To Aggregate or Not? Learning with Separate Noisy Labels

A PREPRINT

Jiaheng Wei* UC Santa Cruz
Zhaowei Zhu* UC Santa Cruz
Tianyi Luo UC Santa Cruz
Ehsan Amid Google Brain
Abhishek Kumar Google Brain
Yang Liu † UC Santa Cruz

ABSTRACT

The rawly collected training data often comes with separate noisy labels collected from multiple imperfect annotators (e.g., via crowdsourcing). A typically way of using these separate labels is to first aggregate them into one and apply standard training methods. The literature has also studied extensively on effective aggregation approaches. This paper revisits this choice and aims to provide an answer to the question of whether one should aggregate separate noisy labels into single ones or use them separately as given. We theoretically analyze the performance of both approaches under the empirical risk minimization framework for a number of popular loss functions, including the ones designed specifically for the problem of learning with noisy labels. Our theorems conclude that label separation is preferred over label aggregation when the noise rates are high, or the number of labelers/annotations is insufficient. Extensive empirical results validate our conclusions.

1 Introduction

Training high-quality deep neural networks for classification tasks typically requires a large quantity of annotated data. The raw training data often comes with separate noisy labels collected from multiple imperfect annotators. For example, the popular data collection paradigm crowdsourcing [10, 16, 27] offers the platform to collect such annotations from unverified crowd; medical records are often accompanied with diagnosis from multiple doctors [1, 45]; news articles can receive multiple checkings (of the article being fake or not) from different experts [34, 37]. This leads to the situation considered in this paper: learning with multiple separate noisy labels.

The most popular approach to learning from the multiple separate labels would be aggregating the given labels for each instance [40, 58, 44, 42, 30], through an expectation-maximization (EM) inference technique. Each instance will then be provided with one single label, and applied with the standard training procedure.

The primary goal of this paper is to revisit the choice of aggregating separate labels and hope to provide practitioners understandings for the following question:

Should the learner aggregate separate noisy labels for one instance into a single label or not?

Our main contributions can be summarized as follows:

- We provide theoretical insights on how separation methods and aggregation ones result in different biases (Theorem 3.4, 4.2, 4.6) and variances (Theorem 3.6, 4.3, 4.7) of the output classifier from training. Our analysis considers both the standard loss functions in use, as well as popular robust losses that are designed for the problem of learning with noisy labels.
- By comparing the analytical proxy of the worst-case performance bounds, our theoretical results reveal that separating multiple noisy labels is preferred over label aggregation when the noise rates are high, or the number of labelers/annotations is insufficient. The results are consistent for both basic loss function ℓ and robust designs, including loss correction and peer loss.
- We carry out extensive experiments using both synthetic and real-world datasets to validate our theoretical findings.

1.1 Related Works

Label separation vs label aggregation Existing works mainly compare the separation with aggregation by empirical results. For example, it has been shown that label separation could be effective in improving model performance and

*Equal contributions.
†Correspondence to yangliu@ucsc.edu. (Paper under review)

may be potentially more preferable than aggregated labels through majority voting [17]. When training with the cross-entropy loss, Sheng et.al [46] observes that label separation reduces the bias and roughness, and outperforms majority-voting aggregated labels. However, it is unclear whether the results hold when robust treatments are employed. Similar problems have also been studied in corrupted label detection with a result leaning towards separation but not proved [64]. Another line of approach concentrates on the end-to-end training scheme or ensemble methods which takes all the separate noisy labels as the input during the training process [63, 12, 43, 5, 54], and learning from separate noisy labels directly.

Learning with noisy labels Popular approaches in learning with noisy labels could be broadly divided into following categories, i.e., (1) Adjusting the loss on noisy labels by: using the knowledge of noise label transition matrix [35, 36, 61, 67, 68]; re-weighting the per-sample loss by down-weighting instances with potentially wrong labels [24, 4, 3, 32, 19]; or refurbishing the noisy labels [41, 28, 55]. (2) Robust loss designs that do not require the knowledge of noise transition matrix [51, 2, 50, 31, 26, 65, 56]. (3) Regularization techniques to prevent deep neural networks from memorizing noisy labels [59, 22, 23, 7, 53]. (4) Dynamical sample selection procedure which behaves like a semi-supervised manner and begins with a clean sample selection procedure, then makes use of the wrongly-labeled samples [21, 6, 29]. For example, several methods [14, 62, 52] adopt a mentor/peer network to select small-loss samples as "clean" ones for the student/peer network. See [13, 48] for a more detailed survey of existing noise-robust techniques.

2 Formulation

Consider an M-class classification task and let $X \in \mathcal{X}$ and $Y \in \mathcal{Y} := \{1, 2, ..., M\}$ denote the input examples and their corresponding labels, respectively. We assume that $(X, Y) \sim \mathcal{D}$, where $\mathcal{D}$ is the joint data distribution. Samples (x, y) are generated according to random variables (X, Y). In the clean and ideal scenario, the learner has access to N training data points $D := \{(x_n, y_n)\}_{n\in[N]}$. Instead of having access to ground truth labels y_ns, we only have access to a set of noisy labels $\{\tilde{y}^\circ_{n,i}\}_{i\in[K]}$ for $n \in [N]$. For ease of presentation, we adopt the decorator $\circ$ to denote separate labels, and $\bullet$ for aggregated labels specified later. Noisy labels $\tilde{y}^\circ_n$s are generated according to the random variable $\widetilde{Y}^\circ$. We consider the class-dependent label noise transition [24, 35] where $\widetilde{Y}^\circ$ is generated according to a transition matrix T° with its entries defined as follows:

$$T^\circ_{k,l} := \mathbb{P}(\widetilde{Y}^\circ = l|Y = k).$$

Most of the existing results on learning with noisy labels have considered the setting where each x_n is paired with only one noisy label $\tilde{y}^\circ_n$. In practice, we often operate in a setting where each data point x_n is associated with multiple separate labels drawn from the same noisy label generation process [11, 25]. We consider this setting and assume that for each x_n, there are K independent noisy labels $\tilde{y}^\circ_{n,1}, ..., \tilde{y}^\circ_{n,K}$ obtained from K annotators.

We are interested in two popular ways to leverage multiple separate noisy labels:

- Keep the separate labels as separate and apply standard learning with noisy labels techniques to each of them.
- Aggregate noisy labels into one label, and then apply standard learning with noisy data techniques.

We will look into each of the above two settings separately and then answer the question:

"Should the learner aggregate multiple separate noisy labels or not?"

2.1 Label Separation

Denote the column vector $\mathbb{P}_{\widetilde{Y}^\circ} := [\mathbb{P}(\widetilde{Y}^\circ = 1), \cdots, \mathbb{P}(\widetilde{Y}^\circ = M)]^\top$ as the marginal distribution of $\widetilde{Y}^\circ$. Accordingly, we can define $\mathbb{P}_Y$ for Y. Clearly, we have the relation: $\mathbb{P}_{\widetilde{Y}^\circ} = T^\circ \cdot \mathbb{P}_Y, \mathbb{P}_Y = (T^\circ)^{-1} \cdot \mathbb{P}_{\widetilde{Y}^\circ}$. Denote by $\rho^\circ_1 := \mathbb{P}(\widetilde{Y}^\circ = 0|Y = 1), \rho^\circ_0 := \mathbb{P}(\widetilde{Y}^\circ = 1|Y = 0)$. The noise transition matrix T has the following form when $M = 2$:

$$T^\circ = \begin{bmatrix} 1 - \rho^\circ_0 & \rho^\circ_0 \\ \rho^\circ_1 & 1 - \rho^\circ_1 \end{bmatrix}.$$

For label separation, we define the per-sample loss function as:

$$\ell(f(x_n), \tilde{y}^\circ_{n,1}, ..., \tilde{y}^\circ_{n,K}) = \frac{1}{K} \sum_{i\in[K]} \ell(f(x_n), \tilde{y}^\circ_{n,i}).$$

For simplicity, we shorthand $\ell(f(x_n), \tilde{y}^\circ_n) := \ell(f(x_n), \tilde{y}^\circ_{n,1}, ..., \tilde{y}^\circ_{n,K})$ for the loss of label separation method when there is no confusion.

2.2 Label Aggregation

The other way to leverage multiple separate noisy labels is generating a single label via label aggregation methods using K noisy ones:

$$\tilde{y}_n^\bullet := \mathsf{Aggregation}(\tilde{y}_{n,1}^\circ, \tilde{y}_{n,2}^\circ, ..., \tilde{y}_{n,K}^\circ),$$

where the aggregated noisy labels $\tilde{y}_n^\bullet$s are generated according to the random variable $\widetilde{Y}^\bullet$. Denote the confusion matrix for this single & aggregated noisy label as $T^\bullet$. Popular aggregation methods include majority vote and EM inference, which are covered by our theoretical insights since our analyses in later sections would be built on the general label aggregation method. For a better understanding, we introduce the majority vote as an example.

Example of Majority Vote Given the majority voted label, we could compute the transition matrix between $\widetilde{Y}^\bullet$ and the true label Y using the knowledge of T°. The lemma below gives the closed form for $T^\bullet$ in terms of T°, when adopting majority vote.

Lemma 2.1. *Assume K is odd and recall that in the binary classification task, $T_{i,j}^\circ = \mathbb{P}(\widetilde{Y}^\circ = j | Y = i)$, the noise transition matrix of the (majority voting) aggregated noisy labels $T_{p,q}^\bullet$ becomes:*

$$T_{p,q}^\bullet = \sum_{i=0}^{\frac{K+1}{2}-1} \binom{K}{i} (T_{p,q}^\circ)^{K-i} (T_{p,1-q}^\circ)^i, \quad p, q \in \{0, 1\}.$$

When $K = 3$, then $T_{1,0}^\bullet = \mathbb{P}(\widetilde{Y}^\bullet = 0 | Y = 1) = (T_{1,0}^\circ)^3 + \binom{3}{1}(T_{1,0}^\circ)^2(T_{1,1}^\circ)$. Note it still holds that $T_{p,q}^\bullet + T_{p,1-q}^\bullet = 1$. For the aggregation method, as illustrated in Figure 1, the x-axis indicates the number of labelers K, and the y-axis denotes the aggregated noise rate given that the overall noise rate is in $[0.2, 0.4, 0.6, 0.8]$. When the number of labelers is large (i.e., $K < 10$) and the noise rate is small, both majority vote and EM label aggregation methods significantly reduce the noise rate. Although the expectation maximization method consumes much more time when generating the aggregated label, it frequently results in a lower aggregated noise rate than majority vote.

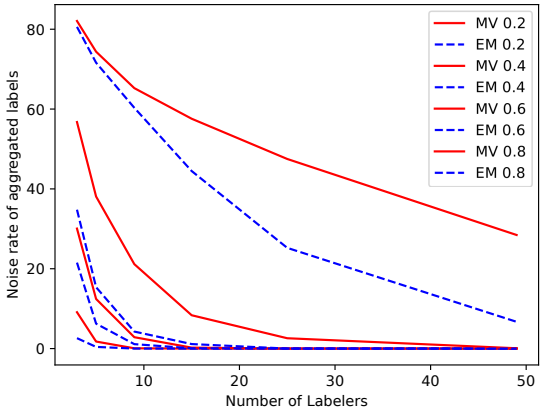

Figure 1: Noise rates of the aggregated labels in synthetic noisy CIFAR-10. MV: majority vote. EM: expectation maximization. 0.2–0.8: Original noise rates before aggregation.

3 Bias and Variance Analyses w.r.t. ℓ-loss

In this section, we provide theoretical insights on how label separation and aggregation methods result in different biases and variances of the classifier prediction, when learning with the standard loss function ℓ.

Suppose the clean training samples $\{(x_n, y_n)\}_{n\in[N]}$ are given by variables (X, Y) such that $(X, Y) \sim \mathcal{D}$. Recall that instead of having access to a set of clean training samples $D = \{(x_n, y_n)\}_{n\in[N]}$, the learner only observes K noisy labels $\tilde{y}_{n,1}^\circ, ..., \tilde{y}_{n,K}^\circ$ for each x_n, denoted by $\widetilde{D}^\circ := \{(x_n, \tilde{y}_{n,1}^\circ, ..., \tilde{y}_{n,K}^\circ)\}_{n\in[N]}$. For separation methods, the noisy training samples are obtained through variables $(X, \widetilde{Y}_1^\circ)$, ..., $(X, \widetilde{Y}_K^\circ)$ where $(X, \widetilde{Y}_i^\circ) \sim \widetilde{\mathcal{D}}^\circ$ for $i \in [K]$. For aggregation methods such as majority vote, we assume the data points and aggregated noisy labels $\widetilde{D}^\bullet := \{(x_n, \tilde{y}_n^\bullet)\}_{n\in[N]}$

are drawn from $(X, \widetilde{Y}^{\bullet}) \sim \widetilde{\mathcal{D}}^{\bullet}$ where $\widetilde{Y}^{\bullet}$ is produced through the majority voting of $\widetilde{Y}_1^{\circ}, ..., \widetilde{Y}_K^{\circ}$. When we mention "noise rate", it usually refers to the average noise: $\mathbb{P}(\widetilde{Y}^{\mathrm{u}} \neq Y)$.

ℓ-risk under the distribution Given the loss ℓ, note that $\ell(f(x_n), \tilde{y}_n^{\circ})$ is denoted as $\ell(f(x_n), \tilde{y}_{n,1}^{\circ}, ..., \tilde{y}_{n,K}^{\circ}) = \frac{1}{K}\sum_{i\in[K]} \ell(f(x_n), \tilde{y}_{n,i}^{\circ})$, we define the empirical ℓ-risk for learning with separated/aggregated labels under noisy labels as: $\hat{R}_{\ell,\widetilde{D}^u}(f) = \frac{1}{N}\sum_{i=1}^{N} \ell\left(f(x_i), \tilde{y}_i^u\right)$, $u \in \{\circ, \bullet\}$ unifies the treatment which is either separation $\circ$ or aggregation $\bullet$. By increasing the sample size N, we would expect $\hat{R}_{\ell,\widetilde{D}^u}(f)$ to be close to the following ℓ-risk under the noisy distribution $\widetilde{\mathcal{D}}^u$: $R_{\ell,\widetilde{\mathcal{D}}^u}(f) = \mathbb{E}_{(X,\widetilde{Y}^u)\sim\widetilde{\mathcal{D}}^u}[\ell(f(X), \widetilde{Y}^u)]$.

3.1 Bias of a Given Classifier w.r.t. ℓ-Loss

We denote by $f^* \in \mathcal{F}$ the optimal classifier obtained through the clean data distribution $(X, Y) \sim \mathcal{D}$ within the hypothesis space $\mathcal{F}$. We formally define the bias of a given classifier $\hat{f}$ as:

Definition 3.1 (Classifier Prediction Bias of ℓ-Loss)**.** Denote by $R_{\ell,\mathcal{D}}(\hat{f}) := \mathbb{E}_{\mathcal{D}}[\ell(\hat{f}(X), Y)]$, $R_{\ell,\mathcal{D}}(f^*) := \mathbb{E}_{\mathcal{D}}[\ell(f^*(X), Y)]$. The bias of classifier $\hat{f}$ writes as: $\mathrm{Bias}(\hat{f}) = R_{\ell,\mathcal{D}}(\hat{f}) - R_{\ell,\mathcal{D}}(f^*)$.

The Bias term quantifies the prediction bias (excess risk) of a given classifier $\hat{f}$ on the clean data distribution $\mathcal{D}$ w.r.t. the optimal achievable classifier f^*, which can be decomposed as [66]

$$\mathrm{Bias}(\hat{f}) = \underbrace{R_{\ell,\mathcal{D}}(\hat{f}) - R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f})}_{\text{Distribution shift}} + \underbrace{R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}) - R_{\ell,\mathcal{D}}(f^*)}_{\text{Estimation error}}. \tag{1}$$

Now we bound the distribution shift and the estimation error in the following two lemmas.

Lemma 3.2 (Distribution shift)**.** *Denote by $p_i := \mathbb{P}(Y = i)$, assume ℓ is upper bounded by $\bar{\ell}$ and lower bounded by $\underline{\ell}$. The distribution shift in Eqn. (1) is upper bounded by*

$$R_{\ell,\mathcal{D}}(\hat{f}) - R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}) \leq \overline{\Delta}_R^{u,1} := \left(\rho_0^u p_0 + \rho_1^u p_1\right) \cdot \left(\bar{\ell} - \underline{\ell}\right). \tag{2}$$

Lemma 3.3 (Estimation error)**.** *Suppose the loss function $\ell(f(x), y)$ is L-Lipschitz for any feasible y. $\forall f \in \mathcal{F}$, with probability at least $1-\delta$, the estimation error is upper bounded by*

$$R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}) - R_{\ell,\mathcal{D}}(f^*) \leq \overline{\Delta}_R^{u,2} := 4L \cdot \mathfrak{R}(\mathcal{F}) + (\bar{\ell} - \underline{\ell}) \cdot \sqrt{\frac{2\log(1/\delta)}{\eta_K^u N}} + \overline{\Delta}_R^{u,1},$$

where $u \in \{\circ, \bullet\}$ denotes either separation or aggregation methods, $\eta_K^{\circ} = \frac{K \cdot \log(\frac{1}{\delta})}{2\left(\log(\frac{K+1}{\delta})\right)^2}$ and $\eta_K^{\bullet} \equiv 1$ indicate the richness factor, which characterizes the effect of the number of labelers, and $\mathfrak{R}(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$.

