# Noisy Label Learning with Instance-Dependent Outliers: Identifiability via Crowd Wisdom

**Tri Nguyen**[*]
School of EECS
Oregon State University
Corvallis, Oregon, USA
nguyetr9@oregonstate.edu

**Shahana Ibrahim**[*]
Department of ECE
University of Central Florida
Orlando, Florida, USA
shahana.ibrahim@ucf.edu

**Xiao Fu**
School of EECS
Oregon State University
Corvallis, Oregon, USA
xiao.fu@oregonstate.edu

## Abstract

The generation of label noise is often modeled as a process involving a probability transition matrix (also interpreted as the *annotator confusion matrix*) imposed onto the label distribution. Under this model, learning the "ground-truth classifier"—i.e., the classifier that can be learned if no noise was present—and the confusion matrix boils down to a model identification problem. Prior works along this line demonstrated appealing empirical performance, yet identifiability of the model was mostly established by assuming an instance-invariant confusion matrix. Having an (occasionally) instance-dependent confusion matrix across data samples is apparently more realistic, but inevitably introduces outliers to the model. Our interest lies in confusion matrix-based noisy label learning with such outliers taken into consideration. We begin with pointing out that under the model of interest, using labels produced by only one annotator is fundamentally insufficient to detect the outliers or identify the ground-truth classifier. Then, we prove that by employing a crowdsourcing strategy involving multiple annotators, a carefully designed loss function can establish the desired model identifiability under reasonable conditions. Our development builds upon a link between the noisy label model and a column-corrupted matrix factorization mode—based on which we show that crowdsourced annotations distinguish nominal data and instance-dependent outliers using a low-dimensional subspace. Experiments show that our learning scheme substantially improves outlier detection and the classifier's testing accuracy.

## 1 Introduction

Deep neural networks can easily overfit to noisy labels due to its excessive expressiveness [1,2]. Many strategies have been proposed to avoid negative impacts of label noise when training neural classifiers; see, e.g., noisy label filtering [3–8], robust loss design [9–13], and noise generation modeling (or loss correction) [14–22]. In the last genre, a noisy class label is modeled as a realization of a categorical distribution, which is the ground-truth label distribution being distorted by a *probability transition matrix*. The transition matrix is interpreted as the "confusion matrix" [9, 14, 18] that can effectively

[*] Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

model the annotators' expertise level and the difficulty of annotating each class/sample, and thus is considered intuitive. Under this model, learning the "label noise-free" target neural classifier boils down to identifying the confusion matrix.

The confusion matrix-based models have proven quite useful in practice—algorithms developed in this line of work often exhibits appealing empirical performance; see, e.g., [9,14–17,19–21,23–31]. In addition, these models admit interesting statistical and algebraic structures, leading to plausible results on identifiability of the confusion matrix and/or the "ground-truth classifier"—i.e., the classifier that can be learned if no noisy annotations were present. However, most of the aforementioned early works considered an instance-invariant confusion matrix—i.e., a confusion matrix is not affected by sample features, but only classes—for analytical and computational simplicity. Considering *instance-dependent* confusion models is more realistic, as the sample characteristics, e.g., lightening and resolution of an image, affect the annotation accuracy [32]. The existence of such (at least occasionally occurred) instance-dependent noisy labels inevitably introduces outliers to the instance-invariant confusion models, leading to performance degradation. In general, learning under instance-dependent confusion matrices is heavily ill-posed. Hence, various problem-specific structures were exploited to add regularization terms and constraints; see, e.g., [27, 28, 33–43]. Nonetheless, unlike the instance-invariant confusion matrix case, identifiability guarantees of the target classifier have been largely under-studied. The lack of theoretical understanding also affects algorithm design—many approaches in this domain had to resort to somewhat ad-hoc treatments with multi-stage training procedures, often involving nontrivial pre- and post-processing; see [27, 28, 33–35, 37, 39, 40].

**Contributions.** To advance understanding, we consider a model where instance-dependent confusion matrices occur occasionally across the samples, and the rest of data share a common *nominal* confusion matrix. This way, the instance-dependent noisy labels can be regarded as outliers. The model is motivated by the fact that only a proportion of all instances may have a labeling difficulty that significantly deviates from the general population [36, 44, 45]. Our contributions are as follows:

*(i) Identifiability via Crowdsourcing.* We first show that using the sparsity prior on outliers is insufficient to identify the model of interest, if only one annotator is present. To circumvent this challenge, we propose to employ a *crowd* of annotators—which is the common practice in data labeling [14, 18, 20, 31]. We show that, by incorporating a carefully designed column sparsity constraint to a *coupled cross-entropy* (CCE) loss from crowdsourcing [20, 31] to integrate the annotators' outputs, the outliers can be provably identified. Consequently, the ground-truth classifier can be learned with generalization guarantees.

*(ii) End-to-end One-Stage Implementation.* Our approach features a one-stage continuous optimization-based implementation. To be specific, our proposed learning loss allows to approximate the column-sparsity constraint using a smoothed nonconvex $\ell_p$ (where $0 < p \leq 1$) function based regularization (see [46]). This way, the overall loss is differentiable and can be readily tackled by off-the-shelf optimizers. This is in contrast to many existing methods that involve multiple stages (e.g., [27, 28, 33–35, 37, 39, 40])—and is arguably easier to implement.

We evaluate the proposed method over a number real datasets that are annotated by machine and human annotators under various conditions and observed nontrivial improvements of testing accuracy.

**Notation.** The notations are summarized in the supplementary material in Sec. A.

## 2 Problem Statement

**The Confusion Model and Learning Goal.** Consider $N$ data items $\{\boldsymbol{x}_n\}_{n=1}^N$ from $K$ classes. Here, $\boldsymbol{x}_n \in \mathbb{R}^D$ represents the feature vector of the $n$th data item. Let $\{y_n\}_{n=1}^N$ be the set of ground-truth labels, where $y_n \in [K]$. Assume that $y_n$'s are unobserved. Instead, we observe the "noisy" version $\{\widehat{y}_n\}_{n=1}^N$. The label $\widehat{y}_n \in [K]$ is noisy due to various reasons, e.g., the lack of expertise of the annotator. In this setting, our goal is to learn a performance-guaranteed classifier using the data items $\{\boldsymbol{x}_n\}_{n=1}^N$, and noisy labels $\{\widehat{y}_n\}_{n=1}^N$. We consider the following expression of $\mathsf{Pr}(\widehat{y}_n = k|\boldsymbol{x}_n)$:

$$\mathsf{Pr}(\widehat{y}_n = k|\boldsymbol{x}_n) = \sum_{k'=1}^{K} \mathsf{Pr}(\widehat{y}_n = k|y_n = k', \boldsymbol{x}_n)\mathsf{Pr}(y_n = k'|\boldsymbol{x}_n). \tag{1}$$

Note that $\mathsf{Pr}(y_n = k|\boldsymbol{x}_n)$ is the *ground-truth label distribution* given the sample $\boldsymbol{x}_n$. This is also the distribution that the *target classifier* wishes to learn from. We represent the *ground-truth classifier* using a function $\boldsymbol{f}^\natural : \mathbb{R}^D \to \mathbb{R}^K$ such that $[\boldsymbol{f}^\natural(\boldsymbol{x}_n)]_k \triangleq \mathsf{Pr}(y_n = k|\boldsymbol{x}_n)$. We call $\boldsymbol{f}^\natural$ the ground-truth classifier as it is the function that we aim to learn—and it can be learned under ideal conditions (e.g., when $N = \infty$ and the learner is a universal approximator), given that no noise is present. We define $\boldsymbol{T}^\natural : \mathbb{R}^D \to \mathbb{R}^{K\times K}$ such that $[\boldsymbol{T}^\natural(\boldsymbol{x}_n)]_{k,k'} \triangleq \mathsf{Pr}(\widehat{y}_n = k|y_n = k', \boldsymbol{x}_n), \forall \boldsymbol{x}_n$. That is, $\boldsymbol{T}^\natural(\boldsymbol{x}_n)$ is the *label transition matrix* or *confusion matrix* of sample $\boldsymbol{x}_n$. Let $[\boldsymbol{g}_n^\natural]_k \triangleq \mathsf{Pr}(\widehat{y}_n = k|\boldsymbol{x}_n)$. Accordingly, the noisy label generation process for $\boldsymbol{x}_n \overset{\text{i.i.d.}}{\sim} \mathcal{D}_{\boldsymbol{x}}$ is modeled as follows:

$$\boldsymbol{g}_n^\natural = \boldsymbol{T}^\natural(\boldsymbol{x}_n)\boldsymbol{f}^\natural(\boldsymbol{x}_n), \tag{2a}$$

$$\widehat{y}_n \sim \mathsf{categorical}(\boldsymbol{g}_n^\natural), \tag{2b}$$

where $\mathsf{categorical}(\boldsymbol{g})$ denotes the $K$-dimensional categorical distribution. Per the physical meaning of $\boldsymbol{T}^\natural(\boldsymbol{x}_n)$ and $\boldsymbol{f}^\natural(\boldsymbol{x}_n)$, we have $\boldsymbol{1}^\top\boldsymbol{T}^\natural(\boldsymbol{x}_n) = \boldsymbol{1}^\top, \boldsymbol{T}^\natural(\boldsymbol{x}_n) \geq \boldsymbol{0}$ and $\boldsymbol{1}^\top\boldsymbol{f}^\natural(\boldsymbol{x}_n) = 1$, $\boldsymbol{f}^\natural(\boldsymbol{x}_n) \geq \boldsymbol{0}$, for all $n$. Under this model, the main goal is to learn $\boldsymbol{f}^\natural$ from $\{\boldsymbol{x}_n\}_{n=1}^N$ and $\{\widehat{y}_n\}_{n=1}^N$.

**Identifiability under Instance-Invariant $\boldsymbol{T}^\natural(\boldsymbol{x})$.** From (2a), it becomes apparent that learning $\boldsymbol{f}^\natural$ is not a straightforward task. Even if $\boldsymbol{g}_n^\natural$ is observed (which is not), it is hard to identify $\boldsymbol{f}^\natural(\boldsymbol{x}_n)$ or $\boldsymbol{T}^\natural(\boldsymbol{x}_n)$ from the product $\boldsymbol{g}_n^\natural = \boldsymbol{T}^\natural(\boldsymbol{x}_n)\boldsymbol{f}^\natural(\boldsymbol{x}_n)$. To tackle the identifiability challenge, a popular approach is to simplify (2) by assuming that all instances have the same confusion matrix, i.e., $\boldsymbol{T}^\natural(\boldsymbol{x}_n) = \boldsymbol{A}^\natural$ with a certain $\boldsymbol{A}^\natural \in \mathbb{R}^{K\times K}$ for all $n$ [16, 18, 19, 21, 47, 48]. Under this assumption, many used the "loss correction" based formulation, e.g., [9, 16, 19, 21, 48]. The loss correction idea modifies the training loss of the classifier by taking the confusion matrix into consideration, e.g.,

$$\underset{\boldsymbol{A}\in\mathcal{A},\boldsymbol{f}\in\mathcal{F}}{\text{minimize}} \; -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}[\widehat{\boldsymbol{y}}_n]_k \log[\boldsymbol{A}\boldsymbol{f}(\boldsymbol{x}_n)]_k, \tag{3}$$

where the objective is a modified *cross-entropy* (CE) loss. The "noise correction" term $\boldsymbol{A}$ and $\boldsymbol{f}$ are used to learn $\boldsymbol{A}^\natural$ and $\boldsymbol{f}^\natural$, respectively, $\widehat{\boldsymbol{y}}_n \in \{0,1\}^K$ denotes the one-hot encoding of the noisy label $\widehat{y}_n$, and $\mathcal{A}$ and $\mathcal{F}$ denote appropriate constraint sets—i.e., $\mathcal{A} = \{\boldsymbol{A} \in \mathbb{R}^{K\times K} \mid \boldsymbol{1}^\top\boldsymbol{A} = \boldsymbol{1}^\top,\ \boldsymbol{A} \geq \boldsymbol{0}\}$ and $\mathcal{F}$ is a certain deep neural network function class whose outputs reside in the probability simplex. Note that the objective is often used together with regularization and additional constraints of $\boldsymbol{A}$ for various purposes; see, e.g., [19, 21, 30]. When $N \to \infty$, the CE term enforces the learned $\boldsymbol{A}$ and $\boldsymbol{f}$ to satisfy $\boldsymbol{g}_n^\natural = \boldsymbol{A}\boldsymbol{f}(\boldsymbol{x}_n)$ for all $n$ [19]. Note that $\widetilde{\boldsymbol{G}} = [\boldsymbol{g}_1^\natural, \ldots, \boldsymbol{g}_N^\natural]$ can be expressed as $\widetilde{\boldsymbol{G}} = \boldsymbol{A}^\natural\left[\boldsymbol{f}^\natural(\boldsymbol{x}_1), \ldots, \boldsymbol{f}^\natural(\boldsymbol{x}_N)\right] = \boldsymbol{A}^\natural\boldsymbol{F}^\natural$, where $\boldsymbol{F}^\natural = [\boldsymbol{f}^\natural(\boldsymbol{x}_1), \ldots, \boldsymbol{f}^\natural(\boldsymbol{x}_N)]$. Consequently, Eq. (3) can be understood as learning $\boldsymbol{A}$ and $\boldsymbol{F}$ such that $\widetilde{\boldsymbol{G}} \approx \boldsymbol{A}\boldsymbol{F}$. As both $\boldsymbol{A}^\natural$ and $\boldsymbol{F}^\natural$ are nonnegative matrices (due to their physical meaning), the identifiability of $\boldsymbol{F}^\natural$ can be connected to uniqueness of the *nonnegative matrix factorization* (NMF) model $\widetilde{\boldsymbol{G}} = \boldsymbol{A}^\natural\boldsymbol{F}^\natural$; see [19, 21, 45].

These are interesting developments, yet the key assumption $\boldsymbol{T}^\natural(\boldsymbol{x}_n) = \boldsymbol{A}^\natural$ for all $n$ appears to be overly stringent. As mentioned, it makes sense to believe that at least a proportion of samples have instance-dependent $\boldsymbol{T}^\natural(\boldsymbol{x}_n)$'s [36, 44, 45, 49]. Not considering such samples may cause performance degradation.

**Instance-Dependent Confusion-Induced Outliers.** When the instance-dependent confusion happens sparingly instead of overwhelmingly, we can re-express $\boldsymbol{T}^\natural(\boldsymbol{x}_n)$ using the following decomposition:

$$\boldsymbol{T}^\natural(\boldsymbol{x}_n) = \boldsymbol{A}^\natural + \boldsymbol{E}^\natural(\boldsymbol{x}_n), \tag{4}$$

where $\boldsymbol{A}^\natural \in \mathbb{R}^{K\times K}$ represents an instance-independent (class-dependent) confusion matrix—which is the nominal confusion matrix that reflects the general annotation difficulty of the dataset. The term $\boldsymbol{E}^\natural(\boldsymbol{x}_n)$ represents the instance-dependent "perturbation". For many $n$'s, $\boldsymbol{E}^\natural(\boldsymbol{x}_n) = \boldsymbol{0}$. When $\boldsymbol{E}^\natural(\boldsymbol{x}_n) \neq \boldsymbol{0}$, we have $(\boldsymbol{A}^\natural + \boldsymbol{E}^\natural(\boldsymbol{x}_n))\boldsymbol{f}^\natural(\boldsymbol{x}_n) = \boldsymbol{A}^\natural\boldsymbol{f}^\natural(\boldsymbol{x}_n) + \boldsymbol{e}_n^\natural$, where $\boldsymbol{e}_n^\natural = \boldsymbol{E}^\natural(\boldsymbol{x}_n)\boldsymbol{f}^\natural(\boldsymbol{x}_n)$. Using (4), the model in (2a) can be expressed as follows:

$$\boldsymbol{g}_n^\natural = \begin{cases} \boldsymbol{A}^\natural\boldsymbol{f}^\natural(\boldsymbol{x}_n) + \boldsymbol{e}_n^\natural, & \forall n \in \mathcal{I} \\ \boldsymbol{A}^\natural\boldsymbol{f}^\natural(\boldsymbol{x}_n), & \forall n \in \mathcal{I}^{\rm c}, \end{cases} \tag{5}$$

where $\mathcal{I}$ and $\mathcal{I}^{\rm c}$ represent the instance index set where $\boldsymbol{E}^\natural(\boldsymbol{x}_n) \neq \boldsymbol{0}$ and its complement, respectively. In other words, $\mathcal{I}$ is the outlier index set.

A remark is that from this point on we will ignore the structure $\boldsymbol{e}_n^\natural = \boldsymbol{E}^\natural(\boldsymbol{x}_n)\boldsymbol{f}^\natural(\boldsymbol{x}_n)$ of the outliers. Disregarding the structure comes with some losses. For example, this way, our method is not able to learn $\boldsymbol{E}^{\natural}(\boldsymbol{x}_n)$ that could be of interest. Nonetheless, considering $\boldsymbol{E}^{\natural}(\boldsymbol{x}_n)$ renders extra modeling and computational burdens. Our treatment simplifies the subsequent analytical and computational developments, particularly, algorithm design. In addition, other types of outliers and anomalies can also be handled by the proposed approach due to the unstructured treatment.

**Challenge of Outlier Detection.** Under the model in (5), the natural idea is to first identify $\mathcal{I}$ and remove the impacts of outliers. Our first attempt is to explicitly model $\boldsymbol{e}_n^{\natural}$ and modify (3) as follows:

$$\begin{aligned} \underset{\boldsymbol{A}\in\mathcal{A},\{\boldsymbol{e}_n\in\mathcal{E}\},\boldsymbol{f}\in\mathcal{F}}{\text{minimize}} \quad & -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}[\widehat{\boldsymbol{y}}_n]_k \log\left[\boldsymbol{A}\boldsymbol{f}(\boldsymbol{x}_n)+\boldsymbol{e}_n\right]_k, \\ \text{subject to} \quad & \sum_{n=1}^{N}\mathbb{1}\left\{\|\boldsymbol{e}_n\|_2>0\right\}\leq C, \end{aligned}$$

where $\mathcal{E}=\{\boldsymbol{e}\in\mathbb{R}^K \mid \boldsymbol{1}^\top\boldsymbol{e}=0\}$ is the constraint set for $\boldsymbol{e}_n$'s (due to the probability simplex constraints on the model parameters in (5)) and $C$ is a scalar that is an estimation of $|\mathcal{I}|$. The idea is to use the prior knowledge that $\boldsymbol{e}_n$ does not occur overwhelmingly as constraint. The hope is that the solution of (6) can detect $\mathcal{I}$, thereby enabling accurate estimation of $\boldsymbol{A}^{\natural}$ and $\boldsymbol{f}^{\natural}$. However, the following fact reveals a conflicting insight:

**Fact 2.1.** *Assume that all $\boldsymbol{f}\in\mathcal{F}$ are universal function approximators, that $\mathsf{rank}(\boldsymbol{A}^{\natural})=K$, and that $N\to\infty$. Suppose that $\mathcal{I}\neq\emptyset$. Optimal solutions of Problem* (6) *can attain the minimal value of the objective function and satisfy the sparsity constraint without detecting any outliers. One such trivial solution $(\boldsymbol{A}^\star,\boldsymbol{f}^\star,\boldsymbol{e}_n^\star)$ is $\boldsymbol{A}^\star=\boldsymbol{I}$, $\boldsymbol{f}^\star(\boldsymbol{x})=\boldsymbol{A}^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_n)+\boldsymbol{e}_n^{\natural}$, and $\boldsymbol{e}_n^\star=\boldsymbol{0}$ for all $n$.*

This fact is easy to show (see Appendix F), yet it highlights a somewhat unexpected issue in confusion matrix-based noisy label learning: If only one nominal confusion matrix $\boldsymbol{A}^{\natural}$ is present (i.e., only one annotator is employed), it does not suffice to recover the ground-truth $\boldsymbol{f}^{\natural}$ when there exists instance-dependent outliers (or any other types of outliers) under the model in (5)—under the condition that $\boldsymbol{f}$ is a universal function approximator. The practical implication is undesirable: as deep neural networks are powerful function approximators and are usually very expressive, Fact 2.1 means that the learned neural networks easily overfit to outliers, even if the sparsity prior on outliers is explicitly used in the loss function.

## 3 Proposed Approach

**Intuition: Exploiting Crowd-induced Subspace.** To get around the issues illustrated in Fact 2.1, our idea is to "create" a low-dimensional subspace where the nominal data reside while the outliers are likely far away from. To this end, we propose a *crowdsourcing* approach. Consider the scenario where $M$ annotators label the dataset $\{\boldsymbol{x}_n\}_{n=1}^{N}$. The noisy label provided by an annotator $m$ for the data item $\boldsymbol{x}_n$ is denoted as $\widehat{y}_n^{(m)}\in[K]$ and follows the generation model as follows:

$$\widehat{y}_n^{(m)}\sim\mathsf{categorical}(\boldsymbol{g}_n^{\natural(m)}),\ \boldsymbol{g}_n^{\natural(m)}=\boldsymbol{A}_m^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_n)+\boldsymbol{e}_n^{\natural(m)},\forall m,n, \tag{6}$$

where $\boldsymbol{A}_m^{\natural}$ is the annotator $m$'s confusion matrix and $\boldsymbol{e}_n^{\natural(m)}$ is the outlier term induced by annotator $m$'s instance-dependent confusion.

