Appendices

A Proof of Theorem 1

Theorem 1. *Assume that the dataset is balanced (each class has the same amount of instances, and c classes in total), and the noise is class-dependent. Given a class transition matrix T_c, such that $T_{c,ij} = P(\bar{Y} = j|Y = i)$. The elements of the corresponding similarity transition matrix T_s can be calculated as*

$$T_{s,00} = \frac{c^2 - c - \left(\sum_j (\sum_i T_{c,ij})^2 - ||T_c||^2_{\text{Fro}}\right)}{c^2 - c}, \qquad T_{s,01} = \frac{\sum_j (\sum_i T_{c,ij})^2 - ||T_c||^2_{\text{Fro}}}{c^2 - c},$$
$$T_{s,10} = \frac{c - ||T_c||^2_{\text{Fro}}}{c}, \qquad T_{s,11} = \frac{||T_c||^2_{\text{Fro}}}{c}.$$

Proof. Assume each class has n samples. $n^2 T_{c,ij} T_{c,i'j'}$ represents the number the kind of data pairs composed by points of $(\bar{Y} = j|Y = i)$ and $(\bar{Y} = j'|Y = i')$. For the first element $T_{s,00}$, $n^2 \sum_{i \neq i'} T_{c,ij} T_{c,i'j'}$ is the number of data pairs with clean similarity labels $H = 0$, while $n^2 \sum_{i \neq i', j \neq j'} T_{c,ij} T_{c,i'j'}$ is the number of data pairs with clean similarity labels $H = 0$ and noisy similarity labels $\bar{H} = 0$. Thus the proportion of these two terms is exact the $T_{s,00} = P(\bar{H} = 0|H = 0)$. The remaining three elements can be represented in the same way. The primal representations are as follows,

$$T_{s,00} = \frac{\sum_{i \neq i', j \neq j'} T_{c,ij} T_{c,i'j'}}{\sum_{i \neq i'} T_{c,ij} T_{c,i'j'}}, \qquad T_{s,01} = \frac{\sum_{i \neq i', j = j'} T_{c,ij} T_{c,i'j'}}{\sum_{i \neq i'} T_{c,ij} T_{c,i'j'}},$$
$$T_{s,10} = \frac{\sum_{i = i', j \neq j'} T_{c,ij} T_{c,i'j'}}{\sum_{i = i'} T_{c,ij} T_{c,i'j'}}, \qquad T_{s,11} = \frac{\sum_{i = i', j = j'} T_{c,ij} T_{c,i'j'}}{\sum_{i = i'} T_{c,ij} T_{c,i'j'}}.$$

Further, note that

$$\sum_{i = i'} T_{c,i,j} T_{c,i',j'} = \sum_{i,j,j'} T_{c,i,j} T_{c,i,j'} = \sum_i (\sum_j T_{c,i,j})(\sum_{j'} T_{c,i,j'}) = c,$$
$$\sum_{i \neq i'} T_{c,i,j} T_{c,i',j'} = \sum_{i \neq i', j, j'} T_{c,i,j} T_{c,i',j'} = \sum_{i \neq i'} (\sum_j T_{c,i,j})(\sum_{j'} T_{c,i,j'}) = (c-1)c,$$
$$\sum_{i = i', j = j'} T_{c,ij} T_{c,i'j'} = ||T_c||^2_{\text{Fro}},$$
$$\sum_{i \neq i', j = j'} T_{c,ij} T_{c,i'j'} = \sum_j (\sum_i T_{c,ij})^2 - ||T_c||^2_{\text{Fro}}.$$

Substituting above equations to the primal representations, we have the Theorem 1 proved. □

B *Pointwise* implies *pairwise*

For an invertible T_c, denote by $\mathbf{v}_j$ the j-th column of T_c and $\mathbf{1}$ the all-one vector. Then,

$$\sum_j (\sum_i T_{c,ij})^2 = \sum_j \langle \mathbf{v}_j, \mathbf{1} \rangle^2 \leq \sum_j ||\mathbf{v}_j||^2 ||\mathbf{1}||^2 = c||T_c||_{\text{Fro}}^2,$$

where we use the Cauchy–Schwarz inequality [Steele, 2004] in the second step. Further, we have

$$\begin{aligned}
T_{s,11} + T_{s,00} &= \frac{||T_c||_{\text{Fro}}^2}{c} + \frac{c^2 - c - \left(\sum_j (\sum_i T_{c,ij})^2 - ||T_c||_{\text{Fro}}^2\right)}{c^2 - c} \\
&= \frac{(c-1)||T_c||_{\text{Fro}}^2 + c^2 - c - \left(\sum_j (\sum_i T_{c,ij})^2 - ||T_c||_{\text{Fro}}^2\right)}{c^2 - c} \\
&= \frac{(c-1)||T_c||_{\text{Fro}}^2 + c^2 - c - \left(\sum_j \langle \mathbf{v}_j, \mathbf{1} \rangle^2 - ||T_c||_{\text{Fro}}^2\right)}{c^2 - c} \\
&\geq \frac{(c-1)||T_c||_{\text{Fro}}^2 + c^2 - c - \left(c||T_c||_{\text{Fro}}^2 - ||T_c||_{\text{Fro}}^2\right)}{c^2 - c} \\
&= 1.
\end{aligned}$$

Thus the learnability of the pointwise classification implies the learnability of the reduced pairwise classification.