Noting that the number of unique instances x_i are the same for both treatments, the duplicated copies of x_i are supposed to introduce at least no less effective samples, i.e., the richness factor satisfies that $\eta_K^u \geq 1$. Thus, we update $\eta_K^{\circ} := \max\{\eta_K^{\circ}, 1\}$, and Figure 2 visualizes the estimated η_K° given different number of labelers as well as δ. It is clear that when the number of labelers is larger, or δ is smaller, $\eta_K^{\circ} > \eta_K^{\bullet}$. Later we shall show how η_K^u influences the bias and variance of the classifier prediction.

To give a more intuitive comparison of the performance of both mechanisms, we adopt the worst-case bias upper bound $\overline{\Delta}_R^u := \overline{\Delta}_R^{u,1} + \overline{\Delta}_R^{u,2}$ from Lemma 3.2 and Lemma 3.3 as a proxy and derive Theorem 3.4.

Theorem 3.4. *Denote by $\alpha_K := (\rho_0^{\circ} p_0 + \rho_1^{\circ} p_1) - (\rho_0^{\bullet} p_0 + \rho_1^{\bullet} p_1)$, $\gamma = \sqrt{\log(1/\delta)/2N}$. The separation bias proxy $\overline{\Delta}_R^{\circ}$ is smaller than the aggregation bias proxy $\overline{\Delta}_R^{\bullet}$ if and only if*

$$\alpha_K \cdot \frac{1}{1 - (\eta_K^{\circ})^{-\frac{1}{2}}} \leq \gamma. \tag{3}$$

Note that α_K and η_K° are non-decreasing w.r.t. the increase of K, in Section 4.3, we will explore how the LHS of Eqn. (3) is influenced by K: a short answer is that the LHS of Eqn. (3) is (generally) monotonically increasing w.r.t. K when K is small, indicating that Eqn. (3) is easier to be achieved given fixed δ, N and a smaller K than a larger one.

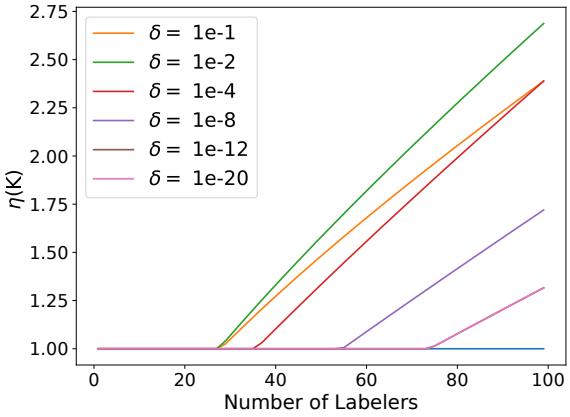

Figure 2: The visualization of estimated η_K° given varied δ.

3.2 Variance of a Given Classifier w.r.t. ℓ-Loss

We now move on to explore the variance of a given classifier when learning with ℓ-loss, prior to the discussion, we define the variance of a given classifier as:

Definition 3.5 (Classifier Prediction Variance of ℓ-Loss)**.** The variance of a given classifier $\hat{f}$ when learned with separation ($\circ$) or aggregation ($\bullet$) is defined as:

$$\mathrm{Var}(\hat{f}) = \mathbb{E}_{(X,\widetilde{Y}^u)\sim\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}(X),\widetilde{Y}^u) - \mathbb{E}_{(X,\widetilde{Y}^u)\sim\widetilde{\mathcal{D}}^u}[\ell(\hat{f}(X),\widetilde{Y}^u)]\right]^2.$$

For $g(x) = x - x^2$, we derive the closed form of Var and the corresponding upper bound as below.

Theorem 3.6. *When $\eta_K^u \geq \frac{2\log(1/\delta)}{N}$, given ℓ is 0-1 loss, we have:*

$$Var(\hat{f}^u) = g(R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u)) \leq \overbrace{g\left(\sqrt{\frac{2\log(1/\delta)}{\eta_K^u N}}\right)}^{\textit{Variance proxy}}. \tag{4}$$

The variance proxy of $Var(\hat{f}^\circ)$ in Eqn. (4) is smaller than that of $Var(\hat{f}^\bullet)$.

Theorem 3.6 provides another view to decide on the choices of separation and aggregation methods, i.e., the proxy of classifier prediction variance. To extend the theoretical conclusions w.r.t. ℓ loss to the multi-class setting, we only need to modify the upper bound of the distribution shift in Eqn. (2), as specified in the following corollary.

Corollary 3.7 (Multi-Class Extension (ℓ-Loss))**.** *In the M-class classification case, the upper bound of the distribution shift in Eqn. (2) becomes:*

$$R_{\ell,\mathcal{D}}(\hat{f}) - R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}) \leq \overline{\Delta}_R^{u,1} := \sum_{j\in[M]} p_j \cdot (1 - T_{j,j}^u) \cdot \left(\overline{\ell} - \underline{\ell}\right). \tag{5}$$

4 Bias and Variance Analyses with Robust Treatments

Intuitively, the learning of noisy labels problem could benefit from more robust loss functions build upon the generic ℓ loss, i.e., backward correction (surrogate loss) [35, 36], and peer loss functions [26]. We move on to explore the best way to learn with multiple copies of noisy labels, when combined with existing robust approaches.

4.1 Backward Loss Correction

When combined with the backward loss correction approach ($\ell \to \ell_\leftarrow$), the empirical ℓ risks become: $\hat{R}_{\ell_\leftarrow,\widetilde{D}^u}(f) = \frac{1}{N}\sum_{i=1}^N \ell_\leftarrow(f(x_i),\tilde{y}_i^u)$, where the corrected loss in the binary case is defined as

$$\ell_\leftarrow(f(x),\tilde{y}^u) = \frac{(1-\rho_{1-\tilde{y}^u})\cdot\ell(f(x),\tilde{y}^u) - \rho_{\tilde{y}^u}\cdot\ell(f(x),1-\tilde{y}^u)}{1-\rho_0^u-\rho_1^u}.$$

Bias of given classifier w.r.t. $\ell_{\leftarrow}$ Suppose the loss function $\ell(f(x), y)$ is L-Lipschitz for any feasible y. Define $L^u_{\leftarrow} := L^u_{\leftarrow 0} \cdot L$, where $L^u_{\leftarrow 0} := \frac{(1+|\rho^u_0 - \rho^u_1|)}{1-\rho^u_0-\rho^u_1}$. Denote by $R_{\ell,\mathcal{D}}(\hat{f})$ the ℓ-risk of the classifier $\hat{f}$ under the clean data distribution $\mathcal{D}$, with $\hat{f} = \hat{f}^u_{\leftarrow} = \operatorname{argmin}_{f\in\mathcal{F}} \hat{R}_{\ell_{\leftarrow},\widetilde{D}^u}(f)$. Lemma 4.1 gives the upper bound of classifier prediction bias when learning with $\ell_{\leftarrow}$ via separation or aggregation methods.

Lemma 4.1. *With probability at least $1-\delta$, we have:*

$$R_{\ell,\mathcal{D}}(\hat{f}^u_{\leftarrow}) - R_{\ell,\mathcal{D}}(f^*) \leq \overline{\Delta}^u_{R_{\leftarrow}} := 4L^u_{\leftarrow} \cdot \mathfrak{R}(\mathcal{F}) + L^u_{\leftarrow 0} \cdot (\overline{\ell} - \underline{\ell}) \cdot \sqrt{\frac{2\log(1/\delta)}{\eta^u_K N}}.$$

Lemma 4.1 offers the upper bound of the performance gap for the given classifier f w.r.t the clean distribution $\mathcal{D}$, comparing to the minimum achievable risk. We consider the bound $\overline{\Delta}^u_{R_{\leftarrow}}$ as a proxy of the bias, and we are interested in the case where training the classifier separately yields a smaller bias proxy compared to that of the aggregation method, formally $\overline{\Delta}^{\circ}_{R_{\leftarrow}} < \overline{\Delta}^{\bullet}_{R_{\leftarrow}}$. For any finite hypothesis class $\mathcal{F} \subset \{f : X \to \{0,1\}\}$, and the sample set $S = \{x_1, ..., x_N\}$, denote by d the VC-dimension of $\mathcal{F}$, we give conditions when training separately yields a smaller bias proxy.

Theorem 4.2. *Denote by $\alpha_K := 1 - L^{\bullet}_{\leftarrow}/L^{\circ}_{\leftarrow}$, $\gamma = 1/\left(1 + \frac{4L}{\overline{\ell}-\underline{\ell}}\sqrt{\frac{d\log(N)}{\log(1/\delta)}}\right)$, where d is the VC-dimension of $\mathcal{F}$. For backward loss correction, the separation bias proxy $\overline{\Delta}^{\circ}_{R_{\leftarrow}}$ is smaller than the aggregation bias proxy $\overline{\Delta}^{\bullet}_{R_{\leftarrow}}$ if and only if*

$$\alpha_K \cdot \frac{1}{1-(\eta^{\circ}_K)^{-\frac{1}{2}}} \leq \gamma. \tag{6}$$

We defer our empirical analysis of the monotonicity of the LHS in Eqn. (6) to Section 4.3 as well, which shares similar monotonicity behavior to learning w.r.t. ℓ.

Variance of given classifiers with Backward Loss Correction Similar to the previous subsection, we now move on to check how separation and aggregation methods result in different variance when training with loss correction.

Theorem 4.3. *When $L^u_{\leftarrow 0}(\eta^u_K)^{-\frac{1}{2}} < \sqrt{\frac{N}{2(\overline{\ell}-\underline{\ell})^2\log(1/\delta)}}$, $Var(\hat{f}^u_{\leftarrow})$ (w.r.t. the 0-1 loss) satisfies:*

$$Var(\hat{f}^u_{\leftarrow}) = g(R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\leftarrow})) \leq \overbrace{g\left(L^u_{\leftarrow 0} \cdot (\overline{\ell} - \underline{\ell}) \cdot \sqrt{\frac{2\log(1/\delta)}{\eta^u_K N}}\right)}^{\text{Variance Proxy}}. \tag{7}$$

The variance proxy of $Var(\hat{f}^{\circ}_{\leftarrow})$ in Eqn. (7) is smaller than that of $Var(\hat{f}^{\bullet}_{\leftarrow})$ if $\sqrt{\eta^{\circ}_K} > \frac{L^{\circ}_{\leftarrow}}{L^{\bullet}_{\leftarrow}}$.

Moving a bit further, when the noise transition matrix is symmetric for both methods, the requirement $\sqrt{\eta^u_K} > \frac{L^{\circ}_{\leftarrow}}{L^{\bullet}_{\leftarrow}}$ could be further simplified as: $\sqrt{\eta^u_K} > \frac{L^{\circ}_{\leftarrow}}{L^{\bullet}_{\leftarrow}} = \frac{1-\rho^{\bullet}_0-\rho^{\bullet}_1}{1-\rho^{\circ}_0-\rho^{\circ}_1}$. For a fixed K, a more efficient aggregation method decreases $\rho^{\bullet}_i$, which makes it harder to satisfy this condition.

Recall $L^u_{\leftarrow} := L^u_{\leftarrow 0} \cdot L$, the theoretical insights of $\ell_{\leftarrow}$ between binary case and the multi-class setting could be bridged by replacing L^u_0 with the multi-class constant specified in the following corollary.

Corollary 4.4 (Multi-Class Extension ($\ell_{\leftarrow}$-Loss))**.** *Given a diagonal-dominant transition matrix T^u, we have*

$$L^u_{\leftarrow 0} = \frac{2\sqrt{M}}{\lambda_{\min}(T^u)},$$

where $\lambda_{\min}(T^u)$ denotes the minimal eigenvalue of the matrix T^u. Particularly, if $T^u_{ii} < 0.5, \forall i \in [M]$, we further have

$$L^u_{\leftarrow 0} = \min\left\{\frac{1}{1-2e^u}, \frac{2\sqrt{M}}{\lambda_{\min}(T^u)}\right\}, \quad \textit{where } e^u := \max_{i\in[M]}(1 - T^u_{ii}).$$

4.2 Peer Loss Functions

Peer Loss function [26] is a family of loss functions that are shown to be robust to label noise, without requiring the knowledge of noise rates. Formally, $\ell_{\looparrowright}(f(x_i), \tilde{y}_i) := \ell(f(x_i), \tilde{y}_i) - \ell(f(x_i^1), \tilde{y}_i^2)$, where the second term checks on mismatched data samples with $(x_i, \tilde{y}_i)$, $(x_i^1, \tilde{y}_i^1)$, $(x_i^2, \tilde{y}_i^2)$, which are randomly drawn from the same data distribution. When combined with the peer loss approach, i.e., $\ell \to \ell_{\looparrowright}$, the two risks become: $\hat{R}_{\ell_{\looparrowright}, \widetilde{D}^u}(f) = \frac{1}{N}\sum_{i=1}^N \ell_{\looparrowright}(f(x_i), \tilde{y}_i^u), u \in \{\circ, \bullet\}$.

Bias of given classifier w.r.t. $\ell_{\looparrowright}$ Suppose the loss function $\ell(f(x), y)$ is L-Lipschitz for any feasible y. Let $L^u_{\looparrowright 0} := 1/(1-\rho_0^u-\rho_1^u), L^u_{\looparrowright} := L^u_{\looparrowright 0} \cdot L$ and $\hat{f}^u_{\looparrowright} = \operatorname{argmin}_{f\in\mathcal{F}} \hat{R}_{\ell_{\looparrowright}, \widetilde{D}^u}(f)$.

Lemma 4.5. *With probability at least $1-\delta$, we have:*

$$R_{\ell,\mathcal{D}}(\hat{f}^u_{\looparrowright}) - R_{\ell,\mathcal{D}}(f^*) \le \overline{\Delta}^u_{R_{\looparrowright}} := 8L^u_{\looparrowright} \cdot \mathfrak{R}(\mathcal{F}) + L^u_{\looparrowright 0} \cdot \sqrt{\frac{2\log(4/\delta)}{\eta_K^u N}} \cdot \left(1 + 2(\bar{\ell} - \underline{\ell})\right).$$

To evaluate the performance of a given classifier yielded by the optimization w.r.t. $\ell_{\looparrowright}$, Lemma 4.5 provides the bias proxy $\overline{\Delta}^u_{R_{\looparrowright}}$ for both treatments. Similarly, we adopt such a proxy to analyze which treatment is more preferable.

Theorem 4.6. *Denote by $\alpha_K := 1 - L^{\bullet}_{\looparrowright}/L^{\circ}_{\looparrowright}$, $\gamma = \frac{1+2(\bar{\ell}-\underline{\ell})}{2L}\sqrt{\frac{\log(4/\delta)}{4d\log(N)}}$, where d denotes the VC-dimension of $\mathcal{F}$. For peer loss, the separation bias proxy $\overline{\Delta}^{\circ}_{R_{\looparrowright}}$ is smaller than the aggregation bias proxy $\overline{\Delta}^{\bullet}_{R_{\looparrowright}}$ if and only if*

$$\alpha_K \cdot \frac{1}{L^{\bullet}_{\looparrowright}/L^{\circ}_{\looparrowright} - (\eta_K^{\circ})^{-\frac{1}{2}}} \le \gamma. \tag{8}$$

Note that the condition in Eqn. (8) shares a similar pattern to that which appeared in the basic loss ℓ and $\ell_{\leftarrow}$. We will empirically illustrate the monotonicity of its LHS in Section 4.3.

Variance of given classifiers with Peer Loss We now move on to check how separation and aggregation methods result in different variances when training with peer loss. Similarly, we can obtain:

Theorem 4.7. *When $\sqrt{\eta_K^u} \ge \sqrt{\frac{2\log(4/\delta)}{N}} \cdot \left(1 + 2(\bar{\ell} - \underline{\ell})\right)$, Var$(\hat{f}^u_{\looparrowright})$ (w.r.t. the 0-1 loss) satisfies:*

$$Var(\hat{f}^u_{\looparrowright}) = g(R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\looparrowright})) \le \overbrace{g\left(L^u_{\looparrowright 0} \cdot \sqrt{\frac{\log(4/\delta)}{2\eta_K^u N}} \cdot \left(1 + 2(\bar{\ell} - \underline{\ell})\right)\right)}^{\text{Variance proxy}}. \tag{9}$$

The variance proxy of Var$(\hat{f}^{\circ}_{\looparrowright})$ in Eqn. (9) is smaller than that of Var$(\hat{f}^{\bullet}_{\looparrowright})$ if $\sqrt{\eta_K^{\circ}} \ge \frac{L^{\circ}_{\looparrowright}}{L^{\bullet}_{\looparrowright}}$.

