On the Robustness of Average Losses for Partial-Label Learning

Jiaqi Lv, Biao Liu, Lei Feng, Ning Xu, Miao Xu, Bo An, Gang Niu, Xin Geng, *Senior Member, IEEE*
Masashi Sugiyama, *Senior Member, IEEE*

Abstract—*Partial-label learning* (PLL) utilizes instances with PLs, where a PL includes several candidate labels but only one is the true label (TL). In PLL, *identification-based strategy* (IBS) purifies each PL on the fly to select the (most likely) TL for training; *average-based strategy* (ABS) treats all candidate labels equally for training and let trained models be able to predict TL. Although PLL research has focused on IBS for better performance, ABS is also worthy of study since *modern IBS behaves like ABS in the beginning of training* to prepare for PL purification and TL selection. In this paper, we analyze why ABS was unsatisfactory and propose how to improve it. Theoretically, we propose two problem settings of PLL and prove that *average PL* losses (APLLs) with *bounded* multi-class losses are *always* robust, while APLLs with *unbounded* losses may be non-robust, which is the first robustness analysis for PLL. Experimentally, we have two promising findings: ABS using bounded losses can match/exceed state-of-the-art performance of IBS using unbounded losses; after using robust APLLs to warm start, IBS can further improve upon itself. Our work draws attention to ABS research, which can in turn boost IBS and push forward the whole PLL.

Index Terms—Partial-label learning, weakly supervised learning, robustness analysis, robust loss.

1 Introduction

Deep neural networks (DNNs) have become the par excellence base model in diverse application domains, which transform the input data (e.g., images) to the specific outputs (e.g., classes). Much of the success in running DNNs is attributed to its internal capability to approximate arbitrarily complex functions mapping input to output [1], [2], [13], [14], as well as an external driving force—labeled training data. It is widely believed that the performance of DNNs is improved as the number of data increases, reaching saturation only when millions of data are available [43], [54], [57], [77]. Their remarkable performance usually comes at a prohibitively high labeling cost, especially when data labeling must be carried out professionally. A shortage of skilled experts, an expensive and time-consuming labeling process, and privacy issues can pose challenges to the acquisition of high-quality labels. As a result, learning with imperfect but inexpensive labels is practically significant.

Crowdsourcing [7] relying on non-expert workers has recently emerged as an attractive surrogate. Unlabeled instances are typically assigned to workers of varying knowledge, and limited by their expertise, they often have difficulty recognizing the exact label from multiple ambiguous categories. Therefore, crowdsourcing platforms naturally allow workers to select several possible labels if they are uncertain about an instance. In this way, an instance is associated with a set of candidate labels where a *fixed but unknown* candidate is the true label. A set of candidate labels is referred to as a *partial label* (PL) for an instance, and the learning paradigm that can handle PLs is termed as *partial-label learning* (PLL) [8], [12], [18], [41], [42], [51], [61], [62], [66], [68], [74], [75], also known as ambiguous label learning [9], [10], [22], [67] and superset learning [27], [36], [37]. PLL attempts to infer the optimal multi-class classifier that is able to accurately predict the true label for unseen instances by fitting PLs, and more ideally, the hypotheses can be modeled by DNNs. PLL problems arise in real-world scenarios [36], [40], [70] as well.

Research on PLL dates back about 20 years. Initially, Jin and Ghahramani [30] built up a maximum likelihood model to reassign class-posterior probabilities to candidate labels iteratively. This work opened up a main research route of PLL that purifies each PL on the fly to select the most likely true label during training [10], [16], [17], [30], [61], [66], which is named the *identification-based strategy* (IBS). Because IBS aims at eliminating the ambiguity [74] between individual instances and their true labels in the training phase, this technique is also commonly known as *disambiguating* [12], [30]. Contrariwise, Hüllermeier and Beringer [26] formalized PLL as a collaborative problem, where all candidate labels

- *J. Lv is with RIKEN Center for Advanced Intelligence Project. E-mail: is.jiaqi.lv@gmail.com*
- *G. Niu is with RIKEN Center for Advanced Intelligence Project; he is also with School of Computer Science and Engineering, Southeast University. E-mail: gang.niu.ml@gmail.com*
- *B. Liu, N. Xu, and X. Geng are with School of Computer Science and Engineering, Southeast University, and Key Lab of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing, China. E-mail: liubiao01@seu.edu.cn; xning@seu.edu.cn; xgeng@seu.edu.cn*
- *L. Feng is with College of Computer Science, Chongqing University. E-mail: lfeng@cqu.edu.cn*
- *M. Xu is with School of Information Technology and Electrical Engineering, The University of Queensland. E-mail: miao.xu@uq.edu.au*
- *B. An is with School of Computer Science and Engineering, Nanyang Technological University. E-mail: boan@ntu.edu.sg*
- *M. Sugiyama is with RIKEN Center for Advanced Intelligence Project; he is also with Graduate School of Frontier Sciences, The University of Tokyo. E-mail: sugi@k.u-tokyo.ac.jp*
  Corresponding Authors: Jiaqi Lv

Manuscript received April 19, 2005; revised August 26, 2015.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3275249

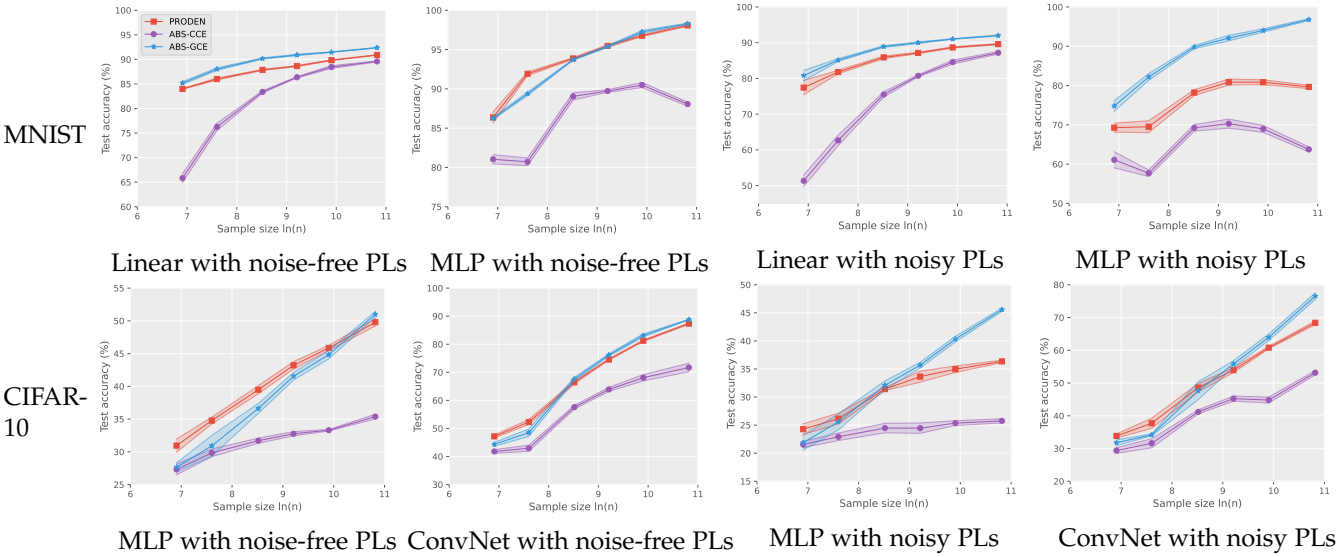

Fig. 1. Comparison of a SOTA IBS method PRODEN [41] and our ABS method with CCE or GCE for PLL under different training dataset sizes. All experiments were repeated 5 times and the standard deviations are presented in shadow. In the upper row, on MNIST, a linear-in-input model (Linear) and a 5-layer multi-layer perceptron (MLP) were trained by stochastic gradient descent [56]. In the bottom row, on CIFAR-10, an MLP-5 and a 12-layer convolutional neural network (ConvNet) [33] were trained by Adam [31]. We can clearly see our ABS method with GCE matches the performance of PRODEN with noise-free PLs, and exceeds PRODEN with noisy PLs, where the true label might not be included in PLs.

contribute to the learning objective *equally*. The idea is that the inductive bias underlying the learning process can benefit disambiguating the given PLs, and let trained models be able to predict the true label of any instance. Such a scheme is called the *average-based strategy* (ABS) [12], [73].

In recent years, the research of PLL has focused on IBS, since it was believed that the performance of IBS is more promising, while little attention has been paid to ABS. This is especially so in the era of deep learning [18], [41], [61]. The pessimism about ABS comes from "memorization" of over-parameterized DNNs, that is, the perfectly fitted DNNs can memorize all training samples, even if their labels are completely arbitrary [15], [71]. ABS is free from identifying the latent true labels during training, and therefore memorize all candidate labels. Then the PLL problem would be degenerated to a multiple-label problem [30], where it is acceptable that an arbitrary candidate label is taken for the "pseudo true label". Then will ABS fail in the true-label prediction and should ABS just be phased out by the times?

In this paper, we argue the research value of ABS by showing its practical potential and theoretical superiority, problematizing the traditional view of ABS and pushing forward PLL as a whole.

Promising experimental findings. Our work is inspired by a set of exploratory experiments. We propose a family of ABS losses named *average partial-label* (APL) losses, which are defined as the average of multi-class losses over all candidate labels. The categorical cross entropy (CCE) loss is the most popular multi-class loss in deep learning nowadays, and we observed that all existing deep IBS methods [18], [41], [61], [67] also adopted the CCE loss. Thus firstly, we trained several standard deep models with the APL loss equipped with the CCE loss on benchmark datasets where true labels were manually corrupted to PLs. In this case, we found that our ABS method with the CCE loss performs poorly as was previously thought, shown in Figure 1. Extending these experiments, we then replaced the CCE loss with another widely-used loss, i.e., the generalized cross entropy (GCE) loss [76], and conducted experiments with early stopping [49]. Surprisingly, we observed that our ABS method with the GCE loss can often be on a par with, or even outperform a SOTA IBS method PRODEN [41], regardless of datasets, models, and optimizers. For the details of data generation processes, please see Section 6.1. These observations challenge common beliefs since one would expect that ABS is incapable of distinguishing true labels, thereby hurting generalization, while the results showed that ABS method with the APL loss can also predict the true label. Therefore, we would like to analyze *what is it that distinguishes ABS methods that perform well from those that do not*, and the answer to this question will hopefully help improve ABS.

Novel robustness analysis. To answer this question, we analyze the *robustness* of our ABS method to PLs, namely, whether the classification error on supervised data of the minimizer of the risk w.r.t. the APL losses is approximated to that of the Bayes classifier (learned using supervised data) [20], [21], [44], [46]. Thanks to the concise form of the APL losses, it is easy to estimate the risk under the APL losses from PLs and carry out *empirical risk minimization* (ERM). Thus we can analyze the robustness through existing mathematical techniques.

Furthermore, we take one more step forward—*unreliable PLL*, which learns from *noisy PLs*, that is, the candidate-label set that *might not* include the true label. As the acquisition of training data expands, noise is inevitable, therefore, it is not an overstatement to say unreliable PLL is imminent in real-world applications. Unfortunately, previous PLL algorithms concentrate on noise-free PLs, and they have not been able to handle the unreliable PLL well, as shown in the two rightmost columns in Figure 1. To avoid confusion, we refer

to the traditional PLL paradigm *reliable PLL*, to reliable PLL and unreliable PLL collectively as "PLL", and to noise-free PL and noisy PL collectively as "PL" in later sections.

To theoretical analyze the cause of success or failure of an ABS method, we formalize two problem settings for the generation processes of PLs, each with several specific instances that differ significantly in their conceptual approaches. With the help of them, we delve into multiple widely-used multi-class loss functions, and formally prove that APL losses with *bounded* loss functions (e.g., GCE) are always robust under mild assumptions on the domination of true labels, while APL losses with *unbounded* loss functions (e.g., CCE) may not be robust. The theoretical results are reconciled with experimental observations in Figure 1. Given that there exists no such analysis for IBS yet, our robustness analysis is novel for not only ABS but also PLL.

ABS improvement to IBS. Moreover, we rethink the existing deep IBS methods. We point out that *all modern IBS methods behave like ABS in the beginning of training* to prepare for PL purification and true-label selection. In other words, they need to use ABS to warm start model training, and use the pretrained model to identify the true label. Thus they will select the correct true labels if ABS can become better, while they have hitherto used unbounded losses throughout the training process. As a consequence, IBS methods can in turn improve upon themselves by our study on ABS: we suggest utilizing bounded losses as a warmup for the first few epochs. We conduct extensive experiments to verify the effectiveness of this improvement.

Contributions. Our contributions can be summarized:

- We establish a theoretically grounded framework for ABS based on a simple yet effective APL loss family, the risk minimization of which is guaranteed to be robust under two problem settings for the data generation processes.
- To the best of our knowledge, we are the first to propose the unreliable PLL paradigm, further developing the practical potentiality of PLL in society. Our ABS method with the APL losses provides an effective baseline for unreliable PLL, and it also works well for reliable PLL without any modifications.
- We redraw the attention of the PLL community to ABS. Our research findings can not only improve ABS, but also enlighten a general principle to incorporate ABS into IBS methods to further enhance their performance, and push forward PLL as a whole.

Paper organization. We recapitulate the related work and discuss the philosophies behind ABS and IBS in Section 2. In Section 3, we give an overview of the problem settings and introduce the importance of robustness analysis. In Section 4, we propose APL losses and formalize generation processes of PLs. We present our main theoretical results and experimental findings in Section 5 and Section 6, respectively. We conclude in Section 7, and defer additional experimental results and all the proofs to the appendix.

2 Related Work

In this section, we review some seminal work in reliable PLL, discuss philosophies behind different technical routes of PLL, and analyze its relation to and difference from other machine learning problems.

Practical reliable PLL. [30] is one of the milestones for reliable PLL. It proposed to disambiguate noise-free PLs by using the expectation-maximization (EM) algorithm. In the E-step, the class-posterior probability is estimated as the normalization of current model predictions. In the M-step, the model parameters are updated in order to minimize the KL divergence between the given estimated probabilities and the model-based distributions. Such a strategy that identifies true labels along with model training is referred to as IBS. Following the milestone work, many IBS algorithms have been developed (e.g., [10], [17], [22], [42], [58], [66], [73], [75]).

Reliable PLL also has been studied along the other research route called ABS, pioneered by Hüllermeier and Beringer [26]. They determined the class label for an unseen instance by voting among the candidate labels of its nearest neighbors. Cour *et al.* [12] proposed a convex loss that distinguishes the averaged output over the candidate labels from outputs over non-candidate labels. It is always believed that ABS is likely to fail as the outputs of pseudo true labels would overwhelm the output of true label. Therefore, the development of the two strategies was not balanced – IBS has been the focus of considerable recent research whilst ABS faded away.

From 2020, deep learning starts injecting new vitality into IBS. Almost at the same time, three works proposed to model classifiers by DNNs. Yao *et al.* [67] adopted ResNet as the backbone together with two specially designed regularizers for partially-labeled image classification. Yao *et al.* [68] used the co-teaching scheme [24] to let two networks interact with each other regarding the confidence levels of the instances. The method proposed by Lv *et al.* [41] progressively identifies true labels based on the memorization effect of DNNs [3], which is flexible on the learning models and loss functions. Later, Feng *et al.* [18] formalized for the first time the generation process of noise-free PLs, based on which they derived two provably consistent algorithms. Wen *et al.* [61] proposed a leveraged weighted loss to trade off the losses on candidate labels and non-candidate ones. Wu and Sugiyama [63] proposed a unified framework includes [18] as a special case. These three works are both compatible with DNNs.

Provable reliable PLL. Although the above practical algorithms have proven empirically successful on specific domains, there is an elusive theoretical gap in the understanding of them. Through the lens of learning theory, some researchers proposed seminal theoretical works in reliable PLL. Liu and Dietterich [37] proposed the *small ambiguity degree condition* to ensure that classification errors on any instance have a probability of being detected. The proof of this theorem requires strict assumptions: the approximation error equals zero and meanwhile the Bayes error equals zero (i.e., the deterministic scenario [48]). Cour *et al.* [12], Feng *et al.* [18], and Wen *et al.* [61] focused on the statistical consistency. They proposed a consistent loss based on some specific data generation process or deterministic scenario assumption, while our findings are general enough to hold under different generation processes and also a stochastic

scenario [48].

Philosophies behind IBS and ABS. IBS iterates between the optimization of a learning model and the identification of the true label. Typically the identified true label has the biggest posterior of all labels, and must be in the candidate-label set. In other words, the "true" one is the "ideal" one. It implies that the true label can be *uniquely determined* given an input—it is satisfied only in the *deterministic scenario* where the class-posterior probability of the true label is equal to 1.

However, the natural world is more like the *stochastic scenario* that possesses some inherent randomness. In this setting, the label is a probabilistic function of the input, indicating that the same input will lead to an ensemble of *unfixed* output labels. ABS essentially gets rid of the deterministic scenario by avoiding recognizing the "ideal" label. The "true" label is considered as an "actually sampled" outcome, and consequently, the philosophy of ABS is compatible with the stochastic scenario. Therefore, it is crucial to design advanced methods and provide theoretical understandings for ABS.

Relevant learning problems. There are some weakly supervised learning problems related to PLL.

- *Complementary-label (CL) learning* [19], [28], [29], [69] learns from weakly-supervised datasets wherein an instance is equipped with a CL. A CL specifies a class that the pattern does NOT belong to, so it can be considered as an extreme noise-free PL case with a fixed number $(k-1)$ of candidate labels. Then from the algorithmic point, reliable PLL algorithms can directly handle the CL learning problem, but not also the other way around.
- *Semi-supervised learning (SSL)* [6], [47], [59], [72], [78] learns from datasets consisting of both labeled and unlabeled data. Since we can regard the universe set of labels as the candidate labels of unlabeled data, SSL has some relation with PLL. However, standard SSL assumes that labeled data are fully supervised, which is different from reliable PLL, where labeled data are still ambiguous.
- *Noisy-label learning (NLL)* [23], [24], [38], [50], [53] learns from noisy supervision where the training data are sampled from a corrupted distribution. Both NLL and PLL should have an underlying transition matrix linking the clean class posterior and the observed class posterior of an instance. Nonetheless, their matrix dimensions are different: the transition matrix is $k \times k$ in NLL and $k \times (2^k - 2)$ in PLL.

3 Preliminaries

In this section, we formally introduce reliable PLL and propose unreliable PLL, and give the definition of robustness.

3.1 Problem Setup

Basic settings. Let us consider a multi-class classification problem of k classes. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the feature space, $\mathcal{Y} = [k] \doteq \{1, 2, \ldots, k\}$ be the label space, and $\mathcal{S} \doteq \{2^{[k]} \backslash \emptyset \backslash \mathcal{Y}\}$ be the *PL space*. $2^{[k]}$ means the collection of all subsets in $[k]$, and $|\mathcal{S}| = 2^k - 2$ because the empty set and the whole label set are excluded. We denote by $p(\boldsymbol{x}, y)$ some probability density of "clean" distribution over $\mathcal{X} \times \mathcal{Y}$. In fully-supervised classification, the goal is a learning model (e.g., a DNN) $f : \mathbb{R}^d \to [k]$ that can make correct prediction on unseen inputs, with a set of i.i.d. supervised training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ sampled from $p(\boldsymbol{x}, y)$. A classifier $f(\boldsymbol{x})$ is routinely assumed to take the following form:

$$f(\boldsymbol{x}) = \arg\max_{i \in \mathcal{Y}} g_i(\boldsymbol{x}),$$

where $g_i(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ outputs a score for class i. In this paper, we concentrate on deep learning: assume the learning model f is a DNN and apply softmax operation to convert scores into a vector of class-posterior probabilities, i.e., $g_i(\boldsymbol{x}) = p(i|\boldsymbol{x}) \in \Delta^{k-1}$ [55], where Δ^{k-1} denotes the k-dimensional simplex.

While in PLL, for the notional clean distribution with probability density $p(\boldsymbol{x}, y)$, we instead observe i.i.d. PL training data $\{(\boldsymbol{x}_i, s_i)\}_{i=1}^n$ from a corrupted version $p(\boldsymbol{x}, s)$ of $p(\boldsymbol{x}, y)$ over $\mathcal{X} \times \mathcal{S}$. The distribution $p(\boldsymbol{x}, s)$ is such that the marginal distribution of instances $p(\boldsymbol{x})$ is unchanged, but the observed label is corrupted to an ambiguous candidate-label set. PLL tries to nonetheless learn the optimal classifier by fitting $\{(\boldsymbol{x}_i, s_i)\}_{i=1}^n$.

The key assumption in reliable PLL is that the PLs are noise-free, which means the latent true label y_i of an instance $\boldsymbol{x}_i$ is always included in its candidate-label set s_i, i.e.,

$$p(y_i \in s_i \mid \boldsymbol{x}_i, s_i) = 1, \ \forall (\boldsymbol{x}_i, y_i) \in p(\boldsymbol{x}, y), \ \forall s_i \in \mathcal{S}.$$

We argue that this assumption is fairly strict since the density $p(\boldsymbol{x}, y)$ of clean distribution is agnostic. For example, annotations of large-scale datasets usually need to be done by distributed workers in crowdsourcing platforms. Requiring crowdsourcing workers to cautiously judge each category to ensure that the correct one must be chosen partially runs counter to the original purpose of reducing labeling costs, and workers are unable to annotate tasks accurately due to the limited knowledge, even when combining labels provided by multiple workers, 100% accuracy is not guaranteed. Another vivid example is medical diagnosis. It is nontrivial for doctors to make an exact judgment on challenging cases with one single exam. Instead, giving several suspected diseases and referring patients to receiving other tests based on those suspicions are less demanding. However, owing to high inter- and intra-observer variability [23], some complicated diseases are less noticeable and often difficult to detect, even for experts, leading to unreliable candidate labels. As the acquisition of training data expands, it is pervasive for label information to be corrupted, but unfortunately, it has never been considered in previous PLL works. Thus we introduce a more general data setting titled *unreliable PLL*:

Definition 1 (unreliable PLL)**.** *Given the joint density $p(\boldsymbol{x}, y, s)$ and its marginal density $p(\boldsymbol{x}, s)$, for any noisy PL data $(\boldsymbol{x}_i, s_i)$ independently sampled from $p(\boldsymbol{x}, s)$, its true label y_i has a probability of $0 \leq \gamma \leq 1$ not being included in the candidate-label set s_i, i.e.,*

$$p(y_i \in s_i \mid \boldsymbol{x}_i, s_i) = 1 - \gamma, \ \forall (\boldsymbol{x}_i, y_i) \sim p(\boldsymbol{x}, y), \ \forall s_i \in \mathcal{S},$$

where γ is called the unreliability rate*. Learning from noisy PL data is called unreliable PLL.*

3.2 Robustness

The ℓ-*risk* of f in fully-supervised learning w.r.t. multi-class loss $\ell : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}^+$ is defined as follows:

$$\mathcal{R}(f;\ell) = \mathbb{E}_{p(\boldsymbol{x},y)}[\ell(f(\boldsymbol{x}),y)].$$

$\mathbb{E}$ denotes the expectation and its subscript indicates the distribution with respect to which the expectation is taken. The *Bayes optimal classifier* that minimizes $\mathcal{R}(f;\ell_{01})$ is given by $f^* = \arg\min_f \mathcal{R}(f;\ell_{01})$, where the optimality is defined over all measurable functions. We denote by $\mathcal{R}^* \doteq \mathcal{R}(f^*)$ the corresponding Bayes risk under the clean distribution. Typically, ℓ is *classification-calibrated* [4], that is, the global minimizer of $\mathcal{R}(f;\ell)$ is the same as that of $\mathcal{R}(f;\ell_{01})$, which can be interpreted as controlling the excess 0-1 risk by controlling the excess ℓ-risk [50].

Denote by $\tilde{\ell} : \mathbb{R}^k \times \mathcal{S} \to \mathbb{R}^+$ a suitably modified ℓ for use with PLs (defined in Section 4.1). Similarly, the *PLL risk* under $p(\boldsymbol{x},s)$ w.r.t. *PLL loss* $\tilde{\ell}$ is defined as

$$\tilde{\mathcal{R}}(f;\tilde{\ell}) = \mathbb{E}_{p(\boldsymbol{x},s)}[\tilde{\ell}(f(\boldsymbol{x}),s)].$$

The aim of PLL is to predict the true label for unseen instances. However, most of the standard learning methods are hard to perform well as they tend to exhibit overfitting on the candidate labels in such scenarios [41].

Constructing *robust* losses from the perspective of the objective function is a powerful means in weakly supervised learning [21], [44], [46]. Its focus is to derive the theoretical guarantee for robust losses so that the learned classifier based on weak supervision approximates the Bayes optimal classifier. Concretely, a loss $\tilde{\ell}$ is robust to PLs (more specifically the risk minimization with $\tilde{\ell}$ is asymptotically robust to PLs) if it guarantees that the *optimal PLL classifier* $\tilde{f}^* = \arg\min_f \tilde{\mathcal{R}}(f;\tilde{\ell})$ converges to the Bayes optimal classifier.

Definition 2 (PL-robustness)**.** *We say that a loss $\tilde{\ell}$ is robust to PL data (PL-robust) if for any $p(\boldsymbol{x},y) \in \mathcal{X}\times\mathcal{Y}$, $\mathcal{R}(\tilde{f}^*) - \mathcal{R}^*$ is bounded.*

$\mathcal{R}(\tilde{f}^*) - \mathcal{R}^*$ is bounded means that $\tilde{f}^*$ learned from PL data has a similar classification error to f^* on the supervised data, i.e., minimizing $\tilde{\mathcal{R}}$ yields an approximate solution that minimizes $\mathcal{R}$. A guarantee of robustness thus sets an analogous calibration theory [4] of PLL. Let $\tilde{\mathcal{R}}^* \doteq \mathcal{R}(\tilde{f}^*)$. Then the robustness condition will often be rewritten as that $\tilde{\mathcal{R}}(f^*) - \tilde{\mathcal{R}}^*$ is bounded, which is slightly weak because it only signifies that f^* is the approximated minimizer of $\tilde{\mathcal{R}}(f^*)$, but does not guarantee the classification performance of $\tilde{f}^*$ on the supervised data.

In statistical learning theory, *consistency* [48] is another important concept. We use the superscript $^\bigstar$ to indicate the optimal solution over a given hypothesis class $\mathcal{F}$, i.e., $f^\bigstar = \arg\min_{f\in\mathcal{F}} \mathcal{R}(f;\ell)$, $\tilde{f}^\bigstar = \arg\min_{f\in\mathcal{F}} \tilde{\mathcal{R}}(f;\tilde{\ell})$. Suppose $\hat{\tilde{f}} = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \tilde{\ell}(f(\boldsymbol{x}_i),s_i)$ is the PLL empirical risk minimizer. The quality of $\hat{\tilde{f}}$ with respect to $f^\bigstar$ is measured by the *estimation error*:

$$\mathcal{R}(\hat{\tilde{f}}) - \mathcal{R}(f^\bigstar) = \underbrace{(\mathcal{R}(\hat{\tilde{f}}) - \mathcal{R}(\tilde{f}^\bigstar))}_{\text{RHS1}} + \underbrace{(\mathcal{R}(\tilde{f}^\bigstar) - \mathcal{R}(f^\bigstar))}_{\text{RHS2}}. \tag{1}$$

If as $n \to \infty$, there is $\mathcal{R}(\hat{\tilde{f}}) \to \mathcal{R}(f^\bigstar)$, we say the PLL is consistent. According to the universal approximation theorem [2], [13] that using a proper DNN, the hypothesis space $\mathcal{F}$ is sufficiently complex to contain the Bayes optimal classifier, we have $f^* = f^\bigstar$, $\tilde{f}^* = \tilde{f}^\bigstar$. Thanks to this, the concepts of robustness and consistency can be well connected in deep learning: RHS2 in Equation (1) is just the robustness measure. Therefore, consistency is a sufficient but not a necessary condition of robustness. Although robustness is a weaker property than consistency, its advantage lies in no need to design an ad-hoc loss for each specific data generation process, which is generally required in the consistent methods. In conclusion, robustness is a common and critical theoretical guarantee in supervised learning, but the mechanism by which it might be achieved remains barely understood in PLL. To the best of our knowledge, this is the first work to analyze the robustness of PLL.

4 Methodology

In this section, we propose a family of APL loss functions for PLs, and introduce the generation processes of PLs.

4.1 A Family of Average PL (APL) Losses

In this paper, we propose a family of loss functions named the average PL (APL) losses following the principled ABS:

$$\tilde{\ell}(f(\boldsymbol{x}),s) = \frac{1}{|s|}\sum\nolimits_{i\in s} \ell(f(\boldsymbol{x}),i), \tag{2}$$

where $|\cdot|$ represents the cardinality. Our learning formulation is built on a simple scheme that combines multiple multi-class losses on the individual candidate. For example, we can use the GCE or CCE loss as the component ℓ. If s is a singleton, the APL loss reduces to the ordinary multi-class loss. The idea of the APL losses comes from a practically motivated process proposed by Feng et al. [18]: they assumed that a candidate-label set is feature-independent and uniformly drawn given a specific true label, i.e., $p(s|y,\boldsymbol{x}) = p(s|y) = \text{const.}$ if $y \in s$, and $p(s|y) = 0$ otherwise. The generation process of noise-free PLs can thus be formalized as $p(s|\boldsymbol{x}) = \sum_y p(s|y)p(y|\boldsymbol{x}) \propto \sum_{y\in s} p(y|\boldsymbol{x})$. Then we could consider replacing the posterior with a loss and obtain $\tilde{\ell}(f(\boldsymbol{x}),s) \propto \sum_{y\in s}\ell(f(\boldsymbol{x}),y)$. It inspires the formula of the APL losses, and the normalization term $1/|s|$ breaks the bias to training data with more candidate labels. The APL losses encourage the larger outputs on candidate labels, while do not explicitly guarantee the true label has the biggest score. Ideally, a"nice" loss ℓ can drive up the output of the true label implicitly resorting to the inductive bias, while a "bad" loss results in an inability to disambiguate. Thus, the issue now is that which multi-class loss functions can bound $\mathcal{R}(\tilde{f}^*) - \mathcal{R}^*$ (or $\tilde{\mathcal{R}}(f^*) - \tilde{\mathcal{R}}^*$), that is, make our ABS method with the APL loss PL-robust.

Let us give an motivating example. $\{z_1, z_2\} \in s$ are two candidate labels of an instance $\boldsymbol{x}$, and z_1 is true. Then the APL loss of f on this sample is $\tilde{\ell}(f(x),s) = \frac{1}{2}[\ell(f(x),z_1) + \ell(f(x),z_2)]$. We would like to increase $g_{z_1}(\boldsymbol{x})$ to get it close to 1 so that $g_{z_2}(\boldsymbol{x})$ is decreased, signifying f successfully remembers the true label without the interference of z_2. Paradoxically, because all candidate labels contribute to

TABLE 1
Bounds of multi-class losses, including the mean absolute error (MAE) loss, the mean square error (MSE) loss, the reverse cross entropy (RCE) loss [60], the generalized cross entropy (GCE) loss, the partially Huberised cross entropy (PCE) loss [46], the categorical cross entropy (CCE) loss, and the focal loss (FL) [35].

Loss function		Bound of loss	Bound of the sum of losses over all classes
MAE	$\ell(f(\boldsymbol{x}),i)=\|\|\boldsymbol{e}^i-f(\boldsymbol{x})\|\|_1$	$0\le\ell(f(\boldsymbol{x}),i)\le 2$	$\sum_{i=1}^k\ell(f(\boldsymbol{x}),i)=2k-2$
MSE	$\ell(f(\boldsymbol{x}),i)=\|\|\boldsymbol{e}^i-f(\boldsymbol{x})\|\|_2^2$	$0\le\ell(f(\boldsymbol{x}),i)\le 2$	$k-1\le\sum_{i=1}^k\ell(f(\boldsymbol{x}),i)\le 2k-2$
RCE	$\ell(f(\boldsymbol{x}),i)=-\sum_{j=1}^k g_j(\boldsymbol{x})\log\boldsymbol{e}_j^i$	$0\le\ell(f(\boldsymbol{x}),i)\le -A, A<0$	$\sum_{i=1}^k\ell(f(\boldsymbol{x}),i)=A-Ak$
GCE	$\ell(f(\boldsymbol{x}),i)=\frac{1-g_i(\boldsymbol{x})^q}{q}$	$0\le\ell(f(\boldsymbol{x}),i)\le\frac{1}{q},\ q\in(0,1]$	$\frac{k-k^{1-q}}{q}\le\sum_{i=1}^k\ell(f(\boldsymbol{x}),i)\le\frac{k-1}{q}$
PCE	$\ell(f(\boldsymbol{x}),i)=\begin{cases}-\tau g_i(\boldsymbol{x})+\log\tau+1, & \text{if } g_i(\boldsymbol{x})\le\frac{1}{\tau},\\ -\log g_i(\boldsymbol{x}) & \text{otherwise}\end{cases}$	$0\le\ell(f(\boldsymbol{x}),i)\le\log\tau+1,\tau>1$	$k\log k\le\sum_{i=1}^k\ell(f(\boldsymbol{x}),i)\le(k-1)(\log\tau+1)$, if $k\le\tau$ $k-\tau+k\log\tau\le\sum_{i=1}^k\ell(f(\boldsymbol{x}),i)\le(k-1)(\log\tau+1)$, if $k>\tau$
CCE	$\ell(f(\boldsymbol{x}),i)=-\log g_i(\boldsymbol{x})$	$\ell(f(\boldsymbol{x}),i)\ge 0$	Unbounded
FL	$\ell(f(\boldsymbol{x}),i)=-(1-g_i(\boldsymbol{x}))^\tau\log g_i(\boldsymbol{x})$	$\ell(f(\boldsymbol{x}),i)>0,\ \tau>0$	Unbounded

minimizing $\tilde{\ell}$, neither $\ell(f(x),z_1)$ nor $\ell(f(x),z_2)$ should be too large. Intuitively, if ℓ has an upper bound, then the value of $\tilde{\ell}$ is acceptable even if $g_{z_2}(\boldsymbol{x})$ is close to 0. But if this is not the case, the optimization algorithm must keep $g_{z_2}(\boldsymbol{x})$ not too small to ensure that $\tilde{\ell}(f(x),z_2)$ is not too small, then memory for the true labels is hindered. The empirical observations in Section 1 also confirm this inference.

We investigate a series of non-negative multi-class loss functions and prove that the APL losses with *bounded* multi-class loss functions [1] are robust to (both noise-free and noisy) PL data in Section 5.

Definition 3. *We say a multi-class loss function is bounded if for any classifier f and input $\boldsymbol{x}\in\mathcal{X}$, the sum of losses over all classes is bounded by certain constants C_1 and C_2:*

$$C_1\le\sum_{i=1}^k\ell(f(\boldsymbol{x}),i)\le C_2. \tag{3}$$

Especially, if $C_1=C_2=C$, i.e., $\sum_{i=1}^k\ell(f(\boldsymbol{x}),i)=C$, the loss function is said to be symmetric.

We should point out that the loss of a bounded loss function on any class is also bounded: $0\le\ell(f(\boldsymbol{x}),i)\le U,\ \forall i\in\mathcal{Y}$, where U is a constant. We examine widely-used loss functions and list their bounds in Table 1. We use a one-hot representation for each label, i.e., if the label $y=i$, its label vector is represented as $\boldsymbol{e}^i$, where the j-th element is given by $\boldsymbol{e}_j^i=1$ if $i=j$, otherwise 0. