1

csk737502

November 2024

1 Introduction

sectionInvestigation of LID Variation for Noisy Clients In this chapter, we will explore the average LID variation of sample feature vectors for clean and noisy clients in a federated learning system during training. The investigation will follow this sequence: first, in section x.1, we will provide an overview, necessary definitions, and basic assumptions; next, in section x.2, we will discuss the variation in distances between sample feature vectors and predefined reference vectors; then, in section x.3, based on the distance variation characteristics obtained in section x.2, we will discuss the average LID variation of sample feature vectors for clean and noisy clients during training; finally, in section x.4, we will experimentally verify some of the hypotheses presented in the previous sections.

1.1 Overview, Definitions, and Assumptions

Overview: A client contains N samples, divided into M categories, with each sample assigned a label indicating its category. To process the client's data, a suitable model is used for machine learning, consisting of l_M layers and trained using a cross-entropy loss function and stochastic gradient descent (SGD). After several rounds of training, a preliminary model is generated.

Objective: Based on the aforementioned client, preliminary model, and training method, we aim to understand the changes in the average local intrinsic dimensionality (LID) of sample feature vectors within the client before and after each subsequent training round.

Definitions:

1. The output generated by sample X through the hth layer of the model before the lth round of training is called the feature vector of sample X in the hth layer during the lth round of training.

2. Let the set of all samples be $\{X_1, X_2, \ldots, X_N\}$, and the set of feature vectors of these samples in the hth layer during the lth round of training be $\{V_1, V_2, \ldots, V_N\}$, where each V_j is the feature vector of X_j in that layer. If the length of the feature vectors is l_V, all feature vectors are distributed in

an l_V-dimensional space, referred to as the feature space of the hth layer of the model during the lth training round. $\{V_1, V_2, \ldots, V_N\}$ are distributed in this feature space, and if a subset of feature vectors $\{V_{S_1}, V_{S_2}, \ldots, V_{S_n}\}$ (where $n \leq N$ and $S_1, S_2, \ldots, S_n \in \{1, 2, \ldots, N\}$) clusters within a finite region and is significantly separated from other category feature vectors, then $\{V_{S_1}, V_{S_2}, \ldots, V_{S_n}\}$ is said to form a cluster in the feature space.

3. In the feature space, feature vectors can be considered as points because mathematically, each vector can represent a position. The numerical values of the feature vector determine the point's location in the feature space. For example, a three-dimensional feature vector $[v_1, v_2, v_3]$ corresponds to a point at coordinates (v_1, v_2, v_3) in the feature space. In this chapter, whenever feature vectors are referred to as points, this definition applies.

4. In the sample set, for any chosen sample A, its feature vector in the hth layer during the lth training round is denoted as y, and the reference vector of the cluster containing this feature vector is denoted as y_s (the specific definition of the reference vector is given in Assumption 2). The distance between A's feature vector in the hth layer and the reference vector is denoted as $\|y - y_s\|$. The change in this distance before and after the lth training round is denoted as $\Delta\|y - y_s\|$.

Basic Assumptions:

Assumption 1: During training, after several rounds of training, feature vectors of different categories gradually form corresponding clusters. For a cluster whose feature vectors correspond to samples belonging to the same actual category, the label categories have only three possible cases:

(a) Sample A's actual category is a, and the label category is b. Such samples are denoted as A_1 type samples.

(b) Sample A's actual category is c, and the label category is a. Such samples are denoted as A_2 type samples.

(c) Sample A's actual category is a, and the label category is a. Such samples are denoted as A_3 type samples.

where $a, b, c \in \{1, 2, \ldots, M\}$ and $a \neq b$, $a \neq c$. This cluster is then called an a type cluster.

It is assumed that each cluster belongs to one category, and there is only one cluster per category, resulting in M clusters with distinct categories.

Additionally:

- For a clean client, all samples are of type A_3.
- For a noisy client, samples include all three types: A_1, A_2, and A_3.

Assumption 2: For an a type cluster in the hth layer of the model ($h \in \{1, 2, \ldots, l_M\}$, $a \in \{1, 2, \ldots, M\}$), after several rounds of training resulting in a preliminary model, a reference vector y_{sa} can be taken. For any other two points A, B in the cluster, with corresponding feature vectors y_A and y_B, and with the ith component of the resulting probability vectors being p_{iA} and p_{iB}, the following holds:

(a) If $p_{iA} > p_{iB}$, then $\|y_A - y_{sa}\| < \|y_B - y_{sa}\|$.

(b) If $p_{iA} < p_{iB}$, then $\|y_A - y_{sa}\| > \|y_B - y_{sa}\|$.

Assumption 3: During training, for any sample A in the dataset, let its feature vector in the hth layer be denoted as y_A; the reference point of the cluster it belongs to is denoted as S, with feature vector y_S. It is assumed that $\Delta\|y_A - y_S\|$ depends only on $\|y_A - y_S\|$, i.e., $\Delta\|y_A - y_S\| = f(\|y_A - y_S\|)$.

1.2 Changes in Distances Between Points and Reference Points in Clusters During Training

Consider a sample point A whose input vector before the lth training round in the hth layer of the model is x, and its output vector is y. The length of vector x is l_X, and the length of vector y is l_Y. The reference vector of the cluster where feature vector y_A is located is denoted as y_S.

In this section, we discuss the sign and magnitude of $\Delta\|y - y_S\|$ during the lth training round. Since l, A, and h are arbitrarily chosen, this can reflect the general behavior of $\Delta\|y - y_S\|$ during training.

Our method is to analyze the influence of changes in y before and after the lth training round on $\Delta\|y - y_S\|$. This influence is examined in terms of its sign and absolute magnitude. Furthermore, it is not necessary to consider the impact of all components of y; it suffices to discuss the effect of any single component. If the impact of this component on $\Delta\|y - y_S\|$ is clear, e.g., it is definitively positive, then the overall effect of y on $\Delta\|y - y_S\|$ follows suit. We proceed to analyze the impact of the change in the jth component of y on $\Delta\|y - y_S\|$ before and after the lth training round.