C Proof of Theorem 2

Theorem 2. *Assume that the dataset is balanced (each class has the same amount of samples), and the noise is class-dependent. When the number of classes $c \geq 8$, the noise rate of noisy similarity labels is lower than that of the noisy class labels.*

Proof. Assume each class has n points. As we state in the proof of Theorem 1, the number of data pairs with clean similarity labels $H = 0$ and noisy similarity labels $\bar{H} = 0$ is $n^2 \sum_{i \neq i', j \neq j'} T_{c,ij} T_{c,i'j'}$. We denote it by N_{00}. Similarly, we have,

$$\begin{aligned}
N_{00} &= n^2 \sum_{i \neq i', j \neq j'} T_{c,ij} T_{c,i'j'}, \qquad & N_{01} &= n^2 \sum_{i \neq i', j = j'} T_{c,ij} T_{c,i'j'}, \\
N_{10} &= n^2 \sum_{i = i', j \neq j'} T_{c,ij} T_{c,i'j'}, \qquad & N_{11} &= n^2 \sum_{i = i', j = j'} T_{c,ij} T_{c,i'j'}.
\end{aligned}$$

The noise rate is the proportion of the number of noisy labels to the number of total labels.

Assume that the number of classes is c. We have

$$S_{noise} = \frac{N_{01} + N_{10}}{N_{00} + N_{01} + N_{10} + N_{11}} = \frac{N_{01} + N_{10}}{c^2 n^2},$$
$$C_{noise} = \frac{n \sum_{i \neq j} T_{c,ij}}{cn}.$$

Let S_{noise} minus C_{noise}, we have

$$\begin{aligned} S_{noise} - C_{noise} &= \frac{n^2 \sum_{i \neq i', j = j'} T_{c,ij} T_{c,i'j'} + n^2 \sum_{i = i', j \neq j'} T_{c,ij} T_{c,i'j'}}{c^2 n^2} - \frac{n \sum_{i \neq j} T_{c,ij}}{cn} \\ &= \frac{\sum_{i \neq i', j = j'} T_{c,ij} T_{c,i'j'} + \sum_{i = i', j \neq j'} T_{c,ij} T_{c,i'j'} - c \sum_{i \neq j} T_{c,ij}}{c^2}. \end{aligned}$$

Let $A = \sum_{i \neq i', j = j'} T_{c,ij} T_{c,i'j'} + \sum_{i = i', j \neq j'} T_{c,ij} T_{c,i'j'} - c \sum_{i \neq j} T_{c,ij}$, we have

$$\begin{aligned} A &= \sum_{i \neq i', j = j'} T_{c,ij} T_{c,i'j'} + \sum_{i = i', j \neq j'} T_{c,ij} T_{c,i'j'} - c \sum_{i \neq j} T_{c,ij} \\ &= \sum_{i \neq i', j = j'} T_{c,ij} T_{c,i'j'} + \sum_{i = i', j \neq j'} T_{c,ij} T_{c,i'j'} - c(\sum_{i,j} T_{c,ij} - \sum_{i=j} T_{c,ij}) \\ &= \sum_{i \neq i', j = j'} T_{c,ij} T_{c,i'j'} + \sum_{i = i', j \neq j'} T_{c,ij} T_{c,i'j'} - c^2 + c \sum_{i=j} T_{c,ij}. \end{aligned}$$

The second equation holds because the row sum of T_c is 1.

For the first term $\sum_{i \neq i', j = j'} T_{c,ij} T_{c,i'j'}$, notice that:

$$\begin{aligned} \sum_{i \neq i', j = j'} T_{c,ij} T_{c,i'j'} &= \sum_j \sum_i T_{c,ij} (\sum_{i' \neq i} T_{c,i'j}) \\ &= \sum_j \sum_i T_{c,ij} (\sum_{i' \neq i} T_{c,i'j} + T_{c,ij} - T_{c,ij}) \\ &= \sum_j \sum_i T_{c,ij} (\sum_{i'} T_{c,i'j} - T_{c,ij}) \\ &= \sum_j \sum_i T_{c,ij} (S_j - T_{c,ij}) \qquad (S_j \text{ is the column sum of the } j-th \text{ column}) \\ &= \sum_j \sum_i T_{c,ij} S_j - T_{c,ij}^2 \\ &= \sum_j S_j \sum_i T_{c,ij} - \sum_j \sum_i T_{c,ij}^2 \\ &= \sum_j S_j^2 - \sum_j \sum_i T_{c,ij}^2. \end{aligned} \tag{1}$$

