1. Investigation of Noise Clients' LID Variations

In this chapter, we will explore the variations in the average Local Intrinsic Dimensionality (LID) of sample feature vectors during the training process of clean and noisy clients in a federated learning system. The exploration will proceed in the following order: First, in Section 1.1, we will outline the background and provide necessary definitions and basic assumptions. Next, in Section 1.2, we will discuss the variation in the distance between the sample feature vectors and predefined benchmark vectors. Then, in Section 1.3, based on the distance variation characteristics obtained in Section 1.2, we will examine the changes in the average LID of the sample feature vectors for clean and noisy clients during training. Finally, in Section 1.4, we will verify some of the hypotheses proposed in the previous sections through experiments.

1.1 Overview, Definitions, and Assumptions

Background: A client contains N samples, which are divided into M categories, and each sample is assigned a label representing its category. To process the client's data, we use an appropriate model for machine learning, with the model having l_M layers, and employ the cross-entropy loss function and stochastic gradient descent (SGD) for training. After several rounds of training, an initial model is generated.

Objective: Based on the above client, initial model, and training method, we aim to understand how the average Local Intrinsic Dimensionality (LID) of the feature vectors of the samples in the client changes before and after each round of training.

Definitions:

1. The output generated by the h-th layer of the model for sample X in the l-th round of training is called the feature vector of sample X at the h-th layer in the l-th round of training.

2. Let the set of all samples be $\{X_1, X_2, \ldots, X_N\}$, and the feature vector set at the h-th layer of the model in the l-th round of training be $\{V_1, V_2, \ldots, V_N\}$, where each V_j is the feature vector of X_j at that layer. If the feature vector has a length of l_V, all feature vectors are distributed in an l_V-dimensional space, referred to as the feature space of the h-th layer of the model in the l-th round of training. If a subset of feature vectors $\{V_{S_1}, V_{S_2}, \ldots, V_{S_n}\}$ (where $n \leq N$ and $S_1, S_2, \ldots, S_n \in \{1, 2, \ldots, N\}$) is concentrated in a limited region and significantly separated from feature vectors of other categories, it is said that $\{V_{S_1}, V_{S_2}, \ldots, V_{S_n}\}$ forms a cluster in the feature space.

3. In the feature space, feature vectors can be viewed as points, because mathematically, each vector represents a position. The values of the vector determine the point' s location in the feature space. For example, a 3-dimensional feature vector $[v_1, v_2, v_3]$

corresponds to a point (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, let sample A have a feature vector y at the h-th layer of the model in the l-th round of training. The benchmark vector of the cluster containing this feature vector is denoted by y_s (the specific definition of the benchmark vector is given in Assumption 2 in the Basic Assumptions). The distance between the feature vector of sample A and the benchmark vector is denoted by $\|y-y_s\|$. The change in this distance before and after the l-th round of training is denoted by $\Delta\|y - y_s\|$.

Basic Assumptions:

Assumption 1: After several rounds of training, the feature vectors of samples from each category gradually form corresponding clusters. The actual category and label category of the samples in a given cluster can only have three possibilities:

(a) Sample A has an actual category a and label category b. This sample is classified as type A_1.

(b) Sample A has an actual category c and label category a. This sample is classified as type A_2.

(c) Sample A has an actual category a and label category a. This sample is classified as type A_3.

Where $a, b, c \in \{1, 2, \ldots, M\}$ and $a \neq b$, $a \neq c$. This cluster is referred to as the a-type cluster.

It is assumed that each cluster has its own category, and each category has exactly one corresponding cluster, resulting in M distinct clusters, each corresponding to a different category.

Additionally, it can be concluded that:

- When the client is a clean client, all samples in the client are of type A_3.
- When the client is a noisy client, the samples include types A_1, A_2, and A_3.

Assumption 2: For the a-type cluster at the h-th layer of the model ($h \in \{1, 2, \ldots, l_M\}$, $a \in \{1, 2, \ldots, M\}$), after several rounds of training, we can extract a benchmark vector y_{sa}. For any two points A and B in this cluster, let their feature vectors be y_A and y_B. The final probability values for these points at the i-th item of the probability vector obtained from subsequent layers of the model are p_{iA} and p_{iB}. These satisfy the following conditions:

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

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

Assumption 3: In each round of training, for any point A in the dataset, its feature vector at the h-th layer is denoted by y_A, and the benchmark point of the cluster containing y_A is denoted by S, with the feature vector of S being 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 the Distance Between Points and the Benchmark Point in the Cluster During Training

Let sample point A have an input vector x at the h-th layer of the model before the l-th round of training, and an output vector y. The length of vector x is l_X, and the length of vector y is l_Y. The benchmark vector of the cluster containing the feature vector y_A is denoted by y_S.