Putting together, we have the following expression:

$$\begin{aligned} \boldsymbol{G}^{\natural}=\begin{bmatrix}\boldsymbol{g}_1^{\natural(1)} & \cdots & \boldsymbol{g}_N^{\natural(1)}\\ \vdots & \ddots & \vdots\\ \boldsymbol{g}_1^{\natural(M)} & \cdots & \boldsymbol{g}_N^{\natural(M)}\end{bmatrix}&=\begin{bmatrix}\boldsymbol{A}_1^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_1)+\boldsymbol{e}_1^{\natural(1)} & \cdots & \boldsymbol{A}_1^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_N)+\boldsymbol{e}_N^{\natural(1)}\\ \vdots & \ddots & \vdots\\ \boldsymbol{A}_M^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_1)+\boldsymbol{e}_1^{\natural(M)} & \cdots & \boldsymbol{A}_M^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_N)+\boldsymbol{e}_N^{\natural(M)}\end{bmatrix}\\ &\Longleftrightarrow \boldsymbol{G}^{\natural}=\boldsymbol{W}^{\natural}\boldsymbol{F}^{\natural}+\boldsymbol{E}^{\natural}, \end{aligned} \tag{7}$$

where $\boldsymbol{F}^{\natural}$ is defined as before, and

$$\boldsymbol{W}^{\natural}=[(\boldsymbol{A}_1^{\natural})^\top,\ldots,(\boldsymbol{A}_M^{\natural})^\top]^\top,\quad \boldsymbol{e}_n^{\natural}=[(\boldsymbol{e}_n^{\natural(1)})^\top,\ldots,(\boldsymbol{e}_n^{\natural(M)})^\top]^\top. \tag{8}$$

Denote $\mathcal{I}^{(m)}$ as the index set where $\boldsymbol{e}_n^{\natural(m)}\neq\boldsymbol{0}$. Then $\mathcal{I}=\mathcal{I}^{(1)}\cup\ldots\cup\mathcal{I}^{(M)}$ stands for the nonzero column support of $\boldsymbol{E}^{\natural}$.

From (7), one can see that now the "data columns", i.e., the columns of the matrix $\boldsymbol{G}^\natural$, live in a high-dimensional space, i.e., $\mathbb{R}^{MK}$. However, the nominal data part $\boldsymbol{W}^\natural \boldsymbol{F}^\natural$ resides in a $K$-dimensional subspace $\mathsf{range}(\boldsymbol{W}^\natural)$ where $K \ll MK$, if multiple annotators are employed. More importantly, $\boldsymbol{e}_j^\natural$ for $j \in \mathcal{I}$ is an $MK$-dimensional vector and is unlikely to be inside $\mathrm{range}(\boldsymbol{W}^\natural)$. With this geometry, it is much more likely that one can separate the outliers from the nominal data.

Next, we will use the above intuition to build a learning loss. We will take into consideration of practical aspects, e.g., missing annotations and finite dataset size $N$.

**Proposed Identification Criterion & Analysis.** In this section, we connect our proposed idea to a practically convenient, end-to-end identification criterion and provide identifiability guarantees for the desired model parameters. We consider the following empirical risk minimization under the crowdsourcing model in (6):

$$\underset{\{\boldsymbol{A}_m \in \mathcal{A}\}, \{\boldsymbol{e}_n^{(m)} \in \mathcal{E}\}, \boldsymbol{f} \in \mathcal{F}}{\text{minimize}} \quad \mathsf{L}_{\mathsf{ce}} \triangleq -\frac{1}{S} \sum_{(m,n) \in \mathcal{S}} \sum_{k=1}^{K} [\widehat{\boldsymbol{y}}_n^{(m)}]_k \log \left[ \boldsymbol{A}_m \boldsymbol{f}(\boldsymbol{x}_n) + \boldsymbol{e}_n^{(m)} \right]_k, \tag{9a}$$

$$\text{subject to} \sum_{n=1}^{N} \mathbb{1} \left\{ \sum_{m=1}^{M} \|\boldsymbol{e}_n^{(m)}\|_2 > 0 \right\} \leq C, \tag{9b}$$

where $\mathcal{S} \subseteq [M] \times [N]$ denotes the set of observed noisy labels indexed by $(m, n)$ with $S = |\mathcal{S}|$ (note that all data items may not be labeled by an annotator), $\widehat{\boldsymbol{y}}_n^{(m)}$ is the one-hot encoding of the annotator-provided label $\widehat{y}_n^{(m)}$, and $C$ is an estimate of $|\mathcal{I}|$. Here, the constraint sets $\mathcal{F}$, $\mathcal{A}$, and $\mathcal{E}$ are as defined before. If $\boldsymbol{e}_n^{(m)}$ is not considered, the objective function (9a) is sometimes referred to as *coupled cross-entropy minimization* (CCEM) in the end-to-end crowdsourcing literature [20, 30, 31, 50]. CCEM has not been used together with the outlier detection constraint (9b)—and the theoretical characterizations of the constrained formulation is unknown.

We will use the following notations in our analysis. We use $\mathcal{I} \subseteq [N]$ to denote index set of the outliers, i.e., $\mathcal{I} = \{n \mid \boldsymbol{e}_n^\natural \neq \boldsymbol{0}\}$. We also consider that the function class in our learning problem, i.e., $\mathcal{F}$, has a complexity measure $\mathscr{R}_{\mathcal{F}}$. In particular, we adopt the so-called spectral-complexity upper bound of the function class $\mathcal{F}$ [51]; see Lemma C.4 in the Appendix.

We first establish that the criterion (9) recovers the ground-truth $\boldsymbol{g}_n^{\natural(m)}$'s with a reasonable accuracy. Specifically, we hope to characterize the average estimation accuracy of $\widehat{\boldsymbol{g}}_n^{(m)}$'s where $\widehat{\boldsymbol{g}}_n^{(m)} = \widehat{\boldsymbol{A}}_m \widehat{\boldsymbol{f}}(\boldsymbol{x}_n) + \widehat{\boldsymbol{e}}_n^{(m)}$ and $\widehat{\boldsymbol{A}}_m$'s, $\widehat{\boldsymbol{e}}_n^{(m)}$'s and $\widehat{\boldsymbol{f}}$ are estimated via solving (9):

**Lemma 3.1.** *Assume that the observed index pairs $(m, n) \in \mathcal{S}$ are sampled uniformly at random with replacement. Also assume $\boldsymbol{f}^\natural \in \mathcal{F}$, $|\mathcal{I}| \leq C \leq N/2$, and that each $[\boldsymbol{A}_m \boldsymbol{f}(\boldsymbol{x}_n) + \boldsymbol{e}_n^{(m)}]_k$, $\forall \boldsymbol{A}_m \in \mathcal{A}, \forall \boldsymbol{e}_n^{(m)} \in \mathcal{E}, \forall \boldsymbol{f} \in \mathcal{F}$ are lower bounded by $(1/\beta)$ for a certain $\beta > 0$. Then, with probability greater than $1 - \delta$, the following holds:*

$$\frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \|\boldsymbol{g}_n^{\natural(m)} - \widehat{\boldsymbol{g}}_n^{(m)}\|_2^2 \leq \epsilon_{\boldsymbol{g}}(S), \quad \epsilon_{\boldsymbol{g}}(S) = \mathcal{O}\left( {}^{\beta \mathfrak{R}_{\mathcal{S}} \log S}/_{\sqrt{S}} + \log(\beta) \sqrt{\log(1/\delta)}/_{\sqrt{S}} \right),$$

*where $\mathfrak{R}_{\mathcal{S}}^2 = K^2 \log(K) + {}^{K \|\boldsymbol{X}\|_{\mathrm{F}}^2 \mathscr{R}_{\mathcal{F}}}/_{S} + CMK \log \sqrt{NM}$ and $\boldsymbol{X} = [\boldsymbol{x}_{n_1}, \ldots, \boldsymbol{x}_{n_S}]$ are the annotated samples, in which $(m_s, n_s) \in \mathcal{S}$ for $s = 1, \ldots, S$.*

The proof is in Appendix C, which uses a similar idea as [31, Proposition 1] but takes into consideration of the outliers. Lemma 3.1 reveals that the criterion (9) essentially recovers the complete matrix $\boldsymbol{G}^\natural$ [cf. Eq, (7)] from the noisy and incomplete observations of its entries, when ${}^{S}/_{C}$ is sufficiently large. To proceed, we will need a suite of assumptions and definitions:

**Assumption 3.2.** The outliers satisfy $\boldsymbol{e}_i^\natural \notin \mathsf{range}(\boldsymbol{W}^\natural), \forall i \in \mathcal{I}$, where $\mathsf{rank}(\boldsymbol{W}^\natural) = K$.

Assumption 3.2 presents the key condition to disassociate instance-dependent and instance-independent confusions. As discussed after Eq. (8), having a larger $M$ would decrease the possibility that any $\boldsymbol{e}_n^\natural$ belongs to $\mathsf{range}(\boldsymbol{W}^\natural)$. We also characterize the "quantity" that $\boldsymbol{e}_n^\natural$ is perturbed from $\mathsf{range}(\boldsymbol{W}^\natural)$ in the following definition:

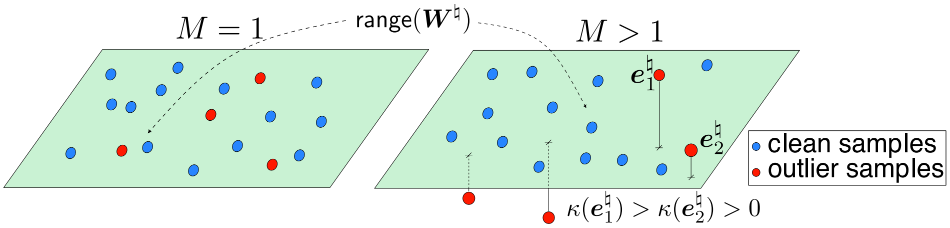
$M = 1$
range($\boldsymbol{W}^\natural$)
$M > 1$
$\boldsymbol{e}_1^\natural$
$\boldsymbol{e}_2^\natural$
● clean samples
● outlier samples
$\kappa(\boldsymbol{e}_1^\natural) > \kappa(\boldsymbol{e}_2^\natural) > 0$

Figure 1: An illustration of outliers and nominal data items w.r.t. $\mathsf{range}(\boldsymbol{W}^\natural)$: (left) when $M = 1$, $\boldsymbol{e}^\natural \in \mathsf{range}(\boldsymbol{W}^\natural)$, (right) while $M > 1$, it is highly likely that $\boldsymbol{e}^\natural \neq \mathsf{range}(\boldsymbol{W}^\natural)$. In addition, the measure $\kappa(\boldsymbol{e})$ is larger for outliers that are farther from $\mathsf{range}(\boldsymbol{W}^\natural)$.

**Definition 3.3.** The *outlier impact score* $\kappa(\boldsymbol{e}_i^\natural)$ of data $i \in \mathcal{I}$ is defined as

$$\kappa(\boldsymbol{e}_i^\natural) \triangleq \min_{\substack{\boldsymbol{W} \in \mathcal{W}; \boldsymbol{h}, \boldsymbol{h}_i \in \mathcal{E}, \\ \mathcal{L} \subset \mathcal{I}^c, |\mathcal{L}| = N - 2|\mathcal{I}|}} \left( \|\boldsymbol{W}^\natural \boldsymbol{F}^\natural(:, \mathcal{L}) - \boldsymbol{W}\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{W}^\natural \boldsymbol{f}_i^\natural + \boldsymbol{e}_i^\natural - \boldsymbol{W}\boldsymbol{h}\|^2 \right) \tag{10}$$

where $\mathcal{W} = \left\{ \boldsymbol{W} = [\boldsymbol{A}_1^\top, \ldots, \boldsymbol{A}_M^\top]^\top \mid \boldsymbol{A}_m \in \mathcal{A} \right\}$.

A larger score $\kappa(\boldsymbol{e}_i^\natural)$ implies it is easier for our criterion to distinguish the outliers from the nominal data. Fig. 1 illustrates the geometry that we rely on to detect outliers. One can notice that when $M = 1$, we always have $\kappa(\boldsymbol{e}_i^\natural) = 0$, as the outlier satisfies $\boldsymbol{e}_i^\natural \in \mathsf{range}(\boldsymbol{W}^\natural)$.

In addition, we consider the following assumptions to assist identifying $\boldsymbol{A}^\natural$ and $\boldsymbol{f}^\natural$:

**Assumption 3.4 (Class Specialists and Anchor Points).** Assume that the following conditions hold:

a. There exists a near-class specialist for each class $k$, i.e., $\forall k \in [K], \exists m_k \in [M]$ such that $\|\boldsymbol{A}_{m_k}^\natural(k,:) - \overline{\boldsymbol{e}}_k^\top\| \leq \xi_1$, where $\overline{\boldsymbol{e}}_k$ is a unit vector.
b. There exists a near-anchor point sample for each class $k$ in the dataset, i.e., $\forall k \in [K], \exists n_k \in \mathcal{I}^c$ such that $\|\boldsymbol{f}^\natural(\boldsymbol{x}_{n_k}) - \overline{\boldsymbol{e}}_k^\top\| \leq \xi_2$, where $\overline{\boldsymbol{e}}_k$ is a unit vector.

Assumption 3.4.a. is sometimes used in the crowdsourcing literature (e.g., [52]) to characterize the expertise of annotators. Assumption 3.4.b. is often seen in loss correction based noisy label learning; see, e.g., [16, 21, 27]. Under Assumptions 3.2 and 3.4, we have the following result:

**Theorem 3.5 (Identifiability and Generalization).** *Let* $(\{\widehat{\boldsymbol{A}}_m\}, \{\widehat{\boldsymbol{e}}_n^{(m)}\}, \widehat{\boldsymbol{f}})$ *be any optimal solution of* (9) *with* $\widehat{\mathcal{I}} = \left\{ n \in [N] \mid \widehat{\boldsymbol{e}}_n = [(\widehat{\boldsymbol{e}}_n^{(1)})^\top, \ldots, (\widehat{\boldsymbol{e}}_n^{(M)})^\top]^\top \neq \boldsymbol{0} \right\}$. *Suppose that the conditions in Lemma 3.1 holds, that we set* $C = |\mathcal{I}|$, *and that Assumption 3.4 holds with* $\xi_1, \xi_2 \leq 1/K$. *Denote* $\sigma$ *as the upper bound* $\sigma_{\max}(\boldsymbol{A}_m^\natural) \leq \sigma, \ \forall m$. *In addition, assume that* $\forall \mathcal{L} \subset \mathcal{I}^c, |\mathcal{L}| = N - 2\,|\mathcal{I}^c|$, $\mathsf{rank}(\boldsymbol{F}^\natural(:, \mathcal{L})) = K$. *Then, for some* $\alpha > 0$, *and* $S > S_0$ *where* $S_0$ *as the smallest integer such that* $\kappa(\boldsymbol{e}_i^\natural) \geq \epsilon_{\boldsymbol{g}}(S_0), \ \forall i \in \mathcal{I}$, *the following result holds with probability greater than* $1 - 2/S - K/T^\alpha$:

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\boldsymbol{x}}} \left[ \min_{\boldsymbol{\Pi}} \|\widehat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x})\|_2^2 \right] \leq K(\eta + \xi_1 + \xi_2),$$

$$\min_{\boldsymbol{\Pi}} \|\widehat{\boldsymbol{A}}_m - \boldsymbol{A}_m^\natural \boldsymbol{\Pi}\|_{\mathrm{F}}^2 = K\sigma^2(\eta + \xi_1 + \xi_2), \ \forall m,$$

*where* $\eta^2 = \mathcal{O}\left( {}^{\beta M T^\alpha}/_{\sqrt{S}} \left( \sqrt{M \log S} + (\|\boldsymbol{X}\|\mathscr{R}_{\mathcal{F}})^{0.25} \right) \right)$, $\boldsymbol{\Pi}$ *a permutation matrix,* $\boldsymbol{X}$ *is defined as before, and* $T = N - |\mathcal{I}|$ . *In addition, we have exact outlier detection, i.e.,* $\widehat{\mathcal{I}} = \mathcal{I}$.

The proof is in Appendix D. We should mention that we set $C = |\mathcal{I}|$ for notation simplicity. With a notation-wise slightly more complicated definition of $\mathcal{L}$, the same proof holds under $C \geq |\mathcal{I}|$, which leads to $T = N - C$ and $\mathcal{I} \subseteq \widehat{\mathcal{I}}$. That is, over-estimated $C$ still enables identifying $\mathcal{I}$ with the price of discarding some nominal samples. While the criterion (9) offers the desired identifiability, it requires the presence of class-specialist annotators and the anchor data points (c.f. Assumption 3.4), which are considered relatively restrictive [19, 31]. In [19], the CE loss was combined with a simplex volume minimization-based regularization to establish identifiability of $\boldsymbol{f}^\natural$ under more relaxed conditions, namely, the *sufficiently scattered condition* (SSC) from the NMF literature [53]. We show that this is also viable in the presence of outliers:

**Theorem 3.6 (Enhanced Identifiability).** *Let* $(\{\widehat{\boldsymbol{A}}_m\}, \{\widehat{\boldsymbol{e}}_n^{(m)}\}, \widehat{\boldsymbol{f}})$ *be any optimal solution of* (9) *with* $\widehat{\boldsymbol{W}} \triangleq [\widehat{\boldsymbol{A}}_1^\top, \ldots, \widehat{\boldsymbol{A}}_M^\top]^\top$ *admitting the minimum volume of* $\text{conv}\{\widehat{\boldsymbol{W}}\}$ *(i.e., the simplex spanned by its columns) over all possible optimal solutions. Assume that we set* $C = |\mathcal{I}|$ *and that* $\boldsymbol{F}^\natural(:, \mathcal{I}^c)$ *satisfies the SSC. Under Assumptions 3.2 and the same conditions used in Lemma 3.1, the following result holds with probability greater than* $1 - \delta$*, when* $S$ *grows to infinity:*

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathcal{D}_{\boldsymbol{x}}} \left[ \|\widehat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x})\|_2^2 \right] = \mathcal{O}(|\mathcal{I}^c|^{-5/8} \, (2\|\boldsymbol{X}(:, \mathcal{I}^c)\|_{\mathrm{F}} \mathscr{R}_{\mathcal{F}})^{\frac{1}{4}} + \sqrt{\log(1/\delta)/|\mathcal{I}^c|}).$$

*In addition, we have* $\|\widehat{\boldsymbol{A}}_m - \boldsymbol{A}_m^\natural \boldsymbol{\Pi}\|_{\mathrm{F}} = 0, \forall m$ *and* $\widehat{\mathcal{I}} = \mathcal{I}$*.*

The proof is relegated to Appendix E, which also holds for $C \geq |\mathcal{I}|$ with slight modifications as discussed before. Theorem 3.6 shows that the target classifier can be accurately estimated in the presence of instance-dependent outliers, without needing anchor samples or class specialists.

**Implementation.** Our learning losses allows relatively easy implementation and optimization. We consider the following regularized criterion:

$$\mathop{\text{minimize}}_{\boldsymbol{A}_m \in \mathcal{A}, \boldsymbol{e}_n^{(m)} \in \mathcal{E}, \boldsymbol{f} \in \mathcal{F}} \quad \mathsf{L}_{\mathsf{ce}} + \mu_1 \mathsf{L}_{\mathsf{outlier}} + \mu_2 \mathsf{L}_{\mathsf{vol}}.$$

Note that we use $\mathsf{L}_{\mathsf{outlier}} = \sum_{n=1}^{N} (\sum_{m=1}^{M} \|\boldsymbol{e}_n^{(m)}\|_2^2 + \zeta)^{\frac{p}{2}}$ where $\zeta > 0$ and $0 < p \leq 1$ to approximate the column-sparsity constraint. The $\ell_2/\ell_p$ nonconvex mixed quasi-norm has proven to be a very effective approximator for column sparsity [54]. In addition, as the function is differentiable, it offers an optimization-friendly surrogate.

For the minimum-volume loss, we use $\mathsf{L}_{\mathsf{vol}} = -\log\det(\boldsymbol{F}\boldsymbol{F}^\top)$, where $\boldsymbol{F} = [\boldsymbol{f}(\boldsymbol{x}_1), \ldots, \boldsymbol{f}(\boldsymbol{x}_N)]$. Note that the term encourages maximizing the volume of $\text{conv}\{\boldsymbol{F}\}$, which in turn minimizes the volume of $\text{conv}\{\boldsymbol{W}\}$ under the model $\widetilde{\boldsymbol{G}} = \boldsymbol{W}\boldsymbol{F}$. Therefore, the volume of $\text{conv}\{\boldsymbol{W}\}$ can be minimized via using either $\log\det(\boldsymbol{W}^\top\boldsymbol{W})$ or $-\log\det(\boldsymbol{F}\boldsymbol{F}^\top)$ as the regularizer. The reason why we choose the latter is because minimizing $\log\det(\boldsymbol{W}^\top\boldsymbol{W})$ is ill-defined—the term encourages a rank-deficient $\boldsymbol{W}$ instead of a small-volume full-rank $\boldsymbol{W}$. The remedy in the literature is to use $\log\det(\boldsymbol{W}^\top\boldsymbol{W} + \gamma\boldsymbol{I})$ with $\gamma > 0$ [54] or adding structural constraints (e.g., diagonal dominance) to $\boldsymbol{W}$ [19], but these require more parameter tuning and more prior knowledge.

