural Information Processing Systems*, 34:26091–26102, 2021.

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

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

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

Ezzeldin, Y. H., Yan, S., He, C., Ferrara, E., and Avestimehr, S. Fairfed: Enabling group fairness in federated learning. *arXiv preprint arXiv:2110.00857*, 2021.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Hu, S., Wu, Z. S., and Smith, V. Provably fair federated learning via bounded group loss. *arXiv preprint arXiv:2203.10190*, 2022.

James, R. G., Ellison, C. J., and Crutchfield, J. P. dit: a Python package for discrete information theory. *The Journal of Open Source Software*, 3(25):738, 2018. doi: https://doi.org/10.21105/joss.00738.

Kamishima, T., Akaho, S., and Sakuma, J. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650. IEEE, 2011.

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

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL `https://proceedings.mlr.press/v54/mcmahan17a.html`.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021a. ISSN 0360-0300. doi: 10.1145/3457607. URL `https://doi.org/10.1145/3457607`.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021b. ISSN 0360-0300. doi: 10.1145/3457607. URL `https://doi.org/10.1145/3457607`.

Papadaki, A., Martinez, N., Bertran, M., Sapiro, G., and Rodrigues, M. Minimax demographic group fairness in federated learning. *arXiv preprint arXiv:2201.08304*, 2022.

Pessach, D. and Shmueli, E. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), feb 2022. ISSN 0360-0300. doi: 10.1145/3494672. URL `https://doi.org/10.1145/3494672`.

Rodríguez-Gálvez, B., Granqvist, F., van Dalen, R., and Seigel, M. Enforcing fairness in private federated learning via the modified method of differential multipliers. *arXiv preprint arXiv:2109.08604*, 2021.

Sah, M. P. and Singh, A. Aggregation techniques in federated learning: Comprehensive survey, challenges and opportunities. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 1962–1967, 2022. doi: 10.1109/ICACITE53722.2022.9823702.

Shi, Y., Yu, H., and Leung, C. A survey of fairness-aware federated learning. *arXiv preprint arXiv:2111.01872*, 2021.

Smith, M., Munoz, C., and Patil, D. J. Big Risks, Big Opportunities: the Intersection of Big Data and Civil Rights, 2016.

Yang, Q. *Federated learning / Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, Han Yu.* Synthesis lectures on artificial intelligence and machine learning, #43. Morgan & Claypool, San Rafael, California, 2020. ISBN 1-68173-698-5.

Zeng, Y., Chen, H., and Lee, K. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2021.

Zhang, D. Y., Kou, Z., and Wang, D. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 1051–1060, 2020. doi: 10.1109/BigData50022.2020.9378043.

A. Additional Results and Proofs for Section 2

Lemma 1. *[Relationship between Global Statistical Parity Gap and* $\mathrm{I}(Z, \hat{Y})$*] Let* Z *and* $\hat{Y}$ *be binary and* $\Pr(Z = 0) = 1 - \Pr(Z = 1) = \alpha$*. The global statistical parity gap* $SP_{global} = |\Pr(\hat{Y} = 1|Z = 1) - \Pr(\hat{Y} = 1|Z = 0)|$ *is bounded by* $\frac{\sqrt{0.5\ \mathrm{I}(Z,\hat{Y})}}{2\alpha(1-\alpha)}$.

Proof. Mutual information can be expressed as KL divergence.

$$\mathrm{I}(Z, \hat{Y}) = D\left(P(\hat{Y}, Z)||P(\hat{Y}), P(Z)\right)$$

Using Pinskers Inequality,

$$d_{tv}(P, Q) \leq \sqrt{0.5D(P, Q)}$$

where, $d_{tv}(P, Q)$ is the total variation between two probability distributions P, Q.

$$\begin{aligned}
d_{tv}\left(\Pr(\hat{Y}, Z), \Pr(\hat{Y})\Pr(Z)\right) &= \frac{1}{2}\sum_{\hat{y},z}\left|\Pr_{\hat{Y},Z}(\hat{y}, z) - \Pr_{\hat{Y}}(\hat{y})\Pr_{Z}(z)\right| \\
&= \sum_{z}\Pr_{Z}(z)\sum_{\hat{y},z}\frac{1}{2}\left|\Pr_{\hat{Y}|Z=z}(\hat{y}) - \Pr_{\hat{Y}}(\hat{y})\right| \\
&= \frac{1}{2}\Pr(Z = 1)\left[|\Pr(\hat{Y} = 1|Z = 1) - \Pr(\hat{Y} = 1)| + |\Pr(\hat{Y} = 0|Z = 1) - \Pr(\hat{Y} = 0)|\right] \\
&+ \frac{1}{2}\Pr(Z = 0)\left[|\Pr(\hat{Y} = 1|Z = 0) - \Pr(\hat{Y} = 1)| + |\Pr(\hat{Y} = 0|Z = 0) - \Pr(\hat{Y} = 0)|\right] \\
&= \frac{1}{2}\alpha(1-\alpha)|SP1| + \frac{1}{2}\alpha(1-\alpha)|SP0| + \frac{1}{2}\alpha(1-\alpha)|SP1| + \frac{1}{2}\alpha(1-\alpha)|SP0| \\
&= \alpha(1-\alpha)|SP1| + \alpha(1-\alpha)|SP0| \qquad (6)
\end{aligned}$$