An additional note: in this subsection, when discussing classifications where a variable equals 0, the result may yield $\Delta\|y - y_S\| = 0$, which would not align with the intended conclusion. However, one component y_j causing $\Delta\|y - y_S\| = 0$ does not imply that every component of y causes $\Delta\|y - y_S\| = 0$. In fact, the probability of every component of y resulting in $\Delta\|y - y_S\| = 0$ is very low; therefore, such cases are not mentioned in the stated conclusion.

Based on the above setup and the basic principles of machine learning, we have:

$$x = (x_1, x_2, \ldots, x_{l_X}) \quad y = (y_1, y_2, \ldots, y_{l_Y}) \quad y_j = \sum_{k=1}^{d} W_{kj} x_k + b_j$$

According to the SGD update rule, after training this sample and updating the weights:

$$W_{kj} = W_{kj} - \eta \frac{\partial L}{\partial W_{kj}} = W_{kj} - \eta \frac{\partial L}{\partial y_j} x_k, \quad b_j = b_j - \eta \frac{\partial L}{\partial y_j}$$

Assume that after training, x_k is updated to $x_k + \Delta x_k$, then the updated y'_j is:

$$\begin{aligned} y'_j &= \sum_{i=1}^{d} \left(W_{ij} - \eta \frac{\partial L}{\partial y_j} x_i \right) (x_i + \Delta x_i) + b_j - \eta \frac{\partial L}{\partial y_j} \\ &= \sum_{i=1}^{d} W_{ij} x_i - \eta \frac{\partial L}{\partial y_j} \sum_{i=1}^{d} x_i^2 + \sum_{i=1}^{d} W_{ij} \Delta x_i - \eta \frac{\partial L}{\partial y_j} \sum_{i=1}^{d} x_i \Delta x_i + b_j - \eta \frac{\partial L}{\partial y_j} \\ &= y_j - \left(\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i \Delta x_i \right) \eta \frac{\partial L}{\partial y_j} + \sum_{i=1}^{d} W_{ij} \Delta x_i \end{aligned}$$

Thus, we have: $y'_j - y_j = -\left(\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i \Delta x_i \right) \eta \frac{\partial L}{\partial y_j} + \sum_{i=1}^{d} W_{ij} \Delta x_i$.

For convenience of notation, let:

1. $\Delta y_j = y'_j - y_j$.
2. $\Delta y_{j1} = -\left(\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i \Delta x_i \right) \eta \frac{\partial L}{\partial y_j}$, called the first change term.

   With a related assumption:

   Assumption 4: $\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i \Delta x_i > 0$
3. $\Delta y_{j2} = \sum_{i=1}^{d} W_{ij} \Delta x_i$, called the second change term.
4. Let the vector of sample A after the last layer of the neural network but before the softmax operation be denoted as Z, where $Z = (z_1, z_2, \ldots, z_{l_Z})$.
5. The probability vector output for sample A before the lth training round is denoted as P, with its ith component as p_i, representing the predicted probability that sample A belongs to category i.

Without loss of generality, assume that the feature vector y of sample A is in an a-type cluster ($a \in \{1, 2, \ldots, M\}$). According to Assumption 1, sample A belongs to one of the three types: A_1, A_2, or A_3.

We will examine each of these possibilities to determine the sign and magnitude of $\Delta \|y - y_S\|$ for each type of sample A during the lth training round.

Theorem 1: When sample A is of type A_1, $\Delta \|y - y_S\| > 0$ after the lth training round, and the smaller $\|y - y_S\|$ is, the larger $\Delta \|y - y_S\|$ becomes.

When point A is of type A_1, it is an a-type point but misclassified as type b. After several training rounds, we have $p_a > p_b \gg p_t$ for any $t \in \{1, 2, \ldots, M\}$ where $t \neq a, b$.

Therefore, $p_a + p_b \approx 1$.

At this point:

$$\frac{\partial L}{\partial y_j} = \frac{\partial L}{\partial p_b}\frac{\partial p_b}{\partial y_j} = -\frac{1}{p_b}\frac{\partial p_b}{\partial y_j} \approx -\frac{1}{1-p_a}\frac{\partial p_b}{\partial y_j}$$

Next, we will consider the effects of the first change term Δy_{j1} and the second change term Δy_{j2} on $\Delta\|y - y_S\|$.

For Δy_{j1}, we have the following intermediate conclusion:

Intermediate Conclusion 1: When point A is of type A_1, if only the first change term exists (i.e., $\Delta y_j = \Delta y_{j1}$), then $\Delta\|y - y_S\| > 0$ and the smaller $\|y - y_S\|$, the larger $\Delta\|y - y_S\|$.

Proof:

There are three possible relationships for $\frac{\partial p_b}{\partial y_j}$ with 0:

- $\frac{\partial p_b}{\partial y_j} > 0$ implies $\frac{\partial L}{\partial y_j} < 0$ (a), leading to $\Delta y_{j1} > 0$ (b), which implies $\Delta p_{b_j} > 0$ (c), resulting in $\Delta p_{a_j} < 0$ (d), leading to $\Delta\|y - y_S\| > 0$ (e). - $\frac{\partial p_b}{\partial y_j} < 0$ implies $\frac{\partial L}{\partial y_j} > 0$ (a), leading to $\Delta y_{j1} < 0$ (b), which implies $\Delta p_{b_j} > 0$ (c), resulting in $\Delta p_{a_j} < 0$ (d), leading to $\Delta\|y - y_S\| > 0$ (e). - $\frac{\partial p_b}{\partial y_j} = 0$ leads to $\Delta\|y - y_S\| = 0$.