In this subsection, we will discuss the sign and magnitude of $\Delta\|y - y_S\|$ during the l-th round of training. Since l, A, and h are arbitrary, this analysis will reflect the general behavior of $\Delta\|y - y_S\|$ during the training process.

Our approach is to analyze the impact of changes in y before and after the l-th round of training on $\Delta\|y - y_S\|$, focusing on two aspects: the sign (positive or negative) and the magnitude. Furthermore, we do not need to discuss the effect of all components of y, but rather focus on a specific component of y. If the effect of this component on $\Delta\|y - y_S\|$ is clear, such as always resulting in a positive change, then the overall impact of y on $\Delta\|y-y_S\|$ will also follow the same trend. Next, we will discuss the impact of the change in the j-th component of y on $\Delta\|y - y_S\|$ before and after the l-th round of training.

A preliminary note: In the discussion of this subsection, it is possible that one of the cases in a classification discussion results in a variable being zero, leading to a conclusion of $\Delta\|y - y_S\| = 0$, which may contradict the expected results. However, if a component y_j causes $\Delta\|y-y_S\| = 0$, it does not imply that every component of y results in $\Delta\|y-y_S\| = 0$. In fact, the probability that every component of y causes $\Delta\|y - y_S\| = 0$ is extremely low, so the situation of $\Delta\|y - y_S\| = 0$ is not addressed in the expected conclusions.

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 and updating the weights for this sam-

ple:

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

Let the updated x_k become $x_k + \Delta x_k$, 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, let us define:

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 variation term.

   We make the following assumption for the first variation term:

   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 variation term.
4. The vector Z is the output of sample A through the final layer of the neural network, but before softmax processing, with $Z = (z_1, z_2, \ldots, z_{l_Z})$.
5. The probability vector output by the neural network for sample A before the l-th round of training is denoted by P, with the i-th component denoted by 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 lies in the a-type cluster in its feature space ($a \in \{1, 2, \ldots, M\}$). According to Assumption 1, sample A belongs to one of the A_1, A_2, or A_3 types.

For these three possibilities, we will discuss the sign and magnitude of $\Delta \|y - y_S\|$ for each type of sample A during the l-th round of training.

Theorem 1: When the sample A is of class A_1, there is always $\Delta \|y - y_s\| > 0$ before and after the l-th round of training, and the smaller $\|y - y_s\|$ is, the larger $\Delta \|y - y_s\|$ becomes.

When point A is of class A_1, it is an A-class point, but it is misclassified as a B-class point. Therefore, after several rounds of training, we have $p_a > p_b \gg p_t$ ($t \in [1, M]$ and $t \neq a, b$).

Thus, $p_a + p_b \approx 1$.

At this point, we have:

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

Next, we consider the impact of the first variation term Δy_{j1} and the second variation term Δy_{j2} on $\Delta\|y - y_s\|$.

For Δy_{j1}, we have **Intermediate Conclusion 1**: When point A is of class A_1, if only the first variation term exists, i.e., $\Delta y_j = \Delta y_{j1}$, then $\Delta||y - y_s|| > 0$ and the smaller $\|y - y_s\|$, the larger $\Delta||y - y_s||$.

The proof is as follows:

There are three possible relations between $\frac{\partial p_b}{\partial y_j}$ and 0:

$\frac{\partial p_b}{\partial y_j} > 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial y_j} < 0 \overset{(b)}{\Longrightarrow} \Delta y_{j1} > 0 \overset{(c)}{\Longrightarrow} \Delta p_{b_j} > 0 \overset{(d)}{\Longrightarrow} \Delta p_{a_j} < 0 \overset{(e)}{\Longrightarrow} \Delta||y - y_s|| > 0$

$\frac{\partial p_b}{\partial y_j} < 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial y_j} > 0 \overset{(b)}{\Longrightarrow} \Delta y_{j1} < 0 \overset{(c)}{\Longrightarrow} \Delta p_{b_j} > 0 \overset{(d)}{\Longrightarrow} \Delta p_{a_j} < 0 \overset{(e)}{\Longrightarrow} \Delta||y - y_s|| > 0$

$\frac{\partial p_b}{\partial y_j} = 0 \overset{\text{Similarly to the above}}{\Longrightarrow} \Delta||y - y_s|| = 0$

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

(c) $\Delta p_{b_j} = \frac{\partial p_b}{\partial y_j}\Delta y_j$, and we know that $\frac{\partial p_b}{\partial y_j}$ and Δy_{j1} have 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 the definition of the reference vector in Hypothesis 2, a decrease in p_a will directly lead to an increase in the distance between the feature vector of sample A and the reference point. Here, Δp_{aj} and Δp_{bj} are the components that y_j contributes to Δp_a and Δp_b.