Theoretical insights of $\ell_{\looparrowright}$ also have the multi-class extensions, we only need to generate $L^u_{\looparrowright 0}$ to the multi-class setting along with additional conditions specified as below:

Corollary 4.8 (Multi-Class Extension ($\ell_{\looparrowright}$-Loss))**.** *Assume $\ell_{\looparrowright}$ is classification-calibrated in the multi-class setting, and the clean label Y has equal prior $P(Y=j) = \frac{1}{M}, \forall j \in [M]$. For the uniform noise transition matrix [56] such that $T^u_{j,i} = \rho_i^u, \forall j \in [M]$, we have: $L^u_{\looparrowright 0} = 1/(1 - \sum_{i\in[M]} \rho_i^u)$.*

4.3 Analysis of the Theoretical Conditions

Recall that the established conditions in Theorems 3.4, 4.2, 4.6 are implicitly relevant to the number of labelers K, and the RHS of Eqns. (3, 6, 8) are constants. We proceed to analyze the monotonicity of the corresponding LHS (in the form of $\alpha_K \cdot \frac{1}{\beta_K - (\eta_K^{\circ})^{-\frac{1}{2}}}$) w.r.t. the increase of K, where $\beta_K = 1$ for ℓ and $\ell_{\leftarrow}$, $\beta_K = L^{\bullet}_{\looparrowright}/L^{\circ}_{\looparrowright}$ for $\ell_{\looparrowright}$. Thus, we have: $O(\text{LHS}) = O(\alpha_K \cdot (\beta_K - O(\frac{\log(K)}{\sqrt{K}}))^{-1})$. We visualize this order under different symmetric T° in Figure 3. It can be observed that, when K is small (e.g., $K \le 5$), the LHS parts of these conditions increase with K, while they may decrease with K if K is sufficiently large. Recall that separation is better if LHS is less than the constant value γ. Therefore, Figure 3 shows the trends that aggregation is generally better than separation when K is sufficiently large.

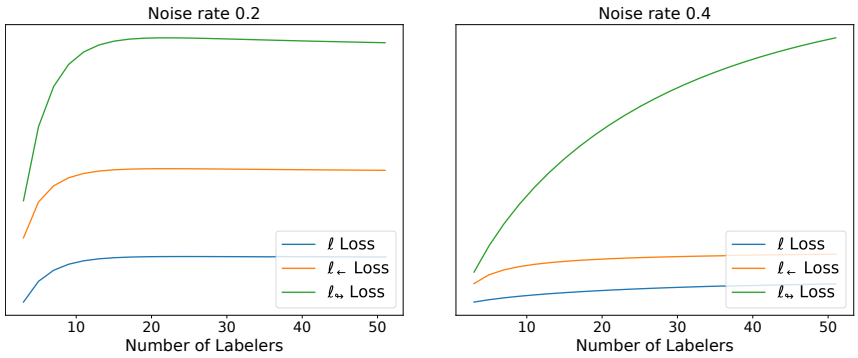

Figure 3: The monotonicity of the LHS in Eqn. (3, 6, 8) w.r.t. the increase of K.

Tightness of the bias proxies In Theorems 3.4, 4.2, 4.6, we view the error bounds $\overline{\Delta}^u_R$, $\overline{\Delta}^u_{R\leftarrow}$, $\overline{\Delta}^u_{R\looparrowright}$ as proxies of the worst-case performance of the trained classifier. For the standard loss function ℓ, it has been proven that [33, 20] under mild conditions of ℓ and $\mathcal{F}$, the lower bound of the performance gap between a trained classifier ($\hat{f}$) and the optimal achievable one (i.e., f^*) $R_{\ell,\mathcal{D}}(\hat{f}) - R_{\ell,\mathcal{D}}(f^*)$ is of the order $O(\sqrt{1/N})$, which is of the same order as that in Theorem 3.4. Noting the behavior concluded from the worst-case bounds may not always hold for each individual case, we further use experiments to validate our analyses in the next section.

5 Experimental Results

In this section, we empirically compare the performance of different treatments on the multiple noisy labels when learning with robust loss functions (CE loss, forward loss correction, and peer loss). We consider several treatments including label aggregation methods (majority vote and EM inference) and the label separation method. Assuming that multiple noisy labels have different weights, EM inference can be used to solve the problem under this assumption by treating the aggregated labels as hidden variables [8, 47, 40, 39]. In the E-step, the probabilities of the aggregated labels are estimated using the weighted aggregation approach based on the fixed weights of multiple noisy labels. In the M-step, EM inference method re-estimates the weights of multiple noisy labels based on the current aggregated labels. This iteration continues until all aggregated labels remain unchanged. As for label separation, we adopted the mini-batch separation method, i.e., each training sample x_n is assigned with K noisy labels in each batch.

5.1 Experiment on Synthetic Noisy Datasets

Experimental results on synthetic noisy UCI datasets [9] We adopt six UCI datasets to empirically compare the performances of label separation and aggregation methods, when learning with CE loss, backward correction [35, 36], and Peer Loss [26]. The noisy annotations given by multiple annotators are simulated by *symmetric label noise*, which assumes $T_{i,j} = \frac{\epsilon}{M-1}$ for $j \neq i$ for each annotator, where ϵ quantifies the overall noise rate of the generated noisy labels. In Figure 4, we adopt two UCI datasets (StatLog: ($M = 6$); Optical: ($M = 10$)) for illustration. From the results in Figure 4, it is quite clear that: the *label separation method outperforms both aggregation methods (majority-vote and EM inference) consistently, and is considered to be more beneficial on such small scale datasets*. Results on additional datasets and more details are deferred to the Appendix.

Experimental results on synthetic noisy CIFAR-10 dataset [18] On CIFAR-10 dataset, we consider two types of simulation for the separate noisy labels: *symmetric label noise* model and *instance-dependent label noise* [6, 67], where ϵ is the average noise rate and different labelers follow different instance-dependent noise transition matrices. For a fair comparison, we adopt the ResNet-34 model [15], the same training procedure and batch-size for all considered treatments on the separate noisy labels.

Figure 5 shares the following insights regarding the preference of the treatments: in the low noise regime or when K is large, aggregating separate noisy labels significantly reduces the noise rates and aggregation methods tend out to have a better performance; while in the high noise regime or when K is small, the performances of separation methods tend out to be more promising. With the increasing of K or ϵ, we can observe a preference transition from label separation to label aggregation methods.

Table 1: The performances of CE/BW/PeerLoss trained on 2 UCI datasets (Breast, and German datasets), with aggregated labels (majority vote, EM inference), and separated labels. We highlight the results with Green (for separation method) and Red (for aggregation methods) if the performance gap is larger than 0.05. (K is the number of labels per training image)

UCI-Breast (symmetric) CE

$\epsilon = 0.2$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	96.05	96.05	96.49	96.93	97.37	97.37
EM	96.93	96.05	96.49	96.93	97.37	97.37
Sep	96.49	95.18	96.49	96.93	97.81	98.25
$\epsilon = 0.4$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	96.05	96.49	95.18	95.18	96.49	96.93
EM	96.05	92.98	89.47	94.30	96.05	96.93
Sep	92.11	94.30	95.61	96.49	96.93	96.93

UCI-German (symmetric) CE

$\epsilon = 0.2$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	69.00	71.50	71.50	73.50	73.00	73.00
EM	58.75	63.50	65.75	66.50	65.50	65.50
Sep	70.00	70.75	66.00	69.75	70.75	69.25
$\epsilon = 0.4$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	65.75	62.25	62.75	68.50	71.75	70.50
EM	61.00	60.00	61.50	54.00	62.00	63.25
Sep	68.25	65.50	65.00	64.50	64.75	69.50

UCI-Breast (symmetric) BW

$\epsilon = 0.2$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	95.61	96.49	96.05	96.93	96.93	96.93
EM	95.61	96.49	96.05	96.93	96.93	96.93
Sep	95.18	93.42	96.49	96.05	97.37	98.25
$\epsilon = 0.4$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	89.91	96.05	94.74	94.30	96.05	96.49
EM	81.14	94.30	92.11	94.74	92.54	96.49
Sep	91.67	93.42	94.30	89.47	92.54	97.37

UCI-German (symmetric) BW

$\epsilon = 0.2$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	72.75	71.50	74.00	75.50	76.50	76.50
EM	62.75	61.50	59.25	64.50	62.50	62.50
Sep	70.50	70.50	73.75	68.25	70.00	72.75
$\epsilon = 0.4$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	65.25	69.50	67.50	69.50	70.50	71.75
EM	57.75	60.25	55.25	53.50	54.00	62.25
Sep	60.25	63.50	63.00	64.25	69.00	64.75

UCI-Breast (symmetric) PeerLoss

$\epsilon = 0.2$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	96.05	96.49	96.49	96.93	96.93	96.93
EM	96.05	96.49	96.49	96.93	96.93	96.93
Sep	94.74	94.30	96.93	96.93	96.93	97.81
$\epsilon = 0.4$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	92.11	95.61	95.18	92.54	96.49	96.05
EM	92.11	92.11	86.40	93.86	95.61	96.93
Sep	92.11	94.30	95.18	95.18	95.61	96.05

UCI-German (symmetric) PeerLoss

$\epsilon = 0.2$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	72.75	71.75	73.00	73.00	72.50	72.50
EM	62.25	64.50	63.75	64.25	62.75	62.75
Sep	70.25	68.00	70.50	70.00	67.00	73.50
$\epsilon = 0.4$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	69.50	66.25	69.50	68.75	69.00	70.00
EM	62.50	61.25	64.25	57.75	59.75	65.00
Sep	64.00	61.25	66.50	68.00	69.25	69.00

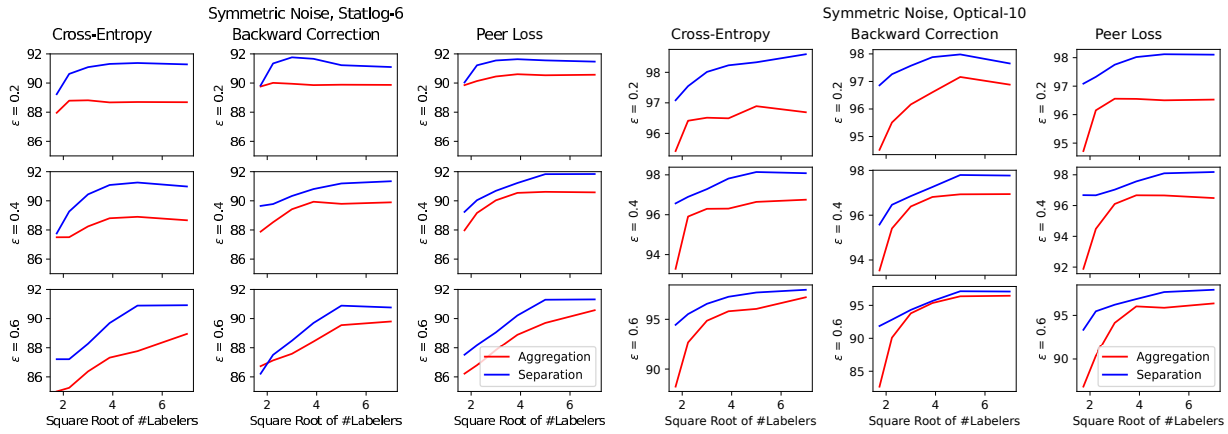

Figure 4: The performances of Cross-Entropy, Backward Loss Correction, and Peer Loss trained on synthetic noisy Statlog-6/Optical-10 aggregated labels (we report the better results between majority vote and EM inference for each K, and noise rate ϵ), and separated labels. X-axis: the value of the number of labelers $\sqrt{K}$; Y axis: the best test accuracy achieved.

5.2 Empirical Verification of the Theoretical Bounds

To verify the comparisons of bias proxies (i.e., Theorem 3.4) through an empirical perspective, we adopt two binary classification UCI datasets for demonstration: Breast and German datasets, as shown in Table 1. Clearly, on these two binary classification tasks, label aggregation methods tend to outperform label separation, and we attribute this phenomenon to the fact that "denoising effect of label aggregation is more significant in the binary case".

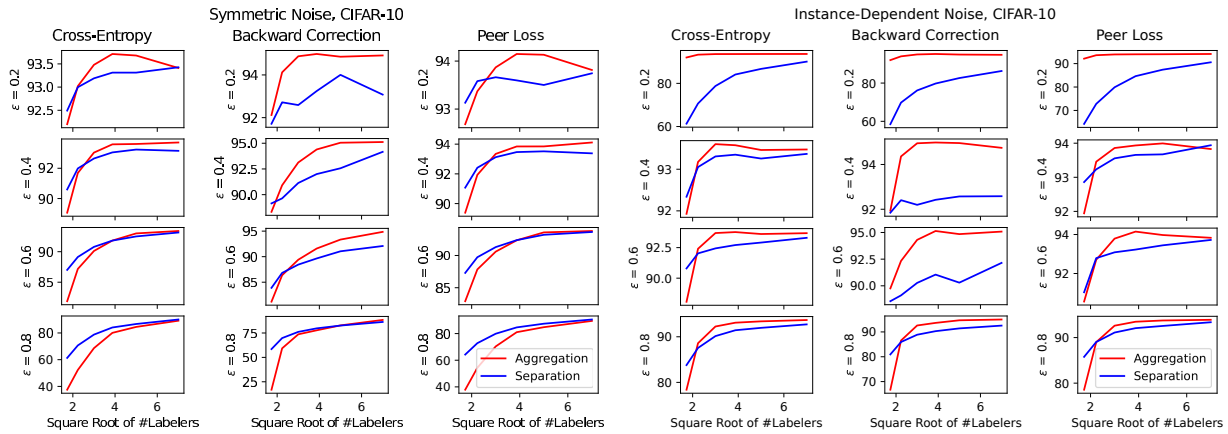

Figure 5: The performances of Cross-Entropy, Backward Loss Correction, and Peer Loss trained on synthetic noisy CIFAR-10 aggregated labels (we report the better results between majority vote, EM inference for each K, and noise rate ϵ), and separated labels. X-axis: the value of $\sqrt{K}$ where K denotes the number of labels per training example; Y axis: the best achieved test accuracy.

Table 2: Empirical verification of Theorem 3.4 on Breast & German UCI datasets.

Dataset	ρ_i°	p_0	N	$(1-\delta, S_K)$
Breast	0.2	0.3726	569	$(0.62, \{K > 49\})$
Breast	0.4	0.3726	569	$(0.62, \{K > 49\})$
German	0.2	0.3	1000	$(0.98, \{K > 15\})$
German	0.4	0.3	1000	$(0.98, \{K > 23\})$

For Theorem 3.4 (CE loss), the condition requires $\alpha_K / \left(1 - (\eta_K^\circ)^{-\frac{1}{2}}\right)$, where $\alpha = (\rho_0^\circ p_0 + \rho_1^\circ p_1) - (\rho_0^\bullet p_0 + \rho_1^\bullet p_1)$, $\gamma = \sqrt{\log(1/\delta)/2N}$. For two binary UCI datasets (Breast & German), the information could be summarized in Table 2, where the column $(1-\delta, S_K)$ means: when the number of annotators belongs to the set S_K, the label separation method is likely to under-perform label aggregation (i.e., majority vote) with probability at least $1-\delta$. For example, in the last row of Table 2, when training on UCI German dataset with CE loss under noise rate 0.4 (the noise rate of separate noisy labels), Theorem 3.4 reveals that with probability at least 0.98, label aggregation (with majority vote) is better than label separation when $K > 23$, which aligns well with our empirical observations (label separation is better only when $K < 15$).

5.3 Experiments on realistic noisy datasets

Note that in real-world scenarios, the label-noise pattern may differ due to the expertise of each human annotator. We further compare the different treatments on two realistic noisy datasets: CIFAR-10N [57], and CIFAR-10H [38]. CIFAR-10N provides each CIFAR-10 train image with 3 independent human annotations, while CIFAR-10H gives ≈ 50 annotations for each CIFAR-10 test image.

In Table 3, we repeat the reproduce of three robust loss functions with three different treatments on the separate noisy labels. We report the best achieved test accuracy for Cross-Entropy/Backward Correction/Peer Loss methods when learning with label aggregation methods (majority-vote and EM inference) and the separation method (soft-label). We observe that the separation method tends to have a better performance than aggregation ones. This may be attributed to the relative high noise rate ($\epsilon \approx 0.18$) in CIFAR-N and the insufficient amount of labelers ($K = 3$). Note that since the noise level in CIFAR-10H is low ($\epsilon \approx 0.07$ wrong labels), label aggregation methods can infer higher quality labels, and thus, result in a better performance than separation methods (Red colored cells in Table 3 and 4).

5.4 Hypothesis Testing

We adopt the paired t-test to show which treatment on the separate noisy labels is better, under certain conditions. In Table 5, we report the statistic and p-value given by the hypothesis testing results. The column "Methods" indicate the two methods we want to compare (A & B). Positive statistics means that A is better than B in the metric of test

Table 3: Experimental results on CIFAR-10N and CIFAR-10H dataset with $K = 3$. We highlight the results with Green (for separation method) and Red (for aggregation methods) if the performance gap is large than 0.05.

CIFAR-10N ($\epsilon \approx 0.18$)	**CE**	**BW**	**PL**
Majority-Vote	89.52	89.23	89.84
EM-Inference	89.19	88.88	88.92
Separation	89.77	89.20	89.97
CIFAR-10H ($\epsilon \approx 0.09$)	**CE**	**BW**	**PL**
Majority-Vote	80.86	82.72	82.11
EM-Inference	80.81	82.43	81.73
Separation	76.75	79.07	78.08

Table 4: Experimental results on CIFAR10-H with $K \geq 5$. We highlight the results with Green (for separation method) and Red (for aggregation methods) if the performance gap is large than 0.05.