(a) $\frac{\partial L}{\partial y_j} = -\frac{1}{1-p_a}\frac{\partial p_b}{\partial y_j}$ and $-\frac{1}{1-p_a} < 0$.

(b) $\Delta y_{j1} = -\left(\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i \Delta x_i\right)\eta\frac{\partial L}{\partial y_j}$, where $\left(\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i \Delta x_i\right) > 0$, and $\eta > 0$.

(c) $\Delta p_{b_j} = \frac{\partial p_b}{\partial y_j}\Delta y_j$, with $\frac{\partial p_b}{\partial y_j}$ and Δy_{j1} having the same sign.

(d) $p_{a_j} \approx 1 - p_{b_j} \Rightarrow \Delta p_{a_j} \approx -\Delta p_{b_j}$.

(e) According to Assumption 2 regarding the reference vector, a decrease in p_a directly results in an increase in the distance between the feature vector of sample A and the reference vector. Here, Δp_{a_j} and Δp_{b_j} are the components of y_j contributing to Δp_a and Δp_b.

Therefore, after the lth training round, Δy_{j1} always causes the feature vector of sample A to move further from the reference vector.

Since $|\frac{\partial L}{\partial y_j}| = \frac{1}{1-p_a}|\frac{\partial p_b}{\partial y_j}|$, $|\frac{\partial L}{\partial y_j}|$ is positively correlated with p_a, implying $|\Delta y_j|$ is positively correlated with p_a.

Thus, as p_a increases, $|\Delta y_j|$ increases $\Rightarrow$ as $\|y - y_S\|$ decreases, $\Delta\|y - y_S\|$ increases.

(f) According to Assumption 2, the larger p_a, the smaller $\|y - y_S\|$.

(g) $\Delta p_a \approx -\frac{\partial p_b}{\partial y_j}\Delta y_j$, so the larger $|\Delta y_j|$, the larger $|\Delta p_a|$. According to Assumption 2, the larger $|\Delta y_j|$, the larger $|\Delta\|y - y_S\|$—.

In summary, when point A is of type A_1, if only the first change term exists, $\Delta\|y - y_S\| > 0$ and the smaller $\|y - y_S\|$, the larger $\Delta\|y - y_S\|$.

For Δy_{j2}, we have:

Intermediate Conclusion 2: When A is of type A_1 and $h > 1$ (when $h = 1$, $\Delta y_{j2} = 0$; this case will be discussed later), if only the second change term exists, i.e., $\Delta y_j = \Delta y_{j2}$, then $\Delta\|y - y_S\| > 0$ and the smaller $\|y - y_S\|$, the larger $\Delta\|y - y_S\|$.

Proof: In this part of the discussion, it is essential to clarify that the vector x defined at the beginning of subsection x.2 is both the input to the hth layer and the output of the $(h-1)$th layer. Therefore, Δx_k also exists before and after the lth training round and can be divided into the first change term Δx_{k1} and the second change term Δx_{k2}.

$$\frac{\partial p_b}{\partial x_k} = \frac{\partial p_b}{\partial y_j}\frac{\partial y_j}{\partial x_k} = W_{kj}\frac{\partial p_b}{\partial y_j}$$

We discuss three cases, considering only the influence of the first change term Δx_{k1} while temporarily ignoring the second change term Δx_{k2}.

(1) When $\frac{\partial p_b}{\partial y_j} > 0$:

- If $W_{kj} > 0$, then $\frac{\partial p_b}{\partial x_k} > 0 \Rightarrow \frac{\partial L}{\partial x_k} < 0$ (a) $\Rightarrow \Delta x_{k1} > 0$ (b) $\Rightarrow W_{kj}\Delta x_k > 0$ when only considering the effect of the first change term on Δx_k. - If $W_{kj} < 0$, then $\frac{\partial p_b}{\partial x_k} < 0 \Rightarrow \frac{\partial L}{\partial x_k} > 0$ (a) $\Rightarrow \Delta x_{k1} < 0$ (b) $\Rightarrow W_{kj}\Delta x_k > 0$ when only considering the effect of the first change term on Δx_k. - If $W_{kj} = 0$, then $W_{kj}\Delta x_k = 0$.

(a) From the discussion in Intermediate Conclusion 1, $\frac{\partial L}{\partial x_k} \approx -\frac{1}{1-p_a}\frac{\partial p_b}{\partial x_k}$, and $-\frac{1}{1-p_a} < 0$.

(b) From the discussion in Intermediate Conclusion 1, Δx_{k1} has the opposite sign of $\frac{\partial L}{\partial x_k}$.

(c) W_{kj} and Δx_{k1} have the same sign.

Since W_{kj} is unlikely to be 0 for all k, we have:

$$\sum_{k=1}^{l_X} W_{kj}\Delta x_k > 0$$

Furthermore, $|\frac{\partial L}{\partial x_k}|$ increases as p_a increases, implying $|\Delta x_{k1}|$ increases as $|\frac{\partial L}{\partial x_k}|$ increases:

$$|\Delta x_{k1}| = \left(\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i \Delta x_i\right)\eta|\frac{\partial L}{\partial x_k}|$$

$|\Delta x_k|$ increases as p_a increases:

$$\Rightarrow \text{Since } p_a \text{ is inversely related to } \|y - y_S\| \text{ (by Assumption 4),}$$

$|\Delta x_k|$ increases as $\|y - y_S\|$ decreases:

$$\Rightarrow \sum_{k=1}^{l_X} W_{kj}\Delta x_k \text{ increases as } \|y - y_S\| \text{ decreases.}$$

$$\Rightarrow \Delta y_j \text{ increases as } \|y - y_S\| \text{ decreases.}$$

From the discussion in Intermediate Conclusion 1 regarding the case where $\frac{\partial p_b}{\partial y_j} > 0$, when $\frac{\partial p_b}{\partial y_j} > 0$, $\Delta y_j > 0$, causing $\Delta\|y - y_S\| > 0$, and an increase in Δy_j results in an increase in $\Delta\|y - y_S\|$.