Due to the symmetry of i and j, for the second term $\sum_{i=i',j\neq j'} T_{c,ij}T_{c,i'j'}$, we have

$$\begin{aligned}\sum_{i=i',j\neq j'} T_{c,ij}T_{c,i'j'} &= \sum_j \sum_i T_{c,ij}(R_i - T_{c,ij}) \qquad (R_i \text{ is the row sum of the } i-th \text{ row, and } R_i = 1)\\ &= \sum_j \sum_i T_{c,ij} - T_{c,ij}^2 \\ &= c - \sum_j \sum_i T_{c,ij}^2. \end{aligned} \tag{2}$$

Therefore, substituting Equation (1) and (2) into A, we have

$$A = \sum_j S_j^2 - \sum_j \sum_i T_{c,ij}^2 + c - \sum_j \sum_i T_{c,ij}^2 - c^2 + c\sum_{i=j} T_{c,ij}.$$

To prove $S_{noise} - C_{noise} \leq 0$ is equivalent to prove $A \leq 0$.

Let $M = c^2 - c$, $N = \sum_j S_j^2 - 2\sum_j \sum_i T_{ij}^2 + c\sum_{i=j} T_{ij}$ (we drop the subscript c in $T_{c,ij}$), and $A = N - M$. Now we utilize the Adjustment method [Su and Xiong, 2015] to scale N. For every iteration, we denote the original N by N_o, and the adjusted N by N_a.

Since $c \geq 8$, there can not exist three columns with column sum bigger than $c/2 - 1$. Otherwise, the sum of the three columns will be bigger than c, which is impossible because the sum of the whole matrix is c.

Therefore, first, we assume that the $j, k-th$ columns have column sum bigger than $c/2 - 1$. Then, for the row i, we add the elements l, which are not in $j, k-th$ columns, to the diagonal element. We have

$$\begin{aligned} N_a - N_o &= (S_i + T_{il})^2 + (S_l + T_{il})^2 + cT_{il} - 2(T_{ii} + T_{il})^2 - S_i^2 - S_l^2 + 2(T_{ii}^2 + T_{il}^2) \\ &= T_{il}(2T_{il} + 2S_i - 2S_l + c - 4T_{ii}) \\ &\geq T_{il}(2T_{il} - 2S_l + c - 2T_{ii}) \qquad (\because S_i \geq T_{ii}) \\ &> T_{il}(2T_{il} - c + 2 + c - 2T_{ii}) \qquad (\because S_l < c/2 - 1) \\ &\geq 0. \qquad (\because T_{ii} \leq 1) \end{aligned}$$

We do such adjustment to every rows, then N_a is getting bigger and the adjusted matrix will only have values on diagonal elements and the $j, k-th$ columns. Since the diagonal elements are dominant in the row, $S_j + S_k < 2c/3 + 2/3$ (because for $i \neq j, k, T_{ij} + T_{ik} < 2/3$).

Assume that the column sum of $k-th$ column is no bigger than that of the $j-th$ column, and thus $S_k < c/3 + 1/3$. Then, for a row i, we add the T_{ik} to T_{ii}. We have

$$\begin{aligned} N_a - N_o &= (S_i + T_{ik})^2 + (S_k + T_{ik})^2 + cT_{ik} - 2(T_{ii} + T_{ik})^2 - S_i^2 - S_k^2 + 2(T_{ii}^2 + T_{ik}^2) \\ &= T_{ik}(2T_{ik} + 2S_i - 2S_k + c - 4T_{ii}) \\ &\geq T_{ik}(2T_{ik} - 2S_k + c - 2T_{ii}) \qquad (\because S_i \geq T_{ii}) \\ &> T_{ik}(2T_{ik} + c/3 - 2/3 - 2T_{ii}) \qquad (\because S_k < c/3 + 1/3) \\ &\geq 0. \qquad (\because c \geq 8, \text{ and } T_{ii} \leq 1) \end{aligned}$$

We do such adjustment to every rows, then N_a is getting bigger and the adjusted matrix will only have values on diagonal elements and the $j-th$ column, which is called final matrix.

Note that if there is only one column with a column sum bigger than $c/2-1$, we can adjust the rest $c-1$ columns as above and then obtain the final matrix as well. If there is no column with a column sum bigger than $c/2-1$, we can adjust all the elements as above and then obtain a *unit matrix*. For the unit matrix, $A = N - M < N_a - M = 0$, the Theorem 2 is proved.