The constraints on $\boldsymbol{A}_m$'s can be handled by adding `softmax` to the columns. The function class $\mathcal{F}$ can be a designated deep neural network model (e.g., ResNet), where the outputs are also constrained by `softmax`. The constraint on $\boldsymbol{e}_n^{(m)}$ can be handled by parameterizing each $\boldsymbol{e}_n^m = \widetilde{\boldsymbol{e}}_n^{(m)} - (\boldsymbol{1}^\top \widetilde{\boldsymbol{e}}_n^{(m)})\boldsymbol{1}/K$, where $\widetilde{\boldsymbol{e}}_n^{(m)}$ is trainable weight vector. More details of the implementation are provided in Appendix G.1. We name our approach as *Crowdsourcing-based Outlier-robust criterion and INstance-dependence-aware deep neural Network learning*, abbreviated as `COINNet`.

## 4 Related Works

Our development is related to works in transition matrix-based [14–19, 21, 23, 27–29, 33, 47, 55, 56] and sample selection-based [28, 33, 35, 37, 39] noisy label learning, confusion identification in crowdsourcing [14, 18, 57], end-to-end crowdsourcing [20, 30, 31, 49] (with some paying particular attention to incomplete annotations [58, 59]), and identifiability of NMF [53, 60]—see detailed discussions of the related work in Appendix B.

A particular connection to highlight is the usage of multiple annotators as a key enabler of our identifiability guarantees. This result (implicitly) aligns with some recently proposed approaches that advocate for the use of multiple labels in learning under label noise [45, 49]. Specifically, the work by [45] established the identifiability of instance-independent label noise using the consensus of at least three similar data items with the same true label. Similarly, the work by [49] leveraged Kruskal's identifiability-based arguments to recommend using at least three noisy labels to establish the identifiability of the instance-dependent noise transition matrix. These works, particularly [49], were more concerned with the theoretical limitations other than realizable methods. They did not connect the existence of repeated noisy labels to crowdsourcing, but used clusterability properties of datasets to support their arguments. In this work, we advocate *diversity in crowd wisdom*—labels from multiple annotators—based on an intuitive geometric characteristic of the model, which leads to clearly realizable implementations.

# 5 Experiments

**Baselines.** The proposed method is compared with a number of existing baselines in noisy label learning. We choose several end-to-end crowdsourcing methods, namely, `GeoCrowdNet(F)&(W)` [31], `TraceReg` [30], `CrowdLayer` [20], and `Max-MIG` [61]. We also employ four different instance-dependent noisy learning approach, namely, `MEIDTM` [36], `PTD` [27], and `BLTM` [33]—which all use a single annotator. In addition, we present the result of `VolMinNet` [19] that uses the instance-independent confusion model and the volume minimization regularization. We also use two noise-robust loss function-based approaches, namely `GCE` [10] and `Reweight` [21]. For baselines that were developed for single annotator cases, we train them using the labels after majority voting. Note that all the confusion matrix-based noise correction methods (including ours) inherently have a permutation ambiguity (cf. the $\boldsymbol{\Pi}$ term in Theorems 3.5 and 3.6). Hence, we report the highest classification accuracy attained among all possible permutation matrices for every method. In practice, $\boldsymbol{\Pi}$ can be removed by additional annotator inspection on several anchor samples, but we skipped these steps to keep the evaluation simple.

## 5.1 Experiments with Machine Annotations

**Dataset.** We consider the CIFAR-10 [62] and the STL-10 datasets [63]—see Appendix G for details.

**Experiment Setup.** To generate multiple noisy labels, we use $M = 3$ machine annotators for each dataset. In order to simulate a wide range of annotation behaviors, we employ different classification and clustering methods such as *$k$-nearest neighbors* (kNN), logistic regression, and convolutional neural networks. Each machine annotator is trained by randomly choosing a subset of the training data. We control the labeling accuracies of these machine annotators by varying the size of the dataset and the number of epochs during their training phase. This results in annotators with different labeling qualities, with their average individual noise rates around 20% (*good*), 40% (*medium*), and 70% (*low*). This way, we set up the following three cases: *(i)* High Noise: Three machine annotators are with low, medium, and good quality, respectively. *(i)* Medium Noise: Two machine annotators are with medium quality and one machine annotator with good quality. *(i)* Low Noise: One machine annotators is with medium quality and two machine annotators with good quality. More details are in Appendix G.2.

**Neural Network and Optimizer Settings.** We use ResNet-34 and ResNet-9 architectures [64] as the backbone to run all methods on CIFAR-10 and STL-10 datasets, respectively. For our proposed approach `COINNet`, we fix $\zeta = 10^{-10}, p = 0.4$, and $\mu_1 = \mu_2 = 0.01$. Adam [65] is used as the optimizer with weight decay of $10^{-4}$, learning rate of 0.01, and batch size of 512.

More details and ablation studies with the hyperparameters are provided in Appendix G.

**Results.** Table 1 presents the average testing accuracy of the methods. One can see that our approach `COINNet` consistently outperforms all baselines in different scenarios under test. Notably, the performance gap between `COINNet` and the second-best method is more significant in the high noise regime. For example, in the CIFAR-10 high noise scenario, `COINNet` shows around 4% improvement over the second-best performing baseline `GeoCrowdNet(F)`. Another key observation is that the instance-dependent modeling based baselines such as `MEIDTM`, `PTD`, and `BLTM`, exhibit much degraded performance in high-noise scenarios, possibly due to their multi-stage procedures that accumulate errors easily under such circumstances. In addition, our approach outperforms the instance-independent confusion-based baselines, such as `VolMinNet`, `Reweight` and `GeoCrowdNet(W)`, by nontrivial margins, showing that modeling instance-dependent outliers is indeed beneficial.

## 5.2 Experiments Using Real Annotations

In this section, we use three datasets to evaluate the algorithms.

**CIFAR-10N.** The first dataset that we use is the CIFAR-10N dataset [66]. The data has $N = 60,000$ samples from $K = 10$ classes and the samples were labeled by $M = 3$ anonymous real annotators. The error rates of the 3 annotators are 17.23%, 18.12%, and 17.64%, respectively.

**LabelMe.** We also test the algorithms over the LabelMe dataset [67,68]. The dataset has $N = 2,688$ samples from $K = 8$ classes, and $M = 59$ anonymous real annotators were involved in labeling the data. The average error rate is 25.95%.

Table 1: Average classification accuracy using machine annotators in CIFAR-10 and STL-10 datasets under different labeling scenarios. **Bold black** represents the best and blue represents the second best.

| **Method** | **CIFAR-10** | | | **STL-10** | | |
|---|---|---|---|---|---|---|
| | High Noise | Medium Noise | Low Noise | High Noise | Medium Noise | Low Noise |
| PTD | $61.98 \pm 0.35$ | $65.71 \pm 0.69$ | $79.21 \pm 0.41$ | $22.94 \pm 2.15$ | $30.92 \pm 5.45$ | $23.98 \pm 11.52$ |
| BLTM | $52.62 \pm 1.12$ | $61.19 \pm 0.45$ | $71.36 \pm 0.48$ | $31.28 \pm 2.78$ | $30.57 \pm 4.34$ | $32.73 \pm 3.45$ |
| VolMinNet | $49.83 \pm 2.66$ | $50.93 \pm 3.34$ | $70.44 \pm 7.32$ | $42.38 \pm 3.64$ | $60.79 \pm 4.67$ | $64.67 \pm 2.93$ |
| Reweight | $59.73 \pm 1.25$ | $67.57 \pm 0.62$ | $77.40 \pm 0.66$ | $48.57 \pm 8.52$ | $54.23 \pm 3.86$ | $54.67 \pm 6.93$ |
| GCE | $52.14 \pm 0.67$ | $63.90 \pm 0.12$ | $73.26 \pm 0.38$ | $58.35 \pm 0.74$ | $60.65 \pm 0.69$ | $61.70 \pm 2.54$ |
| MEIDTM | $51.38 \pm 0.62$ | $56.20 \pm 0.65$ | $71.14 \pm 0.50$ | $55.50 \pm 0.52$ | $60.20 \pm 0.22$ | $59.84 \pm 3.57$ |
| CrowdLayer | $62.76 \pm 3.12$ | $67.24 \pm 2.26$ | $80.00 \pm 3.02$ | $47.83 \pm 5.51$ | $61.46 \pm 4.58$ | $66.64 \pm 2.48$ |
| TraceReg | $64.90 \pm 0.26$ | $71.21 \pm 0.32$ | $81.81 \pm 0.42$ | $50.14 \pm 5.47$ | $64.50 \pm 0.04$ | $65.67 \pm 3.86$ |
| MaxMIG | $51.92 \pm 1.10$ | $68.45 \pm 0.21$ | $83.16 \pm 0.53$ | $67.92 \pm 0.49$ | $71.75 \pm 0.20$ | $74.32 \pm 0.76$ |
| GeoCrowdNet(F) | $67.22 \pm 0.27$ | $71.91 \pm 0.50$ | $82.18 \pm 0.39$ | $64.70 \pm 0.82$ | $66.15 \pm 0.49$ | $68.40 \pm 0.31$ |
| GeoCrowdNet(W) | $64.08 \pm 0.83$ | $70.01 \pm 0.08$ | $81.53 \pm 0.27$ | $49.30 \pm 4.62$ | $61.54 \pm 4.90$ | $68.37 \pm 0.41$ |
| COINNet (Ours) | $\mathbf{71.22} \pm \mathbf{0.72}$ | $\mathbf{73.31} \pm \mathbf{0.09}$ | $\mathbf{84.14} \pm \mathbf{0.38}$ | $\mathbf{70.12} \pm \mathbf{0.48}$ | $\mathbf{73.11} \pm \mathbf{0.37}$ | $\mathbf{76.39} \pm \mathbf{0.58}$ |

Table 2: Average classification accuracy on CIFAR-10N, LabelMe, and ImageNet-15N datasets, labeled by human annotators. **Bold black** represents the best and blue represents the second best.

| **Method/Dataset** | CIFAR-10N | LabelMe | ImageNet-15N |
|---|---|---|---|
| PTD | $89.52 \pm 0.24$ | $84.18 \pm 1.36$ | $65.53 \pm 0.18$ |
| BLTM | $75.68 \pm 0.47$ | $82.10 \pm 0.56$ | $66.57 \pm 0.76$ |
| VolMinNet | $86.58 \pm 0.21$ | $79.97 \pm 0.16$ | $63.11 \pm 1.08$ |
| Reweight | $89.56 \pm 0.30$ | $84.51 \pm 0.50$ | $65.85 \pm 2.93$ |
| GCE | $78.01 \pm 7.23$ | $83.41 \pm 0.59$ | $64.71 \pm 1.38$ |
| MEIDTM | $68.69 \pm 0.31$ | $83.53 \pm 0.21$ | $72.66 \pm 0.58$ |
| CrowdLayer | $87.38 \pm 0.43$ | $82.80 \pm 0.90$ | $61.36 \pm 2.73$ |
| TraceReg | $86.57 \pm 0.24$ | $82.83 \pm 0.23$ | $68.43 \pm 0.12$ |
| MaxMIG | $90.11 \pm 0.09$ | $83.73 \pm 0.84$ | $81.13 \pm 1.42$ |
| GeoCrowdNet(F) | $87.19 \pm 0.37$ | $87.21 \pm 0.39$ | $80.45 \pm 1.77$ |
| GeoCrowdNet(W) | $86.43 \pm 0.44$ | $82.83 \pm 0.75$ | $68.79 \pm 0.27$ |
| COINNet (Ours) | $\mathbf{92.09} \pm \mathbf{0.47}$ | $\mathbf{87.60} \pm \mathbf{0.54}$ | $\mathbf{93.71} \pm \mathbf{3.26}$ |

**ImageNet-15N.** In addition to existing datasets, we also acquire noisy annotations by asking AMT workers to annotate some images from ImageNet. We select $K = 15$ classes and submit randomly chosen images to AMT for labeling. Eventually we collect annotations for $N = 2,514$ images from $M = 100$ anonymous real annotators, which serve as our training set. The average error rate of the annotators is 42.68%. The validation and testing sets have 1,462 and 13,157 images, respectively. We release the code and our acquired noisy annotations at `https://github.com/ductri/COINNet`.

**Settings.** For CIFAR-10N, we employ the ResNet-34 architecture to serve as $\boldsymbol{f}$. For the LabelMe and ImageNet-15N datasets, we employ similar settings as in [20]. Specifically, as the training sets are small, the pretrained VGG-16 [69] and CLIP [70] models are used to first extract image embeddings for the experiments on LabelMe and ImageNet-15N, respectively. The embeddings are then fed to $\boldsymbol{f}$, which is a fully connected neural network with one hidden layer and 128 hidden ReLU units. The same encoders and architecture choices are employed for the all methods under test. For our approach `COINNet`, we set $\mu_1 = \mu_2 = 0.1$ for LabelMe, and $\mu_1 = \mu_2 = 0.01$ for CIFAR-10N and ImageNet-15N.

**Results.** Table 2 presents the average testing accuracy of different methods on the three datasets. The proposed approach, `COINNet`, shows promising results in all cases, clearly outperforming the baselines by a noticeable edge. This is consistent with the machine annotator experiments.

Fig. 2 demonstrates the outlier identification results using the CIFAR-10N dataset. Here, we define the outlier indicator as $s_n = \sum_{m=1}^{M} \|\widehat{\boldsymbol{e}}_n^{(m)}\|_2, \forall n$, where $\widehat{\boldsymbol{e}}_n^{(m)}$ are the learned instance-dependent perturbations in the model (5). The histogram plot of these scores in Fig. 2 clearly shows that our method does output two types of samples, i.e., nominal samples and outliers, based on these perturbations. In the figure, the images on the middle and the right are examples from the low and high outlier indicator value regimes, respectively. The images with high $s_n$ values show more instance-dependent confusion characteristics (such as background noise and blurring) compared to

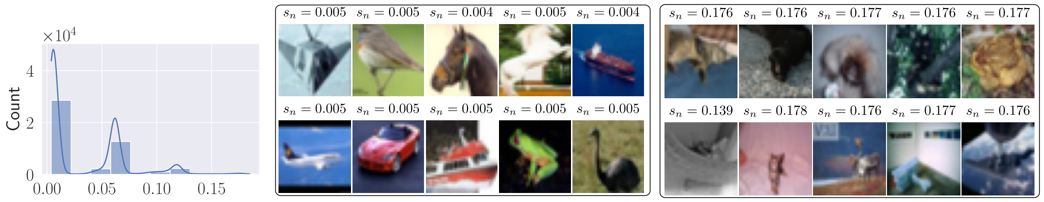

Figure 2: Histogram of the learned outlier indicator values $s_n = \sum_{m=1}^{M} \|\widehat{e}_n^{(m)}\|_2^2$ over all training images in the CIFAR-10N dataset (left), and examples with low (middle) and high (right) $s_n$'s —see more examples in Appendix H.

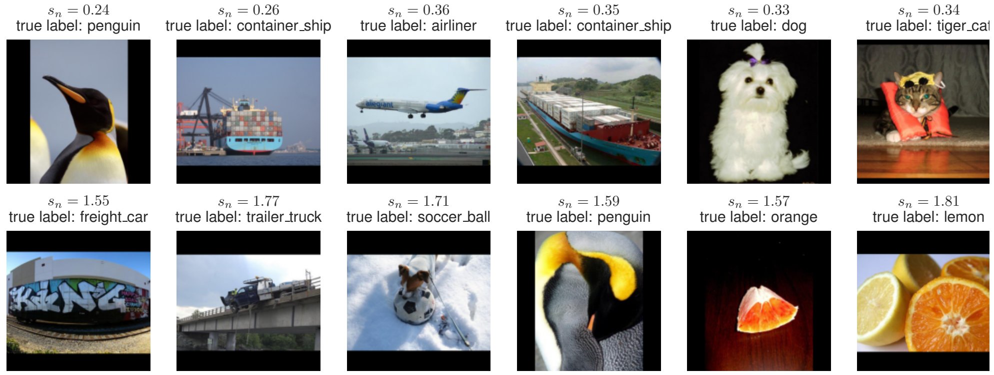

Figure 3: Some examples from ImageNet-15N with low (top) and high (bottom) $s_n$'s

those in the middle. Overall, these results indicate the effectiveness of our outlier-based model in real-world settings.

Fig. 3 shows some examples selected from high-$s_n$ and low-$s_n$ regimes from the results output by `COINNet` in the ImageNet-15N experiment. Similar as before, the images in the first row that have lower $s_n$ scores are visually much easier to recognize. The images in the second row that have about 5 times higher $s_n$ scores are apparently more visually confusing.

More experiments can be found in Appendix H.

## 6 Conclusion

In this work, we considered the noisy label learning problem under a confusion matrix-based model. We developed theory and algorithms in the presence of instance-dependent outliers. Our study revealed that relying solely on single-annotator labels is insufficient for effective outlier detection under the model of interest. We further demonstrated that a crowdsourcing approach, leveraging multiple annotators and a sparsity-constrained loss function, can successfully detect outliers and identify the desired, label noise-free learning system under reasonable conditions. Our analyses and design feature a one-stage differentiable training loss that is optimization-friendly. Empirical results underscore the plausibility of our model and the effectiveness of our proposed method, showing noticeable performance improvements over existing baselines.

**Limitations.** Our work explores a noise transition matrix-based model and treats instance-dependent noisy labels as outliers. The key assumption is that the outliers do not occur overwhelmingly. The assumption is useful, but also could be debatable. In principle, all labels could be generated in an instance-dependent manner. More sophisticated models are needed to deal with this more general case. In addition, our treatment is outlier structure-agnostic. The upshot is that this treatment can also help exclude the negative impacts of other types of outliers. However, the downside is that the structural prior knowledge of the outliers is not fully exploited. If the outlier-generating function could be learned, it can help detect "difficult samples" *before* sending for annotation, which would potentially save resources and reduce annotation errors. We leave this direction for future work.

## Acknowledgment

The work was supported in part by the National Science Foundation (NSF) under Project IIS-2007836.

## References

[1] Devansh Arpit, Stanisław Jastrzundefinedbski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of International Conference on Machine Learning*, page 233–242, 2017.

[2] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *Proceedings of International Conference on Learning Representations*, 2016.

[3] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of International Conference on Machine Learning*, volume 80, pages 2304–2313, 2018.

[4] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of International Conference on Machine Learning*, volume 80, pages 4334–4343, 2018.

[5] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *Proceedings of International Conference on Machine Learning*, pages 7164–7173, 2019.

[6] Eran Malach and Shai Shalev-Shwartz. Decoupling "when to update" from "how to update". In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[7] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, page 8536–8546, 2018.

[8] Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In *Proceedings of International Conference on Machine Learning*, volume 119, pages 10789–10798, 2020.

[9] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[10] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[11] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 322–330, 2019.

[12] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Curriculum loss: Robust learning and generalization against label corruption. In *Proceedings of International Conference on Learning Representations*, 2019.

[13] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. $\mathrm{L}_{DMI}$: A novel information-theoretic loss function for training deep nets robust to label noise. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[14] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics*, pages 20–28, 1979.

[15] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping, 2015.

[16] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

[17] Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, volume 22, pages 2035–2043. 2009.

[18] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(102):1–44, 2016.

[19] Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. In *Proceedings of International Conference on Machine Learning*, pages 6403–6413, 2021.

[20] Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

[21] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:447–461, 2016.

[22] Tri Nguyen, Shahana Ibrahim, and Xiao Fu. Deep clustering with incomplete noisy pairwise annotations: A geometric regularization approach. In *International Conference on Machine Learning*, pages 25980–26007. PMLR, 2023.

[23] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6226–6236, 2020.

[24] Yivan Zhang, Gang Niu, and Masashi Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 12501–12512, 2021.

[25] Kiran K Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional gans to noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[26] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

[27] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *Advances in Neural Information Processing Systems*, volume 33, pages 7597–7610, 2020.

[28] Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *Proceedings of International Conference on Machine Learning*, volume 119, pages 1789–1799, 2020.

[29] David Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in Neural Information Processing Systems*, volume 24, pages 1953–1961, 2011a.

[30] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11236–11245, 2019.

[31] Shahana Ibrahim, Tri Nguyen, and Xiao Fu. Deep learning from crowdsourced labels: Coupled cross-entropy minimization, identifiability, and regularization. In *Proceedings of International Conference on Learning Representations*, 2023.

[32] Kristen Grauman and Bastian Leibe. *Visual Object Recognition*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.

[33] Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. Estimating instance-dependent Bayes-label transition matrix using a deep neural network. In *Proceedings of International Conference on Machine Learning*, 2021.

[34] Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model, 2021.

[35] Zhaowei Zhu, Tongliang Liu, and Yang Liu. A second-order approach to learning with instance-dependent label noise, 2021.

[36] D. Cheng, T. Liu, Y. Ning, N. Wang, B. Han, G. Niu, X. Gao, and M. Sugiyama. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16609–16618, jun 2022.

[37] Shikun Li, Xiaobo Xia, Jiankang Deng, Shiming Ge, and Tongliang Liu. Transferring annotator- and instance-dependent transition matrix for learning from crowds. *arXiv preprint arXiv:2306.03116*, 2023.

[38] Qizhou Wang, Bo Han, Tongliang Liu, Gang Niu, Jian Yang, and Chen Gong. Tackling instance-dependent label noise via a universal probabilistic model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):10183–10191, May 2021.

[39] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *Proceedings of International Conference on Learning Representations*, 2021.

[40] Zhimeng Jiang, Kaixiong Zhou, Zirui Liu, Li Li, Rui Chen, Soo-Hyun Choi, and Xia Hu. An information fusion approach to learning with instance-dependent label noise. In *Proceedings of International Conference on Learning Representations*, 2022.

