Self-cognitive Denoising in the Presence of Multiple Noisy Label Sources

Yi-Xuan Sun [1] Ya-Lin Zhang [1] Bin Han [1] Longfei Li [1] Jun Zhou [1]

Abstract

The strong performance of neural networks typically hinges on the availability of extensive labeled data, yet acquiring ground-truth labels is often challenging. Instead, noisy supervisions from multiple sources, e.g., by multiple well-designed rules, are more convenient to collect. In this paper, we focus on the realistic problem of learning from multiple noisy label sources, and argue that prior studies have overlooked the crucial *self-cognition* ability of neural networks, i.e., the inherent capability of autonomously distinguishing noise during training. We theoretically analyze this ability of neural networks when meeting multiple noisy label sources, which reveals that neural networks possess the capability to recognize both instance-wise noise within each single noisy label source and annotator-wise quality among multiple noisy label sources. Inspired by the theoretical analyses, we introduce an approach named Self-cognitive Denoising for Multiple noisy label sources (SDM), which exploits the self-cognition ability of neural networks to denoise during training. Furthermore, we build a selective distillation module following the theoretical insights to optimize computational efficiency. The experiments on various datasets demonstrate the superiority of our method.

1. Introduction

Neural networks have made remarkable strides across various applications (Goodfellow et al., 2016; He et al., 2016; Tarvainen & Valpola, 2017), and the success hinges significantly on the availability of extensive labeled datasets. However, acquiring accurate ground-truth labels is often costly and time-consuming in realistic scenarios, thus these labels may not always be obtainable. In practice, an alternative is to utilize noisy supervisions from multiple sources, which is more economical and easier to collect (Gao et al., 2022; Zhao et al., 2023). For example, domain experts can create labeling rules that swiftly generate noisy labels for a vast number of samples. Moreover, when these rules are crafted from different perspectives, they can yield diverse sets of labels. This inherent diversity can prove advantageous for training in the absence of ground-truth labels.

In this paper, we address a prevalent and practical challenge: binary classification with multiple noisy label sources, a scenario that falls within the realm of weakly supervised learning (Zhou, 2018). This area of study is characterized by the reliance on several sets of *inaccurate* labels. To illustrate, consider the task of constructing user profiles to predict a user's likelihood to purchase a car. Obtaining ground-truth labels for this task would entail costly and time-consuming investigations. Alternatively, one could apply thoughtfully crafted heuristic rules—for instance, determining whether users have recently engaged in car-related transactions or visited automotive dealerships—to generate multiple potential groups of target users with relative ease. Given this context, it becomes desirable to devise methodologies for harnessing these various noisy label sets effectively.

Existing related methods, called *learning from multiple noisy label sources*, can be generally categorized into two groups, i.e., two-stage approaches (Dawid & Skene, 1979; Whitehill et al., 2009; Welinder et al., 2010; Ibrahim et al., 2019) and end-to-end approaches (Rodrigues & Pereira, 2018; Tanno et al., 2019; Khetan et al., 2017; Guan et al., 2018; Cao et al., 2019; Li et al., 2020c; Gao et al., 2022; Zhao et al., 2023; Ibrahim et al., 2023). The first group of methods often separate the steps of label aggregation and model training. The observed multiple sets of noisy labels are first aggregated into a single set via an estimated probabilistic model. Then, the downstream model is trained with input samples and aggregated labels in a traditional supervised manner. However, these works often neglect the information of input samples when aggregating multiple sets of noisy labels, causing low-quality aggregated labels in practice. The second group of methods often simultaneously learn the annotators' confusions and the classifier in an end-to-end manner. For example, Rodrigues & Pereira (2018) employed a *crowdlayer* to estimate the confusion matrices of multiple annotators and simultaneously update

[1] Ant Group, Hangzhou, China. Correspondence to: Jun Zhou <jun.zhoujun@antgroup.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

the classifier. Tanno et al. (2019) designed a similar criterion to estimate multiple confusion matrices with a trace regularization. Zhao et al. (2023) implicitly estimated the annotator's confusions during training with a Mixture-of-Experts (MoE) model. However, it is often unclear if the estimated confusion matrices can correctly identify the annotators' confusion characteristics.

While prior methods have generally enhanced performance, the *self-cognition* ability of neural networks, i.e., the inherent capability of autonomously distinguishing noise during training, which is crucial for learning from single noisy label source, has been neglected. Specifically, neural networks have been both theoretically and empirically validated to possess the capability of autonomously distinguishing noisy samples during training (Arpit et al., 2017; Li et al., 2020b; Liu et al., 2020; Gui et al., 2021), which has inspired a number of studies (Han et al., 2018; Yu et al., 2019; Wei et al., 2020; Li et al., 2020a) to explore strategies that enable the model to detect and handle noisy samples throughout the training process. Despite its importance, the further exploration of neural networks' *self-cognition* ability in scenarios involving multiple noisy label sources remains limited.

In this paper, we investigate the *self-cognition* capability of neural networks in the presence of multiple noisy label sources. The main contributions of this work are as follows:

- We theoretically reveal that neural networks possess the capability to recognize both the instance-wise noise within each single noisy label source and the annotator-wise quality among multiple noisy label sources.

- Inspired by the theoretical analyses, we propose an approach named Self-cognitive Denoising for Multiple noisy label sources (SDM), which employs a *self-cognition* module to identify both instance-wise noise and annotator-wise quality and adopts a *mutual-denoising* module to aggregate these identifications and accordingly refine the model. Additionally, we design a *selective distillation* module to adaptively distill valuable knowledge from the original model to a more deployment-friendly version.

- We empirically validate the proposed method across various datasets and demonstrate that our method surpasses other competing approaches in performance.

2. Theoretical Insight

2.1. Preliminaries

We focus on the binary classification task with multiple noisy label sources in this paper. Let $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y} = \{0, 1\}$ denote the sample from sample space $\mathcal{X}$ and the associated true label, which is drawn from the true data distribution $p(\boldsymbol{x}, y)$. Suppose that the true label y is determined by the target concept f^*, i.e., $y = f^*(\boldsymbol{x})$. Let $\tilde{y}^1 \in \mathcal{Y}, \ldots, \tilde{y}^s \in \mathcal{Y}$ denote the observed noisy labels from s different sources. We write the training dataset with multiple noisy label sources as $\tilde{D} = \{(\boldsymbol{x}_i, \tilde{y}_i^1, \ldots, \tilde{y}_i^s)\}_{i=1}^n$. Consider the neural network $h_\Theta(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$ with output probability $h_\Theta(\boldsymbol{x}_i) = \hat{p}(\boldsymbol{x}_i)$, where Θ is the parameters of the neural network and $\hat{p}(\boldsymbol{x}_i)$ denotes the predicted probability after the activation function, e.g., sigmoid. The binary cross-entropy loss function ℓ of the pair $(\boldsymbol{x}, \tilde{y})$ when training the neural network h_Θ can be formulated as:

$$\ell(h_\Theta(\boldsymbol{x}), \tilde{y}) = -\big[\tilde{y} \log h_\Theta(\boldsymbol{x}) + (1 - \tilde{y}) \log(1 - h_\Theta(\boldsymbol{x}))\big]$$

Let the classifier induced by h_Θ be $f_\Theta : \mathcal{X} \to \mathcal{Y}$ with predictions $f_\Theta(\boldsymbol{x}_i) = \mathbb{I}\left(h_\Theta\left(\boldsymbol{x}_i\right) > 0.5\right)$, where $\mathbb{I}(\cdot)$ denotes the indicator function. Our goal is to obtain the optimal classifier f_{Θ^*}, which satisfies $f_{\Theta^*}(\boldsymbol{x}) = f^*(\boldsymbol{x})$ for any $\boldsymbol{x}$.

2.2. Theoretical Analyses

In this section, we will analyze the self-cognition ability of neural networks with multiple noisy label sources. Within each source, we assume that the instance-wise noise satisfies the well-used class-dependent assumption (Ghosh et al., 2017; Patrini et al., 2017; Wang et al., 2019; Gui et al., 2021), i.e., $p(\tilde{y}^k|y, \boldsymbol{x}) = p(\tilde{y}^k|y), \forall k \in \{1, \ldots, s\}$. Under this assumption, The noise transition matrix in the k-th source can be formulated as $T^k \in \mathbb{R}^{2\times 2}$, where $T_{ij}^k = p(\tilde{y}^k = j|y = i)$ denotes the probability of an i-th class sample flipped into the j-th class in the k-th source.

Derived from Gui et al. (2021), we give the following theorem to analyze the instance-wise self-cognition ability of neural networks within each single noisy label source:

Theorem 2.1. *(Single noisy label source) Let $h_{\Theta_k^*}$ denote the neural network minimizing the expected loss in the k-th noisy label source, i.e., $\mathbb{E}_{(\boldsymbol{x},\tilde{y}^k)}[\ell(h_\Theta(\boldsymbol{x}), \tilde{y}^k)]$. $(\boldsymbol{x}_1, \tilde{y}^k)$ and $(\boldsymbol{x}_2, \tilde{y}^k)$ are any two samples with the same observed label $\tilde{y}^k$ satisfying that $f^*(\boldsymbol{x}_1) = \tilde{y}^k$ and $f^*(\boldsymbol{x}_2) \neq \tilde{y}^k$. If T^k satisfies that $T_{ii}^k > 0.5, \forall i \in \{0, 1\}$, then we have*

$$\ell(h_{\Theta_k^*}(\boldsymbol{x}_1), \tilde{y}^k) < \ell(h_{\Theta_k^*}(\boldsymbol{x}_2), \tilde{y}^k).$$

Remark. Theorem 2.1 indicates that if diagonal elements of the noise transition matrix T^k are bigger than 0.5, clean samples will have smaller losses than noisy ones for neural network $h_{\Theta_k^*}$. The revelation here is that, when meeting multiple noisy label sources, neural networks still enjoy the instance-wise self-cognition ability within each single source by the small-loss criterion.

Then, we expand the above theorem from the scenario of single source to two sources:

Theorem 2.2. *(Two noisy label sources) Let $h_{\Theta_1^*}$ and $h_{\Theta_2^*}$ denote the neural networks minimizing the expected loss* *in the* 1*-th and* 2*-nd noisy label source respectively, i.e.,* $\mathbb{E}_{(\boldsymbol{x},\tilde{y}^1)}[\ell(h_\Theta(\boldsymbol{x}), \tilde{y}^1)]$ *and* $\mathbb{E}_{(\boldsymbol{x},\tilde{y}^2)}[\ell(h_\Theta(\boldsymbol{x}), \tilde{y}^2)]$*. If* $T^1_{ii} > T^2_{ii} > 0.5$*,* $\forall i \in \{0, 1\}$*, then we have*

$$\mathbb{E}_{(\boldsymbol{x},\tilde{y}^1)}\ell(h_{\Theta^*_1}(\boldsymbol{x}), \tilde{y}^1) < \mathbb{E}_{(\boldsymbol{x},\tilde{y}^2)}\ell(h_{\Theta^*_2}(\boldsymbol{x}), \tilde{y}^2), \quad (1)$$

and

$$\mathbb{E}_{(\boldsymbol{x},y)}\ell(h_{\Theta^*_1}(\boldsymbol{x}), y) < \mathbb{E}_{(\boldsymbol{x},y)}\ell(h_{\Theta^*_2}(\boldsymbol{x}), y). \quad (2)$$

Remark. Compared to Theorem 2.1, Theorem 2.2 offers a novel viewpoint on expected losses, which yields insights at the level of individual annotators. Specifically, when $T^1_{ii} > T^2_{ii} > 0.5$, $\forall i \in \{0, 1\}$, i.e., the 1-th noisy label source is more accurate than the 2-nd, Theorem 2.2 inspires us that: Firstly, Eq (1) indicates that the induced neural networks by the 1-th noisy label source $h_{\Theta^*_1}$ will have smaller expected loss than $h_{\Theta^*_2}$ on the according noisy distribution. It reveals that neural networks tend to fit closer to a higher-quality noisy label source when meeting two different sources, which opens doors for us to discern the annotator-wise quality over two noisy label sources. Secondly, Eq (2) indicates that $h_{\Theta^*_1}$ will also have smaller expected loss than $h_{\Theta^*_2}$ on the true data distribution. It inspires us that the discerned higher-quality noisy label source will induce more precise predictions, thereby serving as helpful guidance for denoising in another source.