Therefore, before and after the l-th round of training, Δy_{j1} always causes the feature vector of sample A to move further away from the reference vector.

Since $\left|\frac{\partial L}{\partial y_j}\right| = \frac{1}{1-p_a}\left|\frac{\partial p_b}{\partial y_j}\right|$, we have $\left|\frac{\partial L}{\partial y_j}\right|$ positively correlated with p_a, which implies that $|\Delta y_j|$ is positively correlated with p_a.

Thus, the larger p_a is, the larger $|\Delta y_j|$ becomes, and consequently, $\|y - y_s\|$ becomes smaller, and $\Delta\|y - y_s\|$ becomes larger.

(f) According to the definition of the reference vector in Hypothesis 2, as p_a increases, $\|y - y_S\|$ becomes smaller.

(g) $\Delta p_a \approx -\frac{\partial p_b}{\partial y_j}\Delta y_j$, so the larger $|\Delta y_j|$, the larger $|\Delta p_a|$ becomes, which implies that

$|\Delta y_j|$ is positively correlated with $|\Delta\|y - y_S\||$.

In conclusion, when point A is of class A_1, if only the first variation term exists, we have $\Delta||y - y_s|| > 0$ and the smaller $\|y - y_s\|$, the larger $\Delta||y - y_s||$.

For Δy_{j2}, we have **Intermediate Conclusion 2**: When point A is of class A_1 and $h > 1$ (when $h = 1$, $\Delta y_{j2} = 0$, which will be discussed later), if only the second variation term exists, i.e., $\Delta y_j = \Delta y_{j2}$, then $\Delta||y-y_s|| > 0$ and the smaller $||y-y_s||$, the larger $\Delta||y-y_s||$.

The proof is as follows:

In this part of the discussion, one important point to note is that the vector x defined at the beginning of section 2.2 is both the input to the h-th layer network and the output of the $(h-1)$-th layer network. Therefore, before and after the l-th round of training, there also exists Δx_k, which can be divided into the first variation term Δx_{k1} and the second variation 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 will consider three cases, and for Δx_k, we only consider the effect of the first variation term Δx_{k1}, temporarily ignoring the second variation term Δx_{k2}.

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

When $w_{kj} > 0$, we have $\frac{\partial p_b}{\partial x_k} > 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial x_k} < 0 \overset{(b)}{\Longrightarrow} \Delta x_{k1} > 0$

$\overset{(c)}{\Longrightarrow}$ When considering only the first variation term, $w_{kj}\Delta x_k > 0$

When $w_{kj} < 0$, we have $\frac{\partial p_b}{\partial x_k} < 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial x_k} > 0 \overset{(b)}{\Longrightarrow} \Delta x_{k1} < 0$

$\overset{(c)}{\Longrightarrow}$ When considering only the first variation term, $w_{kj}\Delta x_k > 0$

When $w_{kj} = 0$, we have $w_{kj}\Delta x_k = 0$.

(a) According to 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 intermediate Conclusion 1, we know that Δx_{k1} has the opposite sign to $\frac{\partial L}{\partial x_k}$.

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

Since w_{kj} is almost never zero, we have $\sum_{k=1}^{l_x} w_{kj}\Delta x_k > 0$

Furthermore, since $\left|\frac{\partial L}{\partial x_k}\right|$ increases as p_a increases, we have: $|\Delta x_{k1}|$ increases as $\left|\frac{\partial L}{\partial x_k}\right|$ increases, i.e., $|\Delta x_k| \approx \left(\sum_{i=1}^{d} x_i^2 + 1 + \sum_{i=1}^{d} x_i\Delta x_i\right)\eta\left|\frac{\partial L}{\partial x_k}\right|$ Thus, $|\Delta x_k|$ increases as p_a increases.

This leads to the conclusion that p_a is negatively correlated with $\|y - y_s\|$ based on Assumption 4: $|\Delta x_k|$ increases as $\|y-y_s\|$ decreases. Therefore, $\sum_{k=1}^{l_x} w_{kj}\Delta x_k$ increases as $\|y-y_s\|$ decreases. Thus, Δy_j increases as $\|y - y_s\|$ decreases.

From the discussion of $\frac{\partial p_b}{\partial y_j} > 0$ in Intermediate Conclusion 1, when $\frac{\partial p_b}{\partial y_j} > 0$, we have $\Delta y_j > 0$, which makes $\Delta\|y - y_s\| > 0$, and as Δy_j increases, $\Delta\|y - y_s\|$ increases.