[41] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. In *Proceedings of International Conference on Learning Representations*, 2021.

[42] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Cicarrelli, Frederik Barkhof, and Daniel Alexander. Disentangling human error from ground truth in segmentation of medical images. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15750–15762. Curran Associates, Inc., 2020.

[43] Hui Guo, Boyu Wang, and Grace Yi. Label correction of crowdsourced noisy annotations with an instance-dependent noise transition model. In *Advances in Neural Information Processing Systems*, 2023.

[44] Qizhou Wang, Jiangchao Yao, Chen Gong, Tongliang Liu, Mingming Gong, Hongxia Yang, and Bo Han. Learning with group noise. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:10192–10200, 05 2021.

[45] Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 12912–12923, 2021.

[46] Xiao Fu, Wing-Kin Ma, José Bioucas-Dias, and Tsung-Han Chan. Semiblind hyperspectral unmixing in the presence of spectral library mismatches. *arXiv preprint arXiv:1507.01661*, 2015.

[47] Qiang Liu, Jian Peng, and Alexander T Ihler. Variational inference for crowdsourcing. In *Advances in neural information processing systems*, volume 25, pages 692–700, 2012.

[48] Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 68–83, 2018.

[49] Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 21475–21496, 2023.

[50] Zhendong Chu, Jing Ma, and Hongning Wang. Learning from crowds by modeling common confusions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[51] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[52] Shahana Ibrahim, Xiao Fu, Nikos Kargas, and Kejun Huang. Crowdsourcing via pairwise co-occurrences: Identifiability and algorithms. In *Advances in Neural Information Processing Systems*, volume 32, pages 7847–7857, 2019.

[53] K. Huang, N. Sidiropoulos, and A. Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Trans. Signal Process.*, 62(1):211–224, 2014.

[54] Xiao Fu, Kejun Huang, Bo Yang, Wing-Kin Ma, and Nicholas D Sidiropoulos. Robust volume minimization-based matrix factorization for remote sensing and document clustering. *IEEE Trans. Signal Process.*, 64(23):6254–6268, 2016.

[55] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[56] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, pages 2424–2432, 2010.

[57] Panagiotis A Traganitis, Alba Pages-Zamora, and Georgios B Giannakis. Blind multiclass ensemble classification. *IEEE Trans. Signal Process.*, 66(18):4737–4752, 2018.

[58] Hansong Zhang, Shikun Li, Dan Zeng, Chenggang Yan, and Shiming Ge. Coupled confusion correction: Learning from crowds with sparse annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16732–16740, 2024.

[59] Shikun Li, Tongliang Liu, Jiyong Tan, Dan Zeng, and Shiming Ge. Trustable co-label learning from multiple noisy annotators. *IEEE Transactions on Multimedia*, 25:1045–1057, 2021.

[60] Nicolas Gillis. *Nonnegative Matrix Factorization*. Society for Industrial and Applied Mathematics, 2020.

[61] Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. Max-Mig: An information theoretic approach for joint learning from crowds. In *Proceedings of International Conference on Learning Representations*, 2019.

[62] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[63] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, volume 15, pages 215–223, 2011.

[64] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[65] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations, ICLR*, 2015.

[66] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

[67] Filipe Rodrigues, Mariana Lourenco, Bernardete Ribeiro, and Francisco C. Pereira. Learning supervised topic models for classification and regression from crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2409–2422, 2017.

[68] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, 2007.

[69] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[70] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[71] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–19, 2022.

[72] Dana Angluin and Philip Laird. Learning from noisy examples. *Mach. Learn.*, 2(4):343–370, apr 1988.

[73] Javed A. Aslam and Scott E. Decatur. On the sample complexity of noise-tolerant learning. *Inf. Process. Lett.*, 57(4):189–195, feb 1996.

[74] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, nov 1998.

[75] Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 489–511, 2013.

[76] Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *Proceedings of International Conference on Machine Learning*, volume 139, pages 825–836, 2021.

[77] Mark A. Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[78] M S Pinsker. The information stability of gaussian random variables and processes (in Russian). *Dokl. Akad. Nauk SSSR*, 133:28–30, 1960.

[79] A.A. Fedotov, P. Harremoes, and F. Topsoe. Refinements of pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6):1491–1498, 2003.

[80] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[81] Andreas Maurer. A vector-contraction inequality for rademacher complexities. *ArXiv*, abs/1605.00251, 2016.

[82] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012.

[83] G. Raskutti, Martin Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls. *IEEE Transactions on Information Theory*, 57:6976 – 6994, 11 2011.

[84] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos. Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain. *IEEE Trans. Signal Process.*, 63(9):2306–2320, May 2015.

[85] Xiao Fu, Kejun Huang, and Nicholas D Sidiropoulos. On identifiability of nonnegative matrix factorization. *IEEE Signal Process. Lett.*, 25(3):328–332, 2018.

[86] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386, 2019.

**Supplementary Material of " Noisy Label Learning with Instance-Dependent Outliers: Identifiability via Crowd Wisdom"**

# A Notation

The notations used in the paper are summarized in Table 3.

Table 3: Definitions of the notations.

| Notation | Definition |
|---|---|
| $x$ | scalar in $\mathbb{R}$ |
| $\boldsymbol{x}$ | vector in $\mathbb{R}^n$, i.e., $\boldsymbol{x} = [x_1, \dots, x_n]^\top$ |
| $[\boldsymbol{x}]_i$ or $\boldsymbol{x}(k)$ | $i$th entry of the vector $\boldsymbol{x}$ |
| $[\boldsymbol{x}]_{i:j}$ | subvector $\boldsymbol{x}$ such that $[\boldsymbol{x}]_{i:j} = [x_i, x_{i+1}, \dots, x_j]^\top$, where $j > i$ |
| $\boldsymbol{X}$ | matrix in $\mathbb{R}^{m \times n}$ |
| $[\boldsymbol{X}]_{i,j}$ or $\boldsymbol{X}(i,j)$ | $(i,j)$th entry of $\boldsymbol{X}$ |
| $\boldsymbol{X}(:,i)$ | $i$th column of $\boldsymbol{X}$ |
| $\boldsymbol{X} \geq \boldsymbol{0}$ | $\boldsymbol{X}(i,j) \geq 0, \forall \, (i,j)$ |
| $\boldsymbol{X}[:]$ | vector representing column-wise stacking of $\boldsymbol{X}$ |
| $\lVert\boldsymbol{x}\rVert_2$ | Euclidean norm of $\boldsymbol{x}$ |
| $\mathsf{nnz}(\boldsymbol{x})$ | number of non-zero entries of $\boldsymbol{x}$ |
| $\lVert\boldsymbol{x}\rVert_0$ | $\ell_0$ norm of $\boldsymbol{x}$; also equal to $\mathsf{nnz}(\boldsymbol{x})$ |
| $\sigma_{\max}(\boldsymbol{X})$ | maximum singular value of $\boldsymbol{X}$ |
| $\sigma_{\min}(\boldsymbol{X})$ | minimum singular value of $\boldsymbol{X}$ |
| $\lVert\boldsymbol{X}\rVert_2$ | spectral norm or 2-norm of $\boldsymbol{X}$, $\lVert\boldsymbol{X}\rVert_2 = \sigma_{\max}(\boldsymbol{X})$ |
| $\lVert\boldsymbol{X}\rVert_{\rm F}$ | Frobenius norm of $\boldsymbol{X}$ |
| $\lVert\boldsymbol{X}\rVert_{2,1}$ | $\ell_{2,1}$ norm of $\boldsymbol{X}$, i.e., $\lVert\boldsymbol{X}\rVert_{2,1} = \sum_{i=1}^{n} \lVert\boldsymbol{X}(:,i)\rVert_2$ |
| $\boldsymbol{\Delta}_K$ | the probability simplex $\{\boldsymbol{x} \in \mathbb{R}^K : \sum_i [\boldsymbol{x}]_i = 1, \boldsymbol{x} \geq \boldsymbol{0}\}$ |
| $\mathsf{cone}(\boldsymbol{X})$ | conic hull of $\boldsymbol{X}$: $\{\boldsymbol{y} \mid \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}, \; \forall \boldsymbol{\theta} \geq \boldsymbol{0}\}$ |
| $\mathsf{convex}(\boldsymbol{X})$ | convex hull of $\boldsymbol{X}$: $\{\boldsymbol{y} \mid \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}, \; \forall \boldsymbol{\theta} \geq \boldsymbol{0}, \boldsymbol{1}^\top \boldsymbol{\theta} = 1\}$ |
| $\dagger$ | pseudo-inverse |
| $\top$ | transpose |
| $\lvert\mathcal{C}\rvert$ | the cardinality of the set $\mathcal{C}$ |
| $[T]$ | $\{1, \dots, T\}$ for an integer $T$ |
| $\boldsymbol{I}_K$ | identity matrix of size $K \times K$ |
| $\boldsymbol{1}$ | all-one vector with proper size |
| $\boldsymbol{0}$ | all-zero vector or matrix with proper size |
| $\overline{\boldsymbol{e}}_i$ | unit vector with the $i$th element being 1 |
| $\mathbb{R}_+^n$ | nonnegative orthant of $\mathbb{R}^n$ |
| $\mathbb{1}[A]$ | Indicator function: $\mathbb{1}[A] = 1$ if the event $A$ happens, otherwise $\mathbb{1}[A] = 0$ |
| $\mathsf{CE}(\boldsymbol{x}, y)$ | cross entropy function: $-\sum_{k=1}^{K} \mathbb{1}[y = k] \log(\boldsymbol{x}(k))$ |

# B Related Work

In this section, we briefly review some of the existing works that is relevant to our proposed method. Noisy label learning has been a prominent research focus within the machine learning community for many years—see a survey of these methods in [71]. Among the numerous approaches proposed, those based on label transition matrix-based models (cf. the model in (2)) have been proven effective and several variants of the model have been introduced; see, e.g., [14–19, 21, 23, 27–29, 33, 47, 52, 55, 56]. A broad classification of them is of instance-independent models and instance-dependent models.

**Instance-independent Model-based Methods.** One of the earliest instance-independent label noise models (i.e., assuming $\boldsymbol{T}^{\natural}(\boldsymbol{x}_n) = \boldsymbol{A}^{\natural}, \forall n$ in (2)) is the *random classification model* (RCN) [72–74], where true labels are flipped with a fixed noise rate independently at random. To extend this modeling to more general scenarios, *class-conditioned model* (CCN) was proposed [9], where the noise rates are assumed to be dependent on the true-class the data item belongs. CCN model has led to the design of several statistically consistent classifiers, as the label transition matrix acts as a correction term for the loss associated with the noisy labels [21, 24, 26, 75]. To provably learn the CCN model, several notions are exploited, e.g., assuming the presence of anchor data items [21], exploiting the clusterability of data items [45], and minimizing the geometric volume of the transition matrix [19].

**Instance-dependent Model-based Approaches.** Instance-dependent label noise models are naturally more plausible. In these models, each data item has a label transition matrix dependent on the data's attributes (cf. Eq. (2)). Recently, a number of methods studied such models by using various simplified representations/parameterizations of the transition matrix $\boldsymbol{T}^{\natural}(\boldsymbol{x}_n)$, e.g., imposing structural assumptions on label transition matrices [27, 41, 43], bounding noise rates [28], modeling the Bayes label transition matrix or posterior transition matrix instead of the true label transition matrix [33, 37, 40], leveraging additional forms of supervision such as confidence scores [76], and crafting specific regularization terms [36, 39, 42]. To be specific, the work in [27] assumed that the label transition matrix is a convex combination of several instance-independent transition matrices and the combination coefficients depend on the instances. The work in [28] assumed a known bounded noise rate and proposed an instance selection-based strategy. The method proposed in [39] also advocated an instance selection scheme, accompanied with a regularization term to combat instance-dependent label noise. In the work [36], the simplification relies on the assumption that many data items that lie close to each other in the feature space have similar label transition matrices. These interesting developments showed empirical progresses, but the theoretical understanding has been limited. Particularly, the aforementioned identifiability issue was not resolved. In addition, most approaches resort to multi-stage training strategies which may be practically inefficient due to potential error propagation. For example, the approach in [33] first performs instance selection, then estimates the (Bayes-label) transition matrices, and finally estimates the classifier, in a three-stage training scheme. Similar multi-stage approaches are employed in [27, 28, 36].

Our approach is also related to sample-selection-based approaches in noisy label learning, e.g., clean sample selection-based methods [28, 33, 35, 37, 39]. However, most of them treat sample selection as a pre-processing step based on strict prior assumptions (e.g., known noise rates) and then estimate the transition matrix using these selected samples. In contrast, our proposed method and identifiability analysis offer a more principled approach by integrating sample selection with the learning of the classifier and transition matrices in an end-to-end fashion.

## C Proof of Lemma 3.1

Lemma 3.1 is re-stated below:

**Lemma C.1.** *Assume that the observed index pairs $(m,n) \in \mathcal{S}$ are sampled uniformly at random with replacement. Also assume $\boldsymbol{f}^{\natural} \in \mathcal{F}$, $|\mathcal{I}| \leq C \leq N/2$, and that each $[\boldsymbol{A}_m \boldsymbol{f}(\boldsymbol{x}_n) + \boldsymbol{e}_n^{(m)}]_k$, $\forall \boldsymbol{A}_m \in \mathcal{A}, \forall \boldsymbol{e}_n^{(m)} \in \mathcal{E}, \forall \boldsymbol{f} \in \mathcal{F}$ are lower bounded by $(1/\beta)$ for a certain $\beta > 0$. Then, with probability greater than $1-\delta$, the following holds:*

$$\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M} \|\boldsymbol{g}_n^{\natural(m)} - \widehat{\boldsymbol{g}}_n^{(m)}\|_2^2 \leq \epsilon_{\boldsymbol{g}}(S), \quad \epsilon_{\boldsymbol{g}}(S) = \mathcal{O}\left( {}^{\beta \mathfrak{R}_{\mathcal{S}} \log S}/_{\sqrt{S}} + \log(\beta) \, {}^{\sqrt{\log(1/\delta)}}/_{\sqrt{S}} \right),$$

*where $\mathfrak{R}_{\mathcal{S}}^2 = K^2 \log(K) + {}^{K\|\boldsymbol{X}\|_{\mathrm{F}}^2 \mathscr{R}_{\mathcal{F}}}/_{S} + CMK \log\sqrt{NM}$ and $\boldsymbol{X} = [\boldsymbol{x}_{n_1}, \ldots, \boldsymbol{x}_{n_S}]$ are the annotated samples, in which $(m_s, n_s) \in \mathcal{S}$ for $s = 1, \ldots, S$.*

The proof is essentially viewing the $\boldsymbol{g}_n^{\natural(m)}$ estimation problem as a quantized matrix completion problem with missing blocks. The key ideas and steps all follow the proof of the end-to-end crowdsourcing work [31, Proposition 1] with the additional consideration of $\boldsymbol{e}_n^{\natural(m)}$. This addition is conceptually not hard to come up with, but the notations can be involved. Thus, we still derive the proof in detail.

**Proof:** Consider the cross entropy-based learning criterion given by

$$\underset{\boldsymbol{A}_m \in \mathcal{A}, \boldsymbol{f} \in \mathcal{F}, \boldsymbol{e}_n^{(m)} \in \mathcal{E}}{\text{minimize}} \ \frac{1}{S} \sum_{(m,n) \in \mathcal{S}} \mathsf{CE}(\boldsymbol{A}_m \boldsymbol{f}(\boldsymbol{x}_n) + \boldsymbol{e}_n^{(m)}, \widehat{y}_n^{(m)}), \tag{11a}$$

$$\text{subject to} \ \sum_{n=1}^{N} \mathbb{1}\left\{ \sum_{m=1}^{M} \|\boldsymbol{e}_n^{(m)}\|_2 > 0 \right\} \leq C. \tag{11b}$$

where $S = |\mathcal{S}|$, and $\mathsf{CE}(\boldsymbol{x}, y) = -\sum_{k=1}^{K} \mathbb{1}[y = k] \log(\boldsymbol{x}(k))$.

Let us denote the objective function in (11a) as follows:

$$\mathsf{L}_{\mathcal{S}}(\boldsymbol{G}; \widehat{\mathcal{Y}}) \triangleq \frac{1}{S} \sum_{(m,n) \in \mathcal{S}} \mathsf{CE}(\boldsymbol{g}_n^{(m)}, \widehat{y}_n^{(m)}), \tag{12}$$

where $\boldsymbol{g}_n^{(m)}$ is given by $\boldsymbol{g}_n^{(m)} = \boldsymbol{A}_m \boldsymbol{f}(\boldsymbol{x}_n) + \boldsymbol{e}_n^{(m)}$, $\boldsymbol{G}$ is given by

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{g}_1^{(1)} & \dots & \boldsymbol{g}_N^{(1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{g}_1^{(M)} & \dots & \boldsymbol{g}_N^{(M)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_1 \boldsymbol{f}(\boldsymbol{x}_1) + \boldsymbol{e}_1^{(1)} & \dots & \boldsymbol{A}_1 \boldsymbol{f}(\boldsymbol{x}_N) + \boldsymbol{e}_N^{(1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{A}_M \boldsymbol{f}(\boldsymbol{x}_1) + \boldsymbol{e}_1^{(M)} & \dots & \boldsymbol{A}_M \boldsymbol{f}(\boldsymbol{x}_N) + \boldsymbol{e}_N^{(M)} \end{bmatrix},$$

and $\widehat{\mathcal{Y}}$ denotes the set of observed noisy labels, i.e., $\{\widehat{y}_n^{(m)}\}_{(m,n) \in \mathcal{S}}$. Let $\{\widehat{\boldsymbol{A}}_m\}$, $\{\widehat{\boldsymbol{e}}_n^{(m)}\}$ and $\widehat{\boldsymbol{f}}$ denote optimal solutions of the learning criterion (11). Let us also define the following:

$$\widehat{\boldsymbol{G}} \triangleq \underset{\boldsymbol{G} \in \mathcal{G}}{\text{arg min}} \ \mathsf{L}_{\mathcal{S}}(\boldsymbol{G}; \widehat{\mathcal{Y}}), \tag{13}$$

where

$\mathcal{G} \triangleq \{\boldsymbol{G} \in \mathbb{R}^{MK \times N} \mid \boldsymbol{g}_n^{(m)} = \boldsymbol{A}_m \boldsymbol{f}(\boldsymbol{x}_n) + \boldsymbol{e}_n^{(m)}, \boldsymbol{f} \in \mathcal{F}, \boldsymbol{e}_n^{(m)} \in \mathcal{E}, \|\boldsymbol{E}\|_0 \leq C, \boldsymbol{A}_m \in \mathcal{A}\}$,

$\mathcal{F} \subset \{\boldsymbol{f} : \mathbb{R}^D \to \mathbb{R}^K\}$ is the neural network function class associated with the classifier function,

$$\mathcal{E} = \{\boldsymbol{e} \in \mathbb{R}^K \mid \boldsymbol{1}^\top \boldsymbol{e} = 0\}, \ \|\boldsymbol{E}\|_0 = \sum_{n=1}^{N} \mathbb{1}\left\{ \sum_{m=1}^{M} \|\boldsymbol{e}_n^{(m)}\|_2 > 0 \right\},$$

$\mathcal{A} = \{\boldsymbol{A} \in \mathbb{R}^{K \times K} \mid \boldsymbol{1}^\top \boldsymbol{A} = \boldsymbol{1}^\top, \ \boldsymbol{A} \geq \boldsymbol{0}\}$.

Let $\boldsymbol{G}^\natural$ represents the ground-truth such that

$$\boldsymbol{G}^\natural = \begin{bmatrix} \boldsymbol{g}_1^{\natural(1)} & \dots & \boldsymbol{g}_N^{\natural(1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{g}_1^{\natural(M)} & \dots & \boldsymbol{g}_N^{\natural(M)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_1^\natural \boldsymbol{f}^\natural(\boldsymbol{x}_1) + \boldsymbol{e}_1^{\natural(1)} & \dots & \boldsymbol{A}_1^\natural \boldsymbol{f}^\natural(\boldsymbol{x}_N) + \boldsymbol{e}_N^{\natural(1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{A}_M^\natural \boldsymbol{f}^\natural(\boldsymbol{x}_1) + \boldsymbol{e}_1^{\natural(M)} & \dots & \boldsymbol{A}_M^\natural \boldsymbol{f}^\natural(\boldsymbol{x}_N) + \boldsymbol{e}_N^{\natural(M)} \end{bmatrix}. \tag{14}$$

We aim to bound $\|\boldsymbol{G}^\natural - \widehat{\boldsymbol{G}}\|_{\mathrm{F}}$. Towards this goal, we adopt the proof strategy based on the viewpoint of matrix completion as employed in [31, 77].