Now we process the final matrix. For simplification, we assume $j = 0$ in the final matrix. We denote the T_{ij} by b_i and T_{ii} by a_i, for $i = \{1, \ldots, c-1\}$. We have

$$\begin{aligned}
N_a &= \sum_i a_i^2 + (1 + \sum_i b_i)^2 + c(\sum_i a_i + 1) - 2(\sum_i a_i^2 + \sum_i b_i^2 + 1) \\
&= (1 + \sum_i b_i)^2 + c\sum_i a_i + c - \sum_i a_i^2 - 2\sum_i b_i^2 - 2 \\
&= 1 + (\sum_i b_i)^2 + 2\sum_i b_i + c\sum_i a_i + c - \sum_i a_i^2 - 2\sum_i b_i^2 - 2 \\
&= (\sum_i b_i)^2 + 2\sum_i b_i - 2\sum_i b_i^2 + c\sum_i a_i - \sum_i a_i^2 + c - 1 \\
&= (\sum_i b_i)^2 + 2\sum_i b_i - 2\sum_i b_i^2 + c\sum_i (1 - b_i) - \sum_i (1 - b_i)^2 + c - 1 \\
&= (\sum_i b_i)^2 + 2\sum_i b_i - 2\sum_i b_i^2 + c^2 - c - c\sum_i b_i - \sum_i (1 - 2b_i + b_i^2) + c - 1 \\
&= (\sum_i b_i)^2 + 4\sum_i b_i - 3\sum_i b_i^2 - c\sum_i b_i + c^2 - c.
\end{aligned}$$

Now we prove $A = N - M \leq N_a - M \leq 0$. Note that

$$\begin{aligned}
N_a - M &= (\sum_i b_i)^2 + 4\sum_i b_i - 3\sum_i b_i^2 - c\sum_i b_i \\
&= (\sum_i b_i)^2 + 3\sum_i b_i - 3\sum_i b_i^2 - (c-1)\sum_i b_i \\
&= (\sum_i b_i)^2 + 3\sum_i b_i - 3\sum_i b_i^2 - (\sum_i (1 - b_i) + \sum_i b_i)\sum_i b_i \\
&= 3\sum_i b_i - 3\sum_i b_i^2 - \sum_i (1 - b_i)\sum_i b_i \\
&= 3\sum_i b_i(1 - b_i) - \sum_i (1 - b_i)\sum_i b_i.
\end{aligned}$$

According to the rearrangement inequality[Hardy et al., 1952], we have

$$\sum_i (1 - b_i)\sum_i b_i \geq (c-1)\sum_i b_i(1 - b_i).$$

Note that $c \geq 8$, thus $3\sum_i b_i(1-b_i) - \sum_i(1-b_i)\sum_i b_i \leq 0$, and $A \leq 0$. Therefore $S_{noise} - C_{noise} \leq 0$, and the equation holds if and only if the noise rate is 0 or every instances have the same noisy class label (i.e., there is one column in the T_c, of which every elements are 1, and the rest elements of the T_c are 0). Above two extreme situations are not considered in this paper. Namely, the noise rate of the noisy similarity labels is lower than that of the noisy class labels. Theorem 2 is proved. □

D Implementation of Class2Simi with *Reweight*

The expected risk for clean pairwise data is

$$R(f) = E_{(X_i,X_j,H_{ij})\sim\mathcal{D}}[\ell(\langle f(X_i), f(X_j)\rangle, H_{ij})],$$

where

$$\begin{aligned}\ell(\langle f(X_i), f(X_j)\rangle, H_{ij}) = &-\sum_{i,j} H_{ij}\log(\langle f(X_i), f(X_j)\rangle) + (1-H_{ij})\log(1-\langle f(X_i), f(X_j)\rangle),\\ &-\sum_{i,j} H_{ij}\log\hat{S}_{ij} + (1-H_{ij})\log(1-\hat{S}_{ij}).\end{aligned}$$

Here, we employ the *importance reweighting* technique to build a *risk-consistent* algorithms. Specifically,

$$\begin{aligned}R(f) &= E_{(X_i,X_j,H_{ij})\sim\mathcal{D}}[\ell(\langle f(X_i), f(X_j)\rangle, H_{ij})]\\ &= \int_{(x_i,x_j)}\sum_k P_{\mathcal{D}}(X_i=x_i, X_j=x_j, H_{ij}=k)\ell(\langle f(X_i), f(X_j)\rangle, H_{ij})d(x_i,x_j)\\ &= \int_{(x_i,x_j)}\sum_k P_{\mathcal{D}_\rho}(X_i, X_j, \bar{H}_{ij}=k)\frac{P_{\mathcal{D}}(X_i, X_j, H_{ij}=k)}{P_{\mathcal{D}_\rho}(X_i, X_j, \bar{H}_{ij}=k)}\ell(\langle f(X_i), f(X_j)\rangle, H_{ij}=k)d(x_i,x_j)\\ &= \int_{(x_i,x_j)}\sum_k P_{\mathcal{D}_\rho}(X_i, X_j, \bar{H}_{ij}=k)\frac{P_{\mathcal{D}}(H_{ij}=k|X_i, X_j)}{P_{\mathcal{D}_\rho}(\bar{H}_{ij}=k|X_i, X_j)}\ell(\langle f(X_i), f(X_j)\rangle, H_{ij}=k)d(x_i,x_j)\\ &= E_{(X_i,X_j,\bar{H}_{ij})\sim\mathcal{D}_\rho}[\bar{\ell}(\langle f(X_i), f(X_j)\rangle, \bar{H}_{ij})],\end{aligned}$$