Furthermore, we expand Theorem 2.2 from the scenario of two sources to multiple sources:

Corollary 2.3. *(Multiple noisy label sources) Let* $h_{\Theta^*_k}$ *denote the neural network minimizing the expected loss in the* k*-th noisy label source among* s *sources, i.e.,* $\mathbb{E}_{(\boldsymbol{x},\tilde{y}^k)}[\ell(h_\Theta(\boldsymbol{x}), \tilde{y}^k)]$*. If* $T^1_{ii} > \cdots > T^s_{ii} > 0.5$*,* $\forall i \in \{0, 1\}$*, then we have*

$$\mathbb{E}_{(\boldsymbol{x},\tilde{y}^1)}\ell(h_{\Theta^*_1}(\boldsymbol{x}), \tilde{y}^1) < \cdots < \mathbb{E}_{(\boldsymbol{x},\tilde{y}^s)}\ell(h_{\Theta^*_s}(\boldsymbol{x}), \tilde{y}^s), \quad (3)$$

and

$$\mathbb{E}_{(\boldsymbol{x},y)}\ell(h_{\Theta^*_1}(\boldsymbol{x}), y) < \cdots < \mathbb{E}_{(\boldsymbol{x},y)}\ell(h_{\Theta^*_s}(\boldsymbol{x}), y). \quad (4)$$

Remark. Corollary 2.3 is a natural extension of Theorem 2.2, which reveals that the inspirations from two noisy label sources still hold when facing the more realistic scenario of multiple noisy label sources. That is, Eq (3) suggests that neural networks have the potential to distinguish the annotator-wise quality among multiple noisy label sources, and Eq (4) implies that the higher-quality noisy label sources can offer constructive guidance for the denoising process in other, lower-quality sources.

2.3. Empirical Verifications

Many previous works (Han et al., 2018; Yu et al., 2019; Li et al., 2020a; Gui et al., 2021) have empirically verified similar results of Theorem 2.1. Furthermore, we will empirically demonstrate that inspirations of Theorem 2.2 and Corollary 2.3 still hold in practice without some basic assumptions.

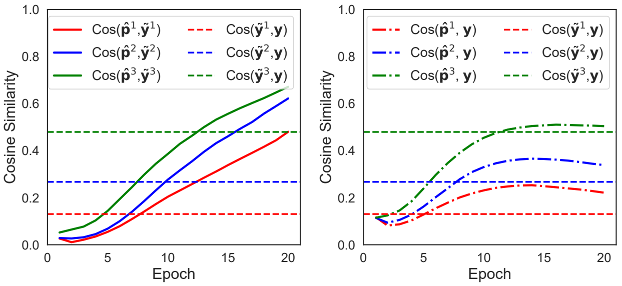


Figure 1. Preliminary studies on the Yelp dataset. $\hat{\boldsymbol{p}}^j$ and $\tilde{\boldsymbol{y}}^j$ denote the predicted probabilities and the noisy labels on the j-th source over training samples respectively. $\boldsymbol{y}$ denotes the ground-truth labels over training samples. **Left**: Verifications of Eq (1) and Eq (3). **Right**: Verifications of Eq (2) and Eq (4).

Theoretically, we analyze the self-cognition ability of individual networks for simplicity. In practice, we unite these individual networks using a multi-tower MLP, which borrows the insight from multi-task learning (Zhang & Yang, 2018) to extract public information from different sources. Meanwhile, we use varying percentages (10%, 30%, 50%) of ground-truth labels to train three LightGBM classifiers (Ke et al., 2017) on the Yelp dataset and consider their predictions as noisy labels from three sources, which breaks the class-dependent assumption. Obviously, the classifier's predictions will be more accurate with more ground-truth labels for training. Under such conditions, we conduct experiments to empirically verify Theorem 2.2 and Corollary 2.3, and the results are illustrated in Figure 1. Instead of loss function in theories, we employ cosine similarity in practice because its normalization term makes it more comparable between different sources. The left part of Figure 1 indicates that neural networks are easier to fit noisy supervisions in a higher-quality source, which empirically verifies the insights from Eq (1) and Eq (3). Meanwhile, the right part of Figure 1 indicates that neural networks' predictions are closer to true label distributions in a higher-quality source, which empirically verifies the insights from Eq (2) and Eq (4). More empirical verifications on different datasets with different types of noise can be found in Appendix C.6.

3. Method

Based on the theoretical insights, we design a *self-cognition* module to identify both instance-wise noise and annotator-wise quality. Aggregating these identifications to assess the reliability of each noisy label, a subsequent *mutual-denoising* module is crafted to refine the training paradigm accordingly. Additionally, a *selective distillation* module is designed to distill valuable knowledge from the original model to a more deployment-friendly and lightweight one.

3.1. Self-cognition

When meeting multiple noisy label sources, it is desirable to distinguish noise from two perspectives: the instance-wise noise within each source, and the annotator-wise quality among different sources. Thanks to the theoretical analyses, i.e., Theorem 2.1 for the former and Eq (3) in Corollary 2.3 for the latter, we can exploit the self-cognition power of neural networks to identify noise from both perspectives.

Specifically, we can begin with a multi-tower neural network that simultaneously learns from various noisy label sources. Let $G_\Theta(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}^s$ denote a multi-tower neural network with parameters Θ and outputs $G_\Theta(\boldsymbol{x}) = [G^1_\Theta(\boldsymbol{x}), \dots, G^s_\Theta(\boldsymbol{x})]^\top = [\hat{p}^1(\boldsymbol{x}), \dots, \hat{p}^s(\boldsymbol{x})]^\top$, where $\hat{p}^j(\boldsymbol{x}) = \sigma(z_j(\boldsymbol{x}))$ is the predicted probability of $\boldsymbol{x}$ for the j-th source, $z_j(\boldsymbol{x})$ is the corresponding output logit and $\sigma(\cdot)$ denotes the sigmoid function. For the noisy dataset $\tilde{D}$, the binary cross entropy loss among different noisy label sources is

$$\mathcal{L}_B = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{s} l(G^j_\Theta(\boldsymbol{x}_i), \tilde{y}^j_i). \tag{5}$$

To distinguish the instance-wise noise within each single noisy label source, we follow the inspiration of Theorem 2.1 to record the instance-wise loss in each source as $l(G^j_{\Theta^t}(\boldsymbol{x}_i), \tilde{y}^j_i)$, which means the loss of i-th sample of j-th source in training epoch t. In this way, the recorded loss information can be gathered into a matrix

$$L^t \in \mathbb{R}^{n\times s} : L^t_{i,j} = l(G^j_{\Theta^t}(\boldsymbol{x}_i), \tilde{y}^j_i). \tag{6}$$

According to Theorem 2.1, the samples with true labels yield smaller loss than those with wrong labels in each noisy label source, which motivates us to select small-loss samples on a column-by-column basis within the loss matrix L^t. Therefore, we calculate a matrix for filtering samples as

$$M^t \in \mathbb{R}^{n\times s} : M^t_{i,j} = \frac{1}{t}\sum_{e=1}^{t}\mathbb{I}(L^e_{i,j} < \tau^e_j), \tag{7}$$

where τ^e_j is an adhoc loss threshold for filtering samples of j-th source in epoch e. We record the averaging matrix during t epochs to obtain more stable results. The element $M^t_{i,j}$ is closer to zero if sample $\boldsymbol{x}_i$ has a larger loss in the j-th source, making it more likely to be filtered out. In practice, we set τ^e_j as the P-th percentile of the whole recorded losses in j-th source, i.e., $\boldsymbol{L}^e_{\cdot,j} = [L^e_{1,j}, \dots, L^e_{n,j}]^\top$. Due to the common class-imbalance problem, we also calculate the thresholds for positive and negative classes respectively.

Building upon the insights provided by Eq (3) in Corollary 2.3, we can capture the relationships between predictions and labels to discern the annotator-wise quality among different noisy label sources. Specifically, in epoch e and j-th source, we calculate the cosine similarity between model's predictions $\boldsymbol{G}^j_{\Theta^e} = [G^j_{\Theta^e}(\boldsymbol{x}_1), \dots, G^j_{\Theta^e}(\boldsymbol{x}_n)]^\top$ and noisy labels $\tilde{\boldsymbol{y}}^j = [\tilde{y}^j_1, \dots, \tilde{y}^j_n]^\top$ as

$$\mathrm{Cos}(\boldsymbol{G}^j_{\Theta^e}, \tilde{\boldsymbol{y}}^j) = \frac{\sum_{i=1}^n G^j_{\Theta^e}(\boldsymbol{x}_i)\cdot \tilde{y}^j_i}{\|\boldsymbol{G}^j_{\Theta^e}\| \cdot \|\tilde{\boldsymbol{y}}^j\|}. \tag{8}$$

The findings derived from Eq (3) in Corollary 2.3 suggest that neural networks tend to fit closer to higher-quality noisy label sources. Consequently, the observed similarity between the model's predictions and noisy labels correlates with the annotator-wise quality, which motivates the following estimation of the annotator-wise quality as

$$\boldsymbol{q}^t \in \mathbb{R}^s : q^t_i = \frac{1}{t}\sum_{e=1}^{t}\frac{\exp\Big(\mathrm{Cos}(\boldsymbol{G}^i_{\Theta^e}, \tilde{\boldsymbol{y}}^i)/T\Big)}{\sum_{j=1}^s \exp\Big(\mathrm{Cos}(\boldsymbol{G}^j_{\Theta^e}, \tilde{\boldsymbol{y}}^j)/T\Big)}, \tag{9}$$

where t denotes the training epoch and T is a hyperparameter of temperature. We also average estimations during t epochs to obtain more stable results. In practice, the self-cognition module is only employed during the initial t_0 epochs to determine M^t and $\boldsymbol{q}^t$. Subsequently, these estimates are held constant to mitigate the over-fitting issue.

3.2. Mutual-denoising

Based on Eq (4) in Corollary 2.3, we know that higher-quality noisy label sources can be beneficial for denoising in other sources. The inherent self-cognition of neural networks—namely, the instance-wise filter matrix M^t that reveals 'which sample is better' and annotator-wise quality vector $\boldsymbol{q}^t$ that reveals 'which source is better'—enables us to identify more reliable noisy labels for the exchange and enhancement of knowledge across different sources.

Specifically, the self-cognition from two perspectives, i.e., the instance-wise filter matrix M^t and annotator-wise quality vector $\boldsymbol{q}^t$, can be aggregated into one weighting matrix to assess the reliability of each noisy label:

$$W^t \in \mathbb{R}^{n\times s} : W^t_{i,j} = \frac{M^t_{i,j}\cdot q^t_j + \epsilon}{\sum_{k=1}^s \Big(M^t_{i,k}\cdot q^t_k + \epsilon\Big)}, \tag{10}$$

where ϵ is a tiny constant for numerical stability. W^t is row-wisely normalized to the range between 0 and 1, and its each element $W^t_{i,j}$ indicates the reliability of the noisy label $\tilde{y}^j_i$. Based on it, we can gather knowledge from the other noisy label sources for the j-th source:

$$\hat{p}^{\neq j,t}(\boldsymbol{x}_i) = \sigma\left(\frac{\sum_{k=1,k\neq j}^s W^t_{i,k}\bar{z}^t_k(\boldsymbol{x}_i)}{\sum_{k=1,k\neq j}^s W^t_{i,k}}\right), \tag{11}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $\bar{z}^t_k(\boldsymbol{x}) = \lambda\bar{z}^{t-1}_k(\boldsymbol{x}) + (1-\lambda)z^t_k(\boldsymbol{x})$ is the Exponential Moving Average (EMA) over output logits, which is a well-used technique (Tarvainen & Valpola, 2017) for aggregating the temporal information. In Eq (11), the EMA output logits from

Algorithm 1 Self-cognitive Denoising for Multiple noisy label sources (SDM)

```
Input: Noisy Dataset $\tilde{D}$. Epochs for self-cognition $t_0$.
Output: Multi-tower neural network $G_\Theta$.
for $t = 1$ to max_epoch do
    if $t = 1$ then
        Initialize the elements of $M^t$, $\boldsymbol{q}^t$ to 1.
    else if $t \leq t_0$ then
        Obtain instance-wise filter matrix $M^t$ by Eq (7).
        Obtain annotator-wise quality vector $\boldsymbol{q}^t$ by Eq (9).
    else
        Set $M^t = M^{t_0}$, $\boldsymbol{q}^t = \boldsymbol{q}^{t_0}$.
    end if
    Gather $M^t$ and $\boldsymbol{q}^t$ into one matrix $W^t$ by Eq (10).
    for $o = 1$ to max_iteration do
        Draw a batch of noisy data $\tilde{B}$ from $\tilde{D}$.
        Compute loss $\mathcal{L}_B$ on $\tilde{B}$ by Eq (5).
        Compute loss $\mathcal{L}_M$ on $\tilde{B}$ by Eq (12).
        Update parameters $\Theta$ with loss $\mathcal{L}$ on $\tilde{B}$ by Eq (13).
    end for
end for
```

different sources are fused by the weighting matrix W^t, enhancing the reliability of the aggregated knowledge denoted by $\hat{p}^{\neq j,t}$. Subsequently, we devise a loss function that facilitates the mutual exchange of valuable knowledge among different noisy label sources as

$$\mathcal{L}_M = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{s}\left(1 - W_{i,j}^t\right)\cdot l\left(G_\Theta^j(\boldsymbol{x}_i), \hat{p}^{\neq j,t}\left(\boldsymbol{x}_i\right)\right). \quad (12)$$

The motivation behind the weighing term $1 - W_{i,j}^t$ is that $W_{i,j}^t$ indicates the reliability of the noisy label $\tilde{y}_i^j$, and $1 - W_{i,j}^t$ can represent how $\boldsymbol{x}_i$ need the knowledge from other sources. Note that we stop gradients for $\hat{p}^{\neq j,t}\left(\boldsymbol{x}_i\right)$ in practice, which means only the gradients from $G_\Theta^j(\boldsymbol{x}_i)$ are used for updating parameters.

During training, the overall loss can be formulated as

$$\mathcal{L} = \mathcal{L}_B + \alpha\mathcal{L}_M, \quad (13)$$

where α is the hyperparameter to balance these two losses. The overall training process of Self-cognitive Denoising for Multiple noisy label sources (SDM) is shown in Algorithm 1, and the ensemble score $\sigma\left(\frac{1}{s}\sum_{j=1}^{s} z_j\left(\boldsymbol{x}\right)\right)$ is used for any test sample $\boldsymbol{x}$ during the inference process.

3.3. Selective Distillation

Owing to the multi-tower architecture of our method, which involves concurrent learning from multiple noisy label sources, there is a potential challenge of computational resource demands when meeting a vast number of such sources. To address this, we also follow the theoretical insights to introduce a selective distillation module, aiming at generating a lightweight single-tower model.

Building upon the insights from theoretical analyses, we can quantify the reliability of each noisy label $\tilde{y}_i^j$ via $W_{i,j}^t$, which motivates us to selectively distill more valuable knowledge from the original model. Specifically, let $g_\theta(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$ denote a single-tower neural network with parameters θ. The distillation loss can be formulated as

$$\mathcal{L}_D = \frac{1}{n}\sum_{i=1}^{n}\left[l\left(g_\theta(\boldsymbol{x}_i), \mathcal{R}\left(\tilde{\boldsymbol{y}}_i, \boldsymbol{w}_i\right)\right) + \beta l\left(g_\theta(\boldsymbol{x}_i), \mathcal{R}\left(\hat{\boldsymbol{p}}_i, \boldsymbol{w}_i\right)\right)\right], \quad (14)$$

where $\tilde{\boldsymbol{y}}_i = [\tilde{y}_i^1, \ldots, \tilde{y}_i^s]^\top$ denotes all noisy labels of $\boldsymbol{x}_i$, $\hat{\boldsymbol{p}}_i = [G_\Theta^1(\boldsymbol{x}_i), \ldots, G_\Theta^s(\boldsymbol{x}_i)]^\top$ denotes the predictions of the trained multi-tower model G_Θ for $\boldsymbol{x}_i$, and $\boldsymbol{w}_i = [W_{i,1}^t, \ldots, W_{i,s}^t]^\top$ represents the weighting vector for $\boldsymbol{x}_i$. Furthermore, the notation $\mathcal{R}(\boldsymbol{a}, \boldsymbol{b})$ describes a weighted random selection process where one element is chosen from the vector $\boldsymbol{a}$ according to the probabilities specified in the associated probability vector $\boldsymbol{b}$. Since the weighting matrix W^t aggregates the instance-wise and annotator-wise cognition, the weighted random selection in Eq (14) will generally provide more reliable signals from both noisy labels and the teacher's predictions. We will later demonstrate in Section 4 that g_θ is capable of producing results on par with G_Θ, while requiring significantly fewer parameters.

4. Experiments

We conduct experiments to answer the following questions:

Q1: Whether SDM can outperform prior methods in the presence of multiple noisy label sources?

Q2: How does SDM perform in different settings, e.g., varying label qualities and number of noisy label sources?

Q3: Can the distilled model g_θ achieve comparable results with fewer parameters?

Q4: Can SDM appropriately estimate instance-wise filter matrix M^t and annotator-wise quality vector $\boldsymbol{q}^t$?

4.1. Datasets

Six benchmark datasets are adopted in experiments, i.e., *Yelp*, *IMDb*, *AgNews (AN)*, *SVHN*, *MNIST*, and *Bank*. Details about these datasets are given in Appendix B.1 due to space constraints. Since these datasets do not have existing noisy labels, we simulate noisy labels from $s = 4$ sources per dataset considering both class-dependent and instance-dependent noise with pre-defined basic label quality $r = 0.05$:

Table 1. Results of learning from multiple noisy labels sources with percentage of AUC.

Type	Class-dependent Noise						Instance-dependent Noise					
Dataset	Yelp	IMDb	AN	SVHN	MNIST	Bank	Yelp	IMDb	AN	SVHN	MNIST	Bank
Single	64.40	56.04	60.91	42.23	59.71	83.07	58.97	50.07	76.85	46.89	95.53	91.61
Major	70.51	57.63	57.44	46.89	79.58	89.93	52.83	53.83	73.63	64.95	95.46	92.50
EBCC	75.49	72.00	67.62	67.62	81.62	88.47	68.36	56.76	77.72	68.08	96.79	93.17
DoctorNet	68.46	61.60	70.35	57.42	63.13	88.87	73.38	64.38	76.26	78.52	97.03	92.80
CrowdLayer	66.98	62.63	69.23	61.58	62.21	88.84	73.55	62.96	87.71	82.67	96.35	93.41
CVL	82.48	68.10	69.54	70.32	**88.65**	86.86	74.73	59.47	73.80	80.16	95.13	93.16
WeaSEL	77.81	68.28	73.35	68.62	85.46	90.61	71.94	63.07	73.48	73.45	94.82	93.05
HE_M	80.15	67.19	68.99	62.12	76.51	85.01	71.60	62.48	83.16	78.16	97.75	92.95
HE_A	80.30	68.77	65.96	58.82	75.78	85.31	71.13	61.72	80.51	78.76	98.00	93.22
SLF	78.52	69.61	69.75	68.81	86.02	85.33	70.46	61.16	80.44	74.47	96.00	93.07
ADMoE	78.20	69.04	75.12	**71.19**	78.44	86.41	75.85	66.07	88.28	79.84	95.40	92.93
SDM (ours)	**85.43**	**75.93**	**81.39**	69.01	87.50	**91.91**	**77.07**	**66.50**	**89.45**	**83.75**	**98.17**	**93.44**

- **Class-dependent:** We swap $1 - r$ of positive samples' labels and the same number of negative samples' labels to generate noisy labels from the 4-th source. Similarly, we repeat the process with $0.9 - r$, $0.8 - r$, and $0.7 - r$ to generate noisy labels from the 3-rd, 2-nd, and 1-st sources respectively. Consequently, the four sources will exhibit a descending order of label quality.

- **Instance-dependent:** Following Zhao et al. (2023), we use r of ground-truth labels to train 4 different classifiers, i.e., Decision Tree, LightGBM, MLP, and Random Forest, and regard their inaccurate predictions as noisy labels from 4 different noisy label sources. In this way, the four sources will naturally have different label quality due to the diversity of the classifiers.

4.2. Setup

For the multi-tower neural networks G_Θ, we use a 3-layer MLP with hidden dimension 128, whose first layer extracts the public features among noisy label sources, and the other two layers are constructed with $s = 4$ towers to model the information of each source. Similarly, we use a simple 3-layer MLP with hidden dimension 128 to construct the single-tower neural network g_θ. More implementation details are given in Appendix B.2. We evaluate the performance with AUC (the Area Under the ROC Curve).

4.3. Compared Methods

We compare our method with various related methods for learning from multiple noisy label sources, including: (1) *Single* means the model is directly learned from one of the noisy label sources; (2) *Major* (Raykar et al., 2009) aggregated the multiple label sets by majority voting; (3) *Enhanced Bayesian Classifier Combination (EBCC)* (Li et al., 2019) aggregated the multiple label sets based on a mean-field variational approach; (4) *DoctorNet* (Guan et al., 2018) automatically learned the weighted vector for annotation integration; (5) *CrowdLayer* (Rodrigues & Pereira, 2018) simultaneously estimated the confusion matrices and updated the neural networks; (6) *Coupled-View deep classifier Learning (CVL)* (Li et al., 2020c) incorporated the idea of multi-view learning, in which the learning view from data was represented by deep neural networks for data classification and the learning view from labels was described by a Naive Bayes classifier for label aggregation; (7) *Weakly Supervised End-to-end Learner (WeaSEL)* (Rühling Cachay et al., 2021) trained a downstream model by maximizing the agreement of its predictions with probabilistic labels generated by another network. (8) HyperEnsemble (Wenzel et al., 2020) trained an individual model for noisy labels from each source, and gathered their outputs by maximizing, i.e., *HE_M* and averaging, i.e., *HE_A*; (9) *Sample-wise Label Fusion (SLF)* (Gao et al., 2022) jointly learned instance-dependent weight vectors and annotator confusion matrices during training; (10) *Anomaly Detection with Mixture-of-Experts (ADMoE)* (Zhao et al., 2023) learned annotators' confusions from noisy label sources without explicit label mapping using a MoE architecture.

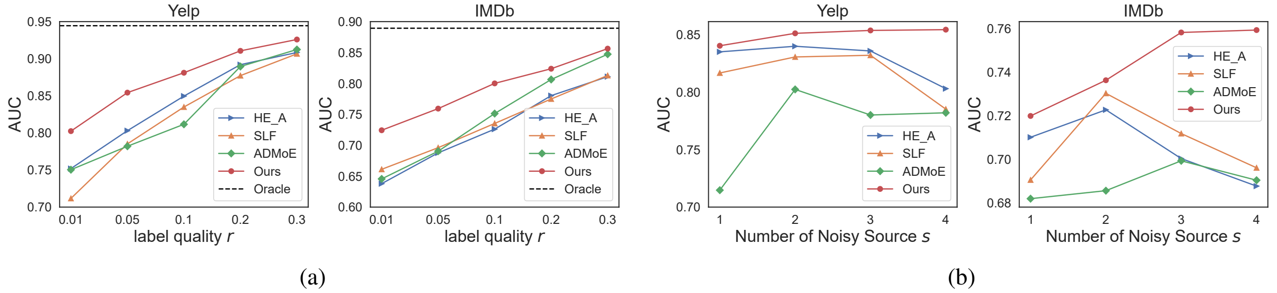
4.4. Results

Q1: Whether SDM can outperform prior methods in the presence of multiple noisy label sources?

The experimental results of the performance by our method and the compared methods on both class-dependent and instance-dependent noise are summarized in Table 1. It indicates that our method surpasses the compared methods on most of the datasets for both noise types.

Q2: How does SDM perform in different settings, e.g., varying label qualities and number of noisy label sources?

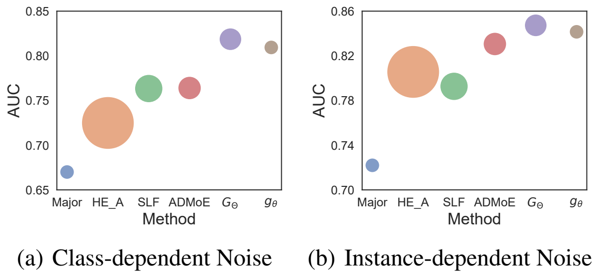
We conduct experiments with various label quality (0.01,



(a)

(b)

Figure 2. Additional experimental results with different (a) label quality r and (b) number of noisy source s.



(a) Class-dependent Noise (b) Instance-dependent Noise

Figure 3. The number of parameters (represented by the area of scatters) and performance of different methods.

0.05, 0.1, 0.2, and 0.3) on Yelp and IMDb datasets with class-dependent noise, and the results are shown in Figure 2(a). It can be found that our method demonstrates consistently superior performance over the compared methods across different levels of label quality. Notably, the advantage of our approach becomes more pronounced as the quality of labels decreases.

We also conduct experiments with various numbers of noisy label source s (from 1 to 4) on Yelp and IMDb datasets with class-dependent noise, and the results are shown in Figure 2(b). It is observed that the performance of the compared methods may deteriorate when they utilize an increasing number of noisy label sources. In contrast, our method shows improved performance with the addition of more noisy label sources, owing to its ability to harness the self-cognition power of neural networks to assess both the instance-wise and annotator-wise reliability of noisy labels.

Q3: Can the distilled model g_θ achieve comparable results with fewer parameters?

To reduce computing resources, we distill the valuable knowledge from the original model G_Θ to a lightweight model g_θ. We conduct experiments to compare their parameters and performance and benchmark them against Major, HE_A, SLF, and ADMoE. The average results over six datasets are demonstrated in Figure 3. The findings reveal that the distilled model g_θ, despite having fewer parameters, achieves performance on par with the original model G_Θ and consistently surpasses the other methods compared. Detailed results for each dataset, also supporting these conclusions, can be found in Appendix C.1.

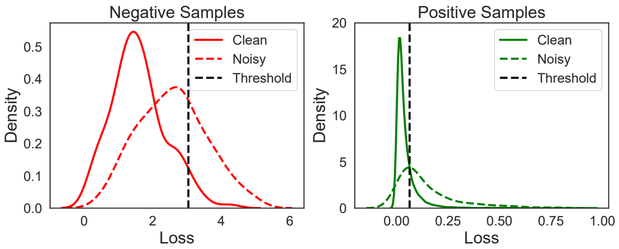
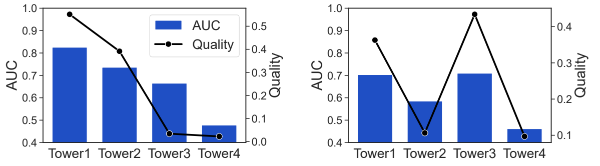


Figure 4. Illustration for the class-wise loss distributions over clean and noisy samples with the associated threshold.



(a) Class-dependent Noise (b) Instance-dependent Noise

Figure 5. Demonstration for the estimated annotator-wise quality vector $\boldsymbol{q}^t$ and the associated performance of each tower.

Q4: Can SDM appropriately estimate instance-wise filter matrix M^t and annotator-wise quality vector $\boldsymbol{q}^t$?

To demonstrate that the filter matrix M^t by Eq (7) indeed filters out noisy samples in each noisy label source, we illustrate class-wise loss distributions of the first tower in epoch t_0 in Figure 4. It is evident that noisy samples typically exhibit higher loss values compared to clean samples, and the majority of samples that are excluded by our designed class-wise thresholds are indeed noisy samples. The complete demonstrations of all four towers are given in Appendix C.2 due to space constraints.

To demonstrate that the quality vector $\boldsymbol{q}^t$ by Eq (9) indeed quantifies the reliability of each noisy label source, we conduct experiments on Yelp dataset and illustrate the estimated

Table 2. Ablation study for the modules in Self-cognitive Denoising for Multiple noisy label sources (SDM).

Type	Class-dependent Noise					Instance-dependent Noise				
Output Tower	Tower1	Tower2	Tower3	Tower4	Ensemble	Tower1	Tower2	Tower3	Tower4	Ensemble
Naive	82.54	73.59	66.48	47.79	80.76	70.30	58.52	70.94	46.15	63.59
+ with M^t	82.16	77.61	71.83	61.72	81.42	71.04	63.42	68.72	63.13	68.58
+ with $\boldsymbol{q}^t$	**83.26**	**82.85**	**79.42**	**73.11**	**85.43**	**71.96**	**71.94**	**76.24**	**73.14**	**77.07**

Table 3. Ablation study for the modules in distilling g_θ.

Ablation	Class-dependent	Instance-dependent
Naive	78.13	71.95
+ with $\boldsymbol{q}^t$	83.83	73.10
+ with distillation	**85.45**	**75.94**

quality vector $\boldsymbol{q}^t$. For comparison, we also illustrate the performance of each tower when training the naive multi-tower model, i.e., by Eq (5) only. Hence, the performance of each tower reflects the reliability of its corresponding noisy label source. The illustration provided in Figure 5 shows that there is a positive correlation between the elements of the estimated quality vector $\boldsymbol{q}^t$ and the reliability of their respective sources. This property will encourage the model to distance itself from potentially detrimental sources and instead focus more on those that are beneficial.

4.5. Ablation Study

In this section, we investigate the contribution of each module in our proposed method on the Yelp dataset.

Effectiveness of the modules in training G_Θ. In our proposed method, we estimate the instance-wise filter matrix M^t and annotator-wise quality vector $\boldsymbol{q}^t$, and then mutual-denoise the model based on them. The ablation study for these two parts is summarized in Table 2, in which the naive method means the model that is trained by Eq (5) only. With mutual-denoising based on instance-wise filter matrix M^t only, i.e., $W^t = M^t$, each tower and the final ensemble scores perform better than the naive model. Combined with the annotator-wise quality vector $\boldsymbol{q}^t$, the model's performance is further improved.

Effectiveness of the modules in distilling g_θ. The ablation study for the proposed selective distillation strategy is shown in Table 3. The naive method means the model is trained with completely random noisy labels only, i.e., $\beta = 0$ and $q_i^t = 1, \forall i$. With the estimated annotator-wise quality vector $\boldsymbol{q}^t$ as sample weights, the model can receive reliable training signals more frequently and perform better than the naive one. Combined with the distillation technique, i.e., $\beta = 1$, the model can learn the valuable knowledge of the multi-tower model G_Θ, which yields a further better performance.

5. Related Works

Learning from single noisy label source is a typical topic in weakly supervised learning (Zhou, 2018; Zhang et al., 2019) with inaccurate supervisions. The early works often tackle noisy labels via robust loss functions (Ghosh et al., 2017; Zhang & Sabuncu, 2018; Wang et al., 2019) or learning a noisy transition matrix (Patrini et al., 2017; Hendrycks et al., 2019; Goldberger & Ben-Reuven, 2017). Recently, a lot of works (Arpit et al., 2017; Li et al., 2020b; Liu et al., 2020; Gui et al., 2021) empirically or theoretically proved the memorization phenomenon of neural networks, i.e., neural networks start to fit correct labels and then overfit incorrect ones. It indicates that the neural networks are able to self-recognize noisy samples during training, which is named the self-cognition ability of the neural networks in this paper. Based on this phenomenon, a lot of works have been proposed to identify and select more accurate samples for training (Han et al., 2018; Yu et al., 2019; Wei et al., 2020; Li et al., 2020a).

Learning from multiple noisy label sources aims to use the wisdom from multiple label sets to obtain a more robust model. Early methods often treat annotation integration and downstream classification as separate tasks. For example, majority voting (Raykar et al., 2009) aggregated all labels via a weighted summation over multiple labels with a constant weight vector. Some other works (Dawid & Skene, 1979; Whitehill et al., 2009; Welinder et al., 2010; Ibrahim et al., 2019) first estimate the annotators' confusions, and then train the downstream classifiers with the integrated labels. However, these works often neglect the information of input samples when aggregating multiple noisy labels, resulting in unsatisfactory aggregated labels in practice. More recently, learning the annotators' confusions and the following classifier in an end-to-end manner has shown improved performance (Rodrigues & Pereira, 2018; Tanno et al., 2019; Guan et al., 2018; Cao et al., 2019; Li et al., 2020c; Gao et al., 2022; Zhao et al., 2023; Ibrahim et al., 2023). For example, Rodrigues & Pereira (2018) proposed a *crowdlayer* to simultaneously estimate the confusion matrices of multiple annotators and update the classifier. Gao et al. (2022) proposed to jointly learn instance-dependent weights and confusion matrices. Zhao et al. (2023) utilized a Mixture-of-Experts (MoE) architecture to implicitly learn annotators' confusions from multiple noisy label sources. However, all these methods neglect the self-cognition power of the neural networks, which is theoretically and empirically proved to be beneficial for learning from multiple noisy label sources in this paper.

Comparison with small-loss-based methods. The small-loss criterion (Gui et al., 2021) generally reflects the model's self-cognition capability at the instance level, and it has been typically studied within the single-source noise learning problem. In the context of multi-source noisy label learning, however, most related works do not focus on the small-loss criterion itself but rather utilize this concept to enhance their schemes' effectiveness and robustness through sample selection. For example, Li et al. (2020c) simply incorporated a small-loss-based co-teaching model (Han et al., 2018) in one view of the proposed multi-view learning. Compared to their key contribution, i.e., leveraging the ideas of multi-view learning into multi-source noisy learning, the utilization of the small-loss criterion is considered trivial and not the central point; Tian et al. (2022) only employed the small-loss-based sample selection during the training of data classifiers. The key contribution of this paper is the proposed mutual correction-based co-training framework, in which the utilization of the small loss criterion is not the key point; Zhang et al. (2024) borrowed the ideas of the small-loss criterion during the distillation of meta sets. The critical point of this article lies in a novel meta-learning-based method for efficiently mitigating the sparse annotation problem, wherein the small loss criterion is only a component of one of the modules, but not the core contribution. Contrastively, our work is centered on the model's self-cognition ability under multiple noisy label sources, both theoretically and methodologically, rather than simply applying off-the-shelf techniques.

6. Conclusion and Future Work

This paper focuses on learning from multiple noisy label sources and presents some novel theoretical analyses about the self-cognition ability of neural networks. The theoretical results inform the development of SDM, an approach that exploits the self-cognition power of neural networks for denoising during training. Additionally, a selective distillation module is designed to obtain a more lightweight model. The experimental results and abundant analyses verify the effectiveness of our method.

Currently, our work is focused on binary classification, which is applicable in many real-world tasks, e.g., anomaly detection and credit risk prediction. However, in the multi-class scenario, the applicability of our theories and the method derived from them is limited, and we need to further explore the necessary theoretical framework and method design. For instance, it is required to make assumptions about noise from different sources by class, and the estimation of annotator-wise quality $\boldsymbol{q}^t$ also needs to be expanded from one dimension to multiple dimensions. We believe our pioneering work in the binary classification scenario will inspire further exploration within the community of NN's self-cognition abilities under multiple noisy label sources, including the theoretical derivation and methodology design for multi-class cases.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Arpit, D., Jastrzkebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., et al. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, 2017.

Cao, P., Xu, Y., Kong, Y., and Wang, Y. Max-mig: an information theoretic approach for joint learning from crowds. *Proceedings of the 7th International Conference on Learning Representations (ICLR'19)*, 2019.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Gao, Z., Sun, F.-K., Yang, M., Ren, S., Xiong, Z., Engeler, M., Burazer, A., Wildling, L., Daniel, L., and Boning, D. S. Learning from multiple annotator noisy labels via sample-wise label fusion. In *Proceedings of the European Conference on Computer Vision (ECCV'22)*, pp. 407–422, 2022.

Ghosh, A., Kumar, H., and Sastry, P. S. Robust loss functions under label noise for deep neural networks. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17)*, 2017.

Goldberger, J. and Ben-Reuven, E. Training deep neural-networks using a noise adaptation layer. In *Proceedings of the 5th International Conference on Learning Representations (ICLR'17)*, 2017.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Guan, M., Gulshan, V., Dai, A., and Hinton, G. Who said what: Modeling individual labelers improves classification. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, 2018.

Gui, X.-J., Wang, W., and Tian, Z.-H. Towards understanding deep learning from noisy labels with small-loss criterion. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI'21)*, pp. 2469–2475, 2021.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. W., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31 (NeurIPS'18)*, 2018.

Han, S., Hu, X., Huang, H., Jiang, M., and Zhao, Y. Adbench: Anomaly detection benchmark. In *Advances in Neural Information Processing Systems 35 (NeurIPS'22)*, pp. 32142–32159, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR'16)*, pp. 770–778, 2016.

Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *Proceedings of the 36th International Conference on Machine Learning (ICML'19)*, pp. 2712–2721, 2019.

Ibrahim, S., Fu, X., Kargas, N., and Huang, K. Crowdsourcing via pairwise co-occurrences: Identifiability and algorithms. In *Advances in Neural Information Processing Systems 32 (NeurIPS'19)*, 2019.

Ibrahim, S., Nguyen, T., and Fu, X. Deep learning from crowdsourced labels: Coupled cross-entropy minimization, identifiability, and regularization. *Proceedings of the 11st International Conference on Learning Representations (ICLR'23)*, 2023.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NIPS'17)*, 2017.

Khetan, A., Lipton, Z. C., and Anandkumar, A. Learning from noisy singly-labeled data. *Proceedings of the 5th International Conference on Learning Representations (ICLR'17)*, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, 2015.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, J., Socher, R., and Hoi, S. C. Dividemix: Learning with noisy labels as semi-supervised learning. *Proceedings of the 8th International Conference on Learning Representations (ICLR'20)*, 2020a.

Li, M., Soltanolkotabi, M., and Oymak, S. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS'20)*, 2020b.

Li, S., Ge, S., Hua, Y., Zhang, C., Wen, H., Liu, T., and Wang, W. Coupled-view deep classifier learning from multiple noisy annotators. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20)*, 2020c.

Li, Y., Rubinstein, B., and Cohn, T. Exploiting worker correlation for label aggregation in crowdsourcing. In *Proceedings of the 36th International Conference on Machine Learning (ICML'19)*, pp. 3886–3895, 2019.

Liu, S., Niles-Weed, J., Razavian, N., and Fernandez-Granda, C. Early-learning regularization prevents memorization of noisy labels. In *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*, 2020.

Markelle, K., Rachel, L., and Kolby, N. The uci machine learning repository, 2013. URL `https://archive.ics.uci.edu`.

Patrini, G., Rozza, A., K. Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017.

Raykar, V. C., Yu, S., Zhao, L. H., Jerebko, A., Florin, C., Valadez, G. H., Bogoni, L., and Moy, L. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of the 26th International Conference on Machine Learning (ICML'09)*, pp. 889–896, 2009.

Rodrigues, F. and Pereira, F. Deep learning from crowds. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18)*, 2018.

Rodrigues, F., Pereira, F., and Ribeiro, B. Gaussian process classification and active learning with multiple annotators. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*, pp. 433–441, 2014.

Rühling Cachay, S., Boecking, B., and Dubrawski, A. End-to-end weak supervision. In *Advances in Neural Information Processing Systems 34 (NeurIPS'21)*, pp. 1845–1857, 2021.

Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D. C., and Silberman, N. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR'19)*, pp. 11244–11253, 2019.

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30 (NIPS'17)*, 2017.

Tian, N., Wu, M., Jiang, J., and Zhang, J. Learning from crowds with mutual correction-based co-training. In *Proceedings of the International Conference on Knowledge Graph*, pp. 257–264, 2022.

Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'19)*, pp. 322–330, 2019.

Wei, H., Feng, L., Chen, X., and An, B. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR'20)*, 2020.

Welinder, P., Branson, S., Perona, P., and Belongie, S. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23 (NIPS'10)*, volume 23, 2010.

Wenzel, F., Snoek, J., Tran, D., and Jenatton, R. Hyperparameter ensembles for robustness and uncertainty quantification. In *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*, 2020.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J., and Ruvolo, P. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22 (NIPS'09)*, 2009.

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I. W., and Sugiyama, M. How does disagreement help generalization against label corruption? In *Proceedings of the 36th International Conference on Machine Learning (ICML'19)*, 2019.

Zhang, H., Li, S., Zeng, D., Yan, C., and Ge, S. Coupled confusion correction: Learning from crowds with sparse annotations. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI'24)*, pp. 16732–16740, 2024.

Zhang, Y. and Yang, Q. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems 31 (NeurIPS'18)*, 2018.

Zhang, Z.-Y., Zhao, P., Jiang, Y., and Zhou, Z.-H. Learning from incomplete and inaccurate supervision. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19)*, pp. 1017–1025, 2019.

Zhao, Y., Zheng, G., Mukherjee, S., McCann, R., and Awadallah, A. Admoe: Anomaly detection with mixture-of-experts from noisy labels. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI'23)*, 2023.

Zhou, Z.-H. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.

A. Proof

Theorem 2.1. *(Single noisy label source) Let $h_{\Theta_k^*}$ denote the neural network minimizing the expected loss in the k-th noisy label source, i.e., $\mathbb{E}_{(\boldsymbol{x},\tilde{y}^k)}[\ell(h_\Theta(\boldsymbol{x}),\tilde{y}^k)]$. $(\boldsymbol{x}_1,\tilde{y}^k)$ and $(\boldsymbol{x}_2,\tilde{y}^k)$ are any two samples with the same observed label $\tilde{y}^k$ satisfying that $f^*(\boldsymbol{x}_1)=\tilde{y}^k$ and $f^*(\boldsymbol{x}_2)\neq\tilde{y}^k$. If T^k satisfies that $T^k_{ii}>0.5$, $\forall i\in\{0,1\}$, then we have*

$$\ell(h_{\Theta_k^*}(\boldsymbol{x}_1),\tilde{y}^k)<\ell(h_{\Theta_k^*}(\boldsymbol{x}_2),\tilde{y}^k).$$

Proof. Let $\tilde{\boldsymbol{d}}^k=[\tilde{d}^k_0,\tilde{d}^k_1]$ denote the one-hot version of $\tilde{y}^k$, i.e., $\tilde{d}^k_{\tilde{y}^k}=1$ and $\tilde{d}^k_{1-\tilde{y}^k}=0$. For binary cross-entropy loss function $\ell(h_\Theta(\boldsymbol{x}),\tilde{y}^k)=-\sum_{i=0}^1\tilde{d}^k_i\log(\hat{p}^k_i(\boldsymbol{x}))$, where $\hat{p}^k_i(\boldsymbol{x})$ is the predicted probability of the i-th class in the k-th noisy label source. Considering the expected loss of noisy data:

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{x},\tilde{y}^k)}[\ell(h_\Theta(\boldsymbol{x}),\tilde{y}^k)]&=-\mathbb{E}_{(\boldsymbol{x},\tilde{y}^k)}\Big[\sum_{i=0}^1\tilde{d}^k_i\log(\hat{p}^k_i(\boldsymbol{x}))\Big]\\
&=-\int_{\boldsymbol{x}\in\mathcal{X}}\sum_{j=0}^1\Big[\sum_{i=0}^1\tilde{d}^k_i\log(\hat{p}^k_i(\boldsymbol{x}))\Big]p(\boldsymbol{x},\tilde{y}^k=j)\mathrm{d}\boldsymbol{x}\\
&=-\int_{\boldsymbol{x}\in\mathcal{X}}\Big[\sum_{j=0}^1\sum_{i=0}^1\tilde{d}^k_i\log(\hat{p}^k_i(\boldsymbol{x}))p(\tilde{y}^k=j|\boldsymbol{x})\Big]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&=-\int_{\boldsymbol{x}\in\mathcal{X}}\Big[\sum_{i=0}^1\Big[\sum_{j=0}^1\tilde{d}^k_i p(\tilde{y}^k=j|\boldsymbol{x})\Big]\log(\hat{p}^k_i(\boldsymbol{x}))\Big]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\
&=-\int_{\boldsymbol{x}\in\mathcal{X}}\Big[\sum_{i=0}^1\mathbb{E}[\tilde{d}^k_i|\boldsymbol{x}]\log(\hat{p}^k_i(\boldsymbol{x}))\Big]p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.
\end{aligned}$$

Therefore, minimizing $\mathbb{E}_{(\boldsymbol{x},\tilde{y}^k)}[\ell(h_\Theta(\boldsymbol{x}),\tilde{y}^k)]$ equals to minimizing $-\sum_{i=0}^1\mathbb{E}[\tilde{d}^k_i|\boldsymbol{x}]\log(\hat{p}^k_i(\boldsymbol{x}))$ for each $\boldsymbol{x}\in\mathcal{X}$. Due to the constraints that $\sum_{i=0}^1\hat{p}^k_i(\boldsymbol{x})=1$ and $0\leq\hat{p}^k_i(\boldsymbol{x})\leq1,\forall i\in\{0,1\}$, we have $-\sum_{i=0}^1\mathbb{E}[\tilde{d}^k_i|\boldsymbol{x}]\log(\hat{p}^k_i(\boldsymbol{x}))$ is minimized when $\hat{p}^k_i(\boldsymbol{x})=\mathbb{E}[\tilde{d}^k_i|\boldsymbol{x}]$, $\forall i\in\{0,1\}$ by Lagrange multiplier method. Furthermore, since $\mathbb{E}[\tilde{d}^k_i|\boldsymbol{x}]=\sum_{j=0}^1\mathbb{I}[i=j]p(\tilde{y}^k=j|\boldsymbol{x})=p(\tilde{y}^k=i|\boldsymbol{x})$, we have $\hat{p}^k_i(\boldsymbol{x})=p(\tilde{y}^k=i|\boldsymbol{x})$. Then we can obtain

$$\begin{aligned}
\hat{p}^k_i(\boldsymbol{x})=p(\tilde{y}^k=i|\boldsymbol{x})&=\sum_{j=0}^1p(\tilde{y}^k=i,y=j|\boldsymbol{x})\\
&=\sum_{j=0}^1p(y=j|\boldsymbol{x})p(\tilde{y}^k=i|y=j,\boldsymbol{x})\\
&=\sum_{j=0}^1p(y=j|\boldsymbol{x})p(\tilde{y}^k=i|y=j)\\
&=p(\tilde{y}^k=i|y=f^*(\boldsymbol{x}))\\
&=T^k_{f^*(\boldsymbol{x})i},
\end{aligned}$$

where the fourth equation is due to the class-dependent noise assumption and the fifth equation is due to that each $\boldsymbol{x}$ has only one true label $f^*(\boldsymbol{x})$. Therefore, the output of $h_{\Theta_k^*}$ satisfies $\hat{p}^k_i(\boldsymbol{x})=T^k_{f^*(\boldsymbol{x})i}$ for $\boldsymbol{x}\in\mathcal{X}$. For any two examples $(\boldsymbol{x}_1,\tilde{y}^k)$ and $(\boldsymbol{x}_2,\tilde{y}^k)$ with the same observed label $\tilde{y}^k$ satisfying that $f^*(\boldsymbol{x}_1)=\tilde{y}^k$ and $f^*(\boldsymbol{x}_2)\neq\tilde{y}^k$, the loss value of $h_{\Theta_k^*}$ on $(\boldsymbol{x}_1,\tilde{y}^k)$ is

$$\ell(h_{\Theta_k^*}(\boldsymbol{x}_1),\tilde{y}^k)=-\log(\hat{p}^k_{\tilde{y}^k}(\boldsymbol{x}_1))=-\log(T^k_{f^*(\boldsymbol{x}_1)\tilde{y}^k}).$$

Similarly, the loss value of $h_{\Theta_k^*}$ on $(\boldsymbol{x}_2,\tilde{y}^k)$ is

$$\ell(h_{\Theta_k^*}(\boldsymbol{x}_2),\tilde{y}^k)=-\log(\hat{p}^k_{\tilde{y}^k}(\boldsymbol{x}_2))=-\log(T^k_{f^*(\boldsymbol{x}_2)\tilde{y}^k}).$$

If T^k satisfies that $T^k_{ii} > 0.5$, $\forall i \in \{0, 1\}$, we have $T^k_{ii} > T^k_{ji}$, $\forall i \neq j$. Since $f^*(\boldsymbol{x}_1) = \tilde{y}^k$ while $f^*(\boldsymbol{x}_2) \neq \tilde{y}^k$, we have $\ell(h_{\Theta^*_k}(\boldsymbol{x}_1), \tilde{y}^k) = -\log(T^k_{f^*(\boldsymbol{x}_1)\tilde{y}^k}) = -\log(T^k_{\tilde{y}^k\tilde{y}^k}) < -\log(T^k_{f^*(\boldsymbol{x}_2)\tilde{y}^k}) = \ell(h_{\Theta^*_k}(\boldsymbol{x}_2), \tilde{y}^k)$. Theorem 2.1 is proved. □

Theorem 2.2. *(Two noisy label sources) Let $h_{\Theta^*_1}$ and $h_{\Theta^*_2}$ denote the neural networks minimizing the expected loss in the 1-th and 2-nd noisy label source respectively, i.e., $\mathbb{E}_{(\boldsymbol{x},\tilde{y}^1)}[\ell(h_\Theta(\boldsymbol{x}), \tilde{y}^1)]$ and $\mathbb{E}_{(\boldsymbol{x},\tilde{y}^2)}[\ell(h_\Theta(\boldsymbol{x}), \tilde{y}^2)]$. If $T^1_{ii} > T^2_{ii} > 0.5$, $\forall i \in \{0, 1\}$, then we have*

$$\mathbb{E}_{(\boldsymbol{x},\tilde{y}^1)}\ell(h_{\Theta^*_1}(\boldsymbol{x}), \tilde{y}^1) < \mathbb{E}_{(\boldsymbol{x},\tilde{y}^2)}\ell(h_{\Theta^*_2}(\boldsymbol{x}), \tilde{y}^2), \tag{1}$$

and

$$\mathbb{E}_{(\boldsymbol{x},y)}\ell(h_{\Theta^*_1}(\boldsymbol{x}), y) < \mathbb{E}_{(\boldsymbol{x},y)}\ell(h_{\Theta^*_2}(\boldsymbol{x}), y). \tag{2}$$

Proof. From the proof of Theorem 2.1, we have the output of $h_{\Theta^*_1}$ satisfies $\hat{p}^1_i(\boldsymbol{x}) = p(\tilde{y}^1 = i|\boldsymbol{x}) = T^1_{f^*(\boldsymbol{x})i}$ for any $\boldsymbol{x} \in \mathcal{X}$. Similarly, the output of $h_{\Theta^*_2}$ satisfies $\hat{p}^2_i(\boldsymbol{x}) = p(\tilde{y}^2 = i|\boldsymbol{x}) = T^2_{f^*(\boldsymbol{x})i}$ for any $\boldsymbol{x} \in \mathcal{X}$. The expected loss value on $(\boldsymbol{x}, \tilde{y}_1)$ is

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{x},\tilde{y}^1)}\ell(h_{\Theta^*_1(\boldsymbol{x})}, \tilde{y}^1) &= -\mathbb{E}_{(\boldsymbol{x},\tilde{y}^1)}\log(\hat{p}^1_{\tilde{y}^1}(\boldsymbol{x})) \\
&= -\mathbb{E}_{(\boldsymbol{x},\tilde{y}^1)}\log(T^1_{f^*(\boldsymbol{x})\tilde{y}^1}) \\
&= -\int_{\boldsymbol{x}\in\mathcal{X}}\sum_{j=0}^{1}\log(T^1_{f^*(\boldsymbol{x})j})p(\boldsymbol{x}, \tilde{y}^1 = j)\mathrm{d}\boldsymbol{x} \\
&= -\int_{\boldsymbol{x}\in\mathcal{X}}\sum_{j=0}^{1}\log(T^1_{f^*(\boldsymbol{x})j})p(\tilde{y}^1 = j|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= -\int_{\boldsymbol{x}\in\mathcal{X}}\sum_{j=0}^{1}T^1_{f^*(\boldsymbol{x})j}\log(T^1_{f^*(\boldsymbol{x})j})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x},
\end{aligned}$$

where the fifth equation is due to $p(\tilde{y}^1 = j|\boldsymbol{x}) = T^1_{f^*(\boldsymbol{x})j}$. Similarly, we have

$$\mathbb{E}_{(\boldsymbol{x},\tilde{y}^2)}\ell(h_{\Theta^*_2(\boldsymbol{x})}, \tilde{y}^2) = -\int_{\boldsymbol{x}\in\mathcal{X}}\sum_{j=0}^{1}T^2_{f^*(\boldsymbol{x})j}\log(T^2_{f^*(\boldsymbol{x})j})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

Construct a function $h(x) = -x\log x - (1 - x)\log(1 - x)$, and the derivative of $h(x)$ is

$$\frac{\partial}{\partial x}h(x) = -1 - \log x + 1 + \log(1 - x) = \log\frac{1 - x}{x}.$$

When $x \in (0.5, 1)$, we have $\frac{\partial}{\partial x}h(x) < 0$, so $h(x)$ is monotonically decreasing. When $f^*(\boldsymbol{x}) = 0$, we have

$$\begin{aligned}
-\sum_{j=0}^{1}T^1_{f^*(\boldsymbol{x})j}\log(T^1_{f^*(\boldsymbol{x})j}) &= -\left[T^1_{00}\log(T^1_{00}) + T^1_{01}\log(T^1_{01})\right] \\
&= -T^1_{00}\log(T^1_{00}) - (1 - T^1_{00})\log(1 - T^1_{00}) \\
&= h(T^1_{00}).
\end{aligned}$$

Similarly, we have $-\sum_{j=0}^{1}T^2_{f^*(\boldsymbol{x})j}\log(T^2_{f^*(\boldsymbol{x})j}) = h(T^2_{00})$. Since $T^1_{00} > T^2_{00} > 0.5$, we know that $h(T^1_{00}) < h(T^2_{00})$, so

$$-\sum_{j=0}^{1}T^1_{f^*(\boldsymbol{x})j}\log(T^1_{f^*(\boldsymbol{x})j}) < -\sum_{j=0}^{1}T^2_{f^*(\boldsymbol{x})j}\log(T^2_{f^*(\boldsymbol{x})j})$$

when $f^*(\boldsymbol{x}) = 0$. By the similar proof process, we can obtain that it also holds when $f^*(\boldsymbol{x}) = 1$. Since the two noisy labels $\tilde{y}^1$ and $\tilde{y}^2$ are always associated with the same sample $\boldsymbol{x}$, we can obtain that

$$-\int_{\boldsymbol{x}\in\mathcal{X}}\sum_{j=0}^{1}T^1_{f^*(\boldsymbol{x})j}\log(T^1_{f^*(\boldsymbol{x})j})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} < -\int_{\boldsymbol{x}\in\mathcal{X}}\sum_{j=0}^{1}T^2_{f^*(\boldsymbol{x})j}\log(T^2_{f^*(\boldsymbol{x})j})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

which equals to Eq (1).

Furthermore, the expected loss value on $(\boldsymbol{x}, y)$ for $h_{\Theta_1^*}$ can be formulated as

$$\begin{aligned}\mathbb{E}_{(\boldsymbol{x},y)}\ell(h_{\Theta_1^*}(\boldsymbol{x}), y) &= -\mathbb{E}_{(\boldsymbol{x},y)}\log(\hat{p}^1_{f^*(\boldsymbol{x})}(\boldsymbol{x})) \\ &= -\mathbb{E}_{(\boldsymbol{x},y)}\log(T^1_{f^*(\boldsymbol{x})f^*(\boldsymbol{x})}) \\ &= -\int_{\boldsymbol{x}\in\mathcal{X}}\sum_{j=0}^{1}\log(T^1_{jj})p(\boldsymbol{x}, y=j)\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Similarly, we have

$$\mathbb{E}_{(\boldsymbol{x},y)}\ell(h_{\Theta_2^*}(\boldsymbol{x}), y) = -\int_{\boldsymbol{x}\in\mathcal{X}}\sum_{j=0}^{1}\log(T^2_{jj})p(\boldsymbol{x}, y=j)\mathrm{d}\boldsymbol{x}.$$

Since $T^1_{jj} > T^2_{jj}, \forall j \in \{0,1\}$, we can obtain that $-\sum_{j=0}^{1}\log(T^1_{jj}) < -\sum_{j=0}^{1}\log(T^2_{jj})$. Due to the fact that $h_{\Theta_1^*}$ and $h_{\Theta_2^*}$ are trained with the example $(\boldsymbol{x}, y)$ from the same distribution, we have that

$$-\int_{\boldsymbol{x}\in\mathcal{X}}\sum_{j=0}^{1}\log(T^1_{jj})p(y=j|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} < -\int_{\boldsymbol{x}\in\mathcal{X}}\sum_{j=0}^{1}\log(T^2_{jj})p(y=j|\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

which equals to Eq (2). □

Corollary 2.3. *(Multiple noisy label sources) Let $h_{\Theta_k^*}$ denote the neural network minimizing the expected loss in the k-th noisy label source among s sources, i.e., $\mathbb{E}_{(\boldsymbol{x},\tilde{y}^k)}[\ell(h_\Theta(\boldsymbol{x}), \tilde{y}^k)]$. If $T^1_{ii} > \cdots > T^s_{ii} > 0.5$, $\forall i \in \{0,1\}$, then we have*

$$\mathbb{E}_{(\boldsymbol{x},\tilde{y}^1)}\ell(h_{\Theta_1^*}(\boldsymbol{x}), \tilde{y}^1) < \cdots < \mathbb{E}_{(\boldsymbol{x},\tilde{y}^s)}\ell(h_{\Theta_s^*}(\boldsymbol{x}), \tilde{y}^s), \tag{3}$$

and

$$\mathbb{E}_{(\boldsymbol{x},y)}\ell(h_{\Theta_1^*}(\boldsymbol{x}), y) < \cdots < \mathbb{E}_{(\boldsymbol{x},y)}\ell(h_{\Theta_s^*}(\boldsymbol{x}), y). \tag{4}$$

Proof. For any $j \in \{1, \ldots, s-1\}$ and any $i \in \{0,1\}$, the transition matrices satisfy that $T^j_{ii} > T^{j+1}_{ii} > 0.5$, so we have that $\mathbb{E}_{(\boldsymbol{x},\tilde{y}^j)}\ell(h_{\Theta_j^*}(\boldsymbol{x}), \tilde{y}^j) < \mathbb{E}_{(\boldsymbol{x},\tilde{y}^{j+1})}\ell(h_{\Theta_{j+1}^*}(\boldsymbol{x}), \tilde{y}^{j+1})$ and $\mathbb{E}_{(\boldsymbol{x},y)}\ell(h_{\Theta_j^*}(\boldsymbol{x}), y) < \mathbb{E}_{(\boldsymbol{x},y)}\ell(h_{\Theta_{j+1}^*}(\boldsymbol{x}), y)$ hold for any $j \in \{1, \ldots, s-1\}$ by Theorem 2.2. These two formulations can be straightforwardly extended to Eq (3) and Eq (4) respectively. □

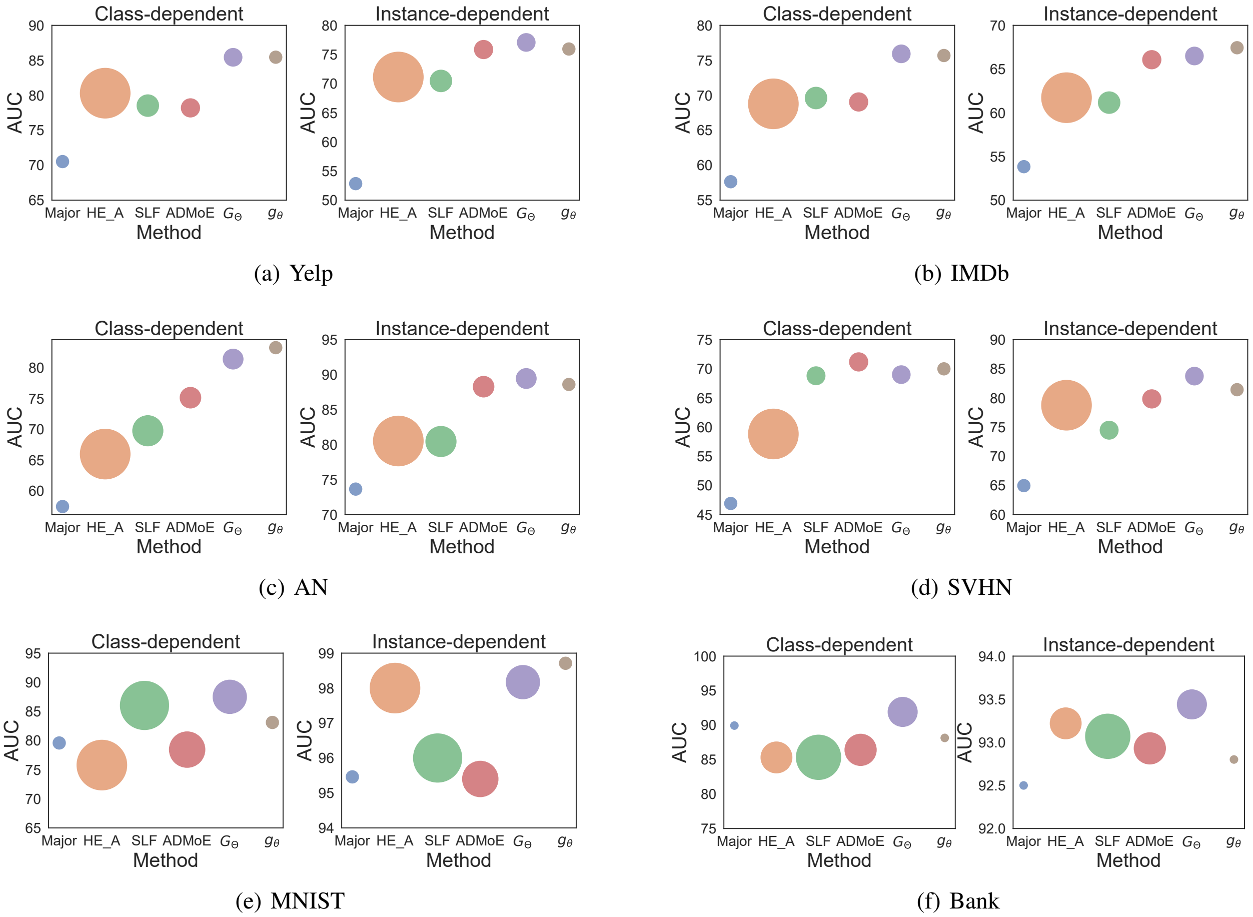
B. Details of Experiments

B.1. Details of Datasets

In experiments, six benchmark datasets are adopted, i.e., three NLP datasets named *Yelp*, *IMDb* and *AgNews*, two CV datasets named *SVHN* and *MNIST*, and a tabular dataset named *Bank*. Following Han et al. (2022); Zhao et al. (2023), we use pre-extracted features and pre-defined binary ground-truth labels for the NLP and CV datasets. For each dataset, we use 70% for training, 25% for testing, and 5% for validation. We list some details of the used datasets as follows:

1. *Yelp*, *IMDb*, and *AgNews*: For these NLP datasets, we utilize BERT (Devlin et al., 2018) to pre-extract features to the dimension of 768. For the Yelp dataset, we regard the reviews of 0 and 1 stars as the positive class, and the reviews of 3 and 4 stars as the negative class. For the IMDb dataset, we use the original ground-truth labels; For the AgNews dataset, we set one of the classes as negative and downsample the remaining classes to 5% of the total instances as positives. All these datasets have 10000 samples, where 500 samples are positive.

2. *SVHN*: For this dataset, we utilize ResNet-18 (He et al., 2016) to pre-extract features to the dimension of 512. We set one of the multi-classes as negative and downsample the remaining classes to 5% of the total instances as positive, constructing 160 positive samples in the total 5208 samples.



(a) Yelp

(b) IMDb

(c) AN

(d) SVHN

(e) MNIST

(f) Bank

Figure 6. Complete results with the number of parameters (represented by the area of scatters) and performance of different methods on six benchmark datasets. The parameters of models vary across datasets due to different dimensions of input features.

3. *MNIST*: For this dataset, we utilize LeNet (LeCun et al., 1998) to pre-extract features to the dimension of 100. There are 7603 samples, where 700 samples are positive.

4. *Bank*: This tabular dataset is from the UCI repository (Markelle et al., 2013), which contains information on various customers for the purpose of predicting their likelihood of conversion. There are 41176 samples with 16 features of different types, e.g., categorical, numerical, and date, where 4639 samples are positive.

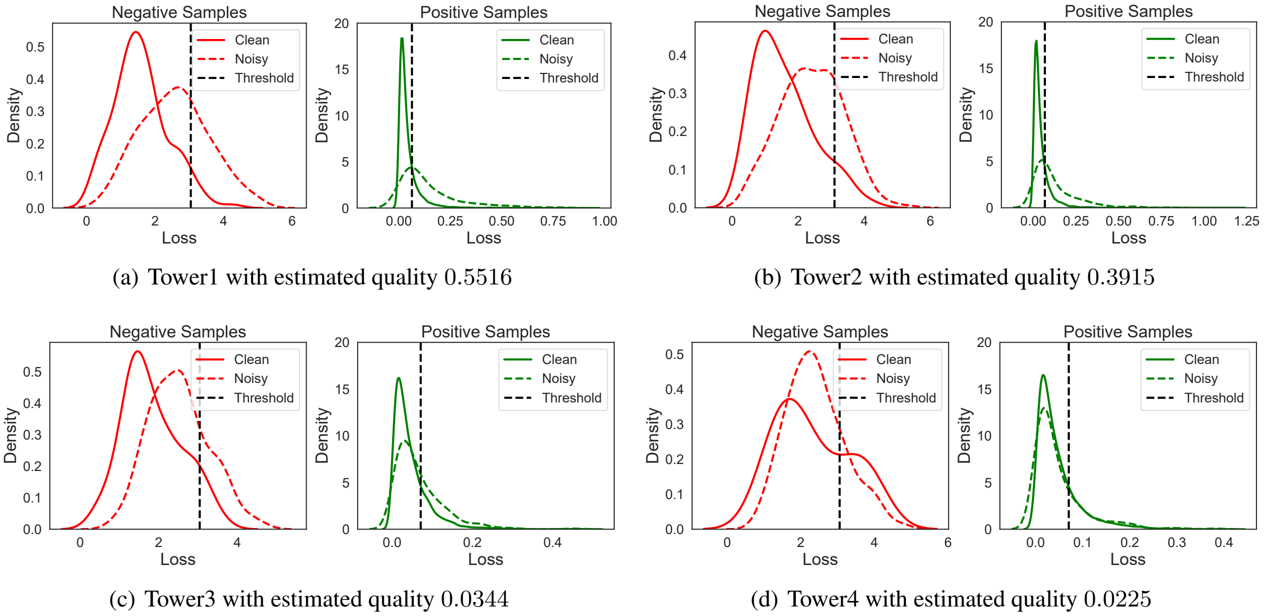
B.2. Details of implementation

During training, we use Adam (Kingma & Ba, 2015) with an initial learning rate of 0.001, a batch size of 256, and training epochs of 100 for both G_Θ and g_θ. We set the hyperparameters $P = 80$, $T = 0.1$, $t_0 = 20$, $\lambda = 0.9$, $\alpha = \beta = 1$ in all the experiments. The analysis of important hyperparameters can be found in Appendix C.3. For a fair comparison, the experiments are also conducted with 100 training epochs and 256 batch size on MLP with hidden dimension 128 for all the compared methods. All results are the averaging results of the last 5 epochs.

C. Additional Experimental Results

C.1. Complete Results for the Distilled Model

To ease the problem of computing resources, we distill the valuable knowledge from the original multi-tower model G_Θ to a lighter single-tower model g_θ. In the main paper, to compare the parameters and performance among these two models and other compared methods, we demonstrate the mean results over six benchmark datasets. Furthermore, we give the



(a) Tower1 with estimated quality 0.5516

(b) Tower2 with estimated quality 0.3915

(c) Tower3 with estimated quality 0.0344

(d) Tower4 with estimated quality 0.0225

Figure 7. Illustration for the class-wise loss distributions over clean and noisy samples with the associated threshold on different towers.

complete demonstrations of each dataset in Figure 6, which indicates that the distilled model g_θ with fewer parameters yields comparable performance and sometimes even better to the original model G_Θ, and outperforms the compared methods on many datasets. Note that the parameters of models vary across datasets due to different dimensions of input features.

C.2. Complete Loss Distributions of Each Tower

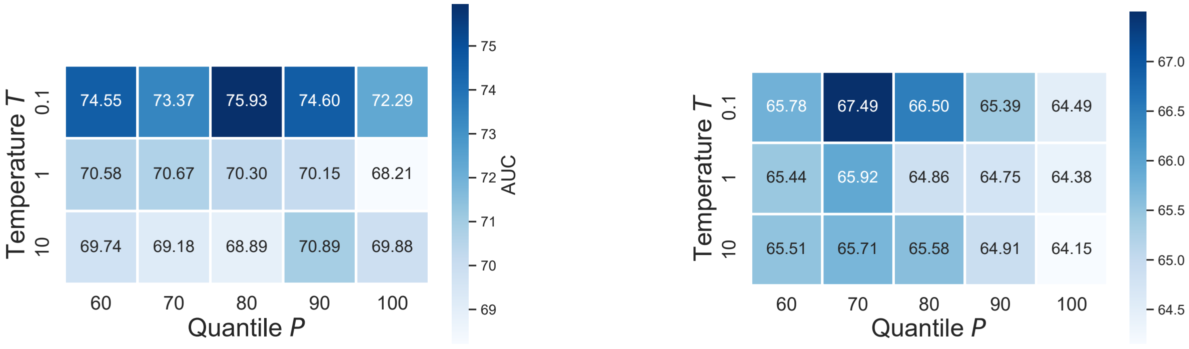
To demonstrate that the filter matrix M^t indeed filters out the noisy samples in each noisy label source, we illustrate the class-wise loss distributions in epoch t_0. In the main paper, we only illustrate the loss distributions of the first tower due to space constraints. Furthermore, we illustrate the complete loss distributions of each tower in Figure 7. It can be found that most of the samples that were filtered out by the first two towers with our designed class-wise thresholds are noisy samples. Nonetheless, the remaining two towers exhibit deficiencies in effectively discriminating noisy samples. This issue stems from their exposure to markedly inferior noisy label sources throughout the training process, which hinders their ability to autonomously identify noisy samples in these sources. Consequently, this introduces the concern that the filter matrix M^t might become redundant or potentially harmful when faced with particularly inferior noisy label sources. Despite this, the calculated annotator-wise quality $\boldsymbol{q}^t$ assigns minimal weights, i.e., 0.0344 and 0.0225, to such sources during the training phase, thereby safeguarding the model's robustness.

C.3. Hyperparameter Analysis

We give the analyses on two important hyperparameters, i.e., the threshold quantile P in estimating instance-wise filter matrix M^t and the temperature in estimating the annotator-wise quality $\boldsymbol{q}^t$. We conduct experiments on the IMDb dataset with both class-dependent and instance-dependent noise, and the results are shown in Figure 8. It can be found that appropriately decreasing temperature T yields better performance. The reason lies in that an appropriately small temperature can amplify the impact of the estimated cosine similarities in Eq (8), which will enforce the model to pay more attention to the helpful sources. As for the threshold quantile P, we find that this hyperparameter exhibits low sensitivity in an appropriate range. We use $P = 80$ and $T = 0.1$ in all the experiments.

C.4. Discussion about Ensemble Strategies

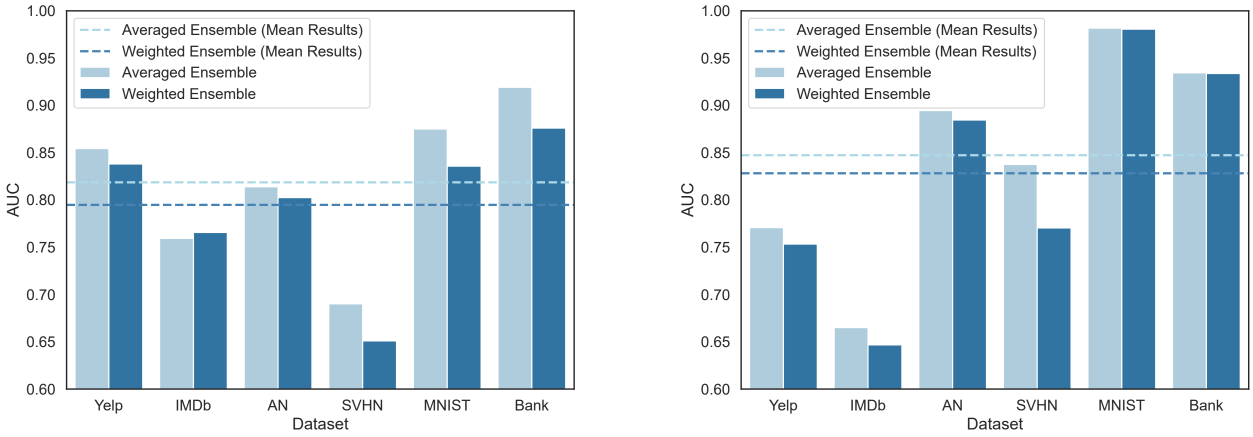
In our proposed method, the averaged ensemble score $\sigma\left(\frac{1}{s}\sum_{j=1}^{s} z_j\left(\boldsymbol{x}\right)\right)$ is used for any test sample $\boldsymbol{x}$ during the inference process. However, it is yet to be investigated whether a weighted ensemble score via $\boldsymbol{q}^t$, i.e., $\sigma\left(\frac{1}{s}\sum_{j=1}^{s} q_j^t z_j(\boldsymbol{x})\right)$, could



(a) Class-dependent Noise

(b) Instance-dependent Noise

Figure 8. Hyperparameter Analysis on the threshold quantile P and temperature T.



(a) Class-dependent Noise

(b) Instance-dependent Noise

Figure 9. Performance on various datasets with different ensemble strategies.

yield improved performance. To answer this, we conduct experiments on various datasets with weighted ensemble scores during the inference process, and the results are illustrated in Figure 9. It is observed that the averaged ensemble yields marginally superior results compared to the weighted ensemble for both class-dependent and instance-dependent noise. The reason lies in that different towers acquire useful knowledge from one another throughout the training process, rendering the initially estimated annotator-wise quality $\boldsymbol{q}^t$ unsuitable for application post-training.

C.5. Results on Real-world Noisy Datasets

In this section, we provide experimental results on two real-world noisy datasets: Music and Sentiment Polarity (Rodrigues et al., 2014). Both datasets were published on Amazon Mechanical Turk for annotation, and separate test sets are provided. We list some details of these two datasets as follows:

- Music: It is a music genre classification dataset, which consists of 1K music pieces with 30 seconds in length. All the music pieces are from 10 music genres (classical, country, disco, hiphop, jazz, rock, blues, reggae, pop, and metal) and are labeled by 44 annotators in total, with an average of 4.2 annotators per piece. We regard the 'blues' as the positive class and others as the negative class.

- Polarity: It contains 5000 sentences from movie reviews extracted from the website RottenTomatoes.com and whose sentiment was classified as positive or negative. The training set are labeled by 203 annotators in total, with an average of 5.5 annotators per instance.

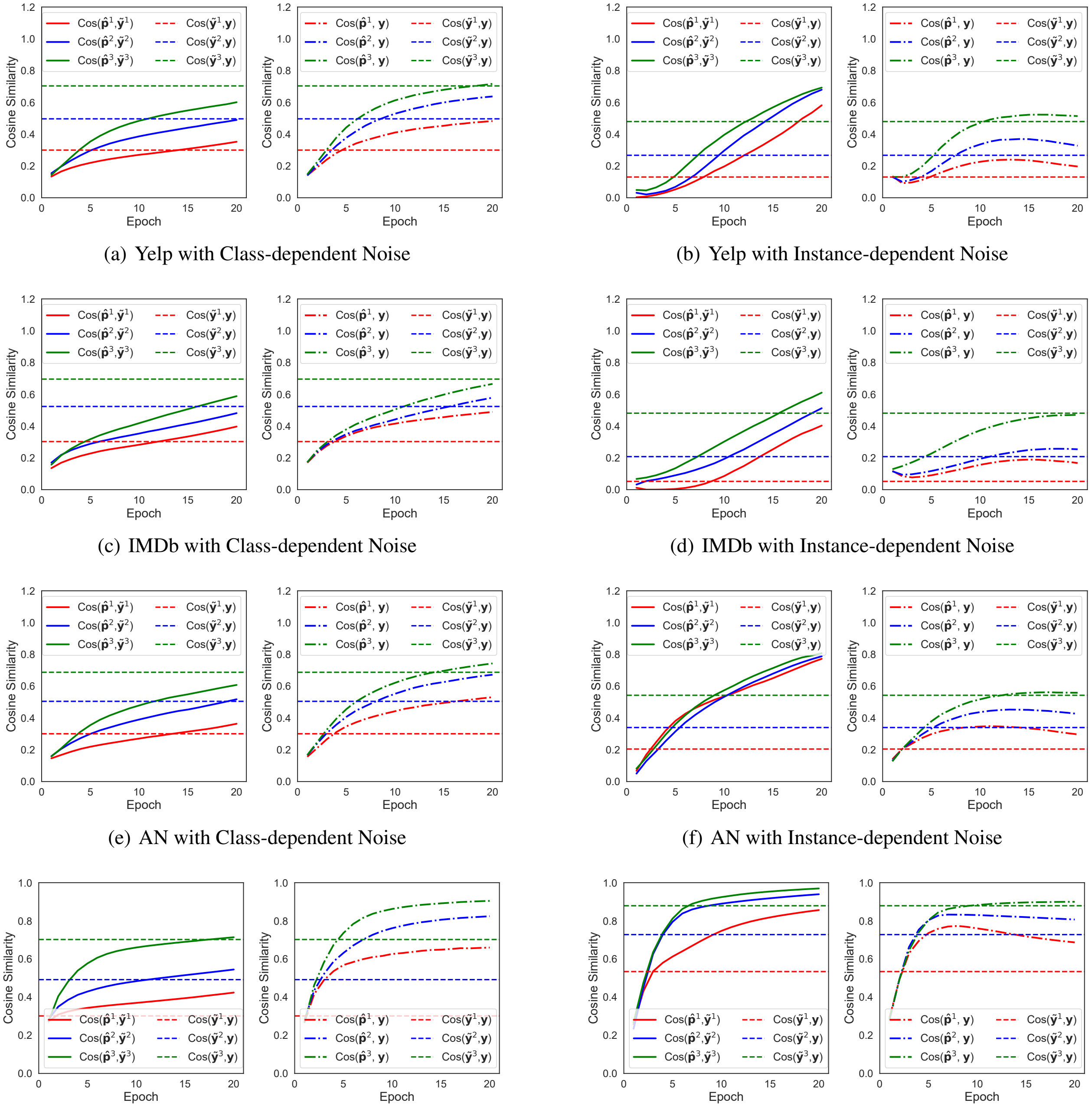
Table 4. Results of real-world noisy datasets with percentage of AUC.

	Single	Major	EBCC	DoctorNet	CrowdLayer	CVL	WeaSEL	HE_M	HE_A	SLF	ADMoE	SDM
Music	68.94	70.95	71.72	71.59	70.56	72.63	71.96	69.98	71.59	72.06	72.74	**74.55**
Polarity	72.17	75.28	75.63	74.93	75.21	75.39	75.45	73.65	75.82	76.13	75.89	**76.43**

For each training instance, we choose labels from 3 annotators. For the instance that does not have 3 annotators, we use the negative label as default. We conduct experiments on these two real-world noisy datasets, with both our methods and the compared methods. The experimental results are summarized in Table 4, which indicates that our method also surpasses the compared methods on the real-world noisy datasets.

C.6. More Empirical verifications for the Theoretical Insights

In the main paper, we empirically verify on the Yelp dataset with instance-dependent noise that the inspirations of Theorem 2.2 and Corollary 2.3 hold in practice. In this section, we aim to provide additional empirical evidence across different datasets featuring both class-dependent and instance-dependent noise, to showcase the broad applicability of our theoretical findings. For the instance-dependent noise, we follow the process in the main paper, i.e., we use varying percentages (10%, 30%, 50%) of ground-truth labels to train three LightGBM classifiers to consider their predictions as noisy labels from three sources. As for the class-dependent noise, we follow the process in the experiment section, i.e., we swap $1 - r$ of positive samples' labels and the same number of negative samples' labels to generate noisy labels. Here, we set $r = 0.7$, $r = 0.5$, $r = 0.3$ for the 1-st, 2-nd, 3-rd noisy label source respectively. We conduct experiments on the Yelp, IMDb, AgNews (AN), and MNIST datasets with both types of noise, and the results are shown in Figure 10. The left part of Figure 10(a)-10(h) jointly indicate that neural networks are easier to fit noisy supervisions in a more accurate source, which empirically verifies the insights from Eq (1) and Eq (3). Meanwhile, the right part of Figure 10(a)-10(h) jointly indicate that neural networks' predictions are closer to true label distributions in a more accurate source, which empirically verifies the insights from Eq (2) and Eq (4).



(a) Yelp with Class-dependent Noise

(b) Yelp with Instance-dependent Noise

(c) IMDb with Class-dependent Noise

(d) IMDb with Instance-dependent Noise

(e) AN with Class-dependent Noise

(f) AN with Instance-dependent Noise

(g) MNIST with Class-dependent Noise

(h) MNIST with Instance-dependent Noise

Figure 10. Empirical verificaitions for the theoretical insights on different datasets with different types of noise.