Let us consider that we have $S$ number of observations $\omega_1, \dots, \omega_S$ such that $\omega_s = (m_s, n_s, \widehat{y}_{n_s}^{(m_s)})$, where each $(m_s, n_s)$ is sampled uniformly at random (with replacement) from $[M] \times [N]$, i.e., the sampling probability for each $(m, n)$ is given by $\pi_{m,n} = \frac{1}{NM}$. In addition, the noisy observations are generated as

$$\widehat{y}_{n_s}^{(m_s)} \sim \mathsf{categorical}(\boldsymbol{g}_{n_s}^{\natural(m_s)}).$$

Let $\mathcal{D}_\omega$ denote the joint probability distribution from where the observations $\omega_s$ are sampled. Then we define the following:

$$\mathsf{L}_{\mathcal{D}}(\boldsymbol{G}, \widehat{\mathcal{Y}}) \triangleq \mathbb{E}_{\omega \sim \mathcal{D}_\omega}[\mathsf{CE}(\boldsymbol{g}_n^{(m)}, \widehat{y}_n^{(m)})].$$

We consider the following set of relations:

$$\begin{aligned}
\mathsf{L}_{\mathcal{D}}(\widehat{\boldsymbol{G}},\widehat{\mathcal{Y}})-\mathsf{L}_{\mathcal{D}}(\boldsymbol{G}^{\natural},\widehat{\mathcal{Y}}) &= \mathsf{L}_{\mathcal{D}}(\widehat{\boldsymbol{G}},\widehat{\mathcal{Y}})-\mathsf{L}_{\mathcal{S}}(\widehat{\boldsymbol{G}};\widehat{\mathcal{Y}})+\mathsf{L}_{\mathcal{S}}(\boldsymbol{G}^{\natural};\widehat{\mathcal{Y}})-\mathsf{L}_{\mathcal{D}}(\boldsymbol{G}^{\natural},\widehat{\mathcal{Y}}) \\
&\quad +\mathsf{L}_{\mathcal{S}}(\widehat{\boldsymbol{G}};\widehat{\mathcal{Y}})-\mathsf{L}_{\mathcal{S}}(\boldsymbol{G}^{\natural};\widehat{\mathcal{Y}}) \\
&\leq \mathsf{L}_{\mathcal{D}}(\widehat{\boldsymbol{G}},\widehat{\mathcal{Y}})-\mathsf{L}_{\mathcal{S}}(\widehat{\boldsymbol{G}};\widehat{\mathcal{Y}})+\mathsf{L}_{\mathcal{S}}(\boldsymbol{G}^{\natural};\widehat{\mathcal{Y}})-\mathsf{L}_{\mathcal{D}}(\boldsymbol{G}^{\natural},\widehat{\mathcal{Y}}) \\
&\leq 2\sup_{\boldsymbol{G}\in\mathcal{G}}\left|\mathsf{L}_{\mathcal{D}}(\boldsymbol{G},\widehat{\mathcal{Y}})-\mathsf{L}_{\mathcal{S}}(\boldsymbol{G};\widehat{\mathcal{Y}})\right|, \qquad (15)
\end{aligned}$$

where the first inequality is obtained since $\widehat{\boldsymbol{G}}$ is the optimal solution as given by (13) and hence satisfies $\mathsf{L}_{\mathcal{S}}(\widehat{\boldsymbol{G}};\widehat{\mathcal{Y}})\leq\mathsf{L}_{\mathcal{S}}(\boldsymbol{G}^{\natural};\widehat{\mathcal{Y}})$. The last inequality is from the given assumption that $\boldsymbol{G}^{\natural}\in\mathcal{G}$ (This is satisfied since it is given that $\boldsymbol{f}^{\natural}\in\mathcal{F}$ and $\|\boldsymbol{E}^{\natural}\|_0\leq C$).

Next, we consider the following set of relations:

$$\begin{aligned}
\mathsf{L}_{\mathcal{D}}(\widehat{\boldsymbol{G}},\widehat{\mathcal{Y}})-\mathsf{L}_{\mathcal{D}}(\boldsymbol{G}^{\natural},\widehat{\mathcal{Y}}) &= \mathbb{E}_{\omega\sim\mathcal{D}_{\omega}}[\mathsf{CE}(\widehat{\boldsymbol{g}}_n^{(m)},\widehat{y}_n^{(m)})-\mathsf{CE}(\boldsymbol{g}_n^{\natural(m)},\widehat{y}_n^{(m)})] \\
&= \mathbb{E}_{(m,n,\widehat{y}_n^m)\sim\mathcal{D}_{\omega}}\left[-\sum_{k=1}^{K}\mathbb{1}[\widehat{y}_n^{(m)}=k]\log[\widehat{\boldsymbol{g}}_n^{(m)}]_k+\sum_{k=1}^{K}\mathbb{1}[\widehat{y}_n^{(m)}=k]\log[\boldsymbol{g}_n^{\natural(m)}]_k\right] \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N}\pi_{m,n}\mathbb{E}_{\widehat{y}_n^m}\left[-\sum_{k=1}^{K}\mathbb{1}[\widehat{y}_n^{(m)}=k]\log[\widehat{\boldsymbol{g}}_n^{(m)}]_k+\sum_{k=1}^{K}\mathbb{1}[\widehat{y}_n^{(m)}=k]\log[\boldsymbol{g}_n^{\natural(m)}]_k\right] \\
&= \frac{1}{NM}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\sum_{k=1}^{K}-[\boldsymbol{g}_n^{\natural(m)}]_k\log[\widehat{\boldsymbol{g}}_n^{(m)}]_k+\sum_{k=1}^{K}[\boldsymbol{g}_n^{\natural(m)}]_k\log[\boldsymbol{g}_n^{\natural(m)}]_k\right) \\
&= \frac{1}{NM}\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}[\boldsymbol{g}_n^{\natural(m)}]_k\log\frac{[\boldsymbol{g}_n^{\natural(m)}]_k}{[\widehat{\boldsymbol{g}}_n^{(m)}]_k} \\
&= \frac{1}{NM}\sum_{m=1}^{M}\sum_{n=1}^{N}\mathsf{D}_{\mathsf{KL}}(\boldsymbol{g}_n^{\natural(m)},\widehat{\boldsymbol{g}}_n^{(m)}) \\
&\geq \frac{1}{2NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\|\boldsymbol{g}_n^{\natural(m)}-\widehat{\boldsymbol{g}}_n^{(m)}\|_1^2 \\
&\geq \frac{1}{2NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\|\boldsymbol{g}_n^{\natural(m)}-\widehat{\boldsymbol{g}}_n^{(m)}\|_2^2 \qquad (16)
\end{aligned}$$

where we applied Pinsker's inequality [78, 79] for the first inequality and the last inequality uses the fact that $\|\boldsymbol{x}\|_1\geq\|\boldsymbol{x}\|_2$ for any vector $\boldsymbol{x}\in\mathbb{R}^K$ having $|[\boldsymbol{x}]_k|\leq 1,\forall k\in[K]$. Here $\mathsf{D}_{\mathsf{KL}}(\boldsymbol{g}_1,\boldsymbol{g}_2)$ denotes the Kullback–Leibler divergence between any $\boldsymbol{g}_1,\boldsymbol{g}_2\in\boldsymbol{\Delta}_K$, i.e., $\mathsf{D}_{\mathsf{KL}}(\boldsymbol{g}_1,\boldsymbol{g}_2)=\sum_{k=1}^{K}[g_1]_k\log\frac{[g_1]_k}{[g_2]_k]}$.

Combining (15) with (16), we get

$$\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\|\boldsymbol{g}_n^{\natural(m)}-\widehat{\boldsymbol{g}}_n^{(m)}\|_2^2\leq 4\sup_{\boldsymbol{G}\in\mathcal{G}}\left|\mathsf{L}_{\mathcal{D}}(\boldsymbol{G},\widehat{\mathcal{Y}})-\mathsf{L}_{\mathcal{S}}(\boldsymbol{G};\widehat{\mathcal{Y}})\right|. \qquad (17)$$

Next, we proceed to characterize the R.H.S. of (17). To achieve this, we invoke the following theorem derived from Theorem 26.5 in [80]:

**Theorem C.2.** *[80, Theorem 26.5] Assume that $\forall(m,n,\widehat{y}_n^{(m)})$ and $\forall\boldsymbol{G}\in\mathcal{G}$, we have $|\mathsf{CE}(\boldsymbol{g}_n^{(m)},\widehat{y}_n^{(m)})|\leq z_{\max}$. Then for any $\boldsymbol{G}\in\mathcal{G}$, the following holds with probability greater than $1-\delta$:*

$$\left|\mathsf{L}_{\mathcal{D}}(\boldsymbol{G},\widehat{\mathcal{Y}})-\mathsf{L}_{\mathcal{S}}(\boldsymbol{G};\widehat{\mathcal{Y}})\right|\leq 2\mathfrak{R}(\ell\circ\mathcal{G}\circ\mathcal{S})+4z_{\max}\sqrt{\frac{2\log(4/\delta)}{S}}, \qquad (18)$$

*where $\ell\circ\mathcal{G}\circ\mathcal{S}$ denotes the set*

$$\ell\circ\mathcal{G}\circ\mathcal{S}\triangleq\left\{\left(\mathsf{CE}(\boldsymbol{g}_{n_1}^{(m_1)},\widehat{y}_{n_1}^{(m_1)}),\ldots,\mathsf{CE}(\boldsymbol{g}_{n_S}^{(m_S)},\widehat{y}_{n_S}^{(m_S)})\right)\mid\boldsymbol{G}\in\mathcal{G}\right\}$$

*and $\mathfrak{R}(\mathcal{X})$ denotes the empirical Rademacher complexity of the set $\mathcal{X}$.*

To apply Theorem C.2, we need to characterize the terms $z_{\max}$ and $\mathfrak{R}(\ell \circ \mathcal{G} \circ \mathcal{S})$. The term $z_{\max}$ can be characterized as below:

$$z_{\max} = \max_{\substack{[\boldsymbol{g}]_k > \frac{1}{\beta} \\ y \in [K]}} \mathsf{CE}(\boldsymbol{g}, y) \leq \max_{\substack{[\boldsymbol{g}]_k > \frac{1}{\beta} \\ y \in [K]}} -\sum_{k=1}^{K} \mathbb{1}[y = k] \log[\boldsymbol{g}]_k \leq \max_{[\boldsymbol{g}]_k > \frac{1}{\beta}} -\log[\boldsymbol{g}]_k = \log(\beta). \tag{19}$$

Note that in the above relation (19), we utilize the given assumption that each $[\boldsymbol{g}]_k \geq 1/\beta, \forall \boldsymbol{g} \in \mathcal{G}$.

To characterize the Rademacher complexity $\mathfrak{R}(\ell \circ \mathcal{G} \circ \mathcal{S})$, we consider its definition as follows [80]:

$$\mathfrak{R}(\ell \circ \mathcal{G} \circ \mathcal{S}) \triangleq \frac{1}{S} \mathbb{E}\left[\sup_{\boldsymbol{G} \in \mathcal{G}} \sum_{s=1}^{S} \sigma_s \mathsf{CE}(\boldsymbol{g}_{n_s}^{(m_s)}, \widehat{y}_{n_s}^{(m_s)})\right], \tag{20}$$

where expectation is w.r.t. the independent Rademacher random variables $\sigma_s \in \{-1, 1\}$. Since $\boldsymbol{g}_n^{(m)}$ are vectors, we utilize the following result to upper bound $\mathfrak{R}(\ell \circ \mathcal{G} \circ \mathcal{S})$:

**Lemma C.3.** *[81] Assume that the function* $\mathsf{CE}(\cdot, y), \forall y$ *has the Lipschitz constant L. Then*

$$\mathbb{E}\left[\mathit{sup}_{\boldsymbol{G} \in \mathcal{G}} \sum_{s=1}^{S} \sigma_s \mathsf{CE}(\boldsymbol{g}_{n_s}^{(m_s)}, \widehat{y}_{n_s}^{(m_s)})\right] \leq \sqrt{2} L \mathbb{E}\left[\mathit{sup}_{\boldsymbol{G} \in \mathcal{G}} \sum_{s=1}^{S} \sum_{k=1}^{K} \sigma_{sk} [\boldsymbol{g}_{n_s}^{(m_s)}]_k\right],$$

*where $\sigma_{sk}$ is an independent (doubly indexed) Rademacher random variable and the expectations are taken w.r.t. the Rademacher random variables.*

To apply Lemma C.3, let us define a vector $\boldsymbol{z} \triangleq \left([\boldsymbol{g}_{n_1}^{(m_1)}]_1, [\boldsymbol{g}_{n_1}^{(m_1)}]_2, \ldots, [\boldsymbol{g}_{n_S}^{(m_S)}]_{K-1}, [\boldsymbol{g}_{n_S}^{(m_S)}]_K\right) \in \mathbb{R}^{SK}$ and the set $\mathcal{Z} \circ \mathcal{S} \triangleq \{\boldsymbol{z} = \left([\boldsymbol{g}_{n_1}^{(m_1)}]_1, [\boldsymbol{g}_{n_1}^{(m_1)}]_2, \ldots, [\boldsymbol{g}_{n_S}^{(m_S)}]_{K-1}, [\boldsymbol{g}_{n_S}^{(m_S)}]_K\right) \mid \boldsymbol{G} \in \mathcal{G}\}$. Using these definitions, let us apply Lemma C.3 in (20) and get the following relation:

$$\begin{aligned}
\mathfrak{R}(\ell \circ \mathcal{G} \circ \mathcal{S}) &\leq \frac{\sqrt{2}\beta}{S} \mathbb{E}\left[\sup_{\boldsymbol{G} \in \mathcal{G}} \sum_{s=1}^{S} \sum_{k=1}^{K} \sigma_{sk} [\boldsymbol{g}_{n_s}^{(m_s)}]_k\right] \\
&= \frac{\sqrt{2}\beta}{S} \mathbb{E}\left[\sup_{\boldsymbol{z} \in \mathcal{Z} \circ \mathcal{S}} \sum_{i}^{SK} \sigma_i [\boldsymbol{z}]_i\right] \\
&= \sqrt{2}\beta K \frac{1}{SK} \mathbb{E}\left[\sup_{\boldsymbol{z} \in \mathcal{Z} \circ \mathcal{S}} \sum_{i=1}^{SK} \sigma_i [\boldsymbol{z}]_i\right] = \sqrt{2}\beta K \mathfrak{R}(\mathcal{Z} \circ \mathcal{S}),
\end{aligned} \tag{21}$$

where $\beta$ is an upper bound of the Lipschitz constant of the cross entropy loss function $\mathsf{CE}(\boldsymbol{g}, y) = -\sum_{k=1}^{K} \mathbb{1}[y = k] \log[\boldsymbol{g}]_k$ when $\boldsymbol{g} \in \boldsymbol{\Delta}_K$ with $[\boldsymbol{x}]_k > (1/\beta)$. $\forall k$.

Next, we will characterize $\mathfrak{R}(\mathcal{Z} \circ \mathcal{S})$ using the covering number of the set $\mathcal{Z} \circ \mathcal{S}$. To define the covering number, let us first consider the concept of $\epsilon$-net covering of a set. The $\epsilon$-net covering of a set $\mathcal{X}$ (denoted as $\overline{\mathcal{X}}_\epsilon$) is defined as the finite subset of $\mathcal{X}$ (i.e.,$\overline{\mathcal{X}}_\epsilon \subseteq \mathcal{X}$) such that for any $\boldsymbol{x} \in \mathcal{X}$, there exists an $\overline{\boldsymbol{x}} \in \overline{\mathcal{X}}_\epsilon$ satisfying $\|\boldsymbol{x} - \overline{\boldsymbol{x}}\|_2 \leq \epsilon$ [82]. The smallest cardinality of the $\epsilon$-nets of $\mathcal{X}$ is known as the covering number of $\mathcal{X}$, which is denoted as $\mathsf{N}(\epsilon, \mathcal{X})$ .

Let us consider a pair of vectors $\boldsymbol{z}_{\mathcal{S}}, \overline{\boldsymbol{z}}_{\mathcal{S}} \in \mathcal{Z} \circ \mathcal{S}$ as below:

$$\begin{aligned}
\boldsymbol{z}_{\mathcal{S}} &= \left([\boldsymbol{g}_{n_1}^{(m_1)}]_1, [\boldsymbol{g}_{n_1}^{(m_1)}]_2, \ldots, [\boldsymbol{g}_{n_S}^{(m_S)}]_{K-1}, [\boldsymbol{g}_{n_S}^{(m_S)}]_K\right) \in \mathbb{R}^{SK},\ \boldsymbol{g}_{n_s}^{(m_s)} = \boldsymbol{A}_{m_s} \boldsymbol{f}(\boldsymbol{x}_{n_s}) + \boldsymbol{e}_{n_s}^{(m_s)}, \\
\overline{\boldsymbol{z}}_{\mathcal{S}} &= \left([\overline{\boldsymbol{g}}_{n_1}^{(m_1)}]_1, [\overline{\boldsymbol{g}}_{n_1}^{(m_1)}]_2, \ldots, [\overline{\boldsymbol{g}}_{n_S}^{(m_S)}]_{K-1}, [\overline{\boldsymbol{g}}_{n_S}^{(m_S)}]_K\right) \in \mathbb{R}^{SK},\ \overline{\boldsymbol{g}}_{n_s}^{(m_s)} = \overline{\boldsymbol{A}}_{m_s} \overline{\boldsymbol{f}}(\boldsymbol{x}_{n_s}) + \overline{\boldsymbol{e}}_{n_s}^{(m_s)},
\end{aligned}$$

where $\overline{\boldsymbol{A}}_{m_s}$ belongs to the $\epsilon$-net covering of $\mathcal{A}$, $\overline{\boldsymbol{f}}$ belongs to the $\epsilon$-net covering of $\mathcal{F}$ , and $\overline{\boldsymbol{e}}_{n_s}^{(m_s)}$ is the appropriate sub-vector belonging to the $\epsilon$-net covering of $\mathcal{E}_0$.

Let us also consider the following definitions:

$$
\begin{aligned}
\boldsymbol{u}_{\mathcal{S}} &= ([\boldsymbol{A}_{m_1}\boldsymbol{f}(\boldsymbol{x}_{n_1})]_1, [\boldsymbol{A}_{m_1}\boldsymbol{f}(\boldsymbol{x}_{n_1})]_2, \dots, [\boldsymbol{A}_{m_S}\boldsymbol{f}(\boldsymbol{x}_{n_S})]_{K-1}, [\boldsymbol{A}_{m_S}\boldsymbol{f}(\boldsymbol{x}_{n_S})]_K) \in \mathbb{R}^{SK},\\
\overline{\boldsymbol{u}}_{\mathcal{S}} &= \big([\overline{\boldsymbol{A}}_{m_1}\overline{\boldsymbol{f}}(\boldsymbol{x}_{n_1})]_1, [\overline{\boldsymbol{A}}_{m_1}\overline{\boldsymbol{f}}(\boldsymbol{x}_{n_1})]_2, \dots, [\overline{\boldsymbol{A}}_{m_S}\overline{\boldsymbol{f}}(\boldsymbol{x}_{n_S})]_{K-1}, [\overline{\boldsymbol{A}}_{m_S}\overline{\boldsymbol{f}}(\boldsymbol{x}_{n_S})]_K\big) \in \mathbb{R}^{SK},\\
\boldsymbol{e}_{\mathcal{S}} &= \Big([\boldsymbol{e}_{n_1}^{(m_1)}]_1, [\boldsymbol{e}_{n_1}^{(m_1)}]_2, \dots, [\boldsymbol{e}_{n_S}^{(m_S)}]_{K-1}, [\boldsymbol{e}_{n_S}^{(m_S)}]_K\Big) \in \mathbb{R}^{SK},\\
\overline{\boldsymbol{e}}_{\mathcal{S}} &= \Big([\overline{\boldsymbol{e}}_{n_1}^{(m_1)}]_1, [\overline{\boldsymbol{e}}_{n_1}^{(m_1)}]_2, \dots, [\overline{\boldsymbol{e}}_{n_S}^{(m_S)}]_{K-1}, [\overline{\boldsymbol{e}}_{n_S}^{(m_S)}]_K\Big) \in \mathbb{R}^{SK},\\
\boldsymbol{e}_{\mathcal{NM}} &= \Big([\boldsymbol{e}_1^{(1)}]_1, [\boldsymbol{e}_1^{(1)}]_2, \dots, [\boldsymbol{e}_N^{(M)}]_{K-1}, [\boldsymbol{e}_N^{(M)}]_K\Big) \in \mathbb{R}^{NMK},\\
\overline{\boldsymbol{e}}_{\mathcal{NM}} &= \Big([\overline{\boldsymbol{e}}_1^{(1)}]_1, [\overline{\boldsymbol{e}}_1^{(1)}]_2, \dots, [\overline{\boldsymbol{e}}_N^{(M)}]_{K-1}, [\overline{\boldsymbol{e}}_N^{(M)}]_K\Big) \in \mathbb{R}^{NMK},
\end{aligned}
$$

Then, we have the following relation due to triangle inequality

$$\|\boldsymbol{z}_{\mathcal{S}} - \overline{\boldsymbol{z}}_{\mathcal{S}}\|_2 \le \|\boldsymbol{u}_{\mathcal{S}} - \overline{\boldsymbol{u}}_{\mathcal{S}}\|_2 + \|\boldsymbol{e}_{\mathcal{S}} - \overline{\boldsymbol{e}}_{\mathcal{S}}\|_2. \tag{22}$$

Next, consider the second term on the R.H.S. of (22):

$$\|\boldsymbol{e}_{\mathcal{S}} - \overline{\boldsymbol{e}}_{\mathcal{S}}\|_2^2 = \sum_{s=1}^{S} \|\boldsymbol{e}_{n_s}^{(m_s)} - \overline{\boldsymbol{e}}_{n_s}^{(m_s)}\|_2^2 \le S\|\boldsymbol{e}_{\mathcal{NM}} - \overline{\boldsymbol{e}}_{\mathcal{NM}}\|_2^2. \tag{23}$$