Therefore, after one round of training, we have $\Delta\|y-y_s\| > 0$, and the smaller $\|y-y_s\|$,

the larger $\Delta\|y - y_s\|$.

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

When $w_{kj} > 0$, we have $\frac{\partial p_b}{\partial x_k} < 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial x_k} > 0 \overset{(b)}{\Longrightarrow} \Delta x_{k1} < 0$

$\overset{(c)}{\Longrightarrow}$ When considering only the first variation term, $w_{kj}\Delta x_k < 0$

When $w_{kj} < 0$, we have $\frac{\partial p_b}{\partial x_k} > 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial x_k} < 0 \overset{(b)}{\Longrightarrow} \Delta x_{k1} > 0$

$\overset{(c)}{\Longrightarrow}$ When considering only the first variation term, $w_{kj}\Delta x_k < 0$

When $w_{kj} = 0$, we have $w_{kj}\Delta x_k = 0$.

(a) From the discussion in Intermediate Conclusion 1, we know

$$\frac{\partial L}{\partial x_k} \approx -\frac{1}{1-p_a}\frac{\partial p_b}{\partial x_k} \quad \text{and} \quad -\frac{1}{1-p_a} < 0$$

(b) From intermediate Conclusion 1, we know that Δx_{k1} has the opposite sign to $\frac{\partial L}{\partial x_k}$.

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

Since w_{kj} is almost never zero, we have $\sum_{k=1}^{l_x} w_{kj}\Delta x_k < 0$

Furthermore, since $\left|\frac{\partial L}{\partial x_k}\right|$ increases as p_a increases, we have:

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

Thus, $|\Delta x_k|$ increases as p_a increases.

This leads to the conclusion that p_a is negatively correlated with $\|y - y_s\|$ based on Assumption 4: $|\Delta x_k|$ increases as $\|y-y_s\|$ decreases. Hence, $-\sum_{k=1}^{l_x} w_{kj}\Delta x_k$ increases as $\|y-y_s\|$ decreases. Thus, $|\Delta y_j|$ increases as $\|y - y_s\|$ decreases.

From the discussion of $\frac{\partial p_b}{\partial y_j} < 0$ in Intermediate Conclusion 1, when $\frac{\partial p_b}{\partial y_j} < 0$, we have $\Delta y_j < 0$, which makes $\Delta\|y - y_s\| > 0$, and as $|\Delta y_{j2}|$ increases, $\Delta\|y - y_s\|$ increases.

Therefore, after one round of training, we have $\Delta\|y-y_s\| > 0$, and the smaller $\|y-y_s\|$, the larger $\Delta\|y - y_s\|$.

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

At this point, we have $\Delta x_{k1} = 0$.

On the basis of the above analysis, we have ignored the second variation term Δx_{k2}. We will now supplement the following considerations:

1. When $h = 2$, $\Delta x_{k2} = 0$, so there is no need to consider the effect of the second variation term.

2. When $h > 2$, based on the discussions in Intermediate Conclusion 1 and Intermediate Conclusion 2, the effects of Δy_{j1} and Δy_{j2} on the sign and magnitude of $\Delta\|y - y_s\|$ are of the same nature (by the same nature, we mean that both terms simultaneously

make $\Delta\|y-y_s\|$ positive or negative, and the effect of the magnitude of $\|y-y_s\|$ on $\Delta\|y-y_s\|$ is also clear). Similarly, the effects of Δx_{k1} and Δx_{k2} on the sign and magnitude of $\Delta\|y-y_s\|$ are also of the same nature.

Thus, we have separately discussed the effects of Δy_{j1} and Δy_{j2} on the sign and magnitude of $\Delta\|y-y_s\|$. We now combine the effects of Δy_{j1} and Δy_{j2}:

1. When $h=1$, $\Delta y_{j2}=0$. In this case, only the first variation term exists, and we have $\Delta\|y-y_s\|>0$, and as $\|y-y_s\|$ decreases, $\Delta\|y-y_s\|$ increases.

2. When $h\geq 2$, the effects of Δy_{j1} and Δy_{j2} on the sign and magnitude of $\Delta\|y-y_s\|$ are of the same nature, and we still have $\Delta\|y-y_s\|>0$, and as $\|y-y_s\|$ decreases, $\Delta\|y-y_s\|$ increases.

Q.E.D.

Theorem 2: When the point A is of class A_2, before and after the l-th round of training, there is always $\Delta\|y-y_s\|<0$, and the smaller $\|y-y_s\|$ is, the smaller $\Delta\|y-y_s\|$ becomes.

When point A is of class A_2, since it is actually a class c point, but is misclassified as a class a point, after several rounds of training, we have $p_a>p_c\gg P_t$ ($t\in\{1,2,\ldots,M\}$, and $t\neq a,c$).