Therefore, after a training session, $\Delta\|y - y_S\| > 0$, and the smaller $\|y - y_S\|$, the larger $\Delta\|y - y_S\|$.

(2) When $\frac{\partial p_b}{\partial y_j} < 0$:

- If $W_{kj} > 0$, then $\frac{\partial p_b}{\partial x_k} < 0 \Rightarrow \frac{\partial L}{\partial x_k} > 0$ (a) $\Rightarrow \Delta x_{k1} < 0$ (b) $\Rightarrow W_{kj}\Delta x_k < 0$ when only considering the effect of the first change term on Δx_k. - If $W_{kj} < 0$, then $\frac{\partial p_b}{\partial x_k} > 0 \Rightarrow \frac{\partial L}{\partial x_k} < 0$ (a) $\Rightarrow \Delta x_{k1} > 0$ (b) $\Rightarrow W_{kj}\Delta x_k < 0$ when only considering the effect of the first change term on Δx_k. - If $W_{kj} = 0$, then $W_{kj}\Delta x_k = 0$.

(a) From the discussion in Intermediate Conclusion 1, $\frac{\partial L}{\partial x_k} \approx -\frac{1}{1-p_a}\frac{\partial p_b}{\partial x_k}$, and $-\frac{1}{1-p_a} < 0$.

(b) From the discussion in Intermediate Conclusion 1, Δx_{k1} has the opposite sign of $\frac{\partial L}{\partial x_k}$.

(c) W_{kj} and Δx_{k1} have opposite signs.

Since W_{kj} is unlikely to be 0 for all k, we have:

$$\sum_{k=1}^{l_X} W_{kj}\Delta x_k < 0$$

Furthermore, $|\frac{\partial L}{\partial x_k}|$ increases as p_a increases, implying $|\Delta x_{k1}|$ increases as $|\frac{\partial L}{\partial x_k}|$ increases:

$$|\Delta x_{k1}| = \left(\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i\Delta x_i\right)\eta|\frac{\partial L}{\partial x_k}|$$

$|\Delta x_k|$ increases as p_a increases:

$$\Rightarrow \text{Since } p_a \text{ is inversely related to } \|y - y_S\| \text{ (by Assumption 4),}$$

$|\Delta x_k|$ increases as $\|y - y_S\|$ decreases:

$$\Rightarrow -\sum_{k=1}^{l_X} W_{kj}\Delta x_k \text{ increases as } \|y - y_S\| \text{ decreases.}$$

$$\Rightarrow |\Delta y_j| \text{ increases as } \|y - y_S\| \text{ decreases.}$$

From the discussion in Intermediate Conclusion 1 regarding the case where $\frac{\partial p_b}{\partial y_j} < 0$, when $\frac{\partial p_b}{\partial y_j} < 0$, $\Delta y_j < 0$, causing $\Delta\|y - y_S\| > 0$, and an increase in $|\Delta y_j|$ results in an increase in $\Delta\|y - y_S\|$.

Therefore, after a training session, $\Delta\|y - y_S\| > 0$, and the smaller $\|y - y_S\|$, the larger $\Delta\|y - y_S\|$.

(3) When $\frac{\partial p_b}{\partial y_j} = 0$:

In this case, $\Delta x_{k1} = 0$.

Now we consider the influence of the second change term Δx_{k2}:

1. When $h = 2$, $\Delta x_{k2} = 0$, so the influence of the second change term does not need to be considered.

2. When $h > 2$, based on the discussions of Δy_{j1} and Δy_{j2}, we know that the influence of the second change term on $\Delta\|y - y_S\|$ in terms of both its sign and magnitude is the same as that of the first change term. Similarly, Δx_{k1} and Δx_{k2} have the same directional effect on $\Delta\|y - y_S\|$.

Given the above analysis, we now consider the combined effects of Δy_{j1} and Δy_{j2}:

1. When $h = 1$, $\Delta y_{j2} = 0$. Only the first change term exists, so $\Delta\|y - y_S\| > 0$, and the smaller $\|y - y_S\|$, the larger $\Delta\|y - y_S\|$.

2. When $h \geq 2$, the effects of Δy_{j1} and Δy_{j2} on the sign and magnitude of $\Delta\|y - y_S\|$ are consistent. Thus, $\Delta\|y - y_S\| > 0$, and the smaller $\|y - y_S\|$, the larger $\Delta\|y - y_S\|$.

Proof completed.

Theorem 2: When point A is of type A_2, after the lth training round, $\Delta\|y - y_S\| < 0$, and the smaller $\|y - y_S\|$, the smaller $\Delta\|y - y_S\|$.

When point A is of type A_2, it is a c-type point but misclassified as type a. After several training rounds, we have $p_a > p_c \gg p_t$ for any $t \in \{1, 2, \ldots, M\}$ where $t \neq a, c$.

Therefore, $p_a + p_c \approx 1$.

At this point:

$$\frac{\partial L}{\partial y_j} = \frac{\partial L}{\partial p_a}\frac{\partial p_a}{\partial y_j} = -\frac{1}{p_a}\frac{\partial p_a}{\partial y_j}$$

Next, we will consider the effects of the first change term Δy_{j1} and the second change term Δy_{j2} on $\Delta\|y - y_S\|$.

For Δy_{j1}, we have the following intermediate conclusion:

Intermediate Conclusion 1: When point A is of type A_2, if only the first change term exists (i.e., $\Delta y_j = \Delta y_{j1}$), then $\Delta\|y - y_S\| < 0$, and the smaller $\|y - y_S\|$, the smaller $\Delta\|y - y_S\|$.