where $\mathcal{D}$ denotes the distribution of clean data; $\mathcal{D}$ denotes the distribution of noisy data, and

$$\bar{\ell}(\langle f(X_i), f(X_j)\rangle, \bar{H}_{ij}) = \frac{P_{\mathcal{D}}(H_{ij}=\bar{H}_{ij}|X_i, X_j)}{P_{\mathcal{D}_\rho}(\bar{H}_{ij}|X_i, X_j)}\ell(\langle f(X_i), f(X_j)\rangle, \bar{H}_{ij}).$$

Empirically, as shown in Figrue 1, we use $\hat{S}_{ij} = f(X_i)^\top f(X_j)$ to measure the similarity of two points in a pair. $P(H_{ij}=1|X_i, X_j)$ and $P(H_{ij}=0|X_i, X_j)$ are approximated by

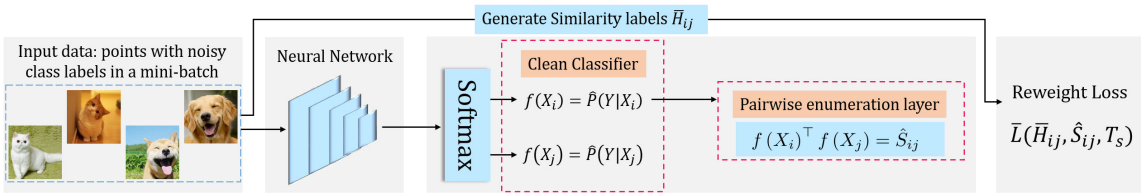

Figure 1: Pipeline of Class2Simi with *Reweight*.

$\hat{S}_{ij}$ and $1 - \hat{S}_{ij}$, respectively. Then $P\left(\bar{H}_{ij} \mid X_i, X_j\right)$ can be approximated according to $P\left(\bar{H}_{ij} \mid X_i, X_j\right) = T_s^\top P\left(H_{ij} \mid X_i, X_j\right)$. Thus a risk-consistent estimator can be built:

$$R_n(f) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha \ell(\langle f(X_i), f(X_j) \rangle, \bar{H}_{ij}),$$

where

$$\alpha = \left\{ \bar{H}_{ij} \frac{\hat{S}_{ij}}{T_{s,11}\hat{S}_{ij} + T_{s,01}(1 - \hat{S}_{ij})} + (1 - \bar{H}_{ij}) \frac{1 - \hat{S}_{ij}}{T_{s,10}\hat{S}_{ij} + T_{s,00}(1 - \hat{S}_{ij})} \right\}.$$

E Proof of Theorem 3

Theorem 3. *Assume the parameter matrices $W_1, \ldots, W_d$ have Frobenius norm at most $M_1, \ldots, M_d$, and the activation functions are 1-Lipschitz, positive-homogeneous, and applied element-wise (such as the ReLU). Assume the transition matrix is given, and the instances X are upper bounded by B, i.e., $\|X\| \leq B$ for all X, and the loss function ℓ is upper bounded by M. Then, for any $\delta > 0$, with probability at least $1 - \delta$,*

$$R(\hat{f}) - R_n(\hat{f}) \leq \frac{(T_{s,11} - T_{s,01}) 2Bc(\sqrt{2d \log 2} + 1) \Pi_{i=1}^{d} M_i}{T_{s,11}\sqrt{n}} + M\sqrt{\frac{\log 1/\delta}{2n}}. \tag{3}$$

Proof. We have defined

$$R(f) = E_{(X_i, X_j, \bar{Y}_i, \bar{Y}_j, \bar{H}_{ij}, T_s) \sim \mathcal{D}_\rho}[\ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})], \tag{4}$$

and

$$R_n(f) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij}), \tag{5}$$

where n is training sample size of the noisy data.

First, we bound the generalization error with Rademacher complexity [Bartlett and Mendelson, 2002].