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 consider the effects of the first variation term Δy_{j1} and the second variation term Δy_{j2} on $\Delta\|y-y_s\|$.

For Δy_{j1}, by **Intermediate Conclusion 1**: when point A is of class A_2, if only the first variation term exists, i.e., $\Delta y_j=\Delta y_{j1}$, we have $\Delta\|y-y_s\|<0$, and the smaller $\|y-y_s\|$ is, the smaller $\Delta\|y-y_s\|$ becomes.

The proof is as follows:

The relationship between $\frac{\partial p_a}{\partial y_j}$ and 0 can have 3 possibilities:

1. $\frac{\partial p_a}{\partial y_j}>0\overset{(a)}{\Longrightarrow}\frac{\partial L}{\partial y_j}<0\overset{(b)}{\Longrightarrow}\Delta y_{j1}>0\overset{(c)}{\Longrightarrow}\Delta p_{a_j}>0\overset{(d)}{\Longrightarrow}\Delta\|y-y_s\|<0.$

2. $\frac{\partial p_a}{\partial y_j}<0\overset{(a)}{\Longrightarrow}\frac{\partial L}{\partial y_j}>0\overset{(b)}{\Longrightarrow}\Delta y_{j1}<0\overset{(c)}{\Longrightarrow}\Delta p_{a_j}>0\overset{(d)}{\Longrightarrow}\Delta\|y-y_s\|<0.$

3. $\frac{\partial p_a}{\partial y_j}=0\overset{\text{Similar to the above}}{\Longrightarrow}\Delta\|y-y_s\|=0$

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

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

(c) $\Delta p_{a_j}=\frac{\partial p_a}{\partial y_j}\Delta y_j$, and since $\frac{\partial p_a}{\partial y_j}$ and Δy_j have the same sign.

(d) According to the definition of the reference vector in Assumption 2, an increase in p_a will directly cause the feature vector of sample A to move closer to the reference point. Here, Δp_{aj}, Δp_{bj} represent the contributions of y_j to Δp_a and Δp_b.

Therefore, before and after the l-th round of training, Δy_{j1} always causes the feature vector of A point to move closer to the reference point.

Also, since $\left|\frac{\partial L}{\partial y_j}\right| = \frac{1}{p_a}\left|\frac{\partial p_a}{\partial y_j}\right|$, we have that $|\Delta y_{j1}|$ is negatively correlated with p_a.

Thus, as p_a increases, $|\Delta y_j|$ decreases, and $\underset{(g)}{\overset{(f)}{\Longrightarrow}}$ $\|y - y_s\|$ decreases, and $\Delta\|y - y_s\|$ decreases.

(f) According to the definition of the reference vector in Assumption 2, as p_a increases, $\|y - y_s\|$ decreases.

(g) $\Delta p_a \approx \frac{\partial p_a}{\partial y_j}\Delta y_j$, hence $|\Delta y_j|$ decreases, $|\Delta p_a|$ decreases, and $\xRightarrow{\text{By Assumption 2}}$ $|\Delta y_j|$ decreases, and $\Delta\|y - y_s\|$ decreases.

In conclusion, when A point is of class A_2, if only the first variation term exists, we have $\Delta\|y - y_s\| < 0$, and the smaller $\|y - y_s\|$ is, the smaller $\Delta\|y - y_s\|$ becomes.

For Δy_{j2}, by **Intermediate Conclusion 2**: when A point is of class A_2 and $h > 1$ (when $h = 1$, $\Delta y_{j2} = 0$, discussed later), if only the second variation term exists, i.e., $\Delta y_j = \Delta y_{j2}$, we have $\Delta\|y - y_s\| < 0$, and the smaller $\|y - y_s\|$ is, the smaller $\Delta\|y - y_s\|$ becomes.

The proof is as follows:

The vector x is both the input to the h-th layer network and the output of the $(h-1)$-th layer network. Therefore, before and after the l-th round of training, Δx_k also exists, and it can be divided into the first variation term Δx_{k_1} and the second variation term Δx_{k_2}.

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

We discuss three cases, and for Δx_k, we only consider the impact of the first variation term Δx_{k1}, temporarily ignoring the second variation term Δx_{k2}, i.e., $\Delta x_k = \Delta x_{k1}$.

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

When $w_{kj} > 0$, we have $\frac{\partial p_a}{\partial x_k} > 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial x_k} < 0 \overset{(b)}{\Longrightarrow} \Delta x_{k1} > 0 \overset{(c)}{\Longrightarrow}$ When only considering the effect of the first variation term on Δx_k, we get $w_{kj}\Delta x_k > 0$.