We also have the following relations for the first term on the R.H.S. of (22) :

$$
\begin{aligned}
\|\boldsymbol{u}_{\mathcal{S}} - \overline{\boldsymbol{u}}_{\mathcal{S}}\|^2 &= \sum_{s=1}^{S}\|\boldsymbol{A}_{m_s}\boldsymbol{f}(\boldsymbol{x}_{n_s}) - \overline{\boldsymbol{A}}_{m_s}\overline{\boldsymbol{f}}(\boldsymbol{x}_{n_s})\|_2^2\\
&= \sum_{s=1}^{S}\|\boldsymbol{A}_{m_s}\boldsymbol{f}(\boldsymbol{x}_{n_s}) - \overline{\boldsymbol{A}}_{m_s}\boldsymbol{f}(\boldsymbol{x}_{n_s}) + \overline{\boldsymbol{A}}_{m_s}\boldsymbol{f}(\boldsymbol{x}_{n_s}) - \overline{\boldsymbol{A}}_{m_s}\overline{\boldsymbol{f}}(\boldsymbol{x}_{n_s})\|_2^2\\
&\le \sum_{s=1}^{S}\big(\|\boldsymbol{A}_{m_s} - \overline{\boldsymbol{A}}_{m_s}\|_{\mathrm{F}}\|\boldsymbol{f}(\boldsymbol{x}_{n_s})\|_2 + \|\overline{\boldsymbol{A}}_{m_s}\|_{\mathrm{F}}\|\boldsymbol{f}(\boldsymbol{x}_{n_s}) - \overline{\boldsymbol{f}}(\boldsymbol{x}_{n_s})\|\big)^2\\
&\le 2\sum_{s=1}^{S}\|\boldsymbol{A}_{m_s} - \overline{\boldsymbol{A}}_{m_s}\|_{\mathrm{F}}^2\|\boldsymbol{f}(\boldsymbol{x}_{n_s})\|_2^2 + 2\sum_{s=1}^{S}\|\overline{\boldsymbol{A}}_{m_s}\|_{\mathrm{F}}^2\|\boldsymbol{f}(\boldsymbol{x}_{n_s}) - \overline{\boldsymbol{f}}(\boldsymbol{x}_{n_s})\|^2\\
&\le 2\sum_{s=1}^{S}\|\boldsymbol{A}_{m_s} - \overline{\boldsymbol{A}}_{m_s}\|_{\mathrm{F}}^2 + 2K\sum_{s=1}^{S}\|\boldsymbol{f}(\boldsymbol{x}_{n_s}) - \overline{\boldsymbol{f}}(\boldsymbol{x}_{n_s})\|^2.
\end{aligned} \tag{24}
$$

In the above set of relations, the first inequality is by triangle inequality. The second inequality uses that fact that $(a+b)^2 \le 2a^2 + 2b^2$. The third inequality uses the fact that the Frobenius norm of $\overline{\boldsymbol{A}}_m(\boldsymbol{x}_n)$'s are bounded by $\sqrt{K}$ and the $\ell_2$ norm of $\boldsymbol{f}(\boldsymbol{x}_n)$ is bounded by 1. Hence, combining (22)-(24), we get that in order to obtain an $\varepsilon$-net covering for the set $\mathcal{Z} \circ \mathcal{S}$ (i.e., $\|\boldsymbol{z} - \overline{\boldsymbol{z}}\|_2 \le \varepsilon$), we only need to show that there exists a $\frac{\varepsilon}{4\sqrt{K}}$-net covering for $\mathcal{F} \circ \mathcal{S}$, an $\frac{\varepsilon}{2\sqrt{S}}$-net covering for $\mathcal{E}_{\mathcal{NM}}$, $\forall m$, and a $\frac{\varepsilon}{4\sqrt{S}}$-net covering for each $\mathcal{A}$'s. Here we have

$$
\begin{aligned}
&\mathcal{F} \circ \mathcal{S} = \{[\boldsymbol{f}(\boldsymbol{x}_1), \dots, \boldsymbol{f}(\boldsymbol{x}_S)] \in \mathbb{R}^{K \times S} \mid \boldsymbol{f} \in \mathcal{F}, (m,n) \in \mathcal{S}\},\\
&\mathcal{E}_{\mathcal{NM}} = \{[\boldsymbol{e}_1^{(1)\top}, \dots, \boldsymbol{e}_N^{(M)\top}]^\top \in \mathbb{R}^{KNM} \mid \boldsymbol{e}_n^{(m)} \in \mathcal{E}, \|\boldsymbol{E}\|_0 \le C, \},
\end{aligned}
$$

$\|\boldsymbol{E}\|_0 = \sum_{n=1}^{N} \mathbb{1}\left\{\sum_{m=1}^{M} \|\boldsymbol{e}_n^{(m)}\|_2 > 0\right\}$. Note that the full rank matrix $\boldsymbol{A}_m \in \mathbb{R}^{K\times K}$ can be represented as a $K^2$-dimensional vector whose Euclidean norm is bounded by $\sqrt{K}$. Hence, the cardinality of the $\frac{\varepsilon}{4\sqrt{S}}$-net covering for $\boldsymbol{A}_m \in \mathbb{R}^{K\times K}$ is at most $\left(\frac{8K\sqrt{SK}}{\varepsilon}\right)^{K^2}$ [80]. We consider the covering number corresponding to the function class $\mathcal{F} \circ \mathcal{S}$. Towards this, we invoke the following lemma:

**Lemma C.4.** *[51, Theorem 3.3] Let fixed nonlinearities $(\sigma_1, \dots, \sigma_L)$ and reference matrices $(M_1, \dots, M_L)$ be given, where $\sigma_i$ is $\rho_i$-Lipschitz and $\sigma_i(0) = 0$. Let spectral norm bounds $(s_1, \dots, s_L)$, and matrix (2,1)-norm bounds $b_1, \dots, b_L$ be given. Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_S$ be the $S$ number of data items. Let $\boldsymbol{h}$ belongs to the following function class:*

$$\mathcal{H} \triangleq \{\boldsymbol{h} : \mathbb{R}^D \to \mathbb{R}^K \mid \boldsymbol{h}(\boldsymbol{x}) = \sigma_L(\boldsymbol{C}_1 \sigma_{L-1}(\boldsymbol{C}_{L-1} \dots \sigma_1(\boldsymbol{C}_1(\boldsymbol{x}) \dots )), \forall x \in \mathbb{R}^D\},$$

*where $\boldsymbol{C}_i \in \mathbb{R}^{d_i \times d_{i-1}}$ with $d_0 = D$ and $d_L = K$. Let $\mathcal{H} \circ \mathcal{S}$ denote the family of matrices obtained by evaluating the data items with all choices of network $\mathcal{H}$:*

$$\mathcal{H} \circ \mathcal{S} \triangleq \{[\boldsymbol{h}(\boldsymbol{x}_1), \dots, \boldsymbol{h}(\boldsymbol{x}_S)] \mid \mathcal{C} = (\boldsymbol{C}_1, \dots, \boldsymbol{C}_L), \|\boldsymbol{C}_i\|_2 \le s_i, \|\boldsymbol{C}_i - \boldsymbol{M}_i\|_{2,1} \le b_i, \boldsymbol{h} \in \mathcal{H}\},$$

*Then for any $\epsilon > 0$,*

$$\log \overline{\mathsf{N}}(\epsilon, \mathcal{H} \circ \mathcal{S}) \le \frac{\|\boldsymbol{X}\|_{\mathrm{F}}^2 \mathscr{R}_{\mathcal{H}}}{\epsilon^2},$$

*where $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_S]$, $\mathscr{R}_{\mathcal{H}} = \log(2H^2) \left(\prod_{j=1}^L s_j^2 \rho_j^2\right) \left(\sum_L^i (b_i/s_i)^{2/3}\right)^3$ is called as the spectral-complexity upper bound of neural network function class $\mathcal{H}$ and $H = \max_\ell d_\ell$.*

Using Lemma C.4, we get the cardinality of the $\frac{\varepsilon}{8\sqrt{SK}}$-net covering for $\mathcal{F} \circ \mathcal{N}$ as below:

$$\overline{\mathsf{N}}\left(\frac{\varepsilon}{8\sqrt{SK}}, \mathcal{F} \circ \mathcal{S}\right) \le \exp\left(\frac{16K\|\boldsymbol{X}_{\mathcal{S}}\|_{\mathrm{F}}^2 \mathscr{R}_{\mathcal{F}}}{\varepsilon^2}\right),$$

where $\boldsymbol{X}_{\mathcal{S}} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_S] \in \mathbb{R}^{d \times S}$ and the parameter $\mathscr{R}_{\mathcal{F}}$ the spectral-complexity upper bound of $\mathcal{F}$.

Next, we proceed to get the cardinality of the $\frac{\varepsilon}{2\sqrt{S}}$-net covering for $\mathcal{E}_{\mathcal{NM}}$. Here, $\mathcal{E}_{\mathcal{NM}}$ is a set of sparse vectors with the number of non-zero elements upper-bounded by $CMK$, i.e., for any $\boldsymbol{e}_{\mathcal{NM}} \in \mathcal{E}_{\mathcal{NM}}$, $\|\boldsymbol{e}_{\mathcal{NM}}\|_0 \le CMK$. To get the covering number of a set of sparse vectors, we invoke the following result:

**Lemma C.5.** *[83] Let $\mathcal{X} = \{\boldsymbol{z} \in \mathbb{R}^N | \|\boldsymbol{z}\|_2 \le q, \|\boldsymbol{z}\|_0 \le C\}$. Suppose that $C \le \frac{N}{2}$. Then, there exists an $\epsilon$-net covering denoted as $\overline{\mathcal{X}}_\epsilon$ such that*

$$\overline{\mathsf{N}}(\epsilon, \mathcal{X}) \le \left(\frac{eN}{C}\frac{q}{\epsilon}\right)^C,$$

*where $e$ is the Euler's number.*

Applying Lemma C.5, we get the cardinality of $\frac{\varepsilon}{2\sqrt{S}}$-net covering of of $\mathcal{E}_{\mathcal{NM}}$ as follows:

$$\begin{aligned}\overline{\mathsf{N}}\left(\frac{\varepsilon}{2\sqrt{S}}, \mathcal{E}_{\mathcal{NM}}\right) &\le \left(\frac{eNMK}{CMK}\frac{4CM\sqrt{S}}{\varepsilon}\right)^{CMK} \\ &= \left(\frac{4eNM\sqrt{S}}{\varepsilon}\right)^{CMK}. \end{aligned} \tag{25}$$

Note that in (25), we used the result that $\|\boldsymbol{e}_{\mathcal{NM}}\|_2 \le \sqrt{2CM} \le 2CM, \forall \boldsymbol{e}_{\mathcal{NM}} \in \mathcal{E}_{\mathcal{NM}}$.

Using the covering number results, the cardinality of the $\varepsilon$-net covering of set $\mathcal{Z} \circ \mathcal{S}$ is bounded by the following:

$$\begin{aligned}\overline{\mathsf{N}}(\varepsilon, \mathcal{Z} \circ \mathcal{S}) &\le \exp\left(\frac{16K\|\boldsymbol{X}_{\mathcal{S}}\|_{\mathrm{F}}^2 \mathcal{R}_{\mathcal{F}}}{\varepsilon^2}\right) \times \left(\frac{4eNM\sqrt{S}}{\varepsilon}\right)^{CMK} \times \left(\frac{8K\sqrt{SK}}{\varepsilon}\right)^{K^2} \\ &= \exp\left(\frac{16K\|\boldsymbol{X}_{\mathcal{S}}\|_{\mathrm{F}}^2 \mathscr{R}_{\mathcal{F}}}{\varepsilon^2} + CMK \log\frac{4eNM\sqrt{S}}{\varepsilon} + K^2 \log\left(\frac{8K\sqrt{SK}}{\varepsilon}\right)\right). \end{aligned} \tag{26}$$

Now that we have characterized $\overline{\mathsf{N}}(\varepsilon, \mathcal{Z} \circ \mathcal{S})$, we invoke the Dudley entropy chaining technique lemma to obtain the Rademacher complexity $\mathfrak{R}(\mathcal{Z} \circ \mathcal{S})$ [80]:

**Lemma C.6.** *Consider set $A \subseteq \mathbb{R}^m$, let $c = \min_{\overline{\boldsymbol{a}}} \max_{\boldsymbol{a}\in A} \|\overline{\boldsymbol{a}} - \boldsymbol{a}\|$. Then for any integer $T > 0$,*

$$\mathfrak{R}(A) \le \frac{c2^{-T}}{\sqrt{m}} + \frac{6c}{m}\sum_{t=1}^{T} 2^{-t}\sqrt{\log \overline{\mathsf{N}}(c2^{-t}, A)}.$$

Since,

$$\|\boldsymbol{z}\| \le \sqrt{S}, \quad \forall \boldsymbol{z} \in \mathcal{Z} \circ \mathcal{S},$$

applying Lemma C.6 on set $\mathcal{Z} \circ \mathcal{S}$ gives

$$\begin{aligned}
&\mathfrak{R}(\mathcal{Z} \circ \mathcal{S}) \\
&\le \frac{\sqrt{S}2^{-T}}{\sqrt{SK}} + \frac{6\sqrt{S}}{SK}\sum_{t=1}^{T} 2^{-t}\sqrt{\log \overline{\mathsf{N}}(\sqrt{S}2^{-t}, \mathcal{Z} \circ \mathcal{S})} \\
&\le \frac{2^{-T}}{\sqrt{K}} + \frac{6}{K\sqrt{S}}\sum_{t=1}^{T} 2^{-t}\sqrt{\frac{16K\|\boldsymbol{X}_{\mathcal{S}}\|_{\mathrm{F}}^2\mathscr{R}_{\mathcal{F}}}{S4^{-t}} + CMK\log\frac{4eNM}{2^{-t}} + K^2\log\left(\frac{8K\sqrt{K}}{2^{-t}}\right)} \\
&\le \frac{2^{-T}}{\sqrt{K}} + \frac{6}{K\sqrt{S}}\sum_{t=1}^{T} \sqrt{\frac{16K\|\boldsymbol{X}_{\mathcal{S}}\|_{\mathrm{F}}^2\mathscr{R}_{\mathcal{F}}}{S} + 4^{-t}\left(CMK\log\frac{4eNM}{2^{-t}} + K^2\log\left(\frac{8K\sqrt{K}}{2^{-t}}\right)\right)} \\
&\le \frac{2^{-T}}{\sqrt{K}} + \frac{6}{K\sqrt{S}}\sum_{t=1}^{T} \sqrt{\frac{16K\|\boldsymbol{X}_{\mathcal{S}}\|_{\mathrm{F}}^2\mathscr{R}_{\mathcal{F}}}{S} + CMK\log(4eNM) + K^2\log\left(8K\sqrt{K}\right)} \\
&= \frac{2^{-T}}{\sqrt{K}} + \frac{6T}{K\sqrt{S}}\sqrt{\frac{16K\|\boldsymbol{X}_{\mathcal{S}}\|_{\mathrm{F}}^2\mathscr{R}_{\mathcal{F}}}{S} + CMK\log(4eNM) + K^2\log\left(8K\sqrt{K}\right)}.
\end{aligned}$$

Choosing $T = \log S/(2\log 2)$, we get the bound

$$\begin{aligned}
\mathfrak{R}(\mathcal{Z} \circ \mathcal{S}) \le &\frac{1}{\sqrt{SK}} \\
&+ \frac{6\log S}{2K\sqrt{S}\log 2}\sqrt{\frac{16K\|\boldsymbol{X}_{\mathcal{S}}\|_{\mathrm{F}}^2\mathscr{R}_{\mathcal{F}}}{S} + CMK\log(4eNM) + K^2\log\left(8K\sqrt{K}\right)}. \qquad (27)
\end{aligned}$$

Combining the upper bound of $\mathfrak{R}(\mathcal{Z} \circ \mathcal{S})$ given by (27) with the upper bound of $\mathfrak{R}(\ell \circ \mathcal{G} \circ \mathcal{S})$ as given by (21) and with the results in (18) and (19), we get that with probability greater than $1 - \delta$,

$$\left|\mathsf{L}_{\mathcal{D}}(\boldsymbol{G}, \widehat{\mathcal{Y}}) - \mathsf{L}_{\mathcal{S}}(\boldsymbol{G}; \widehat{\mathcal{Y}})\right| \le 2\sqrt{2}\beta K\mathfrak{R}(\mathcal{Z} \circ \mathcal{S}) + 4\log(\beta)\sqrt{\frac{2\log(4/\delta)}{S}}, \qquad (28)$$

where $\mathfrak{R}_{\mathcal{S}}(\mathcal{Z} \circ \mathcal{S})$ is upper bounded by (27).

The above relation, combined with (16), implies that with probability greater than $1 - 2\delta$:

$$\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M} \|\boldsymbol{g}_n^{\natural(m)} - \widehat{\boldsymbol{g}}_n^{(m)}\|_2^2 \le 4\sqrt{2}\beta K\mathfrak{R}(\mathcal{Z} \circ \mathcal{S}) + 8\log(\beta)\sqrt{\frac{2\log(4/\delta)}{S}}, \qquad (29)$$

where $\mathfrak{R}(\mathcal{Z} \circ \mathcal{S})$ is upper-bounded in (27).

## D Proof of Theorem 3.5

To obtain the results in Theorem 3.5, we first show that the criterion correctly distinguishes the instance-dependent and instance-independent samples, i.e., $\widehat{\mathcal{I}} = \mathcal{I}$ where we define $\widehat{\mathcal{I}} \triangleq \{i \in [N] \mid \widehat{\boldsymbol{e}}_i \neq \boldsymbol{0}\}$.

Recall $\widehat{\boldsymbol{A}}_m$'s, $\widehat{\boldsymbol{e}}_n^{(m)}$'s and $\widehat{\boldsymbol{f}}$ are defined as in Theorem 3.5, and let $\widehat{\boldsymbol{e}}_n = [(\widehat{\boldsymbol{e}}_n^{(1)})^\top, \ldots, (\widehat{\boldsymbol{e}}_n^{(M)})^\top]^\top$. We first show that $\widehat{\mathcal{I}}^c \subseteq \mathcal{I}^c$. From Lemma 3.1, the following holds:

$$\|\boldsymbol{W}^\natural\boldsymbol{F}^\natural + \boldsymbol{E}^\natural - \widehat{\boldsymbol{W}}\widehat{\boldsymbol{F}} - \widehat{\boldsymbol{E}}\|_{\mathrm{F}} \le \epsilon_{\boldsymbol{g}}(S) \qquad (30)$$

Then the inequality also holds for any submatrix of the left hand side matrix. In particular, for $i \in \mathcal{I}$ and $S > S_0$,

$$
\begin{aligned}
&\epsilon_{\boldsymbol{g}}(S_0) > \epsilon_{\boldsymbol{g}}(S)\\
&\quad \geq \|\boldsymbol{W}^{\natural}\boldsymbol{F}^{\natural}(:, \{i\} \cap \mathcal{I}^c \cap \widehat{\mathcal{I}}^c) + \boldsymbol{E}^{\natural}(:, \{i\} \cap \mathcal{I}^c \cap \widehat{\mathcal{I}}^c) - \widehat{\boldsymbol{W}}\widehat{\boldsymbol{F}}(:, \{i\} \cap \mathcal{I}^c \cap \widehat{\mathcal{I}}^c) - \widehat{\boldsymbol{E}}(:, \{i\} \cap \mathcal{I}^c \cap \widehat{\mathcal{I}}^c)\|_{\mathrm{F}}\\
&\quad = \|\boldsymbol{W}^{\natural}[\boldsymbol{f}_i^{\natural}, \boldsymbol{F}^{\natural}(:, \mathcal{I}^c \cap \widehat{\mathcal{I}}^c)] + [\boldsymbol{e}_i^{\natural}, \boldsymbol{0}, \dots, \boldsymbol{0}] - \widehat{\boldsymbol{W}}[\widehat{\boldsymbol{f}}_i, \widehat{\boldsymbol{F}}(:, \mathcal{I}^c \cap \widehat{\mathcal{I}}^c)] - [\widehat{\boldsymbol{e}}_i, \boldsymbol{0}, \dots, \boldsymbol{0}]\|_{\mathrm{F}}\\
&\quad \geq \|\boldsymbol{W}^{\natural}[\boldsymbol{f}_i^{\natural}, \boldsymbol{F}^{\natural}(:, \mathcal{I}^c \cap \widehat{\mathcal{I}}^c)] + [\boldsymbol{e}_i^{\natural}, \boldsymbol{0}, \dots, \boldsymbol{0}] - \widehat{\boldsymbol{W}}[\widehat{\boldsymbol{f}}_i, \widehat{\boldsymbol{F}}(:, \mathcal{I}^c \cap \widehat{\mathcal{I}}^c)]\|_{\mathrm{F}} - \|[\widehat{\boldsymbol{e}}_i, \boldsymbol{0}, \dots, \boldsymbol{0}]\|_{\mathrm{F}}\\
&\quad \geq \kappa(\boldsymbol{e}_i^{\natural}) - \|[\widehat{\boldsymbol{e}}_i, \boldsymbol{0}, ..., \boldsymbol{0}]\|_{\mathrm{F}} \quad \text{(by Definition 3.3)}\\
&\quad \geq \epsilon_{\boldsymbol{g}}(S_0) - \|[\widehat{\boldsymbol{e}}_i, \boldsymbol{0}, ..., \boldsymbol{0}]\|_{\mathrm{F}} \quad \text{(by the assumption in Theorem 3.5).}
\end{aligned}
$$

Therefore, $\widehat{\boldsymbol{e}}_i \neq \boldsymbol{0}$ and that concludes $\widehat{\mathcal{I}}^c \subseteq \mathcal{I}$. On the other hand, since $\widehat{\boldsymbol{e}}_n^{(m)}$'s minimizes $\sum_{n=1}^{N} \mathbb{1}\left\{\sum_{m=1}^{M} \|\widehat{\boldsymbol{e}}_n^{(m)}\|_2 > 0\right\}$, it holds that $\mathcal{I}^c \subseteq \widehat{\mathcal{I}}^c$. This implies that $\mathcal{I}^c = \widehat{\mathcal{I}}^c$ must hold.