Theorem 4 (Bartlett and Mendelson [2002]). *Let the loss function be upper bounded by M. Then, for any $\delta > 0$, with the probability $1 - \delta$, we have*

$$\sup_{f \in \mathcal{F}} |R(f) - R_n(f)| \leq 2\mathfrak{R}_n(\ell \circ \mathcal{F}) + M\sqrt{\frac{\log 1/\delta}{2n}}, \tag{6}$$

where $\mathfrak{R}_n(\ell \circ \mathcal{F})$ is the Rademacher complexity defined by

$$\mathfrak{R}_n(\ell \circ \mathcal{F}) = E\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})\right], \tag{7}$$

and $\{\sigma_1, \cdots, \sigma_n\}$ are Rademacher variables uniformly distributed from $\{-1, 1\}$.

Before further upper bound the Rademacher complexity $\mathfrak{R}_n(\ell \circ \mathcal{F})$, we discuss the special loss function and its *Lipschitz continuity* w.r.t $h_k(X_i), k = \{1, \ldots, c\}$.

Lemma 1. *Given similarity transition matrix T_s, loss function $\ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})$ is μ-Lipschitz with respect to $h_k(X_i), k = \{1, \ldots, c\}$, and $\mu = (T_{s,11} - T_{s,01})/T_{s,11}$*

$$\left|\frac{\partial \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})}{\partial h_k(X_i)}\right| < \frac{T_{s,11} - T_{s,01}}{T_{s,11}}. \tag{8}$$

Detailed proof of Lemma 1 can be found in Section E.1.

Lemma 1 shows that the loss function is μ-Lipschitz with respect to $h_k(X_i), k = \{1, \ldots, c\}$.

Based on Lemma 1, we can further upper bound the Rademacher complexity $\mathfrak{R}_n(\ell \circ \mathcal{F})$ by the following lemma.

Lemma 2. *Given similarity transition matrix T_s and assume that loss function $\ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})$ is μ-Lipschitz with respect to $h_k(X_i), k = \{1, \ldots, c\}$, we have*

$$\begin{aligned}\mathfrak{R}_n(\ell \circ \mathcal{F}) &= E\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})\right] \\ &\leq \mu c E\left[\sup_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(X_i)\right],\end{aligned} \tag{9}$$

where H is the function class induced by the deep neural network.

Detailed proof of Lemma 2 can be found in Section E.2.

The right-hand side of the above inequality, indicating the hypothesis complexity of deep neural networks and bounding the Rademacher complexity, can be bounded by the following theorem.

Theorem 5. *[Golowich et al., 2018] Assume the Frobenius norm of the weight matrices $W_1, \ldots, W_d$ are at most $M_1, \ldots, M_d$. Let the activation functions be 1-Lipschitz, positive-homogeneous, and applied element-wise (such as the ReLU). Let X is upper bounded by B, i.e., for any X, $\|X\| \leq B$. Then,*

$$E\left[\sup_{h\in H}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(X_i)\right] \leq \frac{B(\sqrt{2d\log 2}+1)\Pi_{i=1}^{d}M_i}{\sqrt{n}}. \tag{10}$$

Combining Lemma 1,2, and Theorem 4, 5, Theorem 3 is proved. □

E.1 Proof of Lemma 1

Recall that

$$\begin{aligned}\ell(f(X_i), f(X_j), T_s, \bar{H}_{ij}=1) &= -\log(\hat{\bar{S}}_{ij}) \\ &= -\log(\hat{S}_{ij}\times T_{s,11} + (1-\hat{S}_{ij})\times T_{s,01}) \\ &= -\log(f(X_i)^\top f(X_j)\times T_{s,11} + (1-f(X_i)^\top f(X_j))\times T_{s,01}),\end{aligned} \tag{11}$$

where

$$\begin{aligned}f(X_i) &= [f_1(X_i), \ldots, f_c(X_i)]^\top \\ &= \left[\left(\frac{\exp(h_1(X))}{\sum_{k=1}^{c}\exp(h_k(X))}\right), \ldots, \left(\frac{\exp(h_c(X))}{\sum_{k=1}^{c}\exp(h_k(X))}\right)\right]^\top.\end{aligned} \tag{12}$$

Take the derivative of $\ell(f(X_i), f(X_j), T_s, \bar{H}_{ij}=1)$ w.r.t. $h_k(X_i)$, we have

$$\frac{\partial\ell(f(X_i), f(X_j), T_s, \bar{H}_{ij}=1)}{\partial h_k(X_i)} = \frac{\partial\ell(f(X_i), f(X_j), T_s, \bar{H}_{ij}=1)}{\partial\hat{\bar{S}}_{ij}}\left[\frac{\partial f(X_i)}{\partial h_k(X_i)}\right]^\top\frac{\partial\hat{\bar{S}}_{ij}}{\partial f(X_i)}, \tag{13}$$

where

$$\begin{aligned}&\frac{\partial\ell(f(X_i), f(X_j), T_s, \bar{H}_{ij}=1)}{\partial\hat{\bar{S}}_{ij}} = -\frac{1}{f(X_i)^\top f(X_j)\times T_{s,11} + (1-f(X_i)^\top f(X_j))\times T_{s,01}}, \\ &\frac{\partial\hat{\bar{S}}_{ij}}{\partial f(X_i)} = f(X_j)\times T_{s,11} - f(X_j)\times T_{s,01}, \\ &\frac{\partial f(X_i)}{\partial h_k(X_i)} = f'(X_i) = [f_1'(X_i), \ldots, f_c'(X_i)]^\top.\end{aligned}$$

Note that the derivative of the softmax function has some properties, i.e., if $m \neq k$, $f_m'(X_i) = -f_m(X_i)f_k(X_i)$ and if $m = k$, $f_k'(X_i) = (1-f_k(X_i))f_k(X_i)$.

