1. 对噪声客户端 LID 变化情况的探究

在本章中，我们将探讨联邦学习系统中的干净客户端和噪声客户端在训练过程中样本特征向量平均 LID 变化的情况。探究的顺序如下：首先，在 x.1 节中，我们将概述背景情况，并给出必要的定义与基本假设；接着，在 x.2 节中，我们将讨论样本点特征向量与预定义基准向量之间距离的变化情况；然后，在 x.3 节中，我们将基于 x.2 节中得到的距离变化特性，讨论训练过程中干净客户端和噪声客户端样本特征向量平均 LID 的变化；最后，在 x.4 节中，我们将通过实验验证前几节中提出的一些假设。

1.1 概述、定义与假设

背景：在一个客户端中，存有 N 个样本，这些样本被划分为 M 个类别，每个样本都被分配了一个标签以表示其所属类别。为了处理客户端的数据，我们使用合适的模型进行机器学习，模型的层数为 l_M，并采用交叉熵损失函数和随机梯度下降 (SGD) 进行训练。经过若干轮训练后，已生成了一个初步的模型。

目标：在上述客户端、初步模型及训练方法的基础上，我们希望了解在随后的每轮训练前后，客户端中样本的特征向量的平均局部内在维度（LID）值的变化情况。

定义：

1. 样本 X 在第 l 轮训练前，通过模型的第 h 层所生成的输出，称为样本 X 在第 l 轮训练中模型第 h 层的特征向量。

2. 记所有样本的集合为 $\{X_1, X_2, ..., X_N\}$，这些样本在第 l 轮训练中模型第 h 层的特征向量集合为 $\{V_1, V_2, ..., V_N\}$，其中每个 V_j 是 X_j 在该层的特征向量。若特征向量长度为 l_V，则所有的特征向量都分布在一个 l_V 维的空间中，称该空间为第 l 轮训练模型第 h 层的特征空间。$\{V_1, V_2, ..., V_N\}$ 分布在该特征空间中，若特征向量集的某个子集 $\{V_{S_1}, V_{S_2}, ..., V_{S_n}\}$ (n ≤ N 且 S_1,S_2,...,S_n ∈ {1, 2,..., N}) 集中在一个有限区域内，且与其他类别的特征向量显著分离，则称 $\{V_{S_1}, V_{S_2}, ..., V_{S_n}\}$ 在特征空间中形成了一个集群。

3. 在特征空间中，特征向量可以被视为点，这是因为在数学上，每个向量可以表示一个位置。特征向量的数值决定了这个点在特征空间中的位置。例如，一个三维特征向量 $[v_1, v_2, v_3]$ 对应特征空间中的一个坐标为 (v_1, v_2, v_3) 的点。在本章节中，但凡称特征向量为点时，均依照此定义。

4. 在样本集合中，任取一个样本 A，样本 A 在第 l 轮训练中模型第 h 层的特征向量为 y，该特征向量所处集群的基准向量为 y_s (基准向量的具体定义见基本假设中的假设 2)。则 A 点第 h 层特征向量与基准向量的距离记作 $\|y - y_s\|$。且第 l 轮训练前后，A 点第 h 层特征向量与基准向量之间的距离变化记作 $\Delta\|y - y_s\|$。

基本假设：

假设 1: 在训练过程中，经过若干轮次的训练后，各类样本的特征向量逐渐形成相应的集群，若某集群中的特征向量所对应的样本的实际类别与标签类别仅有三种可能性：

(a) 样本 A 实际类别为 a，标签类别为 b。把这类样本记作 A_1 类样本；

(b) 样本 A 实际类别为 c，标签类别为 a。把这类样本记作 A_2 类样本。

(c) 样本 A 实际类别为 a，标签类别为 a。把这类样本记作 A_3 类样本；

其中 a,b,c $\in \{1, 2, ..., M\}$ 且 $a \neq b$，$a \neq c$。则称该集群为 a 类集群。

假设每个集群均有所属类别，且每个类别仅有 1 个集群对应；即共形成 M 个个集群且集群类别两两不同。

同时可以得知：

- 当客户端为干净客户端时，其中样本均为 A_3 类样本。
- 当客户端为噪声客户端时，其中样本包括 A_1,A_2,A_3 三类。

假设 2: 针对模型第 h 层的第 a 类集群 ($h \in \{1, 2, \ldots, l_M\}$, $a \in \{1, 2, \ldots, M\}$)，在经过若干轮训练得到初步模型后，我们可以取出一个基准向量为 y_{sa}。在该集群中任取另外两点 A, B，这两点对应的特征向量分别为 y_A 和 y_B。且这两点通过后续模型层得到最终的概率向量的第 i 项分别为 p_{iA} 和 p_{iB}。满足：

(a) 若 $p_{iA} > p_{iB}$，则 $\|y_A - y_{S_a}\| < \|y_B - y_{S_a}\|$。

(b) 若 $p_{iA} < p_{iB}$，则 $\|y_A - y_{S_a}\| > \|y_B - y_{S_a}\|$。

假设 3: 每轮训练中，在数据集中任取一点 A，其在第 h 层的特征向量记作 y_A；y_A 所属集群的基准点记为 S，其特征向量记作 y_S。假设 $\Delta\|y_A - y_S\|$ 仅与 $\|y_A - y_S\|$ 有关，即 $\Delta\|y_A - y_S\| = f(\|y_A - y_S\|)$