When $w_{kj} < 0$, we have $\frac{\partial p_a}{\partial x_k} < 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial x_k} > 0 \overset{(b)}{\Longrightarrow} \Delta x_{k1} < 0 \overset{(c)}{\Longrightarrow}$ When only considering the effect of the first variation term on Δx_k, we get $w_{kj}\Delta x_k > 0$.

When $w_{kj} = 0$, we have $w_{kj}\Delta x_k = 0$.

(a) According to the discussion in 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) According to the discussion in Intermediate Conclusion 1, Δx_{k_1} is opposite in sign to $\frac{\partial L}{\partial x_k}$.

(c) w_{kj} and Δx_{k_1} have the same sign.

Since w_{kj} is almost never zero, we have: $\sum_{k=1}^{l_x} w_{kj}\Delta x_k > 0$.

Furthermore, as $|\frac{\partial L}{\partial x_k}|$ decreases as p_a increases, we have $|\Delta x_{k_1}|$ increases as $|\frac{\partial L}{\partial x_k}|$ increases. Also, $|\Delta x_k|$ decreases as p_a increases: $|\Delta x_k|$ decreases as p_a increases.

By Assumption 4, we have: $|\Delta x_k|$ decreases as $\|y - y_s\|$ decreases.

Thus, $\sum_{k=1}^{l_x} w_{kj}\Delta x_k$ decreases as $\|y-y_s\|$ decreases. $\Rightarrow \Delta y_j$ decreases as $\|y-y_s\|$ decreases.

According to the discussion in Intermediate Conclusion 1 for the case when $\frac{\partial p_a}{\partial y_j} > 0$, we have that when $\frac{\partial p_a}{\partial y_j} > 0$, $\Delta y_j > 0$ causes $\Delta\|y - y_s\| < 0$, and Δy_j decreases, causing $\Delta\|y - y_s\|$ to decrease.

Therefore, after one round of training, $\Delta\|y - y_s\| < 0$, and the smaller $\|y - y_s\|$ is, the smaller $\Delta\|y - y_s\|$ becomes.

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

When $w_{kj} > 0$, we have $\frac{\partial p_a}{\partial x_k} < 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial x_k} > 0 \overset{(b)}{\Longrightarrow} \Delta x_{k1} < 0 \overset{(c)}{\Longrightarrow}$ When only considering the effect of the first variation term on Δx_k, we get $w_{kj}\Delta x_k < 0$.

When $w_{kj} < 0$, we have $\frac{\partial p_a}{\partial x_k} > 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial x_k} < 0 \overset{(b)}{\Longrightarrow} \Delta x_{k1} > 0 \overset{(c)}{\Longrightarrow}$ When only considering the effect of the first variation term on Δx_k, we get $w_{kj}\Delta x_k < 0$.

When $w_{kj} = 0$, we have $w_{kj}\Delta x_k = 0$.

(a) According to the discussion in 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) According to the discussion in Intermediate Conclusion 1, Δx_{k_1} is opposite in sign to $\frac{\partial L}{\partial x_k}$.

(c) w_{kj} and Δx_{k_1} have opposite signs.

Since w_{kj} is almost never zero, we have: $\sum_{k=1}^{l_x} w_{kj}\Delta x_k < 0$.

Furthermore, as $|\frac{\partial L}{\partial x_k}|$ decreases as p_a increases, we have $|\Delta x_{k_1}|$ increases as $|\frac{\partial L}{\partial x_k}|$ increases. Also, $|\Delta x_k|$ decreases as p_a increases: $|\Delta x_k|$ decreases as p_a increases.

By Assumption 4, we have: $|\Delta x_k|$ decreases as $\|y - y_s\|$ decreases.

Thus, $|\sum_{k=1}^{l_x} w_{kj}\Delta x_k|$ decreases as $\|y - y_s\|$ decreases. $\Rightarrow$ $|\Delta y_j|$ decreases as $\|y - y_s\|$ decreases.

According to the discussion in Intermediate Conclusion 1 for the case when $\frac{\partial p_a}{\partial y_j} < 0$, we have that when $\frac{\partial p_a}{\partial y_j} < 0$, $\Delta y_j < 0$ causes $\Delta\|y - y_s\| > 0$, and $|\Delta y_j|$ decreases, causing $\Delta\|y - y_s\|$ to decrease.

Therefore, after one round of training, $\Delta\|y - y_s\| < 0$, and the smaller $\|y - y_s\|$ is, the smaller $\Delta\|y - y_s\|$ becomes.

On the basis of the above analysis, we have ignored the impact of the second variation term on Δx_k. We now make additional considerations:

1. When $h = 1$, $\Delta x_k = 0$, and even the impact of the second variation term on Δy does not need to be considered.