We denote by $Vector_m$ the $m-th$ element in $Vector$ for those complex vectors. Because $0 < f_m(X_i) < 1, \forall m \in \{1, \ldots, c\}$, we have

$$f'_m(X_i) \leq |f'_m(X_i)| < f_m(X_i), \qquad \forall m \in \{1, \ldots, c\}; \tag{14}$$

$$f'(X_i)^\top f(X_j) < f(X_i)^\top f(X_j). \tag{15}$$

Therefore,

$$\begin{aligned}
&\left|\frac{\partial \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij}=1)}{\partial h_k(X_i)}\right| = \left|\frac{\partial \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij}=1)}{\partial \hat{\bar{S}}_{ij}} \left[\frac{\partial f(X_i)}{\partial h_k(X_i)}\right]^\top \frac{\partial \hat{\bar{S}}_{ij}}{\partial f(X_i)}\right| \\
&= \left|\frac{f'(X_i)^\top f(X_j) \times T_{s,11} - f'(X_i)^\top f(X_j) \times T_{s,01}}{f(X_i)^\top f(X_j) \times T_{s,11} + (1 - f(X_i)^\top f(X_j)) \times T_{s,01}}\right| \\
&< \left|\frac{f(X_i)^\top f(X_j) \times T_{s,11} - f(X_i)^\top f(X_j) \times T_{s,01}}{f(X_i)^\top f(X_j) \times T_{s,11} + (1 - f(X_i)^\top f(X_j)) \times T_{s,01}}\right| \\
&< \left|\frac{T_{s,11} - T_{s,01}}{T_{s,11}}\right| \\
&= \frac{T_{s,11} - T_{s,01}}{T_{s,11}}.
\end{aligned} \tag{16}$$

The second inequality holds because of $T_{s,11} > T_{s,01}$ (Detailed proof can be found in Section E.1.1) and Equation (20). The third inequality holds because of $f(X_i)^\top f(X_j) < 1$.

Similarly, we can prove

$$\left|\frac{\partial \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij}=0)}{\partial h_k(X_i)}\right| < \frac{T_{s,11} - T_{s,01}}{T_{s,11}}. \tag{17}$$

Combining Equation (16) and Equation (17), we obtain

$$\left|\frac{\partial \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})}{\partial h_k(X_i)}\right| < \frac{T_{s,11} - T_{s,01}}{T_{s,11}}. \tag{18}$$

E.1.1 Proof of $T_{s,11} > T_{s,01}$

As we mentioned in Section C, we have,

$$\begin{aligned}
N_{00} &= n^2 \sum_{i \neq i', j \neq j'} T_{c,ij} T_{c,i'j'}, \qquad N_{01} = n^2 \sum_{i \neq i', j = j'} T_{c,ij} T_{c,i'j'}, \\
N_{10} &= n^2 \sum_{i = i', j \neq j'} T_{c,ij} T_{c,i'j'}, \qquad N_{11} = n^2 \sum_{i = i', j = j'} T_{c,ij} T_{c,i'j'}, \\
T_{s,01} &= \frac{N_{01}}{N_{00} + N_{01}}, \qquad T_{s,11} = \frac{N_{11}}{N_{10} + N_{11}}, \\
T_{s,11} - T_{s,01} &= \frac{N_{11}N_{00} + N_{11}N_{01} - N_{01}N_{10} - N_{01}N_{11}}{(N_{00} + N_{01})(N_{10} + N_{11})}.
\end{aligned}$$

Let us review the definition of similarity labels: if two instances belong to the same class, they will have similarity label $S = 1$, otherwise $S = 0$. That is to say, for a k-class dataset, only $\frac{1}{k}$ of similarity data has similarity labels $S = 1$, and the rest $1 - \frac{1}{k}$ has similarity labels $S = 0$. We denote the number of data with similarity labels $S = 1$ by N_1, otherwise N_0. Therefore, for the balanced dataset with n samples of each class, $N_1 = cn^2$, and $N_0 = c(c-1)n^2$. Let $A = T_{s,11} - T_{s,01}$, we have

$$\begin{aligned} A &= N_{11}N_{00} - N_{01}N_{10} \\ &= N_{11}N_{00} - (N_0 - N_{00})(N_1 - N_{11}) \\ &= N_{11}N_{00} - N_0N_1 - N_{11}N_{00} + N_{11}N_0 + N_1N_{00} \\ &= N_{11}N_0 - N_{01}N_1 \\ &= c(c-1)n^2N_{11} - cn^2N_{01} \\ &> 0. \end{aligned}$$

The last equation holds because of $(c-1)N_{11} - N_{01} > 0$ according to the rearrangement inequality [Hardy et al., 1952].