where, $\Pr(Z = 0) = 1 - \Pr(Z = 1) = \alpha$ and

$SPi = \Pr(\hat{Y} = i|Z = 1) - \Pr(\hat{Y} = i|Z = 0) = \Pr(\hat{Y} = i|Z = 1) - \Pr(\hat{Y} = i)$.

To complete the proof, we show that $|SP1| = |SP0|$

$$\begin{aligned}
SP1 =& \Pr(\hat{Y} = 1|Z = 1) - \Pr(\hat{Y} = 1) \\
=& \Pr(\hat{Y} = 1|Z = 1) - \left(1 - \Pr(\hat{Y} = 0)\right) \\
=& -1 + \Pr(\hat{Y} = 1|Z = 1) + \Pr(\hat{Y} = 0) \\
=& -\Pr(\hat{Y} = 0|Z = 1) + \Pr(\hat{Y} = 0) = -SP0
\end{aligned}$$

Hence, $|SP1| = |SP0|$ and from (6), we get

$$2\alpha(1-\alpha)|SP_1| \leq \sqrt{0.5MI}$$

Corollary 2. *The statistical parity at each client* k *can be expressed as*

$$|SP_k| \leq \frac{\sqrt{0.5\ \mathrm{I}(Z, \hat{Y}|S = k)}}{2\alpha_k(1-\alpha_k)}$$

where, $\alpha_k = \Pr(Z = 0|S = k) = 1 - \Pr(Z = 1|S = k)$.

□

B. Additional Results and Proofs for Section 3

Theorem 3 (Necessary and Sufficient Condition to Achieve Global Fairness Using Local Fairness)**.** *If local disparity* $\mathrm{I}(Z, \hat{Y}|S)$ *goes to zero, then the global disparity* $\mathrm{I}(Z, \hat{Y})$ *also goes to zero, if and only if the redundant disparity* $\mathrm{Red}(Z{:}\hat{Y}, S) = 0$.

Proof. From the PID of local and global disparity,

$$\mathrm{I}(Z; \hat{Y}) = \mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Red}(Z{:}\hat{Y}, S).$$
$$\mathrm{I}(Z; \hat{Y}|S) = \mathrm{Uni}(Z{:}\hat{Y}|S) + \mathrm{Syn}(Z{:}(\hat{Y}, S)).$$

Therefore if, $\mathrm{I}(Z; \hat{Y}|S) = 0$, then $\mathrm{Uni}(Z{:}\hat{Y}|S) = 0$
Hence,

$$\mathrm{I}(Z; \hat{Y}) = \mathrm{Red}(Z{:}\hat{Y}, S)$$
$$\mathrm{I}(Z; \hat{Y}) = 0 \iff \mathrm{Red}(Z{:}\hat{Y}, S) = 0$$

□

Lemma 3. $Z \perp\!\!\!\perp S \implies \mathrm{Red}(Z{:}\hat{Y}, S) = 0$. *When* Z *and* S *are independent, the redundant disparity is zero.*

Proof. By leveraging the PID of $\mathrm{I}(Z; S)$ and the non-negative property of the PID terms.

$$\mathrm{I}(Z; S) = \mathrm{Uni}(Z{:}S|\hat{Y}) + \mathrm{Red}(Z{:}\hat{Y}, S)$$
$$\mathrm{I}(Z; S) \geq \mathrm{Red}(Z{:}\hat{Y}, S)$$

Hence, $Z \perp\!\!\!\perp S \implies \mathrm{Red}(Z{:}\hat{Y}, S) = 0$. □

Lemma 4. *If synergistic disparity* $\mathrm{Syn}(Z{:}(\hat{Y}, S)) = 0$*, the redundant disparity* $\mathrm{Red}(Z{:}\hat{Y}, S) = 0$ *if* $\hat{Y}$ *and* S *are independent* $\hat{Y} \perp\!\!\!\perp S$ *or* $\mathrm{I}(\hat{Y}; S) = 0$*, even if* $\mathrm{I}(Z; S) > 0$.