Proof:

There are three possible relationships for $\frac{\partial p_a}{\partial y_j}$ with 0:

- $\frac{\partial p_a}{\partial y_j} > 0$ implies $\frac{\partial L}{\partial y_j} < 0$ (a), leading to $\Delta y_{j1} > 0$ (b), which implies $\Delta p_{a_j} > 0$ (c), resulting in $\Delta\|y - y_S\| < 0$ (d). - $\frac{\partial p_a}{\partial y_j} < 0$ implies $\frac{\partial L}{\partial y_j} > 0$ (a), leading to $\Delta y_{j1} < 0$ (b), which implies $\Delta p_{a_j} > 0$ (c), resulting in $\Delta\|y - y_S\| < 0$ (d). - $\frac{\partial p_a}{\partial y_j} = 0$ leads to $\Delta\|y - y_S\| = 0$.

(a) $\frac{\partial L}{\partial y_j} = -\frac{1}{p_a}\frac{\partial p_a}{\partial y_j}$, and $-\frac{1}{p_a} < 0$.

(b) $\Delta y_{j1} = -\left(\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i \Delta x_i\right) \eta \frac{\partial L}{\partial y_j}$, where $\left(\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i \Delta x_i\right) > 0$, and $\eta > 0$.

(c) $\Delta p_{a_j} = \frac{\partial p_a}{\partial y_j}\Delta y_j$, with $\frac{\partial p_a}{\partial y_j}$ and Δy_{j1} having the same sign.

(d) According to Assumption 2 regarding the reference vector, an increase in p_a directly results in a decrease in the distance between the feature vector of sample A and the reference vector. Here, Δp_{a_j} and Δp_{b_j} are the components of y_j contributing to Δp_a and Δp_b.

Therefore, after the lth training round, Δy_{j1} always causes the feature vector of sample A to move closer to the reference vector.

Since $|\frac{\partial L}{\partial y_j}| = \frac{1}{p_a}|\frac{\partial p_a}{\partial y_j}|$, $|\Delta y_{j1}|$ is inversely correlated with p_a.

Thus, as p_a increases, $|\Delta y_j|$ decreases $\Rightarrow$ as $\|y - y_S\|$ decreases, $\Delta\|y - y_S\|$ decreases.

(f) According to Assumption 2, the larger p_a, the smaller $\|y - y_S\|$.

(g) $\Delta p_a \approx \frac{\partial p_a}{\partial y_j}\Delta y_j$, so the smaller $|\Delta y_j|$, the smaller $|\Delta p_a|$. According to Assumption 2, the smaller $|\Delta y_j|$, the smaller $|\Delta\|y - y_S\|$—.

In summary, when point A is of type A_2, if only the first change term exists, $\Delta\|y - y_S\| < 0$ and the smaller $\|y - y_S\|$, the smaller $\Delta\|y - y_S\|$.

For Δy_{j2}, we have the following intermediate conclusion:

Intermediate Conclusion 2: When point A is of type A_2 and $h > 1$ (when $h = 1$, $\Delta y_{j2} = 0$; this case will be discussed later), if only the second change term exists (i.e., $\Delta y_j = \Delta y_{j2}$), then $\Delta\|y - y_S\| < 0$ and the smaller $\|y - y_S\|$, the smaller $\Delta\|y - y_S\|$.

Proof:

The vector x is both the input to the hth layer and the output of the $(h-1)$th layer, so Δx_k exists before and after the lth training round and can be divided into the first change term Δx_{k1} and the second change term Δx_{k2}.

$$\frac{\partial p_a}{\partial x_k} = \frac{\partial p_a}{\partial y_j}\frac{\partial y_j}{\partial x_k} = W_{kj}\frac{\partial p_a}{\partial y_j}$$

We discuss three cases, considering only the influence of the first change term Δx_{k1} while temporarily ignoring the second change term Δx_{k2}, i.e., $\Delta x_k = \Delta x_{k1}$.

(1) When $\frac{\partial p_a}{\partial y_j} > 0$:

- If $W_{kj} > 0$, then $\frac{\partial p_a}{\partial x_k} > 0 \Rightarrow \frac{\partial L}{\partial x_k} < 0$ (a) $\Rightarrow \Delta x_{k1} > 0$ (b) $\Rightarrow W_{kj}\Delta x_k > 0$ when only considering the effect of the first change term on Δx_k. - If $W_{kj} < 0$, then $\frac{\partial p_a}{\partial x_k} < 0 \Rightarrow \frac{\partial L}{\partial x_k} > 0$ (a) $\Rightarrow \Delta x_{k1} < 0$ (b) $\Rightarrow W_{kj}\Delta x_k > 0$ when only considering the effect of the first change term on Δx_k. - If $W_{kj} = 0$, then $W_{kj}\Delta x_k = 0$.

(a) From Intermediate Conclusion 1, $\frac{\partial L}{\partial x_k} \approx -\frac{1}{p_a}\frac{\partial p_a}{\partial x_k}$ and $-\frac{1}{p_a} < 0$.

(b) From Intermediate Conclusion 1, Δx_{k1} has the opposite sign of $\frac{\partial L}{\partial x_k}$.

(c) W_{kj} and Δx_{k1} have the same sign.

Since W_{kj} is unlikely to be zero for all k, we have:

$$\sum_{k=1}^{l_X} W_{kj}\Delta x_k > 0$$

Additionally, $|\frac{\partial L}{\partial x_k}|$ decreases as p_a increases $\Rightarrow$ $|\Delta x_{k1}|$ decreases as $|\frac{\partial L}{\partial x_k}|$ decreases:

$$|\Delta x_{k1}| = \left(\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i\Delta x_i\right)\eta|\frac{\partial L}{\partial x_k}|$$

$|\Delta x_k|$ decreases as p_a increases:

$$\Rightarrow \text{Since } p_a \text{ is inversely related to } \|y - y_S\| \text{ (by Assumption 4),}$$

$|\Delta x_k|$ decreases as $\|y - y_S\|$ decreases:

$$\Rightarrow \sum_{k=1}^{l_X} W_{kj}\Delta x_k \text{ decreases as } \|y - y_S\| \text{ decreases.}$$

$$\Rightarrow \Delta y_j \text{ decreases as } \|y - y_S\| \text{ decreases.}$$

From the discussion in Intermediate Conclusion 1 regarding the case where $\frac{\partial p_a}{\partial y_j} > 0$, when $\frac{\partial p_a}{\partial y_j} > 0$, $\Delta y_j > 0$ results in $\Delta\|y - y_S\| < 0$, and a decrease in Δy_j results in a decrease in $\Delta\|y - y_S\|$.