With $\mathcal{I}^c = \widehat{\mathcal{I}}^c$, we have the following relation from (30):

$$
\frac{1}{MN}\|\boldsymbol{W}^{\natural}\boldsymbol{F}^{\natural}(:, \mathcal{I}^c) - \widehat{\boldsymbol{W}}\widehat{\boldsymbol{F}}(:, \mathcal{I}^c)\|_{\mathrm{F}}^2 \leq 4\sqrt{2}\beta K \mathfrak{R}(\mathcal{Z} \circ \mathcal{S}) + 8\log(\beta)\sqrt{\frac{2\log(4/\delta)}{S}}, \tag{31}
$$

To proceed, we invoke a result from the work [31] considering Assumption 3.4 satisfied, and the set $\mathcal{I}^c$ is uniformly chosen from $[N]$, with probability at least $1 - \dfrac{2}{S} - \dfrac{2K}{N^{\alpha}}$,

$$
\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left[\min_{\boldsymbol{\Pi}}\|\widehat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{\Pi}\boldsymbol{f}^{(\natural}(\boldsymbol{x})\|_2^2\right] \leq \frac{c_3(\zeta + \sqrt{TK\kappa})^2}{T(1-\sqrt{K}\kappa)^2} + 64T^{-5/8}(2\|\boldsymbol{X}\|_{\mathrm{F}}\mathcal{R}_{\mathcal{F}})^{0.25} + 16\sqrt{\frac{2\log(4S)}{N}}
$$

where

$$
\begin{aligned}
\zeta^2 &= 4\sqrt{2}\beta K\mathfrak{R}(\mathcal{Z} \circ \mathcal{S}) + 8\log(\beta)\sqrt{\frac{2\log(4S)}{S}} = \mathcal{O}\left(\frac{\beta\mathfrak{R}_{\mathcal{S}}\log S}{\sqrt{S}} + \log(\beta)\,\frac{\sqrt{\log(1/\delta)}}{\sqrt{S}}\right),\\
\varphi^2 &= MT^{\alpha}\left(4\sqrt{2}\beta K\mathfrak{R}(\mathcal{Z} \circ \mathcal{S}) + 8\log(\beta)\sqrt{\frac{2\log(4S)}{S}} + 4/S\right)\\
\kappa &= \varphi + \xi_1 + \xi_2 + \sqrt{K}\xi_1\xi_2,\\
\boldsymbol{X} &= [\boldsymbol{x}_{n_1}, \dots, \boldsymbol{x}_{n_S}], \quad \text{where } (m_s, n_s) \in \mathcal{S},\\
\mathfrak{R}_{\mathcal{S}}^2 &= K^2\log(K) + \frac{K\|\boldsymbol{X}\|_{\mathrm{F}}^2\mathscr{R}_{\mathcal{F}}}{S} + CMK\log\sqrt{NM},\\
T &= N - |\mathcal{I}|\,.
\end{aligned}
$$

# E Proof of Theorem 3.6

The proof of Theorem 3.6 follows similar steps in finite case presented in Theorem 3.5, except that we have the exact recovery of $\widehat{\boldsymbol{G}}$ when $S \to \infty$. We repeat the derivation here for completeness.

We first show that the criterion correctly distinguishes the instance-dependent and instance-independent labels.

Let ($\widehat{\boldsymbol{A}}_m$'s, $\widehat{\boldsymbol{e}}_n^{(m)}$'s and $\widehat{\boldsymbol{f}}$) be an optimal solutions as state in Theorem E. From Lemma (3.1), the following holds as the number of observations grows $S$ infinitely large:

$$
\widehat{\boldsymbol{g}}_n^{(m)} = \widehat{\boldsymbol{A}}_m\widehat{\boldsymbol{f}}(\boldsymbol{x}_n) + \widehat{\boldsymbol{e}}_n^{(m)} = \boldsymbol{g}_n{}^{\natural}(m) = \boldsymbol{A}_m^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_n) + \boldsymbol{e}_n^{\natural(m)}, \forall n, m. \tag{32}
$$

Consider the following relation:

$$
\boldsymbol{G}^{\natural} = \begin{bmatrix} \boldsymbol{g}_1^{\natural(1)} & \dots & \boldsymbol{g}_N^{\natural(1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{g}_1^{\natural(M)} & \dots & \boldsymbol{g}_N^{\natural(M)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_1^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_1) + \boldsymbol{e}_1^{\natural(1)} & \dots & \boldsymbol{A}_1^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_N) + \boldsymbol{e}_N^{\natural(1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{A}_M^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_1) + \boldsymbol{e}_1^{\natural(M)} & \dots & \boldsymbol{A}_M^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_N) + \boldsymbol{e}_N^{\natural(M)} \end{bmatrix}. \tag{33}
$$

Here, each column of the matrix $\boldsymbol{G}^{\natural}$ is given by

$$[\boldsymbol{G}^{\natural}]_n = \underbrace{\begin{bmatrix} \boldsymbol{A}_1^{\natural} \\ \vdots \\ \boldsymbol{A}_M^{\natural} \end{bmatrix}}_{\boldsymbol{W}^{\natural} \in \mathbb{R}^{MK \times K}} \boldsymbol{f}^{\natural}(\boldsymbol{x}_n) + \underbrace{\begin{bmatrix} \boldsymbol{e}_n^{\natural(1)} \\ \vdots \\ \boldsymbol{e}_n^{\natural(M)} \end{bmatrix}}_{\boldsymbol{e}_n^{\natural} \in \mathbb{R}^{MK}}$$

$$\implies [\boldsymbol{G}^{\natural}]_n = \boldsymbol{W}^{\natural} \boldsymbol{f}^{\natural}(\boldsymbol{x}_n) + \boldsymbol{e}_n^{\natural}, \forall n. \tag{34}$$

Similarly, we can also define

$$\widehat{\boldsymbol{W}} = \begin{bmatrix} \widehat{\boldsymbol{A}}_1 \\ \vdots \\ \widehat{\boldsymbol{A}}_M \end{bmatrix}, \ \widehat{\boldsymbol{e}}_n = \begin{bmatrix} \widehat{\boldsymbol{e}}_n^{(1)} \\ \vdots \\ \widehat{\boldsymbol{e}}_n^{(M)} \end{bmatrix}.$$

From (32), we get that

$$[\boldsymbol{G}^{\natural}]_n = \widehat{\boldsymbol{W}} \widehat{\boldsymbol{f}}(\boldsymbol{x}_n) + \widehat{\boldsymbol{e}}_n, \forall n. \tag{35}$$

We will use contradiction-based argument towards this. Let us start by considering the case where there exists $i^\star \in \mathcal{I}^c$ as well as $i^\star \in \widehat{\mathcal{I}}^c$ holds. Then, combining (34) and (35) we get

$$\widehat{\boldsymbol{W}} \widehat{\boldsymbol{f}}(\boldsymbol{x}_{i^\star}) = \boldsymbol{W}^{\natural} \boldsymbol{f}^{\natural}(\boldsymbol{x}_{i^\star}), \forall i^\star \in \mathcal{I}^c \cap \widehat{\mathcal{I}}^c. \tag{36}$$

We can rewrite the relation (36) as

$$\widehat{\boldsymbol{W}} \widehat{\boldsymbol{F}}(:, \mathcal{I}^c \cap \widehat{\mathcal{I}}^c) = \boldsymbol{W}^{\natural} \boldsymbol{F}^{\natural}(:, \mathcal{I}^c \cap \widehat{\mathcal{I}}^c).$$

Since any submatrix formed by $N - 2|\mathcal{I}^c|$ columns of $\boldsymbol{F}^{\natural} = [\boldsymbol{f}(\boldsymbol{x}_1), \dots, \boldsymbol{f}(\boldsymbol{x}_N)]$ has rank $K$ and that $\mathsf{rank}(\boldsymbol{W}^{\natural}) = K$, we get that there exists a nonsingular matrix $\boldsymbol{Q} \in \mathbb{R}^{K \times K}$ such that

$$\widehat{\boldsymbol{W}} = \boldsymbol{W}^{\natural} \boldsymbol{Q}. \tag{37}$$

Next, we consider the case where there exists $i^\star \in \mathcal{I}$ and $i^\star \in \widehat{\mathcal{I}}^c$. Then, again from (34), (35) we get that

$$\widehat{\boldsymbol{W}} \widehat{\boldsymbol{f}}(\boldsymbol{x}_{i^\star}) = \boldsymbol{W}^{\natural} \boldsymbol{f}^{\natural}(\boldsymbol{x}_{i^\star}) + \boldsymbol{e}_{i^\star}^{\natural} \tag{38}$$

However, by applying the relation (37) in (38), we get

$$\begin{aligned} \boldsymbol{W}^{\natural} \boldsymbol{Q} \widehat{\boldsymbol{f}}(\boldsymbol{x}_{i^\star}) &= \boldsymbol{W}^{\natural} \boldsymbol{f}^{\natural}(\boldsymbol{x}_{i^\star}) + \boldsymbol{e}_{i^\star}^{\natural} \\ \implies \boldsymbol{W}^{\natural} \boldsymbol{Q} \widehat{\boldsymbol{f}}(\boldsymbol{x}_{i^\star}) - \boldsymbol{W}^{\natural} \boldsymbol{f}^{\natural}(\boldsymbol{x}_{i^\star}) &= \boldsymbol{e}_{i^\star}^{\natural} \\ \implies \boldsymbol{W}^{\natural} \boldsymbol{q} &= \boldsymbol{e}_{i^\star}^{\natural}, \end{aligned}$$

where $\boldsymbol{q} = \boldsymbol{Q} \widehat{\boldsymbol{f}}(\boldsymbol{x}_{i^\star}) - \boldsymbol{f}^{\natural}(\boldsymbol{x}_{i^\star})$ is a nonzero vector since it is assumed that $\boldsymbol{e}_{i^\star}^{\natural} \neq \boldsymbol{0}$. The above relation implies that $\boldsymbol{e}_{i^\star}^{\natural} \in \mathrm{range}(\boldsymbol{W}^{\natural})$ must hold which is a contradiction to the assumption that $\boldsymbol{e}_{i^\star}^{\natural} \notin \mathrm{range}(\boldsymbol{W}^{\natural})$ when $i^\star \in \mathcal{I}$. Hence, the relation (38) is feasible only if $\boldsymbol{e}_{i^\star}^{\natural} = \boldsymbol{0}$. It means that if there exists $i^\star \in \widehat{\mathcal{I}}^c$, we should have $i^\star \in \mathcal{I}^c$, leading to the conclusion that $\widehat{\mathcal{I}}^c \subseteq \mathcal{I}^c$.

On the other hand, if $\widehat{\mathcal{I}}^c \subset \mathcal{I}^c$ holds, then

$$\sum_{n=1}^{N} \mathbb{1}\left\{ \sum_{m=1}^{M} \|\widehat{\boldsymbol{e}}_n^{(m)}\|_2 > 0 \right\} > \sum_{n=1}^{N} \mathbb{1}\left\{ \sum_{m=1}^{M} \|\boldsymbol{e}_n^{\natural(m)}\|_2 > 0 \right\}.$$

Since $\widehat{\boldsymbol{e}}_n^{(m)}$'s are the optimal solution of Problem (9) with minimal $\sum_{n=1}^{N} \mathbb{1}\left\{ \sum_{m=1}^{M} \|\widehat{\boldsymbol{e}}_n^{(m)}\|_2 > 0 \right\}$, the following should hold

$$\sum_{n=1}^{N} \mathbb{1}\left\{ \sum_{m=1}^{M} \|\widehat{\boldsymbol{e}}_n^{(m)}\|_2 > 0 \right\} \leq \sum_{n=1}^{N} \mathbb{1}\left\{ \sum_{m=1}^{M} \|\boldsymbol{e}_n^{\natural(m)}\|_2 > 0 \right\},$$

which is a contradiction. This implies that $\mathcal{I} = \widehat{\mathcal{I}}$ must hold.

Since $\mathcal{I} = \widehat{\mathcal{I}}$ holds, we have the following relation from (32):

$$\boldsymbol{G}^{\natural}(:, \mathcal{I}^c) = \widehat{\boldsymbol{G}}(:, \mathcal{I}^c), \tag{39}$$

where we utilized the following notations:

$$\begin{aligned}
\widehat{\boldsymbol{G}}(:, \mathcal{I}^c) &= \widehat{\boldsymbol{W}} \widehat{\boldsymbol{F}}(:, \mathcal{I}^c), \\
\boldsymbol{G}^{\natural}(:, \mathcal{I}^c) &= \boldsymbol{W}^{\natural} \boldsymbol{F}^{\natural}(:, \mathcal{I}^c), \\
\widehat{\boldsymbol{g}}_i &= [\boldsymbol{g}_i^{(1)\top}, \dots, \boldsymbol{g}_i^{(M)\top}]^\top, i \in \mathcal{I}^c, \\
\boldsymbol{W}^{\natural} &= [\boldsymbol{A}_1^{\top\natural}, \dots, \boldsymbol{A}_M^{\top\natural}]^\top \\
\widehat{\boldsymbol{W}} &= [\widehat{\boldsymbol{A}}_1^\top, \dots, \widehat{\boldsymbol{A}}_M^\top]^\top.
\end{aligned}$$

Next, we connect the $\det(\boldsymbol{W}^\top \boldsymbol{W})$ term to our optimal solutions. This part is similar to the classical proofs in [84, 85], with proper modifications. Consider the following result extracted from the proof of [84, Theorem 1]:

**Lemma E.1.** *Suppose a matrix $\boldsymbol{Y} \in \mathbb{R}^{K \times J}$ satisfies $\boldsymbol{Y} \geq \boldsymbol{0}$, $\boldsymbol{1}^\top \boldsymbol{Y} = \boldsymbol{1}^\top$ , $\mathsf{rank}(\boldsymbol{Y}) = K$, and SSC. Then, for any $\widehat{\boldsymbol{Y}} = \boldsymbol{Q}\boldsymbol{Y}$ satisfying $\widehat{\boldsymbol{Y}} \geq \boldsymbol{0}$, $\boldsymbol{1}^\top \widehat{\boldsymbol{Y}} = \boldsymbol{1}^\top$, the following holds:*

$$|\det(\boldsymbol{Q})| \leq 1,$$

*The equality holds only if $\boldsymbol{Q}$ is a permutation matrix.*

Since $\widehat{\boldsymbol{W}}$ is the optimal solution of the criterion given by Theorem 3.6, the following holds:

$$\det(\widehat{\boldsymbol{W}}^\top \widehat{\boldsymbol{W}}) \leq \det(\boldsymbol{W}^{\natural\top} \boldsymbol{W}^{\natural}). \tag{40}$$

We also have $\boldsymbol{W}^{\natural}$ and $\boldsymbol{F}^{\natural}(:, \mathcal{I}^c)$ are of rank $K$. Hence, from (39), we get that $\widehat{\boldsymbol{W}}$ and $\widehat{\boldsymbol{F}}(:, \mathcal{I}^c)$ satisfies $\widehat{\boldsymbol{W}} = \boldsymbol{W}^{\natural} \boldsymbol{Q}^{-1}$, $\widehat{\boldsymbol{F}}(:, \mathcal{I}^c) = \boldsymbol{Q} \boldsymbol{F}^{\natural}(:, \mathcal{I}^c)$, for a certain invertible matrix $\boldsymbol{Q}$. Let us assume that $\boldsymbol{Q}$ is not a permutation matrix, Then, we have

$$\begin{aligned}
\det(\boldsymbol{W}^{\natural\top} \boldsymbol{W}^{\natural}) &= \det(\boldsymbol{Q}^\top \widehat{\boldsymbol{W}}^\top \widehat{\boldsymbol{W}} \boldsymbol{Q}) \\
&= |\det(\boldsymbol{Q})|^2 \det(\widehat{\boldsymbol{W}}^\top \widehat{\boldsymbol{W}}) \\
&< \det(\widehat{\boldsymbol{W}}^\top \widehat{\boldsymbol{W}}),
\end{aligned}$$

where the last inequality is from Lemma E.1 using the SSC condition on $\boldsymbol{F}^{\natural}(:, \mathcal{I}^c)$. Note that the result is a contradiction from (40). Hence, $\boldsymbol{Q}$ must be a permutation matrix. This implies that the optimal solution $\widehat{\boldsymbol{W}}$ and $\widehat{\boldsymbol{F}}(:, \mathcal{I}^c)$ satisfies the following:

$$\widehat{\boldsymbol{W}} = \boldsymbol{W}^{\natural} \boldsymbol{\Pi}, \tag{41a}$$

$$\widehat{\boldsymbol{F}}(:, \mathcal{I}^c) = \boldsymbol{\Pi}^\top \boldsymbol{F}^{\natural}(:, \mathcal{I}^c). \tag{41b}$$

# F Proof of Fact 2.1

For $\mathsf{rank}(\boldsymbol{A}^{\natural}) = K$, when $N \to \infty$, the objective in (6) seeks the optimal solutions $\boldsymbol{A}^\star$, $\boldsymbol{f}^\star$ and $\boldsymbol{e}_n^\star$ such that

$$\boldsymbol{A}^{\natural} \boldsymbol{f}^{\natural}(\boldsymbol{x}) + \boldsymbol{e}_n^{\natural} = \boldsymbol{A}^\star \boldsymbol{f}^\star(\boldsymbol{x}_n) + \boldsymbol{e}_n^\star, \quad n \in [N]. \tag{42}$$

This can be seen from Lemma 3.1 with $M = 1$. One solution that satisfies the equality (42) and the sparsity constraint in (6) is $\boldsymbol{A}^\star = \boldsymbol{I}, \boldsymbol{f}^\star(\boldsymbol{x}) = \boldsymbol{A}^{\natural} \boldsymbol{f}^{\natural}(\boldsymbol{x}_n) + \boldsymbol{e}_n^{\natural}$, and $\boldsymbol{e}_n^\star = \boldsymbol{0}$, which can always hold when $\boldsymbol{f}$ is a universal function approximator.

Table 4: Noise levels and methods used to train machine annotators.

| **Case** | **Noise Level** | **Training Method/Architecture** |
|---|---|---|
| CIFAR-10 | | |
| High Noise | 0.26, 0.69, 0.57 | Logistic Regression, $k$-NN, ResNet34 |
| Medium Noise | 0.26, 0.46, 0.57 | $k$-NN, ResNet34, ResNet34 |
| Low Noise | 0.26, 0.29, 0.57 | $k$-NN, ResNet34, ResNet34 |
| STL-10 | | |
| High Noise | 0.15, 0.25, 0.69 | $k$-NN, ResNet9, ResNet9 |
| Medium Noise | 0.15, 0.36, 0.21 | ResNet9, ResNet9, ResNet9 |
| Low Noise | 0.15, 0.02, 0.21 | ResNet9, ResNet9, ResNet9 |

Table 5: Average classification accuracy on CIFAR-10 using synthetic annotators over 3 random trials.

| **Method** | **Noise rate** $\tau = 0.2$ | | | **Noise rate** $\tau = 0.4$ | | |
|---|---|---|---|---|---|---|
| | $\eta = 0.1$ | $\eta = 0.3$ | $\eta = 0.5$ | $\eta = 0.1$ | $\eta = 0.3$ | $\eta = 0.5$ |
| `CrowdLayer` | $91.01 \pm 0.28$ | $88.49 \pm 0.39$ | $84.59 \pm 0.47$ | $89.00 \pm 0.31$ | $86.47 \pm 0.33$ | $82.39 \pm 0.09$ |
| `TraceReg` | $91.76 \pm 0.38$ | $89.55 \pm 0.20$ | $83.60 \pm 3.14$ | $89.95 \pm 0.19$ | $87.28 \pm 0.08$ | $76.73 \pm 1.24$ |
| `MaxMIG` | $88.85 \pm 0.29$ | $84.99 \pm 0.41$ | $81.96 \pm 0.16$ | $86.27 \pm 0.59$ | $82.69 \pm 0.57$ | $78.60 \pm 0.33$ |
| `GeoCrowdNet(F)` | $91.36 \pm 0.26$ | $89.13 \pm 0.29$ | $85.28 \pm 0.34$ | $89.21 \pm 0.43$ | $86.73 \pm 0.36$ | $83.03 \pm 0.49$ |
| `GeoCrowdNet(W)` | $91.21 \pm 0.28$ | $88.69 \pm 0.29$ | $84.68 \pm 0.44$ | $89.16 \pm 0.14$ | $86.17 \pm 0.39$ | $81.49 \pm 0.31$ |
| `MEIDTM` | $91.27 \pm 0.09$ | $84.50 \pm 4.24$ | $52.77 \pm 4.05$ | $87.74 \pm 0.09$ | $61.33 \pm 3.92$ | $49.88 \pm 1.11$ |
| `PTD` | $79.59 \pm 0.46$ | $69.02 \pm 1.61$ | $52.56 \pm 1.05$ | $70.33 \pm 0.63$ | $50.94 \pm 5.48$ | $37.83 \pm 5.87$ |
| `BLTM` | $73.82 \pm 1.77$ | $66.27 \pm 1.60$ | $53.72 \pm 1.84$ | $66.34 \pm 1.84$ | $55.83 \pm 0.67$ | $46.20 \pm 0.97$ |
| `VolMinNet` | $89.95 \pm 0.23$ | $87.34 \pm 0.26$ | $82.60 \pm 0.50$ | $87.32 \pm 0.38$ | $83.62 \pm 0.14$ | $76.50 \pm 1.13$ |
| `Reweight` | $87.39 \pm 0.82$ | $82.35 \pm 0.79$ | $81.82 \pm 0.10$ | $88.34 \pm 0.44$ | $86.16 \pm 0.55$ | $81.67 \pm 0.61$ |
| `GCE` | $88.75 \pm 0.31$ | $86.98 \pm 0.21$ | 83.72 +/ 0.20 | $86.70 \pm 0.15$ | $84.58 \pm 0.26$ | $80.29 \pm 0.56$ |
| `COINNet (Ours)` | $\mathbf{92.23 \pm 0.36}$ | $\mathbf{90.06 \pm 0.06}$ | $\mathbf{86.47 \pm 0.04}$ | $\mathbf{91.97 \pm 0.02}$ | $\mathbf{87.90 \pm 0.03}$ | $\mathbf{83.45 \pm 0.04}$ |

# G More Details on Experiment Settings and Datasets

## G.1 Implementation

We train with batch size of 512, number of epochs 200, `Adam` optimizer with learning rate of $0.01$ and learning rate scheduler `OneCycleLR` [86]. We initialize $\boldsymbol{A}_m$ using an identity matrix going through softmax and $\boldsymbol{e}_m = \mathbf{0}$ for all $m \in [M]$. We also apply standard data augmentation including random cropping, random flipping, when training on CIFAR-10, CIFAR-10N, and STL-10. The experiment results are averaged over 3-5 random trials. All runs have been conducted using either Nvidia A40 or Nvidia DGX H100 GPU. Each run of `COINNet` on CIFAR-10 takes about 60 minutes using Nvidia A40, and 30 minutes using Nvidia DGX H100 GPU.