CE	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
Majority-Vote	80.69	80.73	81.37	81.79	81.66
EM-Inference	80.97	80.96	81.24	81.01	81.68
Separation	79.65	80.91	81.07	80.78	80.81
BW	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
Majority-Vote	82.51	82.75	83.27	83.59	83.68
EM-Inference	82.30	82.68	82.74	82.89	83.08
Separation	82.14	82.48	81.92	81.72	81.69
PL	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
Majority-Vote	81.84	81.85	82.39	82.98	82.83
EM-Inference	81.89	82.30	82.53	82.86	82.73
Separation	80.25	81.89	81.00	80.71	80.89

accuracy. Given a specific setting, denote by $\text{Acc}_{\text{method}}$ as the list of test accuracy that belongs to this setting (i.e., CIFAR-10N, $K = 3$), including CE, BW, PL loss functions, the basic hypothesis could be summarized as below:

- **Null hypothesis**: there exists zero mean difference between (1) Acc_{MV} and Acc_{EM}; or (2) Acc_{MV} and Acc_{Sep}; or (3) Acc_{EM} and Acc_{Sep};
- **Alternative hypothesis**: there exists non-zero mean difference between (1) Acc_{MV} and Acc_{EM}; or (2) Acc_{MV} and Acc_{Sep}; or (3) Acc_{EM} and Acc_{Sep}.

To clarify, the three cases in the above hypothesis are tested independently. For test accuracy comparisons of CIFAR-10N in Table 3, the setting of hypothesis test is $K = 3$ and the label noise rate is relatively high (18%). All p-values are larger than 0.05, indicating that we should reject the null hypothesis, and we can conclude that the performance of these three methods on CIFAR-10N (high noise, small K) satisfies: EM<MV<Sep.

For CIFAR-10H in Table 3 and 4, all the label noise rate is relatively low. We consider two scenarios ($K < 15$: the number of annotators is small; $K \geq 15$: the number of annotators is large). p-values among MV and EM are always large, which mean that the denoising effect of the advanced label aggregation method (EM) is negligible under CIFAR-10H dataset. However, p-values of remaining settings are larger than 0.05, indicating that we should reject the null hypothesis, and we can conclude that the performance of these 3 methods on CIFAR-10H (low noise, small/large K) satisfies: EM/MV > Sep.

Table 5: Hypothesis testing results of the comparisons between label aggregation methods and the label separation method on realistic noisy datasets.

Setting	Methods	Statistic	p-value
CIFAR-10N ($K = 3$, high noise)	MV & EM	2.650	0.057
CIFAR-10N ($K = 3$, high noise)	MV & Sep	-0.401	0.708
CIFAR-10N ($K = 3$, high noise)	EM & Sep	-2.596	0.060
CIFAR-10H ($K < 15$, low noise)	MV & EM	-0.003	0.998
CIFAR-10H ($K < 15$, low noise)	MV & Sep	2.336	0.033
CIFAR-10H ($K < 15$, low noise)	EM & Sep	2.390	0.030
CIFAR-10H ($K \geq 15$, low noise)	MV & EM	0.805	0.433
CIFAR-10H ($K \geq 15$, low noise)	MV & Sep	4.426	0.000
CIFAR-10H ($K \geq 15$, low noise)	EM & Sep	3.727	0.002

6 Conclusions

When learning with separate noisy labels, we explore the answer to the question "whether one should aggregate separate noisy labels into single ones or use them separately as given". In the empirical risk minimization framework, we theoretically show that label separation could be more beneficial than label aggregation when the noise rates are high or the number of labelers is insufficient. These insights hold for a number of popular loss function including several robust treatments. Empirical results on synthetic and real-world datasets validate our conclusion.

References

[1] Shadi Albarqouni, Christoph Baur, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging*, 35(5):1313–1321, 2016.

[2] Ehsan Amid, Manfred K Warmuth, Rohan Anil, and Tomer Koren. Robust bi-tempered logistic loss based on Bregman divergences. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] Noga Bar, Tomer Koren, and Raja Giryes. Multiplicative reweighting for robust neural network optimization. *arXiv preprint arXiv:2102.12192*, 2021.

[4] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. *Advances in Neural Information Processing Systems*, 30, 2017.

[5] Zhijun Chen, Huimin Wang, Hailong Sun, Pengpeng Chen, Tao Han, Xudong Liu, and Jie Yang. Structured probabilistic end-to-end learning from crowds. In *IJCAI*, pages 1512–1518, 2020.

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

[7] Hao Cheng, Zhaowei Zhu, Xing Sun, and Yang Liu. Demystifying how self-supervised features improve training from noisy labels. *arXiv preprint arXiv:2110.09022*, 2021.

[8] Alexander Philip Dawid and Allan M Skene. 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.

[9] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[10] Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200, 2012.

[11] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

[12] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[13] Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W Tsang, James T Kwok, and Masashi Sugiyama. A survey of label-noise representation learning: Past, present and future. *arXiv preprint arXiv:2011.04406*, 2020.

[14] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, pages 8527–8537, 2018.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Jeff Howe et al. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[17] Panagiotis G Ipeirotis, Foster Provost, Victor S Sheng, and Jing Wang. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, 28(2):402–441, 2014.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[19] Abhishek Kumar and Ehsan Amid. Constrained instance and class reweighting for robust learning under label noise. *arXiv preprint arXiv:2111.05428*, 2021.

[20] Guillaume Lecué and Shahar Mendelson. Sharper lower bounds on the performance of the empirical risk minimization algorithm. *Bernoulli*, pages 605–613, 2010.

[21] Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive early-learning correction for segmentation from noisy annotations. *arXiv preprint arXiv:2110.03740*, 2021.

[22] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

[23] Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. *arXiv preprint arXiv:2202.14026*, 2022.

[24] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.

[25] Yang Liu. Understanding instance-level label noise: Disparate impacts and treatments. In *International Conference on Machine Learning*, pages 6725–6735. PMLR, 2021.

[26] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, pages 6226–6236. PMLR, 2020.

[27] Yang Liu and Mingyan Liu. An online learning approach to improving the quality of crowd-sourcing. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):217–230, 2015.

[28] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020.

[29] Tianyi Luo, Xingyu Li, Hainan Wang, and Yang Liu. Research replication prediction using weakly supervised learning. In *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020.

[30] Tianyi Luo and Yang Liu. Machine truth serum. *arXiv preprint arXiv:1909.13004*, 2019.

[31] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning*, pages 6543–6553. PMLR, 2020.

[32] Negin Majidi, Ehsan Amid, Hossein Talebi, and Manfred K. Warmuth. Exponentiated gradient reweighting for robust training under label noise and beyond. *arXiv preprint arXiv:2104.01493*, 2021.

[33] Shahar Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Transactions on Information Theory*, 54(8):3797–3803, 2008.

[34] Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 258–267, 2015.

[35] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

[36] 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.

[37] Gordon Pennycook and David G Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526, 2019.

[38] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.

[39] Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. An evaluation of aggregation techniques in crowdsourcing. In *International Conference on Web Information Systems Engineering*, pages 1–15. Springer, 2013.

[40] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of machine learning research*, 11(4), 2010.

[41] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[42] 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.

[43] Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[44] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Gaussian process classification and active learning with multiple annotators. In *International conference on machine learning*, pages 433–441. PMLR, 2014.

[45] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

[46] Victor S Sheng, Jing Zhang, Bin Gu, and Xindong Wu. Majority voting and pairing with multiple noisy labeling. *IEEE Transactions on Knowledge and Data Engineering*, 31(7):1355–1368, 2017.

[47] Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. Inferring ground truth from subjective labelling of venus images. *Advances in neural information processing systems*, 7, 1994.

[48] 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*, 2022.

[49] James M Varah. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its applications*, 11(1):3–5, 1975.

[50] Jingkang Wang, Hongyi Guo, Zhaowei Zhu, and Yang Liu. Policy learning using weak supervision. *Advances in Neural Information Processing Systems*, 34, 2021.

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

[52] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13726–13735, 2020.

[53] Hongxin Wei, Lue Tao, Renchunzi Xie, and Bo An. Open-set label noise can improve robustness against inherent label noise. *Advances in Neural Information Processing Systems*, 34, 2021.

[54] Hongxin Wei, Renchunzi Xie, Lei Feng, Bo Han, and Bo An. Deep learning from multiple noisy annotators as a union. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[55] Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, and Yang Liu. Understanding generalized label smoothing when learning with noisy labels. *arXiv preprint arXiv:2106.04149*, 2021.

[56] Jiaheng Wei and Yang Liu. When optimizing f-divergence is robust with label noise. *arXiv preprint arXiv:2011.03687*, 2020.

[57] 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. *arXiv preprint arXiv:2110.12088*, 2021.

[58] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22, 2009.

[59] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2020.

[60] 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. *Advances in Neural Information Processing Systems*, 33:7597–7610, 2020.

[61] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in Neural Information Processing Systems*, 32, 2019.

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

[63] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[64] Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. *arXiv preprint arXiv:2110.06283*, 2022.

[65] Zhaowei Zhu, Tongliang Liu, and Yang Liu. A second-order approach to learning with instance-dependent label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10113–10123, 2021.

[66] Zhaowei Zhu, Tianyi Luo, and Yang Liu. The rich get richer: Disparate impact of semi-supervised learning. *arXiv preprint arXiv:2110.06282*, 2021.

[67] Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. In *International Conference on Machine Learning*, pages 12912–12923. PMLR, 2021.

[68] Zhaowei Zhu, Jialu Wang, and Yang Liu. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. *arXiv preprint arXiv:2202.01273*, 2022.

A Proof Sketch of Core Theorems

We briefly introduce the proof sketch of Lemma 4.1 because it sets up the foundation for the analyses on Backward Loss Correction and it covers the proofs of the standard ℓ loss in Section 3 as a special case.

A.1 Proof of Lemma 4.1

Proof. Our proof can be divided into four steps as follows.

Step 1: Apply Hoeffding's inequality for each group. We divide the noisy train samples $\{(x_n, \tilde{y}^\circ_{n,k})\}_{n\in[N]}$ into K groups, for $k \in [K]$, i.e., $\{(x_n, \tilde{y}^\circ_{n,1})\}_{n\in[N]}$, $\cdots$, $\{(x_n, \tilde{y}^\circ_{n,K})\}_{n\in[N]}$. Note within each group, e.g., group $\{(x_n, \tilde{y}^\circ_{n,1})\}_{n\in[N]}$, all the N training samples are i.i.d. Additionally, training samples between any two different groups are also i.i.d. given feature set $\{x_n\}_{n\in[N]}$. Thus, with one group $\{(x_n, \tilde{y}_{n,1})\}_{n\in[N]}$, w.p. $1-\delta_0$, we have

$$\left|\hat{R}_{\mathbf{1}^\circ_\leftarrow|\text{Group-1}}(f) - R_{\mathbf{1}^\circ_\leftarrow}(f)\right| \leq \left(\overline{\mathbf{1}^\circ_\leftarrow} - \underline{\mathbf{1}^\circ_\leftarrow}\right) \cdot \sqrt{\frac{\log(1/\delta_0)}{2N}}, \forall f.$$

where we have $\overline{\mathbf{1}^\circ_\leftarrow} - \underline{\mathbf{1}^\circ_\leftarrow} := L^\circ_{\leftarrow 0} = \frac{(1+|\rho^\circ_0 - \rho^\circ_1|)}{1-\rho^\circ_0-\rho^\circ_1}$.

Step 2: Adopt the union bound for all groups. Applying the above technique on the other groups and by the union bound, we know that w.p. at least $1-K\delta_0$, $\forall k \in [K]$, each $\hat{R}_{\mathbf{1}^\circ_\leftarrow|\text{Group-k}}(f), k \in [K]$ can be seen as a random variable within range:

$$\left[R_{\mathbf{1}^\circ_\leftarrow}(f) - L^\circ_{\leftarrow 0} \cdot \sqrt{\frac{\log(1/\delta_0)}{2N}}, R_{\mathbf{1}^\circ_\leftarrow}(f) + L^\circ_{\leftarrow 0} \cdot \sqrt{\frac{\log(1/\delta_0)}{2N}}\right].$$

The randomness is from noisy labels $\tilde{y}_{n,k}$.

Step 3: Hoeffding inequality for $\hat{R}_{\mathbf{1}^\circ_\leftarrow|\text{Group-k}}(f), k \in [K]$ These K random variables are i.i.d. when the feature set is fixed. By Hoeffding's inequality, w.p. at least $1-K\delta_0-\delta_1$, $\forall f$, we have

$$\left|\hat{R}_{\mathbf{1}^\circ_\leftarrow}(f) - R_{\mathbf{1}^\circ_\leftarrow}(f)\right| \leq L^\circ_{\leftarrow 0} \cdot \sqrt{\frac{\log(1/\delta_1)\log(1/\delta_0)}{NK}}.$$

Step 4: Rademacher bound on the maximal deviation For $\delta_0 = \delta_1 = \frac{\delta}{K+1}$, with the Rademacher bound on the maximal deviation between risks and empirical ones, for $f^* \in \mathcal{F}$ and the separation method, with probability at least $1-\delta$, we have:

$$\max_{f\in\mathcal{F}} \left|\hat{R}_{\ell_\leftarrow,\widetilde{D}^\circ}(f) - R_{\ell_\leftarrow,\widetilde{\mathcal{D}}^\circ}(f)\right| \leq 2\mathfrak{R}^\circ(\ell_\leftarrow \circ \mathcal{F}) + L^\circ_{\leftarrow 0} \cdot (\overline{\ell} - \underline{\ell}) \cdot \log\left(\frac{K+1}{\delta}\right) \cdot \sqrt{\frac{1}{NK}},$$

$$\max_{f\in\mathcal{F}} \left|\hat{R}_{\ell_\leftarrow,\widetilde{D}^\bullet}(f) - R_{\ell_\leftarrow,\widetilde{\mathcal{D}}^\bullet}(f)\right| \leq 2\mathfrak{R}^\bullet(\ell_\leftarrow \circ \mathcal{F}) + L^\bullet_{\leftarrow 0} \cdot (\overline{\ell} - \underline{\ell}) \cdot \sqrt{\frac{\log(1/\delta)}{2N}},$$

where we define $\overline{\ell}, \underline{\ell}$ as the upper and lower bound of loss function ℓ respectively, and $\mathfrak{R}^u(\ell_\leftarrow \circ \mathcal{F})$ is the Rademacher complexity.

Step 5: Adopt the Lipshitz composition property of Rademacher averages. If ℓ is $L-$Lipshitz, then for separation and aggregation methods, $\ell_\leftarrow$ is $L^u_\leftarrow$ Lipshitz with $L^u_\leftarrow = \frac{(1+|\rho^u_0-\rho^u_1|)L}{1-\rho^u_0-\rho^u_1}$.

Step 6: Triangle inequality Bound with the triangle inequality:

$$\begin{aligned}
&R_{\ell,\mathcal{D}}(\hat{f}^u_\leftarrow) - R_{\ell,\mathcal{D}}(f^*) = \underline{R_{\ell_\leftarrow,\widetilde{\mathcal{D}}^u}(\hat{f}^u_\leftarrow) - R_{\ell_\leftarrow,\widetilde{\mathcal{D}}^u}(f^*)} \\
=&\underline{R_{\ell_\leftarrow,\widetilde{\mathcal{D}}^u}(\hat{f}^u_\leftarrow)} - \hat{R}_{\ell_\leftarrow,\widetilde{D}^u}(\hat{f}^u_\leftarrow) + \hat{R}_{\ell_\leftarrow,\widetilde{D}^u}(f^*) - \underline{R_{\ell_\leftarrow,\widetilde{\mathcal{D}}^u}(f^*)} + \hat{R}_{\ell_\leftarrow,\widetilde{D}^u}(\hat{f}^u_\leftarrow) - \hat{R}_{\ell_\leftarrow,\widetilde{D}^u}(f^*) \\
\leq& 0 + 2\max_{f\in\mathcal{F}} |\hat{R}_{\ell_\leftarrow,\widetilde{D}^u}(f) - R_{\ell_\leftarrow,\widetilde{\mathcal{D}}^u}(f)|.
\end{aligned}$$

Conclusions could be derived then. ∎

B Full Proofs

In this section, we briefly introduce all omitted proofs in the main paper.

We firstly give the proof of Lemma 4.1 because it is beneficial for the proofs in Section 3.