Proof. Interaction information expressed in PID terms:

$$\begin{aligned} \mathrm{I}(Z; \hat{Y}; S) &= \mathrm{I}(Z; \hat{Y}) - \mathrm{I}(Z; \hat{Y}|S) \\ &= \mathrm{Red}(Z{:}\hat{Y}, S) - \text{Syn}(Z; (\hat{Y}, S)) \end{aligned}$$

If synergistic information $\text{Syn}(Z; (\hat{Y}, S)) = 0$, we have:

$$\begin{aligned} \mathrm{I}(Z; \hat{Y}; S) &= \mathrm{I}(Z; \hat{Y}) - \mathrm{I}(Z; \hat{Y}|S) \\ &= \mathrm{Red}(Z{:}\hat{Y}, S) \geq 0 \end{aligned}$$

Since the interaction information is positive and symmetric,

$$\mathrm{I}(\hat{Y}; S) \geq \mathrm{I}(\hat{Y}; S) - \mathrm{I}(\hat{Y}; S|Z) = \mathrm{Red}(Z{:}\hat{Y}, S)$$

and therefore, $\hat{Y} \perp\!\!\!\perp S \implies \mathrm{Red}(Z{:}\hat{Y}, S) = 0$. □

Corollary 1. *The local disparity will always be less than global disparity if* $Z, \hat{Y}, S$ *form a Markov chain* $Z - \hat{Y} - S$.

Proof. By leveraging the PID of $\mathrm{I}(Z; S|\hat{Y})$,

$$\mathrm{I}(Z; S|\hat{Y}) = \mathrm{Uni}(Z{:}S|\hat{Y}) + \mathrm{Syn}(Z{:}(\hat{Y}, S))$$

Hence, $\mathrm{I}(Z; S|\hat{Y}) = 0 \implies \mathrm{Syn}(Z{:}(\hat{Y}, S)) = 0$ □

C. Additional Experimental Results

Dataset. The Adult dataset is a publicly available dataset in the UCI repository based on 1994 U.S. census data (Dua & Graff, 2017). The goal of this dataset is to successfully predict whether an individual earns more or less than 50,000 per year based on features such as occupation, marital status, and education. We select *gender* as a sensitive attribute, with Male as $Z = 1$ and Female as $Z = 0$.

Setup. In our federated learning model, both the server and clients had two hidden layers, each containing 32 hidden units. The activation function used was ReLU, with a binary cross-entropy loss function and Adam optimizer. The first round began with the server initializing the weights of the model and sharing them with all clients. Each client then trained their local model on their designated local dataset, which was carefully divided to observe various disparities. The training process for each client was done using a batch size of 64 and 2 epochs. After training, each client shared their weight parameters with the server. The server then used the FedAvg algorithm to aggregate the weights of all clients and update the model. The updated weights were then shared back to the clients. This process was repeated for several rounds until the loss converged. The final model is then used for our evaluations. The python `dit` package (James et al., 2018) was used to estimate our measures. It includes PID functions that allow us to decompose the global and local disparities into unique, redundant, and masked disparities. We implement the definition of unique information from (Bertschinger et al., 2014).

We include additional results, expanded tables, figures, and details that provide a more comprehensive understanding of our study.

PID of Disparity under Heterogeneous Sensitive Attribute Distribution. We analyze the PID of local and global disparities under different sensitive attribute distributions across clients. We train the model with two clients, each having equal-sized datasets. We use $\alpha = \Pr(Z = 0|S = 0)$ to represent sensitive attribute heterogeneity. Note that for a fixed α, the proportions of sensitive attributes at the other client are fixed. For example since $\Pr(Z = 0) = 0.33$ for the Adult dataset, $\alpha = 0.33$ results in even distribution of sensitive attribute across the two clients. Our findings have been succinctly summarized in Figure 4, providing a clear and concise visual representation of our results. Additionally, for a deeper delve, we have also included an extended table in Table 4,

Table 4. The PID of global and local disparity for varying sensitive attribute heterogeneity α *(Extended Version)*