2. When $h = 2$, the second variation term for $\Delta x_k = 0$ is also zero, so the impact of the second variation term does not need to be considered.

3. When $h > 2$, from the discussion of Δy_{j1} and Δy_{j2}, it can be concluded that the impact of the second variation term on $\|y - y_s\|$ and $\Delta\|y - y_s\|$ is in the same direction as the first variation term.

Q.E.D.

Theorem 3: When the point A is an A_3-type point, $\Delta\|y - y_s\| \to 0$.

Proof: In this case, since the point A is an A-type point and is correctly labeled as an A-type point, we can assume: $p_a \gg p_t \quad (t \in [1, M] \text{ and } t \neq a)$

At this point, we have:

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

Since $p_a - 1 \to 0$ and $p_t \to 0$, each term in the above expression becomes very small and can have both positive and negative signs, easily canceling each other out.

Therefore, it can be concluded that when the point A is an A_3-type point, $\Delta\|y - y_s\| \to 0$ before and after the l-th round of training.

Based on previous discussions, before and after the l-th round of training, for A_1-type points, $\Delta\|y - y_s\| > 0$; for A_3-type points, $\Delta\|y - y_s\| < 0$. To facilitate later calculations, we now discuss the absolute values of the changes in the distances for these two types of points.

Theorem 4: When 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 $\|y_u - y_s\| = \|y_v - y_s\|$, 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 benchmark point, the a-th components of the probability vectors for U and V are equal, and we denote this value as p_a.

Since U belongs to theA_1type point, we have $\left|\frac{\partial L}{\partial y_{u_j}}\right| = \frac{1}{1-p_a}\left|\frac{\partial p_b}{\partial y_{u_j}}\right| = \frac{1}{1-p_a}\left|\frac{\partial p_a}{\partial y_{u_j}}\right|$

Since V belongs to theA_2type point, we have $\left|\frac{\partial L}{\partial y_{v_j}}\right| = \frac{1}{p_a}\left|\frac{\partial p_a}{\partial y_{v_j}}\right|$

Assumption 6: For A_1-type points, after a certain number of training rounds, we have $p_a > p_b \gg p_t$ (where $t \in \{1, 2, ..., M\}$ and $t \neq a, b$). Furthermore, for samples whose feature vectors in the h-th layer belong to the a-class cluster, we have $p_a > \frac{1}{2}$.

For A_3-type points, after a certain number of training rounds, we have $p_a > p_c \gg p_t$ (where $t \in \{1, 2, ..., M\}$ and $t \neq a, c$). Furthermore, for samples whose feature vectors in the h-th layer belong to the a-class cluster, we have $p_a > \frac{1}{2}$.

Then, we have: $\frac{1}{1-p_a} > \frac{1}{p_a}$, and in terms of distance alone, we have $\frac{\partial L}{\partial y_{u_j}} > \frac{\partial L}{\partial y_{v_j}}$.

According to results from 2.1 and 2.3, $\frac{\partial L}{\partial y_{u_j}}$ and $\frac{\partial L}{\partial y_{v_j}}$ affect $|\Delta\|y_u - y_s\||$ and $|\Delta\|y_v - y_s\||$ in the same manner.

Thus, we conclude that: $|\Delta\|y_u - y_s\|| > |\Delta\|y_v - y_s\||$.

Proof Complete.

In summary, before and after the l-th round of training:

(1) For A_1-type points, $\Delta\|y_l - y_{l-1}\| > 0$, and the smaller $\|y_l - y_{l-1}\|$ is, the larger $\Delta\|y_l - y_{l-1}\|$ becomes.

(2) For A_2-type points, $\Delta\|y_l - y_{l-1}\| < 0$, and the larger $\|y_l - y_{l-1}\|$ is, the larger $\Delta\|y_l - y_{l-1}\|$ becomes.

(3) For A_3-type points, $\Delta\|y_l - y_{l-1}\| = 0$.

(4) For A_1-type and A_2-type points that are at the same distance from the benchmark point, the distance change for the A_1-type point is always larger.

1.3 Changes in the Sample Point LID Before and After the l-th Round of Training

In a federated learning system with noise, clients can be divided into clean clients and noisy clients. According to the discussion in Section x.2, all samples in clean clients satisfy Theorem 3; while samples in noisy clients can be divided into three categories, each satisfying Theorem 1, Theorem 2, and Theorem 3, respectively. In this subsection, we will discuss the change in the average LID values of the clients before and after each round of training, after obtaining the initial model through several rounds of training, for these two types of clients.