B.1 Proof of Lemma 4.1

Proof. To apply Hoeffding's inequality on the dataset of the separation method, we divide the noisy train samples $\{(x_n, \tilde{y}^\circ_{n,k})\}_{n\in[N]}$ into K groups, for $k \in [K]$, i.e., $\{(x_n, \tilde{y}^\circ_{n,1})\}_{n\in[N]}, \cdots, \{(x_n, \tilde{y}^\circ_{n,K})\}_{n\in[N]}$. Note within each group, e.g., group $\{(x_n, \tilde{y}^\circ_{n,1})\}_{n\in[N]}$, all the N training samples are i.i.d. Additionally, training samples between any two different groups are also i.i.d. given feature set $\{x_n\}_{n\in[N]}$. Thus, with one group $\{(x_n, \tilde{y}_{n,1})\}_{n\in[N]}$, w.p. $1-\delta_0$, we have

$$\left|\hat{R}_{\mathbf{1}^\circ_{\leftarrow}|\text{Group-1}}(f) - R_{\mathbf{1}^\circ_{\leftarrow}}(f)\right| \leq \left(\overline{\mathbf{1}^\circ_{\leftarrow}} - \underline{\mathbf{1}^\circ_{\leftarrow}}\right) \cdot \sqrt{\frac{\log(1/\delta_0)}{2N}}, \forall f.$$

Note that:

$$(T^u)^{-1} = \frac{1}{1-\rho_0^u-\rho_1^u}\begin{pmatrix} 1-\rho_1^u & -\rho_0^u \\ -\rho_1^u & 1-\rho_0^u \end{pmatrix}, \quad \text{for } u \in \{\circ, \bullet\},$$

we have:

$$\overline{\mathbf{1}^\circ_{\leftarrow}} - \underline{\mathbf{1}^\circ_{\leftarrow}} := L^\circ_{\leftarrow 0} = \frac{(1+|\rho_0^\circ - \rho_1^\circ|)}{1-\rho_0^\circ-\rho_1^\circ}.$$

Applying the above technique on the other groups and by the union bound, we know that w.p. at least $1-K\delta_0$, $\forall k \in [K]$,

$$\hat{R}_{\mathbf{1}^\circ_{\leftarrow}|\text{Group-k}}(f) \in \left[R_{\mathbf{1}^\circ_{\leftarrow}}(f) - L^\circ_{\leftarrow 0} \cdot \sqrt{\frac{\log(1/\delta_0)}{2N}}, R_{\mathbf{1}^\circ_{\leftarrow}}(f) + L^\circ_{\leftarrow 0} \cdot \sqrt{\frac{\log(1/\delta_0)}{2N}}\right].$$

Each $\hat{R}_{\mathbf{1}^\circ_{\leftarrow}|\text{Group-k}}(f), k \in [K]$ can be seen as a random variable within range:

$$\left[R_{\mathbf{1}^\circ_{\leftarrow}}(f) - L^\circ_{\leftarrow 0} \cdot \sqrt{\frac{\log(1/\delta_0)}{2N}}, R_{\mathbf{1}^\circ_{\leftarrow}}(f) + L^\circ_{\leftarrow 0} \cdot \sqrt{\frac{\log(1/\delta_0)}{2N}}\right].$$

The randomness is from noisy labels $\tilde{y}_{n,k}$. Recall that the samples between different groups are i.i.d. given $\{x_n\}_{n\in[N]}$. Then the above K random variables are i.i.d. when the feature set is fixed. By Hoeffding's inequality, w.p. at least $1-K\delta_0-\delta_1$, $\forall f$, we have

$$\left|\hat{R}_{\mathbf{1}^\circ_{\leftarrow}}(f) - R_{\mathbf{1}^\circ_{\leftarrow}}(f)\right| \leq 2 \cdot L^\circ_{\leftarrow 0} \cdot \sqrt{\frac{\log(1/\delta_0)}{2N}} \cdot \sqrt{\frac{\log(1/\delta_1)}{2K}} = L^\circ_{\leftarrow 0} \cdot \sqrt{\frac{\log(1/\delta_1)\log(1/\delta_0)}{NK}}.$$

For $\delta_0 = \delta_1 = \frac{\delta}{K+1}$, with the Rademacher bound on the maximal deviation between risks and empirical ones, for $f^* \in \mathcal{F}$ and the separation method, with probability at least $1-\delta$, we have:

$$\begin{aligned}
\max_{f\in\mathcal{F}}\left|\hat{R}_{\ell_\leftarrow,\widetilde{D}^\circ}(f) - R_{\ell_\leftarrow,\widetilde{\mathcal{D}}^\circ}(f)\right| &\leq 2\mathfrak{R}^\circ(\ell_\leftarrow \circ \mathcal{F}) + L^\circ_{\leftarrow 0} \cdot (\overline{\ell} - \underline{\ell}) \cdot \log\left(\frac{K+1}{\delta}\right) \cdot \sqrt{\frac{1}{NK}}, \\
\max_{f\in\mathcal{F}}\left|\hat{R}_{\ell_\leftarrow,\widetilde{D}^\bullet}(f) - R_{\ell_\leftarrow,\widetilde{\mathcal{D}}^\bullet}(f)\right| &\leq 2\mathfrak{R}^\bullet(\ell_\leftarrow \circ \mathcal{F}) + \left(\overline{\ell^\bullet_{\leftarrow}} - \underline{\ell^\bullet_{\leftarrow}}\right) \cdot \sqrt{\frac{\log(1/\delta)}{2N}} \\
&= 2\mathfrak{R}^\bullet(\ell_\leftarrow \circ \mathcal{F}) + L^\bullet_{\leftarrow 0} \cdot (\overline{\ell} - \underline{\ell}) \cdot \sqrt{\frac{\log(1/\delta)}{2N}},
\end{aligned}$$

where we define $\overline{\ell}, \underline{\ell}$ as the upper and lower bound of loss function ℓ respectively, and:

$$\begin{aligned}
\mathfrak{R}^\circ(\ell_\leftarrow \circ \mathcal{F}) &:= \mathbb{E}_{x_i, \tilde{y}^\circ_{i,1}, \ldots, \tilde{y}^\circ_{i,K}, \epsilon_i}\left[\sup_{f\in\mathcal{F}} \frac{1}{NK}\sum_{i=1}^N\sum_{j=1}^K \epsilon_i \ell_\leftarrow(f(x_i), \tilde{y}^\circ_{i,j})\right] \\
&\leq \frac{1}{K}\sum_{j=1}^K \mathbb{E}_{x_i, \tilde{y}^\circ_{i,j}, \epsilon_i}\left[\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N \epsilon_i \ell_\leftarrow(f(x_i), \tilde{y}^\circ_{i,j})\right],
\end{aligned}$$

$$\mathfrak{R}^{\bullet}(\ell_{\leftarrow} \circ \mathcal{F}) := \mathbb{E}_{x_i, \tilde{y}^{\bullet}, \epsilon_i} \left[\sup_{f \in \mathcal{F}} \frac{1}{N} \epsilon_i \ell_{\leftarrow}(f(x_i), \tilde{y}^{\bullet}) \right].$$

Note that we assume the noisy labels given by the K labelers follow the same noise transition matrix, if ℓ is $L-$Lipshitz, then for separation and aggregation methods, $\ell_{\leftarrow}$ is $L^u_{\leftarrow}$ Lipshitz for $u \in \{\circ, \bullet\}$ respectively, where $L^u_{\leftarrow} = \frac{(1+|\rho^u_0 - \rho^u_1|)L}{1-\rho^u_0-\rho^u_1} \leq \frac{2L}{1-\rho^u_0-\rho^u_1}$. By the Lipshitz composition property of Rademacher averages, we have: $\mathfrak{R}^u(\ell_{\leftarrow} \circ \mathcal{F}) \leq L^u_{\leftarrow} \cdot \mathfrak{R}(\mathcal{F})$. Thus, we have:

$$\max_{f \in \mathcal{F}} |\hat{R}_{\ell_{\leftarrow}, \widetilde{D}^{\circ}}(f) - R_{\ell_{\leftarrow}, \widetilde{\mathcal{D}}^{\circ}}(f)| \leq 2L^{\circ}_{\leftarrow} \mathfrak{R}(\mathcal{F}) + \frac{(1 + |\rho^{\circ}_0 - \rho^{\circ}_1|) \cdot (\overline{\ell} - \underline{\ell})}{1 - \rho^{\circ}_0 - \rho^{\circ}_1} \cdot \log(\frac{K+1}{\delta}) \cdot \sqrt{\frac{1}{NK}}, \tag{10}$$

$$\max_{f \in \mathcal{F}} |\hat{R}_{\ell_{\leftarrow}, \widetilde{D}^{\bullet}}(f) - R_{\ell_{\leftarrow}, \widetilde{\mathcal{D}}^{\bullet}}(f)| \leq 2L^{\bullet}_{\leftarrow} \mathfrak{R}(\mathcal{F}) + \frac{(1 + |\rho^{\bullet}_0 - \rho^{\bullet}_1|) \cdot (\overline{\ell} - \underline{\ell})}{1 - \rho^{\bullet}_0 - \rho^{\bullet}_1} \cdot \sqrt{\frac{\log(1/\delta)}{2N}}. \tag{11}$$

Assume $f^* \leftarrow \min_{f \in \mathcal{F}} R_{\ell, \mathcal{D}}(f)$, for separation methods, we further have:

$$\begin{aligned}
&R_{\ell, \mathcal{D}}(\hat{f}^{\circ}_{\leftarrow}) - R_{\ell, \mathcal{D}}(f^*) = \underline{R_{\ell_{\leftarrow}, \widetilde{\mathcal{D}}^{\circ}}(\hat{f}^{\circ}_{\leftarrow}) - R_{\ell_{\leftarrow}, \widetilde{\mathcal{D}}^{\circ}}(f^*)} \\
&= \underline{R_{\ell_{\leftarrow}, \widetilde{\mathcal{D}}^{\circ}}(\hat{f}^{\circ}_{\leftarrow})} - \hat{R}_{\ell_{\leftarrow}, \widetilde{D}^{\circ}}(\hat{f}^{\circ}_{\leftarrow}) + \hat{R}_{\ell_{\leftarrow}, \widetilde{D}^{\circ}}(f^*) - \underline{R_{\ell_{\leftarrow}, \widetilde{\mathcal{D}}^{\circ}}(f^*)} + \hat{R}_{\ell_{\leftarrow}, \widetilde{D}^{\circ}}(\hat{f}^{\circ}_{\leftarrow}) - \hat{R}_{\ell_{\leftarrow}, \widetilde{D}^{\circ}}(f^*) \\
&\leq 0 + 2 \max_{f \in \mathcal{F}} |\hat{R}_{\ell_{\leftarrow}, \widetilde{D}^{\circ}}(f) - R_{\ell_{\leftarrow}, \widetilde{\mathcal{D}}^{\circ}}(f)| \\
&\leq 4L^{\circ}_{\leftarrow} \mathfrak{R}(\mathcal{F}) + 2L^{\circ}_{\leftarrow} \cdot (\overline{\ell} - \underline{\ell}) \cdot \log(\frac{K+1}{\delta}) \cdot \sqrt{\frac{1}{NK}}.
\end{aligned}$$

Similarly, for aggregation methods, we have:

$$\begin{aligned}
&R_{\ell, \mathcal{D}}(\hat{f}^{\bullet}_{\leftarrow}) - R_{\ell, \mathcal{D}}(f^*) = \underline{R_{\ell_{\leftarrow}, \widetilde{\mathcal{D}}^{\bullet}}(\hat{f}^{\bullet}_{\leftarrow}) - R_{\ell_{\leftarrow}, \widetilde{\mathcal{D}}^{\bullet}}(f^*)} \\
&= \underline{R_{\ell_{\leftarrow}, \widetilde{\mathcal{D}}^{\bullet}}(\hat{f}^{\bullet}_{\leftarrow})} - \hat{R}_{\ell_{\leftarrow}, \widetilde{D}^{\bullet}}(\hat{f}^{\bullet}_{\leftarrow}) + \hat{R}_{\ell_{\leftarrow}, \widetilde{D}^{\bullet}}(f^*) - \underline{R_{\ell_{\leftarrow}, \widetilde{\mathcal{D}}^{\bullet}}(f^*)} + \hat{R}_{\ell_{\leftarrow}, \widetilde{D}^{\bullet}}(\hat{f}^{\bullet}_{\leftarrow}) - \hat{R}_{\ell_{\leftarrow}, \widetilde{D}^{\bullet}}(f^*) \\
&\leq 0 + 2 \max_{f \in \mathcal{F}} |\hat{R}_{\ell_{\leftarrow}, \widetilde{D}^{\bullet}}(f) - R_{\ell_{\leftarrow}, \widetilde{\mathcal{D}}^{\bullet}}(f)| \\
&\leq 4L^{\bullet}_{\leftarrow} \mathfrak{R}(\mathcal{F}) + 2L^{\bullet}_{\leftarrow} \cdot (\overline{\ell} - \underline{\ell}) \cdot \sqrt{\frac{\log(1/\delta)}{2N}}.
\end{aligned}$$

Note that $\eta^{\circ}_K = \frac{K \cdot \log(\frac{1}{\delta})}{2\big(\log(\frac{K+1}{\delta})\big)^2}$ and $\eta^{\bullet}_K \equiv 1$, we then have:

$$R_{\ell, \mathcal{D}}(\hat{f}^{u}_{\leftarrow}) - R_{\ell, \mathcal{D}}(f^*) \leq \underbrace{4L^u_{\leftarrow} \mathfrak{R}(\mathcal{F}) + \frac{L^u_{\leftarrow} \cdot (\overline{\ell} - \underline{\ell})}{L} \cdot \sqrt{\frac{2\log(1/\delta)}{\eta^u_K N}}}_{\text{Defined as: } \overline{\Delta}^u_{R\leftarrow}}.$$

□

B.2 Proof of Theorem 4.2

Proof. The proof is straightforward if we proceed the proof of Lemma 4.1 with below discussions. With the knowledge of noise rates for both methods, remember that $L^u_{\leftarrow} = \frac{(1+|\rho^u_0-\rho^u_1|)L}{1-\rho^u_0-\rho^u_1}$, we have:

$$
\begin{aligned}
\overline{\Delta}^{\circ}_{R\leftarrow} < \overline{\Delta}^{\bullet}_{R\leftarrow} \implies & 2L^{\circ}_{\leftarrow}\mathfrak{R}(\mathcal{F}) + \frac{L^{\circ}_{\leftarrow}}{L}\cdot(\overline{\ell}-\underline{\ell})\cdot\log(\frac{K+1}{\delta})\cdot\sqrt{\frac{1}{NK}} \\
& < 2L^{\bullet}_{\leftarrow}\mathfrak{R}(\mathcal{F}) + \frac{L^{\bullet}_{\leftarrow}}{L}\cdot(\overline{\ell}-\underline{\ell})\cdot\sqrt{\frac{\log(1/\delta)}{2N}} \\
\implies & 2\cdot\frac{L^{\circ}_{\leftarrow}-L^{\bullet}_{\leftarrow}}{\overline{\ell}-\underline{\ell}}\cdot L\cdot\mathfrak{R}(\mathcal{F}) < L^{\bullet}_{\leftarrow}\cdot\sqrt{\frac{\log(1/\delta)}{2N}} - L^{\circ}_{\leftarrow}\cdot\log(\frac{K+1}{\delta})\cdot\sqrt{\frac{1}{NK}} \\
\implies & 2\cdot\frac{L^{\circ}_{\leftarrow}-L^{\bullet}_{\leftarrow}}{\overline{\ell}-\underline{\ell}}\cdot L\cdot\mathfrak{R}(\mathcal{F}) < \left(L^{\bullet}_{\leftarrow} - L^{\circ}_{\leftarrow}\cdot\sqrt{\frac{1}{\eta^{\circ}_K}}\right)\sqrt{\frac{\log(1/\delta)}{2N}} \\
\implies & 2\sqrt{\frac{2N}{\log(1/\delta)}}\frac{L^{\circ}_{\leftarrow}-L^{\bullet}_{\leftarrow}}{\overline{\ell}-\underline{\ell}}\cdot L\cdot\mathfrak{R}(\mathcal{F}) < L^{\bullet}_{\leftarrow} - L^{\circ}_{\leftarrow}\cdot\sqrt{\frac{1}{\eta^{\circ}_K}}.
\end{aligned}
$$

For any finite concept class $\mathcal{F}\subset\{f: X\to\{0,1\}\}$, and the sample set $S=\{x_1,...,x_N\}$, the Rademacher complexity is upper bounded by $\sqrt{\frac{2d\log(N)}{N}}$ where d is the VC dimension of $\mathcal{F}$. To achieve $\overline{\Delta}^{\circ}_{R\leftarrow} < \overline{\Delta}^{\bullet}_{R\leftarrow}$, we simply need to find the condition of K (or η°_K) that satisfies the below in-equation:

$$
\begin{aligned}
\overline{\Delta}^{\circ}_{R\leftarrow} < \overline{\Delta}^{\bullet}_{R\leftarrow} \implies & 2\sqrt{\frac{2N}{\log(1/\delta)}}\frac{L^{\circ}_{\leftarrow}-L^{\bullet}_{\leftarrow}}{\overline{\ell}-\underline{\ell}}\cdot L\cdot\sqrt{\frac{2d\log(N)}{N}} < \left(L^{\bullet}_{\leftarrow} - L^{\circ}_{\leftarrow}\cdot\sqrt{\frac{1}{\eta^{\circ}_K}}\right) \\
\implies & 4\frac{L^{\circ}_{\leftarrow}-L^{\bullet}_{\leftarrow}}{\overline{\ell}-\underline{\ell}}\cdot L\cdot\sqrt{\frac{d\log(N)}{\log(1/\delta)}} < \left(L^{\bullet}_{\leftarrow} - L^{\circ}_{\leftarrow}\cdot\sqrt{\frac{1}{\eta^{\circ}_K}}\right) \\
\implies & (L^{\circ}_{\leftarrow}-L^{\bullet}_{\leftarrow})\cdot \underbrace{\frac{4}{\overline{\ell}-\underline{\ell}}\cdot L\cdot\sqrt{\frac{d\log(N)}{\log(1/\delta)}}}_{\text{denoted as } \alpha_{\leftarrow}, \text{which is a function of } N,\delta,d,L} < \left(L^{\bullet}_{\leftarrow} - L^{\circ}_{\leftarrow}\cdot\sqrt{\frac{1}{\eta^{\circ}_K}}\right) \\
\implies & (L^{\circ}_{\leftarrow}-L^{\bullet}_{\leftarrow})\cdot\alpha_{\leftarrow} < \left(L^{\bullet}_{\leftarrow} - L^{\circ}_{\leftarrow}\cdot\sqrt{\frac{1}{\eta^{\circ}_K}}\right) \\
\implies & (L^{\circ}_{\leftarrow}-L^{\bullet}_{\leftarrow})\cdot\alpha_{\leftarrow} < (L^{\bullet}_{\leftarrow} - L^{\circ}_{\leftarrow}) + \left(1-\sqrt{\frac{1}{\eta^{\circ}_K}}\right)\cdot L^{\circ}_{\leftarrow} \\
\implies & \alpha_{\leftarrow}+1 < \left(1-\sqrt{\frac{1}{\eta^{\circ}_K}}\right)\cdot\frac{L^{\circ}_{\leftarrow}}{L^{\circ}_{\leftarrow}-L^{\bullet}_{\leftarrow}} \\
\implies & \frac{1}{\gamma} < \left(1-(\eta^{\circ}_K)^{-\frac{1}{2}}\right)\cdot\frac{L^{\circ}_{\leftarrow}}{L^{\circ}_{\leftarrow}-L^{\bullet}_{\leftarrow}} \\
\implies & \alpha_K\cdot\frac{1}{1-(\eta^{\circ}_K)^{-\frac{1}{2}}} \le \gamma,
\end{aligned}
$$

where we denote by $\alpha_K := 1 - L^{\bullet}_{\leftarrow}/L^{\circ}_{\leftarrow}$, $\gamma = 1/(1+\frac{4L}{\overline{\ell}-\underline{\ell}}\sqrt{\frac{d\log(N)}{\log(1/\delta)}})$. □

B.3 Proof of Theorem 4.3

Proof.

$$\begin{aligned}
&\mathrm{Var}(\hat{f}^u_{\leftarrow}) = \mathbb{E}_{(X,\widetilde{Y}^u)\sim\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u) - \mathbb{E}_{(X,\widetilde{Y}^u)\sim\widetilde{\mathcal{D}}^u}[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)]\right]^2\\
=&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\left[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)\right]^2 + \left[\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)]\right]^2 - 2\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)]\right]\\
=&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)\right]^2 + \left[\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)]\right]^2 - \mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[2\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)]\right]\\
=&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)\right]^2 - \left[\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)]\right]^2\\
=&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)\right]^2 - (R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\leftarrow}))^2.
\end{aligned}$$

A special case is the 0-1 loss, i.e., $\ell(\cdot) = \mathbf{1}(\cdot)$, we then have:

$$\begin{aligned}
\mathrm{Var}(\hat{f}^u_{\leftarrow}) =&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)\right]^2 - (R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\leftarrow}))^2\\
=&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u_{\leftarrow}(X),\widetilde{Y}^u)\right] - (R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\leftarrow}))^2\\
=&R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\leftarrow}) - (R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\leftarrow}))^2,
\end{aligned}$$

where $R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\leftarrow}) \in [0,1]$ and $g(a) = a - a^2$ is monotonically increasing when $a < \frac{1}{2}$. Note that: $R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\leftarrow}) < L^u_{\leftarrow 0}\cdot(\overline{\ell}-\underline{\ell})\cdot\sqrt{\frac{\log(1/\delta)}{2\eta^u_K N}}$, when

$$L^u_{\leftarrow 0}\cdot(\overline{\ell}-\underline{\ell})\cdot\sqrt{\frac{\log(1/\delta)}{2\eta^u_K N}} \le \frac{1}{2} \Longleftrightarrow L^u_{\leftarrow 0}(\eta^u_K)^{-\frac{1}{2}} < \sqrt{\frac{N}{2(\overline{\ell}-\underline{\ell})^2\log(1/\delta)}},$$

we have: $\mathrm{Var}(\hat{f}^u_{\leftarrow}) \le g\left(\frac{L^u_{\leftarrow}\cdot(\overline{\ell}-\underline{\ell})}{L}\cdot\sqrt{\frac{2\log(1/\delta)}{\eta^u_K N}}\right)$.

To achieve: $g\left(\frac{L^\circ_{\leftarrow}\cdot(\overline{\ell}-\underline{\ell})}{L}\cdot\sqrt{\frac{2\log(1/\delta)}{\eta^\circ_K N}}\right) \le g\left(\frac{L^\bullet_{\leftarrow}\cdot(\overline{\ell}-\underline{\ell})}{L}\cdot\sqrt{\frac{2\log(1/\delta)}{\eta^\bullet_K N}}\right)$, we simply need:

$$L^\circ_{\leftarrow 0}\cdot(\overline{\ell}-\underline{\ell})\cdot\sqrt{\frac{2\log(1/\delta)}{\eta^\circ_K N}} \le L^\bullet_{\leftarrow 0}\cdot(\overline{\ell}-\underline{\ell})\cdot\sqrt{\frac{2\log(1/\delta)}{\eta^\bullet_K N}} \Longleftrightarrow \sqrt{\eta^u_K} > \frac{L^\circ_{\leftarrow}}{L^\bullet_{\leftarrow}}.$$

□

B.4 Proof for Corollary 4.4

For a general matrix $U = (T^u)^{-1}$, we firstly note

$$\begin{aligned}
\overline{\mathbf{1}^u_{\leftarrow}} - \underline{\mathbf{1}^u_{\leftarrow}} =& \max_{i,j\in[M]} U^u_{ij} - \min_{i,j\in[M]} U^u_{ij}\\
\le&|\max_{i,j\in[M]} U^u_{ij}| + |\min_{i,j\in[M]} U^u_{ij}|\\
\le&|\max_{i\in[M]}\sum_{j\in[M],U_{ij}>0} U^u_{ij}| + |\min_{i\in[M]}\sum_{j\in[M],U_{ij}<0} U^u_{ij}|.
\end{aligned}$$

Recall $T^u\mathbf{1} = \mathbf{1} \Rightarrow \mathbf{1} = (T^u)^{-1}\mathbf{1}$. We know the above maximum and minimum take the same i. Then

$$\begin{aligned}
\overline{\mathbf{1}^u_{\leftarrow}} - \underline{\mathbf{1}^u_{\leftarrow}} \le&|\max_{i\in[M]}\sum_{j\in[M],U_{ij}>0} U^u_{ij}| + |\min_{i\in[M]}\sum_{j\in[M],U_{ij}<0} U^u_{ij}|\\
=&\|U^u\|_\infty\\
\overset{(a)}{\le}&\frac{1}{\min_{i\in[M]}\left(T^u_{ii} - \sum_{j\neq i} T^u_{ij}\right)}\\
\le&\frac{1}{1-2e^u}, \quad e^u := \max_{i\in[M]}(1-T^u_{ii}), \quad e^u < 0.5.
\end{aligned}$$

Now we prove the inequality (a) [49]. Let ν satisfy

$$\|(T^u)^{-1}\|_\infty = \|(T^u)^{-1}\nu\|_\infty / \|\nu\|_\infty$$

and let $\mu = (T^u)^{-1}\nu$. Then

$$\|(T^u)^{-1}\|_\infty = \|\mu\|_\infty / \|\nu\|_\infty$$

To bound $\|\mu\|$, we choose i such that $\mu_i = \|\mu\|_\infty$. Then

$$T^u_{ii}\mu_i = \nu_i - \sum_{j\neq i} T^u_{ij}\mu_j,$$

which further gives

$$|T^u_{ii}|\|\mu\|_\infty \leq |\nu_i| + \sum_{j\neq i} |T^u_{ij}||\mu_j| \leq |\nu_i| + \|\mu\|_\infty \sum_{j\neq i} |T^u_{ij}|.$$

Therefore,

$$\|\mu\|_\infty \leq \frac{|\nu_i|}{T^u_{ii} - \sum_{j\neq i} T^u_{ij}},$$

and

$$\|(T^u)^{-1}\|_\infty = \|\mu\|_\infty / \|\nu\|_\infty \leq \frac{1}{T^u_{ii} - \sum_{j\neq i} T^u_{ij}}.$$

On the other hand, denoting by $\|U\|_{\max} := \max_{i,j\in[M]} |U_{ij}|$, from eigenvalues, we know

$$\overline{\mathbf{1}^u_{\leftarrow}} - \underline{\mathbf{1}^u_{\leftarrow}} \leq \|U^u\|_\infty \leq \sqrt{M}\lambda_{\max}(U) = \frac{\sqrt{M}}{\lambda_{\min}(T^u)}.$$

where $\lambda_{\min}(T^u)$ denotes the minimal eigenvalue of the matrix T^u. Therefore,

$$\overline{\mathbf{1}^u_{\leftarrow}} - \underline{\mathbf{1}^u_{\leftarrow}} = L^\circ_{\leftarrow 0} = \min\{\frac{1}{1-2e^u_{i_{\max}}}, \frac{\sqrt{M}}{\lambda_{\min}(T^u)}\},$$

where $e^u := \max_{i\in[M]}(1 - T^u_{ii}), \quad e^u < 0.5$, and $\lambda_{\min}(T^u)$ denotes the minimal eigenvalue of the matrix T^u.

B.5 Proof of Lemma 3.2

Proof. Note that for $\hat{f} = \hat{f}^u$, we have:

$$\begin{aligned} &R_{\ell,\mathcal{D}}(\hat{f}^u) - \min_{f\in\mathcal{F}} R_{\ell,\mathcal{D}}(f) = R_{\ell,\mathcal{D}}(\hat{f}^u) - R_{\ell,\mathcal{D}}(f^*) \\ &= \underbrace{R_{\ell,\mathcal{D}}(\hat{f}^u) - R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u)}_{\text{Distribution shift}} + \underbrace{R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u) - \min_{f\in\mathcal{F}} R_{\ell,\widetilde{\mathcal{D}}^u}(f) + \min_{f\in\mathcal{F}} R_{\ell,\widetilde{\mathcal{D}}^u}(f) - R_{\ell,\mathcal{D}}(f^*)}_{\text{Estimation error}}. \end{aligned} \tag{12}$$

The term of distribution shift can be upper bounded by:

$$
\begin{aligned}
&R_{\ell,\mathcal{D}}(\hat{f}^u) - R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u) \\
=&\mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\ell(\hat{f}^u(X),Y)\right] - \mathbb{E}_{(X,\widetilde{Y}_i^u)\sim\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u(X),\widetilde{Y}_i^u)\right] \\
\leq&\max_{f\in\mathcal{F}}\left|\mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\ell(f(X),Y)\right] - \mathbb{E}_{(X,\widetilde{Y}_i^u)\sim\widetilde{\mathcal{D}}^u}\left[\ell(f(X),\widetilde{Y}_i^u)\right]\right| \\
=&\max_{f\in\mathcal{F}}\Big|\mathbb{E}_{(X,Y=1)\sim\mathcal{D}}\left[\ell(f(X),1)\right] + \mathbb{E}_{(X,Y=0)\sim\mathcal{D}}\left[\ell(f(X),0)\right] \\
&- \mathbb{E}_{(X,\widetilde{Y}_i^u)\sim\widetilde{\mathcal{D}}^u,Y=1}\left[\ell(f(X),\widetilde{Y}_i^u)\right] - \mathbb{E}_{(X,\widetilde{Y}_i^u)\sim\widetilde{\mathcal{D}}^u,Y=0}\left[\ell(f(X),\widetilde{Y}_i^u)\right]\Big| \\
=&\max_{f\in\mathcal{F}}\Big|\mathbb{E}_{(X,Y=1)\sim\mathcal{D}}\left[\ell(f(X),1)\right] + \mathbb{E}_{(X,Y=0)\sim\mathcal{D}}\left[\ell(f(X),0)\right] \\
&- \mathbb{E}_{(X,\widetilde{Y}_i^u=1)\sim\widetilde{\mathcal{D}}^u,Y=1}\left[\ell(f(X),1)\right] - \mathbb{E}_{(X,\widetilde{Y}_i^u=0)\sim\widetilde{\mathcal{D}}^u,Y=1}\left[\ell(f(X),0)\right] \\
&- \mathbb{E}_{(X,\widetilde{Y}_i^u=1)\sim\widetilde{\mathcal{D}}^u,Y=0}\left[\ell(f(X),1)\right] - \mathbb{E}_{(X,\widetilde{Y}_i^u=0)\sim\widetilde{\mathcal{D}}^u,Y=0}\left[\ell(f(X),0)\right]\Big| \\
=&\max_{f\in\mathcal{F}}\Big|\underline{\mathbb{E}_{(X,Y=1)\sim\mathcal{D}}\left[\ell(f(X),1)\right]} + \underline{\mathbb{E}_{(X,Y=0)\sim\mathcal{D}}\left[\ell(f(X),0)\right]} \\
&- \underline{\mathbb{E}_{(X,Y=1)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}_i^u=1|Y=1)\cdot\ell(f(X),1)\right]} - \mathbb{E}_{(X,Y=1)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}_i^u=0|Y=1)\cdot\ell(f(X),0)\right] \\
&- \mathbb{E}_{(X,Y=0)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}_i^u=1|Y=0)\cdot\ell(f(X),1)\right] - \underline{\mathbb{E}_{(X,Y=0)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}_i^u=0|Y=0)\cdot\ell(f(X),0)\right]}\Big|.
\end{aligned}
$$

Combine similar terms, we then have:

$$
\begin{aligned}
=&\max_{f\in\mathcal{F}}\Big|\mathbb{E}_{(X,Y=1)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}_i^u=0|Y=1)\cdot\ell(f(X),1)\right] + \mathbb{E}_{(X,Y=0)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}_i^u=1|Y=0)\cdot\ell(f(X),0)\right] \\
&- \mathbb{E}_{(X,Y=1)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}_i^u=0|Y=1)\cdot\ell(f(X),0)\right] - \mathbb{E}_{(X,Y=0)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}_i^u=1|Y=0)\cdot\ell(f(X),1)\right]\Big| \\
=&\max_{f\in\mathcal{F}}\left|\mathbb{E}_{(X,Y=1)\sim\mathcal{D}}\left[\rho_1^u\cdot(\ell(f(X),1)-\ell(f(X),0))\right] + \mathbb{E}_{(X,Y=0)\sim\mathcal{D}}\left[\rho_0^u\cdot(\ell(f(X),0)-\ell(f(X),1))\right]\right| \\
\leq&\max_{f\in\mathcal{F}}\left|\mathbb{E}_{(X,Y=1)\sim\mathcal{D}}\left[\rho_1^u\cdot(\overline{\ell}-\underline{\ell})\right] + \mathbb{E}_{(X,Y=0)\sim\mathcal{D}}\left[\rho_0^u\cdot(\overline{\ell}-\underline{\ell})\right]\right| \\
=&(p_1\rho_1^u + p_0\rho_0^u)\cdot(\overline{\ell}-\underline{\ell}).
\end{aligned}
$$

Thus, we have:

$$
R_{\ell,\mathcal{D}}(\hat{f}) - R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}) \leq \overline{\Delta}_R^{u,1} := (\rho_0^u p_0 + \rho_1^u p_1)\cdot(\overline{\ell}-\underline{\ell}).
$$

□

B.6 Proof of Lemma 3.3

Proof. For the term Estimation error, we have:

$$
\begin{aligned}
&R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}) - R_{\ell,\mathcal{D}}(f^*) \\
=&\underbrace{R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u) - \min_{f\in\mathcal{F}} R_{\ell,\widetilde{\mathcal{D}}^u}(f) + \min_{f\in\mathcal{F}} R_{\ell,\widetilde{\mathcal{D}}^u}(f) - R_{\ell,\mathcal{D}}(f^*)}_{\text{Estimation error}} \\
\leq&\underbrace{R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u) - \min_{f\in\mathcal{F}} R_{\ell,\widetilde{\mathcal{D}}^u}(f)}_{\text{Error 1}} + \underbrace{|\min_{f\in\mathcal{F}} R_{\ell,\widetilde{\mathcal{D}}^u}(f) - R_{\ell,\mathcal{D}}(f^*)|}_{\text{Error 2}}
\end{aligned}
$$

The upper bound of Error 1 could be derived directly from the proof of Lemma 4.1: since the loss function makes no use of loss correction, the L-Lipschitz constant does not have to multiply with the constant and $L_{\leftarrow}^u \to L$. Besides, the constant for the variance term (square term) reduces to $(\overline{\ell}-\underline{\ell})$. Thus, we have:

$$
\text{Error 1} \leq 4L\mathfrak{R}(\mathcal{F}) + (\overline{\ell}-\underline{\ell})\cdot\sqrt{\frac{2\log(1/\delta)}{\eta_K^u N}},\quad \forall f\in\mathcal{F}.
$$

For the term Error 2, the upper bound could be derived with the same procedure as adopted in the proof of Lemma 3.2. Thus, we obtain:

$$R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}) - R_{\ell,\mathcal{D}}(f^*) \leq \underbrace{4L\mathfrak{R}(\mathcal{F}) + (\overline{\ell} - \underline{\ell}) \cdot \sqrt{\frac{2\log(1/\delta)}{\eta_K^u N}} + \overline{\Delta}_R^{u,1}}_{\text{Defined as: } \overline{\Delta}_R^{u,2}}.$$

□

B.7 Proof of Theorem 3.4

Proof. To achieve a smaller upper bound for the separation method, mathematically, we want:

$$\begin{aligned}
&4L\mathfrak{R}(\mathcal{F}) + (\overline{\ell} - \underline{\ell}) \cdot \sqrt{\frac{2\log(1/\delta)}{\eta_K^\circ N}} + 2(\rho_0^\circ p_0 + \rho_1^\circ p_1) \cdot (\overline{\ell} - \underline{\ell}) \\
\leq &4L\mathfrak{R}(\mathcal{F}) + (\overline{\ell} - \underline{\ell}) \cdot \sqrt{\frac{2\log(1/\delta)}{\eta_K^\bullet N}} + 2(\rho_0^\bullet p_0 + \rho_1^\bullet p_1) \cdot (\overline{\ell} - \underline{\ell}),
\end{aligned}$$

which is equivalent to prove:

$$\begin{aligned}
&\sqrt{\frac{\log(1/\delta)}{2N}}((\eta_K^\circ)^{-\frac{1}{2}} - 1) \cdot (\overline{\ell} - \underline{\ell}) \leq [(\rho_0^\bullet p_0 + \rho_1^\bullet p_1) - (\rho_0^\circ p_0 + \rho_1^\circ p_1)] \cdot (\overline{\ell} - \underline{\ell}) \\
\Longleftrightarrow &\sqrt{\frac{\log(1/\delta)}{2N}}(1 - (\eta_K^\circ)^{-\frac{1}{2}}) \geq \underbrace{[(\rho_0^\circ p_0 + \rho_1^\circ p_1) - (\rho_0^\bullet p_0 + \rho_1^\bullet p_1)]}_{\text{De-noising effect of aggregation} \geq 0}. \qquad (13)
\end{aligned}$$

Eqn. (13) then requires: $\sqrt{\frac{\log(1/\delta)}{2N}} \geq \frac{(\rho_0^\circ p_0 + \rho_1^\circ p_1) - (\rho_0^\bullet p_0 + \rho_1^\bullet p_1)}{(1-(\eta_K^\circ)^{-\frac{1}{2}})}$, which is mentioned as $\alpha_K \cdot \frac{1}{1-(\eta_K^\circ)^{-\frac{1}{2}}} \leq \gamma$, where $\alpha_K := (\rho_0^\circ p_0 + \rho_1^\circ p_1) - (\rho_0^\bullet p_0 + \rho_1^\bullet p_1)$, $\gamma = \sqrt{\log(1/\delta)/2N}$.

□

B.8 Proof of Theorem 3.6

Proof. For $u \in \{\circ, \bullet\}$, we have:

$$\begin{aligned}
\text{Var}(\hat{f}^u) =&\mathbb{E}_{(X,\widetilde{Y}^u)\sim\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u(X), \widetilde{Y}^u) - \mathbb{E}_{(X,\widetilde{Y}^u)\sim\widetilde{\mathcal{D}}^u}[\ell(\hat{f}^u(X), \widetilde{Y}^u)]\right]^2 \\
=&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\left[\ell(\hat{f}^u(X), \widetilde{Y}^u)\right]^2 + \left[\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(\hat{f}^u(X), \widetilde{Y}^u)]\right]^2 - 2\ell(\hat{f}^u(X), \widetilde{Y}^u)\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(\hat{f}^u(X), \widetilde{Y}^u)]\right] \\
=&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u(X), \widetilde{Y}^u)\right]^2 + \left[\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(\hat{f}^u(X), \widetilde{Y}^u)]\right]^2 - \mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[2\ell(\hat{f}^u(X), \widetilde{Y}^u)\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(\hat{f}^u(X), \widetilde{Y}^u)]\right] \\
=&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u(X), \widetilde{Y}^u)\right]^2 - \left[\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(\hat{f}^u(X), \widetilde{Y}^u)]\right]^2 \\
=&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u(X), \widetilde{Y}^u)\right]^2 - (R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u))^2.
\end{aligned}$$

A special case is the 0-1 loss, i.e., $\ell(\cdot) = \mathbf{1}(\cdot)$, we then have:

$$\begin{aligned}
\text{Var}(\hat{f}^u) =&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u(X), \widetilde{Y}^u)\right]^2 - (R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u))^2 \\
=&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u(X), \widetilde{Y}^u)\right] - (R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u))^2 \\
=&R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u) - (R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u))^2 = g\left(R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u)\right).
\end{aligned}$$

where $R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u) \in [0,1]$ and $g(a) = a - a^2$ is monotonically increasing when $a < \frac{1}{2}$. Thus, when

$$R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u) \leq \underbrace{(\overline{\ell} - \underline{\ell})}_{\text{reduces to 1}} \cdot \sqrt{\frac{\log(1/\delta)}{2\eta_K^u N}} \leq \frac{1}{2} \Longleftrightarrow \eta_K^u \geq \frac{2\log(1/\delta)}{N},$$

we could derive $\text{Var}(\hat{f}^u) \leq g(\sqrt{\frac{2\log(1/\delta)}{\eta_K^u N}})$.

□

B.9 Proof of Corollary 3.7

Proof. In the multi-class extension, the only difference is the upper bound of Distribution Shift term in Eqn. (12), which now becomes:

$$
\begin{aligned}
& R_{\ell,\mathcal{D}}(\hat{f}^u) - R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u) \\
=& \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\ell(\hat{f}^u(X),Y)\right] - \mathbb{E}_{(X,\widetilde{Y}^u)\sim\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u(X),\widetilde{Y}^u)\right] \\
\leq & \max_{f\in\mathcal{F}}\left|\mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\ell(f(X),Y)\right] - \mathbb{E}_{(X,\widetilde{Y}^u)\sim\widetilde{\mathcal{D}}^u}\left[\ell(f(X),\widetilde{Y}^u)\right]\right| \\
= & \max_{f\in\mathcal{F}}\left|\left[\sum_{j\in[M]}\mathbb{E}_{(X,Y=j)\sim\mathcal{D}}\left[\ell(f(X),j)\right]\right] - \left[\sum_{j\in[M]}\mathbb{E}_{(X,\widetilde{Y}^u)\sim\widetilde{\mathcal{D}}^u,Y=j}\left[\ell(f(X),\widetilde{Y}^u)\right]\right]\right| \\
= & \max_{f\in\mathcal{F}}\left|\left[\sum_{j\in[M]}\mathbb{E}_{(X,Y=j)\sim\mathcal{D}}\left[\ell(f(X),j)\right]\right] - \left[\sum_{k\in[M]}\sum_{j\in[M]}\mathbb{E}_{(X,Y=j)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}^u=k|Y=j)\cdot\ell(f(X),k)\right]\right]\right| \\
= & \max_{f\in\mathcal{F}}\left|\left[\sum_{j\in[M]}\mathbb{E}_{(X,Y=j)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}^u\neq j|Y=j)\cdot\ell(f(X),j)\right]\right]\right. \\
& \qquad \left. - \left[\sum_{k\in[M],k\neq j}\sum_{j\in[M]}\mathbb{E}_{(X,Y=j)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}^u=k|Y=j)\cdot\ell(f(X),k)\right]\right]\right| \\
= & \max_{f\in\mathcal{F}}\left|\sum_{j\in[M]}\mathbb{E}_{(X,Y=j)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}^u\neq j|Y=j)\cdot\ell(f(X),j) - \sum_{k\in[M],k\neq j}\mathbb{P}(\widetilde{Y}^u=k|Y=j)\cdot\ell(f(X),k)\right]\right| \\
\leq & \max_{f\in\mathcal{F}}\left|\sum_{j\in[M]}\mathbb{E}_{(X,Y=j)\sim\mathcal{D}}\left[\mathbb{P}(\widetilde{Y}^u\neq j|Y=j)\cdot\left(\overline{\ell}-\underline{\ell}\right)\right]\right| \qquad \text{(Assumed uniform prior)} \\
= & \sum_{j\in[M]}\mathbb{P}(Y=j)\cdot(1-T^u_{j,j})\left(\overline{\ell}-\underline{\ell}\right).
\end{aligned}
$$

□

B.10 Proof of Lemma 4.5

Proof. The proof of Lemma 4.5 builds on Theorem 7 in [26]: The performance bound for aggregation methods is the special case of Theorem 7 in [26] (adopting $\alpha^* = 1$ defined in [26]). As for that of separation methods, the incurred difference lies in the appearance of the weight of sample complexity η_K°. Thus, we have:

$$
\begin{aligned}
& R_{\ell,\mathcal{D}}(\hat{f}^u_{\looparrowright}) - R_{\ell,\mathcal{D}}(f^*) \leq \frac{1}{1-\rho_0^u-\rho_1^u}\left(8L\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{2\log(4/\delta)}{\eta_K^u N}}\left(1+2(\overline{\ell}-\underline{\ell})\right)\right) \\
\Longleftrightarrow & R_{\ell,\mathcal{D}}(\hat{f}^u_{\looparrowright}) - R_{\ell,\mathcal{D}}(f^*) \leq \overline{\Delta}^u_{R\looparrowright},
\end{aligned}
$$

where $\overline{\Delta}^u_{R\looparrowright} := 8L^u_{\looparrowright}\mathfrak{R}(\mathcal{F}) + L^u_{\looparrowright 0}\sqrt{\frac{2\log(4/\delta)}{\eta_K^u N}}\left(1+2(\overline{\ell}-\underline{\ell})\right)$.

□

B.11 Proof of Theorem 4.6

Proof. Denote by $\overline{\Delta}^u_{R\looparrowright} := \frac{8L\mathfrak{R}(\mathcal{F})}{1-\rho_0^u-\rho_1^u} + \frac{4\sqrt{\frac{\log(4/\delta)}{2\eta_K^u N}}\left(1+2(\bar{\ell}-\underline{\ell})\right)}{1-\rho_0^u-\rho_1^u}$, in order to achieve $\overline{\Delta}^\circ_{R\looparrowright} < \overline{\Delta}^\bullet_{R\looparrowright}$, we require $\overline{\Delta}^\circ_{R\looparrowright} < \overline{\Delta}^\bullet_{R\looparrowright}$, which is equivalent to:

$$\frac{8L\mathfrak{R}(\mathcal{F})}{1-\rho_0^\circ-\rho_1^\circ} + \frac{4\sqrt{\frac{\log(4/\delta)}{2\eta_K^\circ N}}\left(1+2(\bar{\ell}-\underline{\ell})\right)}{1-\rho_0^\circ-\rho_1^\circ} < \frac{8L\mathfrak{R}(\mathcal{F})}{1-\rho_0^\bullet-\rho_1^\bullet} + \frac{4\sqrt{\frac{\log(4/\delta)}{2N}}\left(1+2(\bar{\ell}-\underline{\ell})\right)}{1-\rho_0^\bullet-\rho_1^\bullet},$$

which is further equivalent to:

$$\frac{8L\mathfrak{R}(\mathcal{F})}{1-\rho_0^\circ-\rho_1^\circ} - \frac{8L\mathfrak{R}(\mathcal{F})}{1-\rho_0^\bullet-\rho_1^\bullet} < \frac{4\sqrt{\frac{\log(4/\delta)}{2N}}\left(1+2(\bar{\ell}-\underline{\ell})\right)}{1-\rho_0^\bullet-\rho_1^\bullet} - \frac{4\sqrt{\frac{\log(4/\delta)}{2\eta_K^\circ N}}\left(1+2(\bar{\ell}-\underline{\ell})\right)}{1-\rho_0^\circ-\rho_1^\circ}.$$

Note that both $1-\rho_0^\circ-\rho_1^\circ$ and $1-\rho_0^\bullet-\rho_1^\bullet$ are positive, the above requirement then reduces to:

$$\begin{aligned}
&[(\rho_0^\circ+\rho_1^\circ)-(\rho_0^\bullet+\rho_1^\bullet)]8L\mathfrak{R}(\mathcal{F}) < \left[(1-\rho_0^\circ-\rho_1^\circ)-(1-\rho_0^\bullet-\rho_1^\bullet)\sqrt{\frac{1}{\eta_K^\circ}}\right] 4\sqrt{\frac{\log(4/\delta)}{2N}}\left(1+2(\bar{\ell}-\underline{\ell})\right)\\
\iff &\frac{[(\rho_0^\circ+\rho_1^\circ)-(\rho_0^\bullet+\rho_1^\bullet)]8L\mathfrak{R}(\mathcal{F})}{4\sqrt{\frac{\log(4/\delta)}{2N}}\left(1+2(\bar{\ell}-\underline{\ell})\right)} < (1-\rho_0^\circ-\rho_1^\circ)-(1-\rho_0^\bullet-\rho_1^\bullet)\sqrt{\frac{1}{\eta_K^\circ}}.
\end{aligned}$$

Note that for any finite concept class $\mathcal{F}\subset\{f: X\to\{0,1\}\}$, and the sample set $S=\{x_1,...,x_N\}$, the Rademacher complexity is upper bounded by $\sqrt{\frac{2d\log(N)}{N}}$ where d is the VC dimension of $\mathcal{F}$, a more strict condition to get becomes:

$$\sqrt{\frac{1}{\eta_K^\circ}} < \frac{(1-\rho_0^\circ-\rho_1^\circ)}{(1-\rho_0^\bullet-\rho_1^\bullet)} - \frac{[(\rho_0^\circ+\rho_1^\circ)-(\rho_0^\bullet+\rho_1^\bullet)]8L\sqrt{\frac{2d\log(N)}{N}}}{4(1-\rho_0^\bullet-\rho_1^\bullet)\sqrt{\frac{\log(4/\delta)}{2N}}\left(1+2(\bar{\ell}-\underline{\ell})\right)}.$$

Denote by $\alpha_K := 1-L^\bullet_{\looparrowright}/L^\circ_{\looparrowright}$, $\gamma = \frac{1+2(\bar{\ell}-\underline{\ell})}{2L}\sqrt{\frac{\log(4/\delta)}{4d\log(N)}}$. The above condition is sastisfied if and only if

$$\alpha_K\cdot\frac{1}{L^\bullet_{\looparrowright}/L^\circ_{\looparrowright}-(\eta_K^\circ)^{-\frac{1}{2}}} \le \gamma.$$

□

B.12 Proof of Theorem 4.7

Proof. Similar to the proof of Theorem 3.6, for $u\in\{\circ,\bullet\}$, we have:

$$\mathrm{Var}(\hat{f}^u_{\looparrowright}) = \mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u_{\looparrowright}(X),\widetilde{Y}^u)\right]^2 - (R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\looparrowright}))^2.$$

A special case is the 0-1 loss, i.e., $\ell(\cdot)=\mathbf{1}(\cdot)$, we then have:

$$\begin{aligned}
\mathrm{Var}(\hat{f}^u_{\looparrowright}) =&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u_{\looparrowright}(X),\widetilde{Y}^u)\right]^2 - (R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\looparrowright}))^2\\
=&\mathbb{E}_{\widetilde{\mathcal{D}}^u}\left[\ell(\hat{f}^u_{\looparrowright}(X),\widetilde{Y}^u)\right] - (R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\looparrowright}))^2\\
=&R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\looparrowright}) - (R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\looparrowright}))^2 = g\left(R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\looparrowright})\right)
\end{aligned}$$

where $R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\looparrowright})\in[0,1]$ and $g(a)=a-a^2$ is monotonically increasing when $a<\frac{1}{2}$. Note that:

$$R_{\ell,\widetilde{\mathcal{D}}^u}(\hat{f}^u_{\looparrowright}) < \frac{1}{1-\rho_0^u-\rho_1^u}\sqrt{\frac{\log(4/\delta)}{2\eta_K^u N}}\left(1+2(\bar{\ell}-\underline{\ell})\right),$$

when

$$\frac{1}{1-\rho_0^u-\rho_1^u}\sqrt{\frac{\log(4/\delta)}{2\eta_K^u N}}\left(1+2(\bar{\ell}-\underline{\ell})\right)\leq\frac{1}{2}$$
$$\Longleftrightarrow\sqrt{\eta_K^u}\geq\sqrt{\frac{2\log(4/\delta)}{N}}\frac{1+2(\bar{\ell}-\underline{\ell})}{1-\rho_0^u-\rho_1^u},$$

we have: $\mathrm{Var}(\hat{f}^u_{\leftarrow})\leq g\left(\sqrt{\frac{\log(4/\delta)}{2\eta_K^u N}}\frac{1+2(\bar{\ell}-\underline{\ell})}{1-\rho_0^u-\rho_1^u}\right)$. To achieve: $\mathrm{Var}(\hat{f}^{\circ}_{\looparrowright})<\mathrm{Var}(\hat{f}^{\bullet}_{\looparrowright})$, we simply need:

$$\sqrt{\frac{\log(4/\delta)}{2\eta_K^{\circ} N}}\frac{1+2(\bar{\ell}-\underline{\ell})}{1-\rho_0^{\circ}-\rho_1^{\circ}}\leq\sqrt{\frac{\log(4/\delta)}{2\eta_K^{\bullet} N}}\frac{1+2(\bar{\ell}-\underline{\ell})}{1-\rho_0^{\bullet}-\rho_1^{\bullet}}\Longleftrightarrow\sqrt{\eta_K^{\circ}}\geq\frac{L^{\circ}_{\looparrowright 0}}{L^{\bullet}_{\looparrowright 0}}.$$

□

B.13 Proof of Corollary 4.8

Proof. Regarding the multi-class extension of Lemma 4.5, the only different thing lies in the constant: $L^u_{\looparrowright 0}$. The following Lemma B.1 helps us find out the multi-class form of $L^u_{\looparrowright 0}$.

Lemma B.1. *Assume the clean label Y has equal prior $P(Y=j)=\frac{1}{M},\forall j\in[M]$. For the uniform noise transition matrix [56] such that $T^u_{i,j}=\rho^u_i,\forall j\in[M]$, the expected $\ell_{\looparrowright}$ in the multi-class setting is invariant to label noise up to an affine transformation:*

$$\mathbb{E}_{(X,\widetilde{Y}^u)\sim\widetilde{\mathcal{D}}^u}[\ell_{\looparrowright}(f(X),\widetilde{Y}^u)]=\left(1-\sum_{j\in[M]}\rho^u_j\right)\mathbb{E}_{\mathcal{D}}[\ell_{\looparrowright}(f(X),Y)].\tag{14}$$

Proof of Lemma B.1 Recall that $\mathcal{D}$ and $\widetilde{\mathcal{D}}^u$ refer to the joint distribution over (X,Y) and $(X,\widetilde{Y}^u)$, respectively. We further denote the marginal distributions of X, Y, and $\widetilde{Y}^u$ by $\mathcal{D}_X$, $\mathcal{D}_Y$, and $\widetilde{\mathcal{D}}_{\widetilde{Y}^u}$, respectively. Let $X_p\sim\mathcal{D}_X$, $\widetilde{Y}^u_p\sim\widetilde{\mathcal{D}}_{\widetilde{Y}^u}$ be the random variables corresponding to the peer samples. The peer loss function is defined as

$$\ell_{\looparrowright}(f(x_n),\tilde{y}^u_n)=\ell(f(x_n),\tilde{y}^u_n)-\ell(f(x_{p,n}),\tilde{y}^u_{p,n}),\tag{15}$$

where $(x_n,\tilde{y}^u_n)$ is a normal training sample pair, $x_{p,n}$ and $\tilde{y}^u_{p,n}$ are corresponding peer samples.

Taking expectation for (15) yields

$$\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell_{\looparrowright}(f(X),\widetilde{Y}^u)]=\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(f(X),\widetilde{Y}^u)]-\mathbb{E}_{\widetilde{\mathcal{D}}_{\widetilde{Y}^u}}\left[\mathbb{E}_{\mathcal{D}_X}[\ell(f(X_p),\widetilde{Y}^u_p)]\right].\tag{16}$$

The first term in (16) is

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(f(X),\widetilde{Y}^u)]\\
&=\sum_{j\in[M]}\sum_{i\in[M]}T^u_{ij}\cdot\mathbb{P}(Y=i)\cdot\mathbb{E}_{\mathcal{D}|Y=i}[\ell(f(X),j)]\\
&=\sum_{j\in[M]}\left[T^u_{jj}\cdot\mathbb{P}(Y=j)\cdot\mathbb{E}_{\mathcal{D}|Y=j}[\ell(f(X),j)]+\sum_{i\in[M],i\neq j}T^u_{ij}\cdot\mathbb{P}(Y=i)\cdot\mathbb{E}_{\mathcal{D}|Y=i}[\ell(f(X),j)]\right]\\
&=\sum_{j\in[M]}\left[\left(1-\sum_{i\neq j,i\in[M]}T^u_{ji}\right)\cdot\mathbb{P}(Y=j)\cdot\mathbb{E}_{\mathcal{D}|Y=j}[\ell(f(X),j)]+\sum_{i\in[M],i\neq j}T^u_{ij}\cdot\mathbb{P}(Y=i)\cdot\mathbb{E}_{\mathcal{D}|Y=i}[\ell(f(X),j)]\right].
\end{aligned}$$

Accordingly, noting X_p and $\widetilde{Y}_p^u$ are independent, the second term in (16) is

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\mathcal{D}}_{\widetilde{Y}^u}}\left[\mathbb{E}_{\mathcal{D}_X}[\ell(f(X_p),\widetilde{Y}_p^u)]\right] = \sum_{j\in[M]}\mathbb{P}(\widetilde{Y}_p^u=j)\cdot\mathbb{E}_{\mathcal{D}_X}[\ell(f(X_p),j)] \\
&= \sum_{j\in[M]}\sum_{i\in[M]} T_{ij}^u\cdot\mathbb{P}(Y_p=i)\cdot\mathbb{E}_{\mathcal{D}_X}[\ell(f(X),j)] \\
&= \sum_{j\in[M]}\left[T_{jj}^u\cdot\mathbb{P}(Y_p=j)\cdot\mathbb{E}_{\mathcal{D}_X}[\ell(f(X),j)] + \sum_{i\in[M],i\neq j} T_{ij}^u\cdot\mathbb{P}(Y_p=i)\cdot\mathbb{E}_{\mathcal{D}_X}[\ell(f(X),j)]\right] \\
&= \sum_{j\in[M]}\left[\left(1-\sum_{i\neq j,i\in[M]} T_{ji}^u\right)\cdot\mathbb{P}(Y_p=j)\cdot\mathbb{E}_{\mathcal{D}_X}[\ell(f(X),j)] + \sum_{i\in[M],i\neq j} T_{ij}^u\cdot\mathbb{P}(Y_p=i)\cdot\mathbb{E}_{\mathcal{D}_X}[\ell(f(X),j)]\right].
\end{aligned}$$

In this case, we have $\rho_i^u = T_{ji}^u, \forall j\in[M], j\neq i$. The first term becomes

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell(f(X),\widetilde{Y}^u)] \\
&= \sum_{j\in[M]}\left[\left(1-\sum_{i\neq j,i\in[M]}\rho_i^u\right)\cdot\mathbb{P}(Y=j)\cdot\mathbb{E}_{\mathcal{D}|Y=j}[\ell(f(X),j)] + \sum_{i\in[M],i\neq j}\rho_j^u\cdot\mathbb{P}(Y=i)\cdot\mathbb{E}_{\mathcal{D}|Y=i}[\ell(f(X),j)]\right] \\
&= \sum_{j\in[M]}\left[\left(1-\sum_{i\in[M]}\rho_i^u\right)\cdot\mathbb{P}(Y=j)\cdot\mathbb{E}_{\mathcal{D}|Y=j}[\ell(f(X),j)] + \sum_{i\in[M]}\rho_j^u\cdot\mathbb{P}(Y=i)\cdot\mathbb{E}_{\mathcal{D}|Y=i}[\ell(f(X),j)]\right] \\
&= \left[\left(1-\sum_{i\in[M]}\rho_i^u\right)\cdot\mathbb{E}_{\mathcal{D}}[\ell(f(X),Y)] + \sum_{j\in[M]}\rho_j^u\cdot\mathbb{E}_{\mathcal{D}_X}[\ell(f(X),j)]\right].
\end{aligned}$$

The second term becomes

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\mathcal{D}}_{\widetilde{Y}^u}}\left[\mathbb{E}_{\mathcal{D}_X}[\ell(f(X_p),\widetilde{Y}_p^u)]\right] \\
&= \sum_{j\in[M]}\left[\left(1-\sum_{i\neq j,i\in[M]}\rho_i^u\right)\cdot\mathbb{P}(Y_p=j)\cdot\mathbb{E}_{\mathcal{D}_X}[\ell(f(X),j)] + \sum_{i\in[M],i\neq j}\rho_j^u\cdot\mathbb{P}(Y_p=i)\cdot\mathbb{E}_{\mathcal{D}_X}[\ell(f(X),j)]\right] \\
&= \sum_{j\in[M]}\left[\left(1-\sum_{i\in[M]}\rho_i^u\right)\cdot\mathbb{P}(Y_p=j)\cdot\mathbb{E}_{\mathcal{D}_X}[\ell(f(X),j)] + \sum_{i\in[M]}\rho_j^u\cdot\mathbb{P}(Y_p=i)\cdot\mathbb{E}_{\mathcal{D}_X}[\ell(f(X),j)]\right] \\
&= \left(1-\sum_{i\in[M]}\rho_i^u\right)\cdot\mathbb{E}_{\mathcal{D}_Y}\left[\mathbb{E}_{\mathcal{D}_X}[\ell(f(X_p),Y_p)]\right] + \sum_{j\in[M]}\rho_j^u\cdot\mathbb{E}_{\mathcal{D}_X}[\ell(f(X),j)].
\end{aligned}$$

Comparing the above two terms we have:

$$\mathbb{E}_{\widetilde{\mathcal{D}}^u}[\ell_{\curvearrowright}(f(X),\widetilde{Y}^u)] = \left(1-\sum_{i\in[M]}\rho_i^u\right)\mathbb{E}_{\mathcal{D}}[\ell_{\curvearrowright}(f(X),Y)]. \tag{17}$$

Thus, substituting $L_{\curvearrowright 0}^u := \frac{1}{1-\rho_0^u-\rho_1^u}$ by $\frac{1}{1-\sum_{i\in[M]}\rho_i^u}$, the proof of Corollary 4.8 is finished if we repeat the corresponding proof of the binary task.

□

C Additional Results and Details

D Experiment Details

D.1 Experiment Details on UCI Datasets

Datasets In this paper, we conducted the experiments on two binary (Breast and German) and two multiclass (StatLog and Optical) UCI classification datasets. As for the splitting of training and testing, the original settings are used when training and testing files are provided. The remaining datasets only give one data file. We adopt 50/50 splitting for the testing results' statistical significance as more data is distributed to testing dataset. More specifically, the numbers of (training, testing) samples in Breast, German, StatLog, and Optical datasets are (285, 284), (500, 500), (4435, 2000), and (3823, 1797).

Generating the noisy labels on UCI datasets For each UCI dataset adopted in this paper, the label of each sample in the training dataset will be flipped to the other classes with the probability ϵ (noise rate). For the multiclass classification datasets, the specific label which will be flipped to is randomly selected with the equal probabilities. For binary and multiclass classification datasets, (0.1, 0.2, 0.3, 0.4) and (0.2, 0.4, 0.6, 0.8) are used as different lists of noise rates respectively.

Implementation details We implemented a simple two-layer ReLU Multi-Layer Perceptron (MLP) for the classification task on these four UCI datasets. The Adam optimizer is used with a learning rate of 0.001 and the batch size is 128.

D.2 Experiment Details on CIFAR-10 Datasets

The generation of symmetric noisy dataset is adopted from [56]. As for the instance-dependent label noise, the generating algorithm follows the state-of-the-art method [60]. Both cases adopt noise rates: $[0.2, 0.4, 0.6, 0.8]$. The basic hyper-parameters settings for all methods are listed as follows: mini-batch size (128), optimizer (SGD), initial learning rate (0.1), momentum (0.9), weight decay (0.0005), number of epochs (120) and learning rate decay (0.1 at 50 epochs). Standard data augmentation is applied to each dataset. All experiments run on 8 Nvidia RTX A5000 GPUs.

D.3 Details Results on CIFAR-10 Dataset

Table 6 includes all the detailed accuracy values appeared in Figure 5. The results on synthetic noisy CIFAR-10 dataset aligns well with the theoretical observations: label separation is preferred over label aggregation when the noise rates are high, or the number of labelers/annotations is insufficient.

CIFAR-10, Symmetric CE

$\epsilon = 0.2$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	92.21	92.98	93.54	93.43	93.73	93.40
EM	92.08	92.93	93.54	93.64	93.35	93.37
Sep	92.52	92.89	93.35	93.15	93.42	93.40
$\epsilon = 0.4$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	89.09	91.59	93.18	93.43	93.26	93.44
EM	88.83	91.02	92.54	93.45	93.69	93.68
Sep	90.61	91.95	92.70	92.92	93.32	93.13
$\epsilon = 0.6$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	81.85	87.33	89.88	91.88	92.96	93.40
EM	81.04	85.91	89.76	91.57	92.55	93.10
Sep	87.00	89.19	90.70	91.97	92.40	93.17
$\epsilon = 0.8$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	20.94	44.62	70.91	79.61	84.83	89.09
EM	37.91	50.78	67.19	75.26	82.97	87.97
Sep	61.47	70.10	79.61	83.93	86.82	90.04

CIFAR-10, Instance CE

$\epsilon = 0.2$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	91.99	93.29	93.57	93.47	93.68	93.60
EM	91.92	93.21	93.55	93.61	93.44	93.44
Sep	92.36	92.97	93.43	93.24	93.33	93.35
$\epsilon = 0.4$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	87.14	91.15	93.10	93.15	93.23	93.48
EM	88.07	92.40	93.70	93.58	93.74	93.53
Sep	90.83	91.90	92.63	92.46	93.08	93.26
$\epsilon = 0.6$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	49.22	83.95	89.45	91.60	92.88	93.65
EM	78.34	88.79	91.95	92.97	93.46	93.65
Sep	83.79	87.55	90.15	91.58	91.86	92.74
$\epsilon = 0.8$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	14.59	25.25	34.47	57.99	57.51	87.08
EM	20.03	26.54	65.16	80.10	88.59	92.14
Sep	26.16	28.89	50.35	74.15	71.39	87.54

CIFAR-10, Symmetric BW

$\epsilon = 0.2$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	92.08	94.09	94.92	94.90	94.79	94.90
EM	92.13	93.08	94.90	94.91	94.90	94.86
Sep	91.74	92.61	92.75	92.59	94.44	92.97
$\epsilon = 0.4$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	88.28	91.11	92.73	94.60	94.62	94.81
EM	87.41	90.23	92.83	94.77	94.80	95.18
Sep	89.14	89.68	91.07	92.46	92.26	94.24
$\epsilon = 0.6$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	81.21	86.29	89.51	91.33	93.52	94.81
EM	78.13	84.33	89.44	91.17	92.45	94.60
Sep	83.84	87.05	88.10	89.80	90.95	92.11
$\epsilon = 0.8$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	16.43	60.97	71.11	77.86	82.72	88.41
EM	10.00	45.97	66.02	74.37	80.08	87.42
Sep	58.48	69.86	76.03	79.79	82.60	86.31

CIFAR-10, Instance BW

$\epsilon = 0.2$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	92.03	93.87	95.12	95.11	94.97	94.75
EM	91.93	94.39	94.90	94.84	95.05	94.54
Sep	91.93	92.07	92.70	91.75	93.02	92.47
$\epsilon = 0.4$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	86.61	90.64	93.00	94.73	94.72	94.72
EM	89.83	92.04	94.74	95.00	94.94	94.80
Sep	88.86	87.89	92.09	89.92	91.05	91.96
$\epsilon = 0.6$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	43.78	82.59	88.56	91.47	93.27	95.06
EM	44.92	87.33	91.39	93.58	94.72	94.99
Sep	80.88	86.22	88.45	90.69	91.16	92.61
$\epsilon = 0.8$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	16.00	25.03	33.80	67.91	68.52	86.49
EM	16.06	22.73	53.96	76.24	86.74	92.02
Sep	27.84	26.68	32.72	37.27	54.41	83.37

CIFAR-10, Symmetric PeerLoss

$\epsilon = 0.2$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	92.69	93.35	93.90	94.12	94.15	93.81
EM	92.39	93.25	93.76	93.93	93.52	93.77
Sep	93.15	93.51	93.77	93.51	93.56	93.73
$\epsilon = 0.4$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	89.40	91.88	93.42	93.84	93.83	94.04
EM	89.23	91.41	93.06	93.83	93.85	94.11
Sep	91.08	92.38	93.17	93.40	93.56	93.37
$\epsilon = 0.6$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	82.88	87.95	90.42	92.31	93.61	93.79
EM	81.64	86.45	90.09	91.98	93.23	93.58
Sep	87.28	89.80	91.19	92.42	93.18	93.65
$\epsilon = 0.8$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	21.82	48.71	72.81	80.32	85.27	89.38
EM	38.29	52.63	68.70	77.42	83.94	88.45
Sep	64.32	72.52	80.31	84.65	87.40	90.56

CIFAR-10, Instance PeerLoss

$\epsilon = 0.2$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	92.13	93.53	94.00	93.78	94.13	94.08
EM	91.93	93.51	93.78	93.88	94.03	93.82
Sep	92.86	93.23	93.56	93.72	93.63	93.95
$\epsilon = 0.4$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	88.15	91.61	93.21	93.64	93.84	93.69
EM	90.59	92.60	93.95	94.02	94.06	93.68
Sep	91.06	92.70	93.22	92.92	93.65	93.67
$\epsilon = 0.6$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	60.66	84.99	90.30	91.93	93.16	93.81
EM	78.53	89.11	92.44	93.17	93.96	93.85
Sep	85.76	89.07	91.05	92.22	92.45	93.39
$\epsilon = 0.8$	$K = 3$	$K = 5$	$K = 9$	$K = 15$	$K = 25$	$K = 49$
MV	14.35	24.83	40.49	65.47	69.28	88.05
EM	26.52	28.43	66.72	80.71	89.40	92.41
Sep	33.87	37.49	57.36	77.43	80.51	89.15

Table 6: The performances of CE/BW/PeerLoss trained on (Left half: symmetric noise; right half: instance noise) CIFAR-10 aggregated labels (majority vote, EM inference), and separated labels. (Different number of labels per training image)