α	$\mathrm{I}(Z; S)$	Local	Global	Unique	Redundant	Masked	$\mathrm{I}(\hat{Y}; S)$	Accuracy
0.1	0.1877	0.0262	0.0342	0.0000	0.0342	0.0262	0.0080	86.54%
0.2	0.0575	0.0336	0.0364	0.0064	0.0301	0.0273	0.0028	86.95%
0.3	0.0032	0.0363	0.0365	0.0332	0.0032	0.0031	0.0002	86.86%
0.33	0.0000	0.0340	0.0340	0.0340	0.0000	0.0000	0.0000	87.34%
0.4	0.0154	0.0311	0.0319	0.0186	0.0133	0.0125	0.0009	86.70%
0.5	0.0957	0.0368	0.0413	0.0023	0.0390	0.0345	0.0045	86.77%
0.6	0.2613	0.0242	0.0346	0.0000	0.0346	0.0242	0.0104	86.61%
0.66	0.4392	0.0185	0.0325	0.0000	0.0325	0.0185	0.0140	86.36%

Observing Levels of Masked Disparity. We aim to gain a deeper understanding of the circumstances leading to masked disparities. Through scenario 3, we showed how high masked disparities can occur. However, the level of synergy portrayed in the example may not always be present in reality. We attempt to quantify this using a metric *synergy level*. The synergy level (λ) measures how closely the model prediction $\hat{Y}$ aligns with the XOR of Z and S. A value of 1 represents perfect alignment, while a value of 0 indicates independence. To achieve a high synergy level, we apply the method outlined in Scenario 3. To decrease the level, we randomly switch data points between clients until the synergy level reaches 0. The number of data points switched controls the level of synergy, ranging from 0 to 1. We conduct experiments with varying levels of synergy to observe the impact on masked disparity. The results are summarized in Fig. 5 and an extended Table 5.

Multiple Client Case. Here, we examine scenarios involving multiple clients. Observations are similar to the two-client case we previously studied. To observe a high unique disparity, sensitive attributes need to be identically distributed across clients. To observe the redundant disparity, there must be some dependency between clients and a specific sensitive attribute, meaning certain demographic groups are known to belong to a specific client. The masked disparity can be observed when there is a high level of synergy or XOR behavior between variables Z and S. Note that since S is no longer binary, we can convert its decimal value to binary and then take the XOR.

Table 5. PID of global and local disparity under varying synergy levels λ *(Extended Version)*.

λ	I(Z;S)	Local	Global	Unique	Redundant	Masked	I($\hat{Y};S$)	Accuracy	I($Z;\hat{Y}\|S=0$)	I($Z;\hat{Y}\|S=1$)
0	0.0035	0.0402	0.0373	0.0338	0.0035	0.0063	0.0005	85.12%	0.0196	0.0608
0.25	0.0113	0.0486	0.0419	0.0308	0.0111	0.0178	0.0009	85.54%	0.0819	0.0152
0.5	0.0299	0.0536	0.0335	0.0127	0.0208	0.0410	0.0033	85.24%	0.1056	0.0017
0.75	0.0846	0.0932	0.0366	0.0023	0.0343	0.0909	0.0068	85.26%	0.0024	0.1840
1	0.2409	0.1644	0.0149	0.0000	0.0150	0.1644	.0201	84.30%	0.0839	0.2450

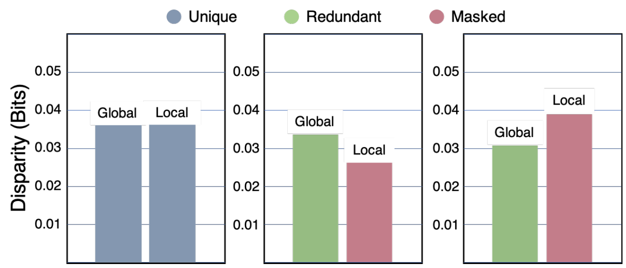

Figure 7. Plot demonstrating scenarios with unique, redundant, and masked disparities for the Adult dataset 5 client case. Difficulty in splitting to achieve pure redundant and masked disparity due to the proportion of labels in the dataset.

We experiment, with $K = 5$ clients. We examine the three disparities. To observe the unique disparity by randomly distributing the data among clients. For redundant disparity, we divide the data such that the first two clients have mostly females and the remaining three clients have mostly males. For masked disparity, we distribute the data similarly to scenario 3 (see Fig. 7).

An extended table has been included to provide a more in-depth analysis of the PID of global and local disparity for various sensitive attribute distributions across 10 clients.

Table 6. PID of global and local disparity for various sensitive attribute distributions across 10 clients.

α	Unique	Redundant	Masked	Global	Local	Accuracy
0.25	0.0219	0.0190	0.0178	0.0409	0.0409	84.85%
0.33	0.0376	0.000	0.0000	0.0376	0.0376	85.58%
0.4	0.0268	0.0141	0.0137	0.0410	0.0405	84.85%
0.45	0.0107	0.0289	0.0270	0.039	0.0377	84.85%