1.2 训练时集群中各点与基准点距离的变化

设某一样本点 A 在第 l 轮训练前的模型第 h 层的输入向量为 x，输出向量为 y。其中，向量 x 的长度为 l_X，向量 y 的长度为 l_Y。特征向量 y_A 所在集群基准向量为 y_S。

在本小节中，我们讨论第 l 轮训练中 $\Delta\|y - y_S\|$ 的正负及大小。由于 l、A、h 都是任意取值的，因此这能够反映训练过程中 $\Delta\|y - y_S\|$ 的一般情况。

我们采用的方法是分析第 l 轮训练前后 y 的变化对 $\Delta\|y - y_S\|$ 的影响，这里的影响分为两方面，即正负和绝对值的大小。更进一步，我们无需讨论 y 中所有项的影响，而可以选择其中任意一项进行讨论。如果该项对 $\Delta\|y - y_S\|$ 的影响是明确的，比如一定导致其为正，那么整体 y 对 $\Delta\|y - y_S\|$ 的影响也是如此。接下来我们讨论 y 的第 j 项在第 l 轮训练前后的变化对 $\Delta\|y - y_S\|$ 的影响。

再做一点提前的补充：在本子节的讨论中，可能会出现分类讨论时其中一种分类为一个变量等于 0，这种分类推导出的结果是 $\Delta\|y-y_S\|=0$，会与预设结论不符合。但 y 中的一项 y_j 导致 $\Delta\|y-y_S\|=0$ 并不意味着 y 中的每一项都使 $\Delta\|y-y_S\|=0$。事实上，y 中每一项均使 $\Delta\|y-y_S\|=0$ 的概率极低，故而在预设结论中并不会提及 $\Delta\|y-y_S\|=0$ 的情况。

根据上述设定及机器学习基本原理，有：

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

根据 SGD 更新原理，在这个样本进行训练并更新权重后：

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

再设训练后 x_k 更新成了 $x_k+\Delta x_k$，则更新后的 y_j'：

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

则有：$y_j'-y_j=-\left(\sum_{i=1}^{d}x_i^2+1+\sum_{i=1}^{d}x_i\Delta x_i\right)\eta\frac{\partial L}{\partial y_j}+\sum_{i=1}^{d}w_{ij}\Delta x_i$

为方便表述，设：

1. $\Delta y_j=y_j'-y_j$。
2. $\Delta y_{j1}=-\left(\sum_{i=1}^{d}x_i^2+1+\sum_{i=1}^{d}x_i\Delta x_i\right)\eta\frac{\partial L}{\partial y_j}$，称为第一变化项。
   并针对第一变化项作出假设：
   假设 4： $\sum_{i=1}^{d}x_i^2+1+\sum_{i=1}^{d}x_i\Delta x_i>0$
3. $\Delta y_{j2}=\sum_{i=1}^{d}w_{ij}\Delta x_i$，称为第二变化项。
4. 样本 A 经过神经网络最后一层但未进行 softmax 处理的向量为 Z，且有 $Z=(z_1,z_2,\ldots,z_{l_Z})$。
5. 样本 A 在第 l 轮训练前经过神经网络输出的概率向量记为 P，其中第 i 项记为 p_i，表示表示样本 A 属于第 i 类别的预测概率。

不失一般性的，设样本 A 的特征向量 y 在其特征空间中处在 a 类集群 ($a \in \{1,2,...,M\}$)。根据假设 1 可知，样本 A 是 A_1,A_2,A_3 三类样本中的一类。

对于这 3 种可能性，我们依次给出每类样本 A 在第 l 轮训练中，$\Delta\|y-y_S\|$ 的正负与大小。

定理 1：当样本 A 是 A_1 类样本时，在第 l 轮训练前后，总是有 $\Delta\|y-y_s\|>0$ 且 $\|y-y_s\|$ 越小，$\Delta\|y-y_s\|$ 越大。

当 A 点为 A_1 类点时，因为其本身是 a 类点，但被误判为 b 类点。所以经由若干轮训练后，有 $p_a > p_b >> p_t$ ($t \in [1,M]$ 且 $t \neq a,b$)

故 $p_a + p_b \approx 1$。

此时：

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

接下来我们分别考虑第一变化项 Δy_{j1} 和第二变化项 Δy_{j2} 对 $\Delta\|y-y_s\|$ 的影响。

对于 Δy_{j1}，有**中间结论 1：**当 A 点为 A_1 类点时，若仅存在第一变化项，即 $\Delta y_j = \Delta y_{j1}$，有 $\Delta||y-y_s|| > 0$ 且 $||y-y_s||$ 越小，$\Delta||y-y_s||$ 越大。

证明如下：

$\frac{\partial p_b}{\partial y_j}$ 与 0 的关系有 3 种可能：

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

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

$\frac{\partial p_b}{\partial y_j} = 0 \xRightarrow{\text{与上同理}} \Delta||y-y_s|| = 0$

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

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

(c) $\Delta p_{b_j} = \frac{\partial p_b}{\partial y_j}\Delta y_j$，且知 $\frac{\partial p_b}{\partial y_j}$ 和 Δy_{j_1} 同号。

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

(e) 根据假设 2 中对基准向量的定义，p_a 减小，会直接导致样本 A 的特征向量与基准点间的距离增大。此处的 Δp_{aj},Δp_{bj} 是 y_j 对 Δp_a 和 Δp_b 贡献的分量。