E.2 Proof of Lemma 2

$$\begin{aligned} & E\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})\right] \\ &= E\left[\sup_{g} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})\right] \\ &= E\left[\sup_{\text{argmax}\{h_1,\ldots,h_c\}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})\right] \\ &= E\left[\sup_{\max\{h_1,\ldots,h_c\}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})\right] \\ &\leq E\left[\sum_{k=1}^{c} \sup_{h_k\in H} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})\right] \\ &= \sum_{k=1}^{c} E\left[\sup_{h_k\in H} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(X_i), f(X_j), T_s, \bar{H}_{ij})\right] \\ &\leq \mu c E\left[\sup_{h_k\in H} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h_k(X_i)\right] \\ &= \mu c E\left[\sup_{h\in H} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(X_i)\right], \end{aligned}$$

where the first three equations hold because given T_s, f and $\max\{h_1, \ldots, h_c\}$ give the same constraint on $h_j(X_i), j = \{1, \ldots, c\}$; the sixth inequality holds because of the Talagrand Contraction Lemma [Ledoux and Talagrand, 2013].

F Further details on experiments

F.1 Network structure and optimization

Note that for *CIFAR-10*, we use ResNet-26 with shake-shake regularization [Gastaldi, 2017] **except** the experiment on noisy T_c in Figure 4, where we use ResNet-32 with pre-activation [He et al., 2016] for shorter training time. In stage 1, We use the same optimization method as *Forward* to learn the transition matrix $\hat{T}_c$. In stage 2, we use Adam optimizer with an initial learning rate 0.001. On *MNIST*, the batch size is 128 and the learning rate decays every 5 epochs by a factor of 0.1 with 30 epochs in total. On *CIFAR-10*, the batch size is 512 and the learning rate decays every 40 epochs by a factor of 0.1 with 200 epochs in total. On *CIFAR-100*, the batch size is 512 and the learning rate decays every 40 epochs by a factor of 0.1 with 120 epochs in total. On *News20*, the batch size is 128 and the learning rate decays every 5 epochs by a factor of 0.1 with 30 epochs in total. On *Clothing1M**, the batch size is 32 and the learning rate drops every 5 epochs by a factor of 0.1 with 10 epochs in total.

F.2 Symmetric and asymmetric noise settings

Symmetric noise setting is defined as follow, where c is the number of classes.

$$\text{Sym-}\rho\text{:} \quad T = \begin{bmatrix} 1-\rho & \frac{\rho}{C-1} & \cdots & \frac{\rho}{C-1} & \frac{\rho}{C-1} \\ \frac{\rho}{C-1} & 1-\rho & \frac{\rho}{C-1} & \cdots & \frac{\rho}{C-1} \\ \vdots & & \ddots & & \vdots \\ \frac{\rho}{C-1} & \cdots & \frac{\rho}{C-1} & 1-\rho & \frac{\rho}{C-1} \\ \frac{\rho}{C-1} & \frac{\rho}{C-1} & \cdots & \frac{\rho}{C-1} & 1-\rho \end{bmatrix}. \tag{19}$$

The asymmetric noise setting is set as follow,

Listing 1: Asymmetric noise (transition matrix) generation.

```
def AsymTransitionMatrixGenerate(NoiseRate=0.3,
    NumClasses=10, seed=1):
    np.random.seed(seed)
    t = np.random.rand(NumClasses, NumClasses)
    i = np.eye(NumClasses)
    t = t + Coef * NumClasses * i
    for a in range(NumClasses):
    t[a] = t[a] / t[a].sum()
return t
```

Coef is set to 1.70, 1.20, 0.60, 0.24 at the rate 0.2, 0.3, 0.4, 0.6, respectively.

References

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002. 7, 8

Xavier Gastaldi. Shake-shake regularization. *arXiv preprint arXiv:1705.07485*, 2017. 12

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *COLT*, 2018. 9

Godfrey Harold Hardy, John Edensor Littlewood, George Pólya, György Pólya, DE Littlewood, et al. *Inequalities*. Cambridge university press, 1952. 5, 11

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645. Springer, 2016. 12

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013. 12

J Michael Steele. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004. 2

Yong Su and Bin Xiong. *Methods and Techniques for Proving Inequalities: In Mathematical Olympiad and Competitions*, volume 11. World Scientific Publishing Company, 2015. 4