Therefore, after a training session, $\Delta\|y - y_S\| < 0$ and the smaller $\|y - y_S\|$, the smaller $\Delta\|y - y_S\|$.

(2) When $\frac{\partial p_a}{\partial y_j} < 0$:

- If $W_{kj} > 0$, then $\frac{\partial p_a}{\partial x_k} < 0 \Rightarrow \frac{\partial L}{\partial x_k} > 0$ (a) $\Rightarrow \Delta x_{k1} < 0$ (b) $\Rightarrow W_{kj}\Delta x_k < 0$ when only considering the effect of the first change term on Δx_k. - If $W_{kj} < 0$, then $\frac{\partial p_a}{\partial x_k} > 0 \Rightarrow \frac{\partial L}{\partial x_k} < 0$ (a) $\Rightarrow \Delta x_{k1} > 0$ (b) $\Rightarrow W_{kj}\Delta x_k < 0$ when only considering the effect of the first change term on Δx_k. - If $W_{kj} = 0$, then $W_{kj}\Delta x_k = 0$.

(a) From Intermediate Conclusion 1, $\frac{\partial L}{\partial x_k} \approx -\frac{1}{p_a}\frac{\partial p_a}{\partial x_k}$ and $-\frac{1}{p_a} < 0$.

(b) From Intermediate Conclusion 1, Δx_{k1} has the opposite sign of $\frac{\partial L}{\partial x_k}$.

(c) W_{kj} and Δx_{k1} have opposite signs.

Since W_{kj} is unlikely to be zero for all k, we have:

$$\sum_{k=1}^{l_X} W_{kj}\Delta x_k < 0$$

Additionally, $|\frac{\partial L}{\partial x_k}|$ decreases as p_a increases $\Rightarrow$ $|\Delta x_{k1}|$ decreases as $|\frac{\partial L}{\partial x_k}|$ decreases:

$$|\Delta x_{k1}| = \left(\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i \Delta x_i\right)\eta|\frac{\partial L}{\partial x_k}|$$

$|\Delta x_k|$ decreases as p_a increases:

$$\Rightarrow \text{Since } p_a \text{ is inversely related to } \|y - y_S\| \text{ (by Assumption 4),}$$

$|\Delta x_k|$ decreases as $\|y - y_S\|$ decreases:

$$\Rightarrow -\sum_{k=1}^{l_X} W_{kj}\Delta x_k \text{ decreases as } \|y - y_S\| \text{ decreases.}$$

$$\Rightarrow |\Delta y_j| \text{ decreases as } \|y - y_S\| \text{ decreases.}$$

From the discussion in Intermediate Conclusion 1 regarding the case where $\frac{\partial p_a}{\partial y_j} < 0$, when $\frac{\partial p_a}{\partial y_j} < 0$, $\Delta y_j < 0$ results in $\Delta \|y - y_S\| < 0$, and a decrease in $|\Delta y_j|$ results in a decrease in $\Delta \|y - y_S\|$.

Therefore, after a training session, $\Delta \|y - y_S\| < 0$ and the smaller $\|y - y_S\|$, the smaller $\Delta \|y - y_S\|$.

Considering the influence of the second change term Δx_{k2}:

1. When $h = 1$, $\Delta x_k = 0$, so the influence of the second change term on Δy does not need to be considered.

2. When $h = 2$, the second change term of Δx_k is zero, so the influence of the second change term does not need to be considered.

3. When $h > 2$, the influence of the second change term on $\Delta \|y - y_S\|$ is consistent with the first change term in terms of both its sign and the effect on the magnitude.

Proof completed.

Theorem 3: When point A is of type A_3, $\Delta \|y - y_S\| \to 0$.

Proof: In this case, since point A is an a-type point and is correctly labeled as an a-type point, we can assume: $p_a \gg p_t$ for any $t \in [1, M]$ where $t \neq a$.

At this point:

$$\frac{\partial L}{\partial y_j} = \frac{\partial L}{\partial z_a}\frac{\partial z_a}{\partial y_j} + \sum_{t=1, t\neq a}^{M} \frac{\partial L}{\partial z_t}\frac{\partial z_t}{\partial y_j} = \frac{\partial z_a}{\partial y_j}(p_a - 1) + \sum_{t=1, t\neq a}^{M} p_t \frac{\partial z_t}{\partial y_j}$$

Since $p_a - 1 \to 0$ and $p_t \to 0$, each term in this expression is small and can have positive or negative signs that cancel out.

Thus, when point A is of type A_3, $\Delta \|y - y_S\| \to 0$.

Proof completed.

Based on the previous discussions, after the lth training round, for A_1 type points, $\Delta \|y - y_S\| > 0$; for A_2 type points, $\Delta \|y - y_S\| < 0$; and for A_3 type points, $\Delta \|y - y_S\| = 0$. To facilitate further calculations, we discuss the absolute magnitude of distance changes for these two types of points.

Theorem 4: Given that y_u is the feature vector of an A_1 type sample and y_v is the feature vector of an A_2 type sample, and if $\|y_u - y_S\| = \|y_v - y_S\|$, then $|\Delta \|y_u - y_S\|| > |\Delta \|y_v - y_S\||$.