故在第 l 轮训练前后，Δy_{j_1} 总使得样本 A 的特征向量向远离基准向量移动。

又因为 $|\frac{\partial L}{\partial y_j}| = \frac{1}{1-p_a}|\frac{\partial p_b}{\partial y_j}|$，所以 $|\frac{\partial L}{\partial y_j}|$ 和 p_a 正相关，$\xRightarrow{(b)}$ $|\Delta y_j|$和p_a 正相关。

所以，p_a 越大，$|\Delta y_j|$ 越大 $\xRightarrow[(g)]{(f)}$ $||y-y_s||$ 越小，$\Delta||y-y_s||$ 越大。

(f) 根据假设 2 中对基准向量的定义，p_a 越大，$\|y-y_S\|$ 越小。

(g)$\Delta p_a \approx -\frac{\partial p_b}{\partial y_j}\Delta y_j$, 故 $|\Delta y_j|$ 越大，$|\Delta p_a|$ 越大 $\xRightarrow{\text{根据假设 2}}$ $|\Delta y_j|$ 越大，$|\Delta\|y-y_S\||$ 越大。

综上，当 A 点为 A_1 类点时，若仅存在第一变化项，有 $\Delta||y-y_s|| > 0$ 且 $||y-y_s||$ 越小，$\Delta||y-y_s||$ 越大。

对于 Δy_{j2}，有**中间结论 2**：当 A 点为 A_1 类点且 h>1(h=1 时 $\Delta y_{j2}=0$，后续再讨论) 时，若仅存在第二变化项，即 $\Delta y_j=\Delta y_{j2}$，有 $\Delta||y-y_s||>0$ 且 $||y-y_s||$ 越小，$\Delta||y-y_s||$ 越大。

证明如下：

在这部分讨论中，需要明确的一点是 x.2 子节开头定义的向量 x 既是第 h 层网络的输入，同时也是第 (h-1) 层网络的输出，因此在第 l 轮训练前后也存在 Δx_k，并且也可以分为第一变化项 Δx_{k_1} 和第二变化项 Δx_{k_2} 两部分。

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

我们分三种情况讨论，且对于 Δx_k，只考虑第一变化项 Δx_{k1} 的影响，暂时忽略第二变化项 Δx_{k2}。

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

当 $w_{kj}>0$ 时，$\frac{\partial p_b}{\partial x_k}>0\overset{(a)}{\Longrightarrow}\frac{\partial L}{\partial x_k}<0\overset{(b)}{\Longrightarrow}\Delta x_{k1}>0\overset{(c)}{\Longrightarrow}$ 在仅考虑第一变化项对 Δx_k 的影响时 $w_{kj}\Delta x_k>0$

当 $w_{kj}<0$ 时，$\frac{\partial p_b}{\partial x_k}<0\overset{(a)}{\Longrightarrow}\frac{\partial L}{\partial x_k}>0\overset{(b)}{\Longrightarrow}\Delta x_{k1}<0\overset{(c)}{\Longrightarrow}$ 在仅考虑第一变化项对 Δx_k 的影响时 $w_{kj}\Delta x_k>0$

当 $w_{kj}=0$ 时，$w_{kj}\Delta x_k=0$

(a) 根据中间结论 1 中的讨论，$\frac{\partial L}{\partial x_k}\approx-\frac{1}{1-p_a}\frac{\partial p_b}{\partial x_k}$ 且 $-\frac{1}{1-p_a}<0$。

(b) 根据中间结论 1 中的讨论，Δx_{k_1} 与 $\frac{\partial L}{\partial x_k}$ 符号相反。

(c) w_{kj} 与 Δx_{k_1} 同号。

因为 w_{kj} 几乎不可能均为 0，故有: $\sum_{k=1}^{l_x}w_{kj}\Delta x_k>0$

更进一步而言，$|\frac{\partial L}{\partial x_k}|$ 随 p_a 增大而增大 $\xrightarrow[|\Delta x_{k_1}|随|\frac{\partial L}{\partial x_k}|增大而增大]{|\Delta x_{k_1}|=\left(\sum_{i=1}^{d}x_i^2+1+\sum_{i=1}^{d}x_i\Delta x_i\right)\eta|\frac{\partial L}{\partial x_k}|}$ $|\Delta x_k|$ 随 p_a 增大而增大 $\xrightarrow[p_a与||y-y_s||负相关]{根据假设\ 4}$ $|\Delta x_k|$ 随 $\|y-y_s\|$ 减小而增大 $\Rightarrow\sum_{k=1}^{l_x}w_{kj}\Delta x_k$ 随$\|y-y_s\|$ 减小而增大。 $\Rightarrow\Delta y_j$随$\|y-y_s\|$减小而增大。

根据中间结论 1 中对 $\frac{\partial p_b}{\partial y_j}>0$ 情况的讨论，当 $\frac{\partial p_b}{\partial y_j}>0$ 时，$\Delta y_j>0$ 使得 $\Delta\|y-y_s\|>0$ 且 Δy_j 增大使得 $\Delta\|y-y_s\|$ 增大。

因此，一次训练后 $\Delta\|y-y_s\|>0$ 且 $\|y-y_s\|$ 越小，$\Delta\|y-y_s\|$ 越大。

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

当 $w_{kj}>0$ 时，$\frac{\partial p_b}{\partial x_k}<0\overset{(a)}{\Longrightarrow}\frac{\partial L}{\partial x_k}>0\overset{(b)}{\Longrightarrow}\Delta x_{k1}<0\overset{(c)}{\Longrightarrow}$ 在仅考虑第一变化项对 Δx_k 的影响时 $w_{kj}\Delta x_k<0$