## G.2 Machine Annotations

For machine annotator experiments, we consider the CIFAR-10 [62] and the STL-10 datasets [63]. The CIFAR-10 dataset consists of $60,000$ labeled color images of animals, vehicles and so on, each having a size of $32 \times 32$ and belonging $K = 10$ different classes. The images are split into training and testing sets with size 50,000 and 10,000, respectively. The STL-10 dataset consists of 13,000 labeled images from 10 different classes, similar to CIFAR-10. Of these, 5,000 images are designated for training and 8,000 for testing.

Table 4 shows the details regarding the individual label noise rates and training methods in our experiments with machine annotators.

## G.3 Real Annotations

For real annotation experiments, we use 2 popular public datasets: CIFAR-10N [66] and LabelMe [67, 68], and construct the ImageNet-15N dataset with a higher number of annotators. All three datasets are labeled by anonymous annotators from the AMT platform. The CIFAR-10N is a “noisy" version of the popular CIFAR-10 dataset, containing 50,000 images for training, and a separate

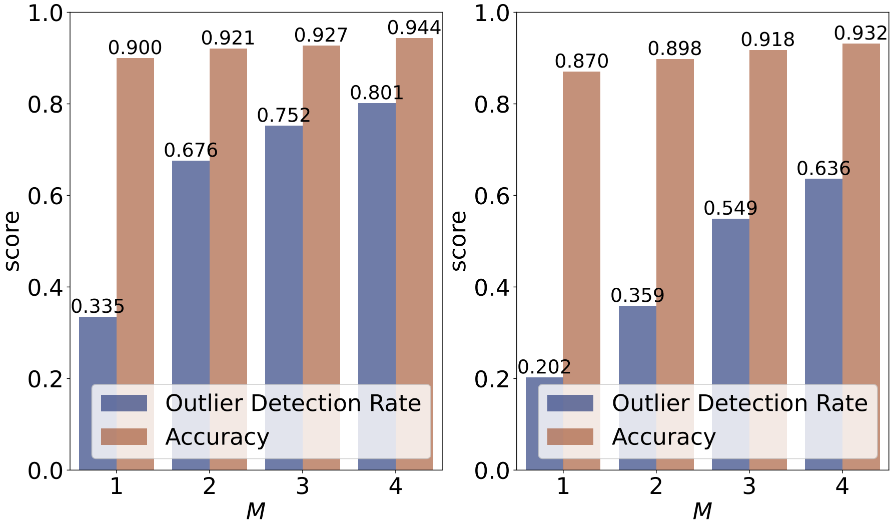

Figure 4: Performance of COINNet on CIFAR-10 with synthetic labels against different number of annotators; left: $\tau = 0.2, \eta = 0.1$, right: $\tau = 0.4, \eta = 0.1$.

10,000 images for testing. We randomly split the training set into 47,500 and 2,500 to use as train and validation set for all methods. Every image in CIFAR-10N has 3 labels.

The LabelMe dataset comprises 2,688 images from eight distinct classes: highway, inside city, tall building, street, forest, coast, mountain, and open country. Out of the available images, 1,000 have been annotated by 59 AMT workers. In total, approximately 2,547 image annotations were collected, with labeling accuracy ranging from 0% to 100% and an average accuracy of 69.20%. To enhance the training dataset, standard augmentation techniques such as rescaling, cropping, and horizontal flips were applied, resulting in a training set of 10,000 images annotated by the same 59 workers—see more details in [67]. The validation set comprises 500 images, while the remaining 1,188 images are reserved for testing.

We select 15 classes from the ImageNet dataset with the intention of including easy-to-confuse classes. These classes are: dog, leopard, sports_car, tiger_cat, airship, aircraft_carrier, trailer_truck, orange, penguin, lemon, soccer_ball, airliner, freight_car, container_ship, passenger_car. We collect annotations for $N = 2,514$ images from $M = 100$ anonymous real annotators. These serve as our training set. The average error rate of the annotators is 42.68%. The validation and testing sets have 1,462 and 13,157 images, respectively. We release the noisy annotations at https://github.com/ductri/COINNet.

# H More Experiments

## H.1 Experiments with Synthetic Annotators

**Noisy Label Generation.** Here, we consider labels corrupted by synthetically generated annotator confusions. To generate the confusion matrices, we utilize the following strategy. We control the instance-independent label noise rate using a parameter $\tau \in (0, 1)$. By employing this parameter, we construct the ground-truth confusion matrix $\boldsymbol{A}^{\natural}$, where the diagonal entries of $\boldsymbol{A}^{\natural}$ are set as $[\boldsymbol{A}^{\natural}]_{k,k} = 1 - \tau, \forall k$, and the off-diagonal entries are set as $[\boldsymbol{A}^{\natural}]_{k,j} = \frac{\tau}{K-1}, \forall k \neq j$. First, we randomly select $1 - \eta$ fraction of the data samples and generate their labels using the instance-independent confusion matrix $\boldsymbol{A}^{\natural}$. For the remaining $\eta$ percent of the samples, we adopt an instance-dependent noise generation process, by following the strategy outlined in [27]. Specifically, for such “outliers", the label confusions vary depending on the features of such instances.

Table 6: Average classification accuracy on CIFAR-10 and STL-10 datasets using machine annotators; results are averaged over 3 random trials.

| **Method** | **CIFAR-10** | | | **STL-10** | | |
|---|---|---|---|---|---|---|
| | High Noise | Medium Noise | Low Noise | High Noise | Medium Noise | Low Noise |
| COINNet, $\mu_1 = 0.01, \mu_2 = 0.01$ | **71.22 ± 0.72** | **73.31 ± 0.09** | **84.14 ± 0.38** | **70.12 ± 0.48** | **73.11 ± 0.37** | **76.39 ± 0.58** |
| COINNet, $\mu_1 = 0.01, \mu_2 = 0.00$ | 71.04 ± 0.33 | 73.11 ± 0.31 | 83.61 ± 0.68 | 58.90 ± 3.44 | 64.47 ± 3.75 | 72.79 ±2.88 |
| COINNet, $\mu_1 = 0.00, \mu_2 = 0.01$ | 71.17 ± 0.46 | 73.18 ± 0.91 | 83.51 ± 0.47 | 66.77 ± 4.81 | 72.40 ± 0.57 | 76.24 ± 0.52 |

Table 7: Average classification accuracy on CIFAR-10 using synthetic annotators; results are averaged over 3 random trials.

| **Method** | **Noise rate** $\tau = 0.2$ | | | **Noise rate** $\tau = 0.4$ | | |
|---|---|---|---|---|---|---|
| | $\eta = 0.1$ | $\eta = 0.3$ | $\eta = 0.5$ | $\eta = 0.1$ | $\eta = 0.3$ | $\eta = 0.5$ |
| COINNet, $\mu_1 = 0.01, \mu_2 = 0.01$ | **92.23 ± 0.36** | **90.06 ± 0.06** | **86.47 ± 0.04** | **91.97 ± 0.02** | 87.90 ± 0.03 | **83.45 ± 0.04** |
| COINNet, $\mu_1 = 0.01, \mu_2 = 0.00$ | **92.23 ± 0.93** | 89.64 ± 0.16 | 86.40 ± 0.18 | 91.16 ± 0.07 | **87.98 ± 0.15** | 83.15 ± 0.16 |
| COINNet, $\mu_1 = 0.00, \mu_2 = 0.01$ | 91.87 ± 0.04 | 89.65 ± 0.13 | 85.44 ± 0.31 | 90.21 ± 0.22 | 87.01 ± 0.48 | 66.50 ± 1.44 |

Table 8: Average classification accuracy over 3 random trials on real datasets.

| **Method** | CIFAR-10N |
|---|---|
| COINNet, $\mu_1 = 0.01, \mu_2 = 0.01$ | 92.09 ± 0.47 |
| COINNet, $\mu_1 = 0.01, \mu_2 = 0.00$ | **92.26 ± 0.24** |
| COINNet, $\mu_1 = 0.00, \mu_2 = 0.01$ | 91.51 ± 0.60 |
| **Method** | LabelMe |
| COINNet, $\mu_1 = 0.10, \mu_2 = 0.10$ | **87.60 ± 0.54** |
| COINNet, $\mu_1 = 0.10, \mu_2 = 0.00$ | 83.64 ± 1.29 |
| COINNet, $\mu_1 = 0.00, \mu_2 = 0.10$ | 84.85 ± 0.22 |

We generate synthetic noisy labels for CIFAR-10 dataset with $M = 3$. CIFAR-10 contains 50,000 images of size $32 \times 32$ as training set, and 10,000 images as test set. We keep the test set untouched while randomly splitting the train set into 2 parts (47,500; 2,500) to serve as training and validation set. We use the same neural network architecture and the optimization settings as used in machine annotator case. The results are presented in Table 5. One can observe that our approach consistently perform well in all scenarios.

To demonstrate the effectiveness of the advocated crowdsourcing approach in outlier detection, we vary the number of annotators $M$ and present the average outlier detection rate and the testing accuracy, where

$$\text{Outlier Detection Rate} = \frac{|\mathcal{I} \cap \{i \mid \|\widehat{\boldsymbol{e}}_i\|_2 \in \text{top } |\mathcal{I}| \text{ values among all } \|\widehat{\boldsymbol{e}}_n\|_2, n \in [N]\}|}{|\mathcal{I}|}.$$

We observe that increasing the number of annotators $M$ shows improvement over both outlier detection and the final accuracy score as shown in Fig. 4.

### H.2 Ablation Study

**Effect of regularization terms.** Tables 6, 7 and 8 show the effectiveness of having both the proposed regularization terms (i.e., sparsity and volume).

**Effect of missing annotations.** Table 9 and Table 10 show results under different levels of missing annotations under various parameters settings. Table 11 and Table 12 present results where only a single label is available for each image. In all four cases, the proposed method demonstrates superior robustness against the negative effect of missing annotations relative to `MaxMIG` and `GeoCrowdNet(F)`, two competitive baselines.

**Effect of initialization.** In Table 13, we test the performance of our approach using different initialization strategies. In particular, we conduct the machine annotator experiment on CIFAR-10 using the following initialization strategies: the confusion matrices are (i) initialized using identity matrices, and (ii) initialized by the `GeoCrowdNet(F)` method. We also include the performance of `GeoCrowdNet(F)` for reference. We observe a slight improvement (around 0.1-0.4%) in accuracy when using initialization from `GeoCrowdNet(F)` with the cost of training `GeoCrowdNet(F)` for 10 epochs.

Table 9: Average classification accuracy v.s. missing rate on CIFAR-10 using synthetic annotators; $M = 3, \tau = 0.2, \eta = 0.3$; results are averaged over 3 random trials.

| **missing rate** | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|---|---|---|---|---|---|
| `MaxMIG` | 85.01 ± 0.26 | 84.32 ± 0.20 | 83.90 ± 0.59 | 83.17 ± 0.70 | 81.80 ± 0.25 |
| `GeoCrowdNet(F)` | 83.19 ± 0.63 | 83.03 ± 0.41 | 81.83 ± 0.17 | 81.46 ± 0.47 | 81.28 ± 0.56 |
| `COINNet(Ours)` | **89.41** ± **0.13** | **89.22** ± **0.21** | **88.98** ± **0.55** | **88.53** ± **0.40** | **87.54** ± **0.16** |

Table 10: Average classification accuracy v.s. missing rate on CIFAR-10 using synthetic annotators; $M = 3, \tau = 0.2, \eta = 0.5$; results are averaged over 3 random trials.

| **missing rate** | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|---|---|---|---|---|---|
| `MaxMIG` | 80.97 ± 0.77 | 79.04 ± 0.90 | 79.46 ± 0.62 | 77.78 ± 0.72 | 76.15 ± 1.81 |
| `GeoCrowdNet(F)` | 80.32 ± 0.55 | 79.42 ± 0.27 | 78.81 ± 0.27 | 76.93 ± 0.39 | 76.01 ± 0.42 |
| `COINNet(Ours)` | **86.23** ± **0.14** | **84.98** ± **0.64** | **84.24** ± **0.71** | **83.79** ± **0.40** | **81.49** ± **0.59** |

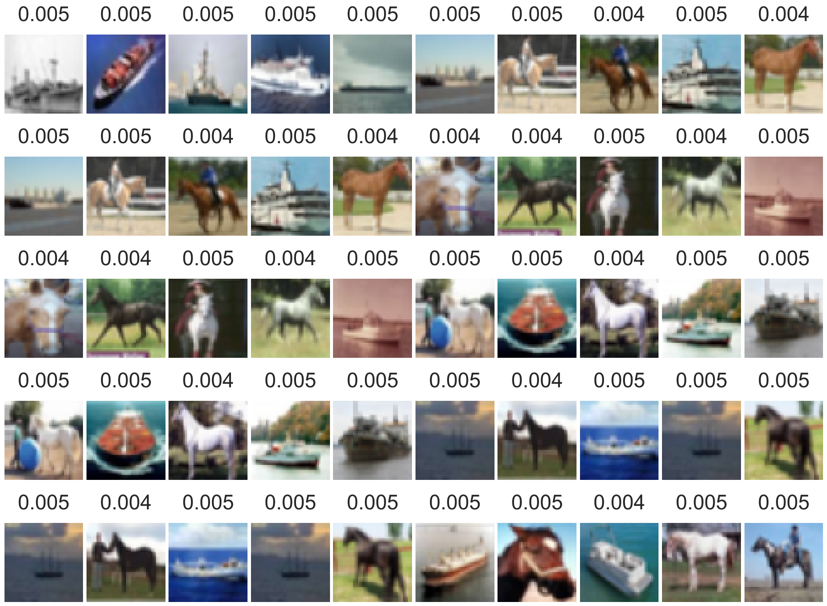

Figure 5: Some examples from the CIFAR-10N dataset learned with low outlier scores $s_n$'s.

### H.3 Outlier Detection in CIFAR-10N Dataset

Following the result in Fig. 2, we show more examples of nominal data items and outlier data items as separated by our approach in Fig. 5 and Fig. 6, respectively. Clearly, more instance-varying confusing characteristics are present in the latter case.

Table 11: Average classification accuracy in the scenario where each image of CIFAR-10 is labeled by only one randomly chosen annotator from $M = 3$ synthetic annotators; results are averaged over 3 random trials.

| | $\tau = 0.2, \eta = 0.3$ | $\tau = 0.2, \eta = 0.5$ |
|---|---|---|
| `MaxMIG` | $81.36 \pm 0.81$ | $70.78 \pm 4.66$ |
| `GeoCrowdNet(F)` | $80.44 \pm 0.18$ | $75.31 \pm 0.33$ |
| `COINNet(Ours)` | $\mathbf{87.18 \pm 0.29}$ | $\mathbf{80.96 \pm 0.38}$ |

Table 12: Average classification accuracy in the scenario where each image of CIFAR-10 is labeled by only one randomly chosen annotator from $M = 3$ machine annotators; results are averaged over 3 random trials—see the annotation generation settings as in Sec. 5.1.

| | **High Noise** | **Medium Noise** | **Low Noise** |
|---|---|---|---|
| `MaxMIG` | $58.58 \pm 0.64$ | $67.64 \pm 0.25$ | $77.01 \pm 0.25$ |
| `GeoCrowdNet(F)` | $55.25 \pm 1.09$ | $66.91 \pm 0.53$ | $75.48 \pm 0.28$ |
| `COINNet(Ours)` | $\mathbf{59.76 \pm 0.62}$ | $\mathbf{70.94 \pm 0.32}$ | $\mathbf{80.12 \pm 0.44}$ |

Table 13: Average classification accuracy with different initializations for the confusion matrices. We use machine annotations with the same setting as described in Sec. 5.1 in the manuscript. Results are averaged over 3 random trials.

| **Initialization Strategy for $A_m$'s** | **High Noise** | **Medium Noise** | **Low Noise** |
|---|---|---|---|
| Identity matrix | $70.37 \pm 0.61$ | $73.33 \pm 0.46$ | $83.25 \pm 0.39$ |
| `GeoCrowdNet(F)` after training 10 epochs | $\mathbf{71.75 \pm 0.73}$ | $\mathbf{73.48 \pm 0.27}$ | $\mathbf{84.26 \pm 0.13}$ |
| Current setting (close to an identity matrix) | $71.22 \pm 0.72$ | $73.31 \pm 0.09$ | $84.14 \pm 0.38$ |

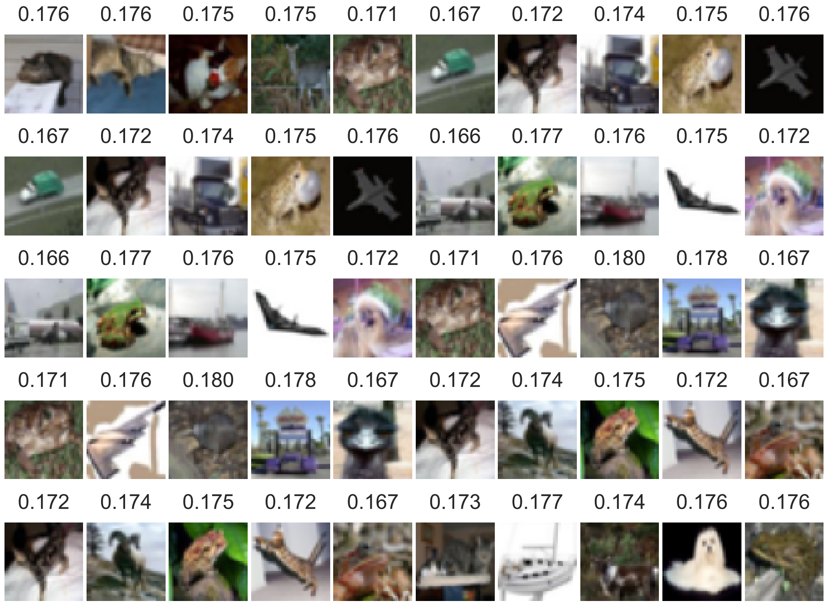

Figure 6: Some examples from the CIFAR-10N dataset learned with high outlier scores $s_n$'s

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims have been discussed and supported by our analysis in Section 2 and 3, particularly by Fact 2.1, Theorem 3.5 and Theorem 3.6 with empirical evidence in Section 5

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations at the end of Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have presented all assumptions we have used in Section 3. All proof are presented in the Appendix, particularly in Appendix D, Appendix E, Appendix F.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have described in details experiment settings as well as the implementation techniques in Section 3, Section 5 and Appendix G. We also include our code in this submission as supplementary material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have included code, including instruction how to run in this submission. It is as almost the same code that we use to to produce all experiment results, except removing all identity related information. All data we used are publicly accessible.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe these in Appendix G

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experiments are repeated at least 3 times. All numbers are accompanied with standard deviation, as described in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have included these information in Appendix G.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have made to our best effort to remove all identity-related information in both the main paper supplementary material.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work tackles the noisy labels problem which is ubiquitous, especially in the big data era, and hence bring positive impact to society.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not see any potential harm in publishing data and trained models of the proposed method.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited all datasets used in Section 5.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There is no new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: There is no human subjects or crowdsourcing experiments conducted in the work. All datasets are from published sources.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: As mentioned above, our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.