Proof: Since $\|y_u - y_S\| = \|y_v - y_S\|$, by the properties of the reference point, the probability vector's ath component for U and V are equal, denoted as p_a.

$$\text{Since } U \text{ is an } A_1 \text{ type point, } \left|\frac{\partial L}{\partial y_{u_j}}\right| = \frac{1}{1 - p_a}\left|\frac{\partial p_b}{\partial y_{u_j}}\right| = \frac{1}{1 - p_a}\left|\frac{\partial p_a}{\partial y_{u_j}}\right|$$

$$\text{Since } V \text{ is an } A_2 \text{ type point, } \left|\frac{\partial L}{\partial y_{v_j}}\right| = \frac{1}{p_a}\left|\frac{\partial p_a}{\partial y_{v_j}}\right|$$

Assumption 6: For A_1 type points, after several rounds of training, $p_a > p_b \gg p_t$ for any $t \in \{1, 2, \ldots, M\}$ where $t \neq a, b$. Furthermore, for samples whose feature vectors are in the a-type cluster at the hth layer, $p_a > \frac{1}{2}$. For A_2 type points, after several rounds of training, $p_a > p_c \gg p_t$ for any $t \in \{1, 2, \ldots, M\}$ where $t \neq a, c$. Additionally, for samples whose feature vectors are in the a-type cluster at the hth layer, $p_a > \frac{1}{2}$.

Since $\frac{1}{1-p_a} > \frac{1}{p_a}$, considering only the distances, $\frac{\partial L}{\partial y_{u_j}} > \frac{\partial L}{\partial y_{v_j}}$.

According to discussions in subsections 2.1 and 2.3, $\frac{\partial L}{\partial y_{u_j}}$ and $\frac{\partial L}{\partial y_{v_j}}$ influence $|\Delta\|y_u - y_S\||$ and $|\Delta\|y_v - y_S\||$ in the same way.

Thus, $|\Delta\|y_u - y_S\|| > |\Delta\|y_v - y_S\||$.

Proof completed.

Summary of the Results After the lth Training Round:

1. For A_1 type points, $\Delta\|y_l - y_{l-1}\| > 0$ and the smaller $\|y_l - y_{l-1}\|$, the larger $\Delta\|y_l - y_{l-1}\|$.
2. For A_2 type points, $\Delta\|y_l - y_{l-1}\| < 0$ and the larger $\|y_l - y_{l-1}\|$, the larger $\Delta\|y_l - y_{l-1}\|$.
3. For A_3 type points, $\Delta\|y_l - y_{l-1}\| = 0$.
4. For points with the same distance from the reference point, the magnitude of the distance change for A_1 type points is always greater than that for A_2 type points.

1.3 Variation in LID of Sample Points Before and After the lth Training Round

In a federated learning system with noise, clients can be categorized as clean clients or noisy clients. Based on the discussions in subsection x.2, all samples in a clean client satisfy Theorem 3, while samples in a noisy client can be categorized into three types that satisfy Theorem 1, Theorem 2, and Theorem 3. In this subsection, we will discuss the average LID variation before and after each training round for these two types of clients after obtaining a preliminary model through several training rounds.

In client O, there are N samples divided into M categories.

Let $lid(x)$ denote the LID value of the xth sample in O before the lth training round, and $lid(O)$ represent the average LID of client O before the lth training round:

$$lid(O) = \frac{1}{N} \sum_{k=1}^{N} lid(x)$$

Let $lid'(x)$ denote the LID value of the xth sample in O after the lth training round, and $lid'(O)$ represent the average LID of client O after the lth training round:

$$lid'(O) = \frac{1}{N} \sum_{k=1}^{N} lid'(x)$$

Consider a sample T in O, and let $lid(T)$ and $lid'(T)$ denote its LID values before and after the lth training round, respectively.

Take two spheres centered at T with radii r_1 and r_2. Let N_1 and N_2 denote the number of points within the spheres of radius r_1 and r_2 before the lth training round, respectively. Let N_1' and N_2' denote the number of points within the spheres of radius r_1 and r_2 after the lth training round, respectively.

According to the definition of LID:

- $\left(\frac{r_2}{r_1}\right)^{lid(T)} = \frac{N_2}{N_1} \Rightarrow lid(T) = \frac{\ln N_2 - \ln N_1}{\ln r_2 - \ln r_1}$
- $\left(\frac{r_2}{r_1}\right)^{lid'(T)} = \frac{N_2'}{N_1'} \Rightarrow lid'(T) = \frac{\ln N_2' - \ln N_1'}{\ln r_2 - \ln r_1}$

To simplify notation, let:

$$\Delta lid(T) = lid'(T) - lid(T) \quad \Delta lid(O) = lid'(O) - lid(O)$$

Theorem 5: If O is a clean client, then $\Delta lid(O) = 0$.

Proof: When O is a clean client, all samples satisfy Theorem 3. Since the positions of feature vectors near T remain relatively fixed, we have: $N_1 = N_2$ and $N_1' = N_2'$.

Thus:

$$lid'(T) = lid(T) \Rightarrow \Delta lid(T) = lid'(T) - lid(T) = 0 \overset{a}{\Longrightarrow} \Delta lid(O) = 0$$

(a) Since T was an arbitrarily chosen sample, all samples in O satisfy $\Delta lid = 0$.

Proof completed.

Theorem 6: If O is a noisy client, then $\Delta lid(O) > 0$.

Proof:

Proof: Based on the four properties derived in subsection x.2, for a sample T within O, there are three possible outcomes for changes in the distance between T and its cluster's reference vector S before and after the lth training round: 1. $\Delta\|y_l - y_{l-1}\| > 0$ and the smaller $\|y_l - y_{l-1}\|$, the larger $\Delta\|y_l - y_{l-1}\|$. 2. $\Delta\|y_l - y_{l-1}\| < 0$ and the larger $\|y_l - y_{l-1}\|$, the larger $\Delta\|y_l - y_{l-1}\|$. 3. $\Delta\|y_l - y_{l-1}\| = 0$.