当 $w_{kj}<0$ 时，$\frac{\partial p_b}{\partial x_k}>0\overset{(a)}{\Longrightarrow}\frac{\partial L}{\partial x_k}<0\overset{(b)}{\Longrightarrow}\Delta x_{k1}>0\overset{(c)}{\Longrightarrow}$ 在仅考虑第一变化项对 Δx_k 的影响时 $w_{kj}\Delta x_k<0$

当 $w_{kj}=0$ 时，$w_{kj}\Delta x_k=0$

(a) 根据中间结论 1 中的讨论，$\frac{\partial L}{\partial x_k}\approx-\frac{1}{1-p_a}\frac{\partial p_b}{\partial x_k}$ 且 $-\frac{1}{1-p_a}<0$。

(b) 根据中间结论 1 中的讨论，Δx_{k_1} 与 $\frac{\partial L}{\partial x_k}$ 符号相反。

(c) w_{kj} 与 Δx_{k_1} 异号。

因为 w_{kj} 几乎不可能均为 0，故有: $\sum_{k=1}^{l_x} w_{kj}\Delta x_k < 0$

更进一步而言，$|\frac{\partial L}{\partial x_k}|$ 随 p_a 增大而增大 $\xrightarrow[|\Delta x_{k_1}|随|\frac{\partial L}{\partial x_k}|增大而增大]{|\Delta x_{k_1}|=\left(\sum_{i=1}^{d} x_i^2+1+\sum_{i=1}^{d} x_i\Delta x_i\right)\eta|\frac{\partial L}{\partial x_k}|}$ $|\Delta x_k|$ 随 p_a 增大而增大 $\xrightarrow[p_a与\|y-y_s\|负相关]{根据假设\ 4}$ $|\Delta x_k|$ 随 $\|y-y_s\|$ 减小而增大 $\Rightarrow$ -$\sum_{k=1}^{l_x} w_{kj}\Delta x_k$ 随$\|y-y_s\|$ 减小而增大。 $\Rightarrow |\Delta y_j|$随$\|y-y_s\|$减小而增大。

根据中间结论 1 中对 $\frac{\partial p_b}{\partial y_j} < 0$ 情况的讨论，当 $\frac{\partial p_b}{\partial y_j} < 0$ 时，$\Delta y_j < 0$ 使得 $\Delta\|y-y_s\| > 0$ 且 $|\Delta y_{j2}|$ 增大使得 $\Delta\|y-y_s\|$ 增大。

因此，一次训练后 $\Delta\|y-y_s\| > 0$ 且 $\|y-y_s\|$ 越小，$\Delta\|y-y_s\|$ 越大。

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

此时 $\Delta x_{k1} = 0$

在以上分析的基础上，我们忽略了第二变化项 Δx_{k2}。我们现在做补充考虑:

1. 当 h=2 时，$\Delta x_{k2} = 0$，不需要考虑第二变化项的影响。

2. 当 h>2 时，根据中间结论 1 和中间结论 2 的讨论，Δy_{j1} 和 Δy_{j2} 对 $\Delta\|y-y_s\|$ 的正负和大小的影响是同一性质的 (这里的同一性质指同时使 $\Delta\|y-y_s\|$ 为正或为负且 $\|y-y_s\|$ 的大小对 $\Delta\|y-y_s\|$ 大小的影响也是明确的)。同理可知，Δx_{k1} 和 Δx_{k2} 对 $\Delta\|y-y_s\|$ 的正负和大小的影响也是同一性质的。

至此，我们分别讨论了 Δy_{j1} 和 Δy_{j2} 对 $\Delta\|y-y_s\|$ 正负及大小的影响。我们在此基础上综合考虑 Δy_{j1} 和 Δy_{j2}:

1. 当 $h = 1$ 时，$\Delta y_{j2} = 0$。此时仅存在第一变化项，有 $\Delta\|y-y_s\| > 0$ 且 $\|y-y_s\|$ 越小，$\Delta\|y-y_s\|$ 越大。

2. 当 $h \geq 2$ 时，Δy_{j1} 和 Δy_{j2} 对 $\Delta\|y-y_s\|$ 正负及大小的影响性质是相同的，依然有 $\Delta\|y-y_s\| > 0$ 且 $\|y-y_s\|$ 越小，$\Delta\|y-y_s\|$ 越大。

证毕。

定理 2: 当 A 点是 A_2 类点时，在第 l 轮训练前后，总是有 $\Delta\|y-y_s\| < 0$ 且 $\|y-y_s\|$ 越小，$\Delta\|y-y_s\|$ 越小。

当 A 点为 A_2 类点时，因为其本身是 c 类点，但被误判为 a 类点。所以经由若干轮训练后，有 $p_a > p_c >> P_t$ ($t \in 1, 2, ..., M$ 且 $t \neq a, c$)

故 $p_a + p_c \approx 1$。

此时:

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

接下来我们分别考虑第一变化项 Δy_{j1} 和第二变化项 Δy_{j2} 对 $\Delta\|y-y_s\|$ 的影响。

对于 Δy_{j1}，有**中间结论 1:** 当 A 点为 A_2 类点时，若仅存在第一变化项，即 $\Delta y_j = \Delta y_{j1}$，有 $\Delta||y-y_s|| < 0$ 且 $||y-y_s||$ 越小，$\Delta||y-y_s||$ 越小。