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

Let the LID value of the x-th sample in O before the l-th round of training be $\text{lid}(x)$, and the average LID value of O before the l-th round of training be $\text{lid}(O)$. Then,

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

Let the LID value of the x-th sample in O after the l-th round of training be $\text{lid}'(x)$, and the average LID value of O after the l-th round of training be $\text{lid}'(O)$. Then,

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

Next, consider a specific sample T in O. Let the LID values before and after the l-th round of training be $\text{lid}(T)$ and $\text{lid}'(T)$, respectively.

Take two balls around point T with radii r_1 and r_2. The number of points within the

ball of radius r_1 before the l-th round of training is N_1, and the number of points within the ball of radius r_2 is N_2. After the l-th round of training, the number of points within the ball of radius r_1 is N_1', and the number of points within the ball of radius r_2 is N_2'.

Based on the definition of LID:

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

For convenience, define:

$$\Delta\text{lid}(T) = \text{lid}'(T) - \text{lid}(T), \quad \Delta\text{lid}(O) = \text{lid}'(O) - \text{lid}(O).$$

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

Proof: When O is a clean client, all its samples satisfy Theorem 2. Since the positions of the feature vectors around T are relatively fixed, we have $N_1 = N_2$ and $N_1' = N_2'$.

Thus,

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

(a) Since T is an arbitrary sample, all samples in O satisfy $\Delta\text{lid} = 0$.

Q.E.D.

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

Proof: According to the four properties obtained in Section x.2, for sample T, the feature vectors around the reference vector S of its cluster can have three possibilities before and after the l-th round of training: 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 the change in distance and the distance, for A_1-class points, $\Delta r = f(r)$; for A_3-class points, $-\Delta r = h(r)$. Thus, we know that $f(r) > 0$ and $h(r) > 0$.

Next, we discuss the change in the LID value of the reference vector S before and after the l-th round of training.

Assume:

1. The proportion of A_1, A_2, and A_3 points around S before the l-th round of training is $c : b : a$, with $c + b + a = 1$.

2. The LID of S before the l-th round of training is lid, and the LID of point T after the l-th round of training is lid$'$.

3. Consider two balls with radii r_1 and r_2 around S. The number of points within the ball of radius r_1 before the l-th round of training is N_1, and the number of points within the ball of radius r_2 is N_2. After the l-th round of training, the number of points within the ball of radius r_1 is N_1', and the number of points within the ball of radius r_2 is N_2'.

According to the definition of LID:

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

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

$$N_2' = N_2 + \frac{4\pi r_2^2 h(r_1)}{\frac{4\pi r_2^3}{3}} b N_2 - \frac{4\pi r_2^2 f(r_2)}{\frac{4\pi r_2^3}{3}} c N_2$$

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

$$= \frac{ln\frac{N_2}{N_1} + ln\frac{[1+3\frac{h(r_2)}{r_2}b - 3\frac{f(r_2)}{r_2}c]}{1+3\frac{h(r_1)}{r_1}b - 3\frac{f(r_1)}{r_1}c}}{lnr_2 - lnr_1} = lid + \frac{ln(1 + \frac{3\frac{h(r_2)}{r_2}b - 3\frac{f(r_2)}{r_2}c - 3\frac{h(r_1)}{r_1}b + 3\frac{f(r_1)}{r_1}c}{1+3\frac{h(r_1)}{r_1}b - \frac{f(r_1)}{r_1}c})}{lnr_2 - lnr_1}$$

Assumption 5: The numbers of A_1-class points and A_2-class points around S are roughly equal, i.e., $b \approx c$.

Then, we have:

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

Since the change magnitude for A_1-class points is larger than for A_3-class points when the distance from the reference point is equal, we have $f(r_2) > h(r_2)$.

Also, because $r_1 < r_2$, we have $f(r_1) > f(r_2)$ and $h(r_1) < h(r_2)$.

Thus, let $f(r_1) = f(r_2) + \Delta f = f + \Delta f$, $h(r_1) = h(r_2) - \Delta h = h - \Delta h$, with $\Delta f, \Delta h > 0$.

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

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

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

$$\Rightarrow \mathrm{lid}'-\mathrm{lid}>0$$

Let us briefly discuss Assumption 5. Although Assumption 5 appears strict, in fact, during the subsequent inequality derivations, there is a considerable amount of scaling, so it is actually sufficient for the number of A_2-class points to not be significantly greater than that of A_1-class points.

Therefore, the LID value of S shows an increasing trend before and after the l-th round of training. Within the same cluster, the dimensional characteristics are relatively uniform, and the change in S's LID can reflect the change in T's LID. Since T is an arbitrary sample point, the LID value of any sample point shows an increasing trend before and after the l-th round of training.

Since the LID of each sample point shows an increasing trend before and after the l-th round of training, the average LID also increases before and after the l-th round of training.

Q.E.D.