Let $r = \|y_l - y_{l-1}\|$ and $\Delta r = \Delta\|y_l - y_{l-1}\|$.

Considering only the relationship between distance change and initial distance: - For A_1 type points, $\Delta r = f(r)$, - For A_2 type points, $-\Delta r = h(r)$, where $f(r) > 0$ and $h(r) > 0$.

We examine the change in the LID value of S before and after the lth training round.

Let:

1. The proportion of A_1, A_2, and A_3 type points near S before the lth training round be $c : b : a$, with $c + b + a = 1$.

2. The LID of S before the lth training round be lid, and the LID after the lth training round be lid'.
3. N_1 and N_2 denote the number of points within spheres of radii r_1 and r_2 centered at S before the lth training round, respectively, and N_1' and N_2' denote the number of points after the lth training round.

According to the definition of LID:

$$\left(\frac{r_2}{r_1}\right)^{lid} = \frac{N_2}{N_1} \Rightarrow lid = \frac{\ln N_2 - \ln N_1}{\ln r_2 - \ln r_1}$$

$$N_1' = N_1 + \frac{4\pi r_1^2 h(r_1)}{\frac{4\pi r_1^3}{3}} bN_1 - \frac{4\pi r_1^2 f(r_1)}{\frac{4\pi r_1^3}{3}} cN_1$$

$$N_2' = N_2 + \frac{4\pi r_2^2 h(r_2)}{\frac{4\pi r_2^3}{3}} bN_2 - \frac{4\pi r_2^2 f(r_2)}{\frac{4\pi r_2^3}{3}} cN_2$$

$$\Rightarrow lid' = \frac{\ln N_2 \left(1 + \frac{3h(r_2)}{r_2} b - \frac{3f(r_2)}{r_2} c\right) - \ln N_1 \left(1 + \frac{3h(r_1)}{r_1} b - \frac{3f(r_1)}{r_1} c\right)}{\ln r_2 - \ln r_1}$$

$$= \frac{\ln \frac{N_2}{N_1} + \ln \frac{1+3\frac{h(r_2)}{r_2}b-3\frac{f(r_2)}{r_2}c}{1+3\frac{h(r_1)}{r_1}b-3\frac{f(r_1)}{r_1}c}}{\ln r_2 - \ln r_1} = lid + \frac{\ln\left(1 + \frac{\frac{h(r_2)}{r_2} - \frac{f(r_2)}{r_2} - \frac{h(r_1)}{r_1} + \frac{f(r_1)}{r_1}}{\frac{1}{3b} + \frac{h(r_1)}{r_1} - \frac{f(r_1)}{r_1}}\right)}{\ln r_2 - \ln r_1}$$

Assumption 5: The number of A_1 and A_2 type points near S is approximately equal, i.e., $b \approx c$.

Then:

$$lid' = lid + \frac{\ln\left(1 + \frac{\frac{h(r_2)}{r_2} - \frac{f(r_2)}{r_2} - \frac{h(r_1)}{r_1} + \frac{f(r_1)}{r_1}}{\frac{1}{3b} + \frac{h(r_1)}{r_1} - \frac{f(r_1)}{r_1}}\right)}{\ln r_2 - \ln r_1}$$

Since the magnitude of distance change is greater for A_1 type points than for A_2 type points at equal distances, $f(r_2) > h(r_2)$.

Also, since $r_1 < r_2$, $f(r_1) > f(r_2)$ and $h(r_1) < h(r_2)$.

$$\Rightarrow f(r_1) = f(r_2) + \Delta f = f + \Delta f, \quad h(r_1) = h(r_2) - \Delta h = h - \Delta h, \quad \Delta f, \Delta h > 0$$

$$\Rightarrow \frac{h(r_2)}{r_2} - \frac{f(r_2)}{r_2} - \frac{h(r_1)}{r_1} + \frac{f(r_1)}{r_1} = \frac{\Delta h + \Delta f}{r_1} + (f - h)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) > 0$$

$$\Rightarrow \ln\left(1 + \frac{\frac{h(r_2)}{r_2} - \frac{f(r_2)}{r_2} - \frac{h(r_1)}{r_1} + \frac{f(r_1)}{r_1}}{\frac{1}{3b} + \frac{h(r_1)}{r_1} - \frac{f(r_1)}{r_1}}\right) > 0$$

$$\Rightarrow \frac{\ln\left(1+\frac{3\frac{h(r_2)}{r_2}b-3\frac{f(r_2)}{r_2}c-3\frac{h(r_1)}{r_1}b+3\frac{f(r_1)}{r_1}c}{1+3\frac{h(r_1)}{r_1}b-\frac{f(r_1)}{r_1}c}\right)}{\ln r_2-\ln r_1} > 0$$

$$\Rightarrow lid' - lid > 0$$

A brief discussion on Assumption 5: While Assumption 5 may seem strict, the subsequent inequality derivations involve significant approximation, so it suffices if A_2 type points are not far more numerous than A_1 type points.

Thus, the LID value at S tends to increase after the lth training round. Within the same cluster, the dimensionality conditions are relatively uniform, so the change in the LID value at S can reflect the change for sample T. Since T was chosen arbitrarily, the LID value of any sample tends to increase after the lth training round.

Since each sample's LID value increases after the lth training round, the average LID value also increases.

Proof completed.

1.4 Verification of Assumptions

In the discussions above, we made six assumptions. In this subsection, we verify some of these assumptions using two experiments. The first experiment verifies Assumption 1, and the second experiment verifies Assumptions 4 and 5.

1.4.1 Experiment 1: Verification of Assumption 1

Objective: To verify that, after several rounds of training in a federated learning system, feature vectors gradually form clusters corresponding to their categories, and for noisy clients, samples can be divided into three types: A_1, A_2, and A_3.