证明如下：

$\frac{\partial p_a}{\partial y_j}$ 与 0 的关系有 3 种可能性：

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

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

$\frac{\partial p_a}{\partial y_j} = 0 \overset{\text{与上同理}}{\Longrightarrow} \Delta||y - y_s|| = 0$

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

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

(c) $\Delta p_{a_j} = \frac{\partial p_a}{\partial y_j}\Delta y_j$，且知 $\frac{\partial p_a}{\partial y_j}$ 和 Δy_j 同号。

(d) 根据假设 2 中对基准向量的定义，p_a 增大，会直接导致样本 A 的特征向量与基准点间的距离减小。此处的 Δp_{aj},Δp_{bj} 是 y_j 对 Δp_a 和 Δp_b 贡献的分量。

故在第 l 轮训练前后，Δy_{j1} 总使得 A 点的特征向量向接近基准点移动。

又因为 $|\frac{\partial L}{\partial y_j}| = \frac{1}{p_a}|\frac{\partial p_a}{\partial y_j}|$，所以 $|\Delta y_{j_1}|$ 和 p_a 负相关

所以，p_a 越大，$|\Delta y_j|$ 越小 $\underset{(g)}{\overset{(f)}{\Longrightarrow}}$ $||y - y_s||$ 越小，$\Delta||y - y_s||$ 越小。

(f) 根据假设 2 中对基准向量的定义，p_a 越大，$\|y - y_S\|$ 越小。

(g)$\Delta p_a \approx \frac{\partial p_a}{\partial y_j}\Delta y_j$, 故 $|\Delta y_j|$ 越小，$|\Delta p_a|$ 越小 $\overset{\text{根据假设 2}}{\Longrightarrow}$ $|\Delta y_j|$ 越小，$\Delta\|y - y_S\|$ 越小。

综上，当 A 点为 A_2 类点时，若仅存在第一变化项，有 $\Delta||y - y_s|| > 0$ 且 $||y - y_s||$ 越小，$\Delta||y - y_s||$ 越小。

对于 Δy_{j2}，有**中间结论 2**：当 A 点为 A_2 类点且 h>1(h=1 时 $\Delta y_{j2} = 0$，后续再讨论) 时，若仅存在第二变化项，即 $\Delta y_j = \Delta y_{j2}$，有 $\Delta||y - y_s|| < 0$ 且 $||y - y_s||$ 越小，$\Delta||y - y_s||$ 越小。

证明如下：

向量 x 既是第 h 层网络的输入，同时也是第 (h-1) 层网络的输出，因此在第 l 轮训练前后也存在 Δx_k，并且也可以分为第一变化项 Δx_{k_1} 和第二变化项 Δx_{k_2} 两部分。

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

我们分三种情况讨论，且对于 Δx_k，只考虑第一变化项 Δx_{k1} 的影响，暂时忽略第二变化项 Δx_{k2}, 即 $\Delta x_k = \Delta x_{k1}$。

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

当 $w_{kj} > 0$ 时，$\frac{\partial p_a}{\partial x_k} > 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial x_k} < 0 \overset{(b)}{\Longrightarrow} \Delta x_{k1} > 0 \overset{(c)}{\Longrightarrow}$ 在仅考虑第一变化项对 Δx_k 的影响时 $w_{kj}\Delta x_k > 0$

当 $w_{kj} < 0$ 时，$\frac{\partial p_a}{\partial x_k} < 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial x_k} > 0 \overset{(b)}{\Longrightarrow} \Delta x_{k1} < 0 \overset{(c)}{\Longrightarrow}$ 在仅考虑第一变化项对 Δx_k 的影响时 $w_{kj}\Delta x_k > 0$

当 $w_{kj} = 0$ 时，$w_{kj}\Delta x_k = 0$

(a) 根据中间结论 1 中的讨论，$\frac{\partial L}{\partial x_k} \approx -\frac{1}{p_a}\frac{\partial p_a}{\partial x_k}$ 且 $-\frac{1}{p_a} < 0$。

(b) 根据中间结论 1 中的讨论，Δx_{k_1} 与 $\frac{\partial L}{\partial x_k}$ 符号相反。

(c) w_{kj} 与 Δx_{k_1} 同号。

因为 w_{kj} 几乎不可能均为 0，故有: $\sum_{k=1}^{l_x} w_{kj}\Delta x_k > 0$。

更进一步而言，$|\frac{\partial L}{\partial x_k}|$ 随 p_a 增大而减小 $\xrightarrow[|\Delta x_{k_1}|\text{随}|\frac{\partial L}{\partial x_k}|\text{增大而增大}]{|\Delta x_{k_1}|=\left(\sum_{i=1}^{d} x_i^2+1+\sum_{i=1}^{d} x_i\Delta x_i\right)\eta|\frac{\partial L}{\partial x_k}|}$ $|\Delta x_k|$ 随 p_a 增大而减小 $\xrightarrow[p_a\text{与}\|y-y_s\|\text{负相关}]{\text{根据假设 4}}$ $|\Delta x_k|$ 随 $\|y-y_s\|$ 减小而减小 $\Rightarrow \sum_{k=1}^{l_x} w_{kj}\Delta x_k$ 随$\|y-y_s\|$ 减小而减小。 $\Rightarrow \Delta y_j$随$\|y-y_s\|$减小而减小。

根据中间结论 1 中对 $\frac{\partial p_a}{\partial y_j} > 0$ 情况的讨论，当 $\frac{\partial a_b}{\partial y_j} > 0$ 时，$\Delta y_j > 0$ 使得 $\Delta\|y-y_s\| < 0$ 且 Δy_j 减小使得 $\Delta\|y-y_s\|$ 减小。

因此，一次训练后 $\Delta\|y-y_s\| < 0$ 且 $\|y-y_s\|$ 越小，$\Delta\|y-y_s\|$ 越小。

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

当 $w_{kj} > 0$ 时，$\frac{\partial p_a}{\partial x_k} < 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial x_k} > 0 \overset{(b)}{\Longrightarrow} \Delta x_{k1} < 0 \overset{(c)}{\Longrightarrow}$ 在仅考虑第一变化项对 Δx_k 的影响时 $w_{kj}\Delta x_k < 0$

当 $w_{kj} < 0$ 时，$\frac{\partial p_a}{\partial x_k} > 0 \overset{(a)}{\Longrightarrow} \frac{\partial L}{\partial x_k} < 0 \overset{(b)}{\Longrightarrow} \Delta x_{k1} > 0 \overset{(c)}{\Longrightarrow}$ 在仅考虑第一变化项对 Δx_k 的影响时 $w_{kj}\Delta x_k < 0$

当 $w_{kj} = 0$ 时，$w_{kj}\Delta x_k = 0$

(a) 根据中间结论 1 中的讨论，$\frac{\partial L}{\partial x_k} \approx -\frac{1}{p_a}\frac{\partial p_a}{\partial x_k}$ 且 $-\frac{1}{p_a} < 0$。

(b) 根据中间结论 1 中的讨论，Δx_{k_1} 与 $\frac{\partial L}{\partial x_k}$ 符号相反。

(c) w_{kj} 与 Δx_{k_1} 异号。

因为 w_{kj} 几乎不可能均为 0，故有: $\sum_{k=1}^{l_x} w_{kj}\Delta x_k < 0$。

更进一步而言，$|\frac{\partial L}{\partial x_k}|$ 随 p_a 增大而减小 $\xrightarrow[|\Delta x_{k_1}|\text{随}|\frac{\partial L}{\partial x_k}|\text{增大而增大}]{|\Delta x_{k_1}|=\left(\sum_{i=1}^{d} x_i^2+1+\sum_{i=1}^{d} x_i\Delta x_i\right)\eta|\frac{\partial L}{\partial x_k}|}$ $|\Delta x_k|$ 随 p_a 增大而减小 $\xrightarrow[p_a\text{与}\|y-y_s\|\text{负相关}]{\text{根据假设 4}}$ $|\Delta x_k|$ 随 $\|y-y_s\|$ 减小而减小 $\Rightarrow |\sum_{k=1}^{l_x} w_{kj}\Delta x_k|$ 随$\|y-y_s\|$ 减小而减小。 $\Rightarrow |\Delta y_j|$随$\|y-y_s\|$减小而减小。

根据中间结论 1 中对 $\frac{\partial p_a}{\partial y_j} < 0$ 情况的讨论，当 $\frac{\partial p_a}{\partial y_j} < 0$ 时，$\Delta y_j < 0$ 使得 $\Delta\|y-y_s\| > 0$ 且 $|\Delta y_j|$ 减小使得 $\Delta\|y-y_s\|$ 减小。

因此，一次训练后 $\Delta\|y-y_s\| < 0$ 且 $\|y-y_s\|$ 越小，$\Delta\|y-y_s\|$ 越小。

在以上分析的基础上，我们忽略了第二变化项对 Δx_k 的影响。我们现在做补充考虑:

1. 当 h=1 时，$\Delta x_k = 0$，甚至连 Δy 都不需要考虑第二变化项的影响。
2. 当 h=2 时，$\Delta x_k = 0$ 的第二变化项为 0，也不需要考虑第二变化项的影响。
3. 当 h>2 时,通过对 Δy_{j1} 和 Δy_{j2} 的讨论可知,第二变化项对于 $\|y-y_s\|$ 和 $\Delta\|y-y_s\|$ 的影响，均与第一变化项的影响同一方向。

证毕

定理 3：当 A 点为 A_3 类点时，$\Delta\|y-y_s\|\to 0$。

证明：在这种情况下，因为 A 点是 a 类点, 且被正确标记为 a 类点。故可假设 $p_a >> p_t(t\in[1,M]且t\neq a)$

此时，

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

因为 $p_a-1\to 0$ 且 $p_t\to 0$，所以上式中每一项都很小且正负均有可能，极易相抵。

因此可认为当 A 点为 A_3 类点时，第 l 轮训练前后，$\Delta\|y-y_s\|\to 0$

通过之前的讨论，在 l 轮训练前后，对于 A_1 类点，$\Delta\|y-y_s\|>0$；对于 A_3 类点，$\Delta\|y-y_s\|<0$。所以，为了方便后续计算，我们讨论一下这两类点距离变化的绝对值的大小。

定理 4：当 y_u 为 A_1 类样本的特征向量，当 y_v 为 A_2 类样本的特征向量，且有 $\|y_u-y_s\|=\|y_v-y_s\|$，则 $|\Delta\|y_u-y_s\||>|\Delta\|y_v-y_s\||$。

证明：因为 $\|y_u-y_s\|=\|y_v-y_s\|$，所以由基准点的性质，U 与 V 概率向量的第 a 项相等，设为 p_a。

因为 U 属于A_1类点，故 $\left|\frac{\partial L}{\partial y_{u_j}}\right|=\frac{1}{1-p_a}\left|\frac{\partial p_b}{\partial y_{u_j}}\right|=\frac{1}{1-p_a}\left|\frac{\partial p_a}{\partial y_{u_j}}\right|$

因为 V 属于A_2类点，故 $\left|\frac{\partial L}{\partial y_{v_j}}\right|=\frac{1}{p_a}\left|\frac{\partial p_a}{\partial y_{v_j}}\right|$

假设 6: 对于 A_1 类点,经过一定轮次的训练后,有 $p_a>p_b\gg p_t(t\in\{1,2,...,M\}且t\neq a,b)$。更进一步的，对于第 h 层特征向量处于 a 类集群中的样本，均有 $p_a>1/2$。对于 A_3 类点，经过一定轮次的训练后，有 $p_a>p_c\gg p_t(t\in\{1,2,...,M\}且t\neq a,c)$。更进一步的，对于第 h 层特征向量处于 a 类集群中的样本，均有 $p_a>1/2$。

则 $\frac{1}{1-p_a}>\frac{1}{p_a}$，在仅考虑距离的情况下，有 $\frac{\partial L}{\partial y_{u_j}}>\frac{\partial L}{\partial y_{v_j}}$。

根据 2.1 与 2.3，$\frac{\partial L}{\partial y_{u_j}}$ 与 $\frac{\partial L}{\partial y_{v_j}}$ 以同样的方式影响 $|\Delta\|y_u-y_s\||$ 与 $|\Delta\|y_v-y_s\||$。

故有，$|\Delta\|y_u-y_s\||>|\Delta\|y_v-y_s\||$

证毕

综上所述，在第 l 轮训练前后：

(1) 对于 A_1 类点, $\Delta\|y_l-y_{l-1}\|>0$ 且 $\|y_l-y_{l-1}\|$ 越小, $\Delta\|y_l-y_{l-1}\|$ 越大。

(2) 对于 A_2 类点, $\Delta\|y_l-y_{l-1}\|<0$ 且 $\|y_l-y_{l-1}\|$ 越大, $\Delta\|y_l-y_{l-1}\|$ 越大。

(3) 对于 A_3 类点, $\Delta\|y_l-y_{l-1}\|=0$。

(4) 对于与基准点距离相同的 A_1 类点与 A_2 类点，总是 A_1 类点的距离变化幅度更大。

1.3 样本点 LID 在第 l 轮训练前后的的变化

在一个存在噪声的联邦学习系统中，客户端可分为干净客户端和噪声客户端。根据 x.2 中的讨论，干净客户端中样本全部满足定理 3；噪声客户端中样本则可分为三类，分别满足定理 1，定理 2，定理 3。在本子节中，我们分别讨论这两类客户端在经过若干轮次的训练得到初步模型后，每轮训练前后客户端平均 LID 值的变化。

在客户端 O 中，存在 N 个样本，这些样本被划分为 M 个类别。

记 O 中第 x 个样本在第 l 轮训练前的 LID 值为 lid(x)，O 在第 l 轮训练前的平均 LID 值为 lid(O)，则有 $lid(O) = \frac{1}{N}\sum_{k=1}^{N} lid(x)$。

再记 O 中第 x 个样本在第 l 轮训练后的 LID 值为 lid'(x)，O 在第 l 轮训练后的平均 LID 值为 lid'(O)，则有 $lid'(O) = \frac{1}{N}\sum_{k=1}^{N} lid'(x)$。

讨论 O 中某一样本 T，设其在第 l 轮训练前后 LID 值分别为 lid(T) 及 lid'(T)。

取 T 点附近半径为 r_1 和 r_2 的两个球。第 l 轮训练前半径为 r_1 的球内点数量为 N_1，半径为 r_2 的球内点数量为 N_2。第 l 轮训练后半径为 r_1 的球内点数量为 N_1'，半径为 r_2 的球内点数量为 N_2'。

根据 LID 的定义：

- $(\frac{r_2}{r_1})^{lid(T)} = \frac{N_2}{N_1} \Rightarrow lid(T) = \frac{lnN_2 - lnN_1}{lnr_2 - lnr_1}$
- $(\frac{r_2}{r_1})^{lid'(T)} = \frac{N_2'}{N_1'} \Rightarrow lid'(T) = \frac{lnN_2' - lnN_1'}{lnr_2 - lnr_1}$

为方便表述，令 $\Delta lid(T) = lid'(T) - lid(T) \quad \Delta lid(O) = lid'(O) - lid(O)$

定理 5：若 O 为干净客户端，$\Delta lid(O) = 0$。

证明：当 O 为干净客户端时，其中样本全部满足定理 2。因为 T 的特征向量附近的特征向量位置相对固定，故有 $N_1 = N_2$且$N_1' = N_2'$。

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

(a) 因为 T 为任取样本，故 O 中所有样本满足 $\Delta lid = 0$。

证毕。

定理 6：若 O 为噪声客户端，$\Delta lid(O) > 0$。

证明：根据 x.2 中得到的 4 条性质，对于样本 T，其所处集群的基准向量 S 附近特征向量，在第 l 轮训练前后有三种可能：1.$\Delta\|y_l - y_{l-1}\| > 0$ 且 $\|y_l - y_{l-1}\|$ 越小, $\Delta\|y_l - y_{l-1}\|$ 越大。2.$\Delta\|y_l - y_{l-1}\| < 0$ 且 $\|y_l - y_{l-1}\|$ 越大, $\Delta\|y_l - y_{l-1}\|$ 越大。3.$\Delta\|y_l - y_{l-1}\| = 0$。

令 $r = \|y_l - y_{l-1}\|, \Delta r = \Delta\|y_l - y_{l-1}\|$

只考虑距离变化与距离之间的关系, 对于 A_1 类点 $\Delta r = f(r)$；对于 A_3 类点 $-\Delta r = h(r)$。则可知 $f(r) > 0, h(r) > 0$。

我们讨论基准向量 S 的 LID 值在第 l 轮前后的变化。

设:

1. 第 l 轮训练前 S 附近 A_1,A_2,A_3 这三类点的比例为 c:b:a 且 c+b+a = 1。

2. 第 l 轮训练前 S 的 LID 为 lid，第 l 轮训练后 T 点的 LID 为 lid'。

3. 取 S 附近半径为 r_1 和 r_2 的两个球。第 l 轮训练前半径为 r_1 的球内点数量为 N_1，半径为 r_2 的球内点数量为 N_2。第 l 轮训练后半径为 r_1 的球内点数量为 N_1'，半径为 r_2 的球内点数量为 N_2'

根据 LID 的定义：$(\frac{r_2}{r_1})^{lid} = \frac{N_2}{N_1} \Rightarrow lid = \frac{lnN_2 - lnN_1}{lnr_2 - lnr_1}$

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

$$N_2' = N_2 + \frac{4\pi r_2^2 h(r_1)}{\frac{4\pi r_2^3}{3}} bN_2 - \frac{4\pi r_2^2 f(r_2)}{\frac{4\pi r_2^3}{3}} cN_2$$

$$\Rightarrow lid' = \frac{lnN_2\left(1 + \frac{3h(r_2)}{r_2}b - \frac{3f(r_2)}{r_2}c\right) - lnN_1\left(1 + \frac{3h(r_1)}{r_1}b - \frac{3f(r_1)}{r_1}c\right)}{lnr_2 - lnr_1}$$

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

假设 5: 在 S 附近，A_1 类点与 A_2 类点数量相近，即 b≈c。
则有，

$$lid' = lid + \frac{ln(1 + \frac{\frac{h(r_2)}{r_2} - \frac{f(r_2)}{r_2} - \frac{h(r_1)}{r_1} + \frac{f(r_1)}{r_1}}{\frac{1}{3b} + \frac{h(r_1)}{r_1} - \frac{f(r_1)}{r_1}})}{lnr_2 - lnr_1}$$

因为与基准点距离相等时，A_1 类点变化幅度大于 A_3 类点，则有 $f(r_2) > h(r_2)$。
又因为 $r_1 < r_2$, 所以 $f(r_1) > f(r_2)$且$h(r_1) < h(r_2)$
$\Rightarrow$ 记 $f(r_1) = f(r_2) + \Delta f = f + \Delta f, h(r_1) = h(r_2) - \Delta h = h - \Delta h$ 且 $\Delta f, \Delta h > 0$
$\Rightarrow \frac{h(r_2)}{r_2} - \frac{f(r_2)}{r_2} - \frac{h(r_1)}{r_1} + \frac{f(r_1)}{r_1} = \frac{h}{r_2} - \frac{f}{r_2} - \frac{h-\Delta h}{r_1} + \frac{f+\Delta f}{r_1} = \frac{\Delta h + \Delta f}{r_1} + (f-h)(\frac{1}{r_1} - \frac{1}{r_2}) > 0$
$\Rightarrow ln(1 + \frac{\frac{h(r_2)}{r_2} - \frac{f(r_2)}{r_2} - \frac{h(r_1)}{r_1} + \frac{f(r_1)}{r_1}}{\frac{1}{3b} + \frac{h(r_1)}{r_1} - \frac{f(r_1)}{r_1}}) > 0 \Rightarrow \frac{ln(1 + \frac{3\frac{h(r_2)}{r_2}b - 3\frac{f(r_2)}{r_2}c - 3\frac{h(r_1)}{r_1}b + 3\frac{f(r_1)}{r_1}c}{1+3\frac{h(r_1)}{r_1}b - \frac{f(r_1)}{r_1}c})}{lnr_2 - lnr_1} > 0$
$\Rightarrow lid' - lid > 0$

在此稍微讨论一下假设 5。假设 5 看似严格，但事实上，后续不等式推导过程中有较多放缩，故实则 A_2 类点不远多于 A_1 类点即可。

所以 S 的 LID 值在第 l 轮训练前后有上升趋势。在同一集群内，维度情况相对统一，可以用 S 的 LID 变化情况反映样本 T 的 LID 变化情况。因为 T 点为任取的一点，故任一样本点的 LID 值在第 l 轮前后有上升趋势。

因为每一个样本点的 LID 在第 l 轮训练前后均有上升趋势，故平均 LID 值在第 l 轮训练前后上升。

证毕。

1.4 验证假设

在上述论证中，我们前后做出了 6 个假设。在本子节中，我们用两个实验来验证部分假设，其中第 1 个实验验证假设 1；第 2 个实验验证假设 4 与假设 5。

1.4.1 实验 1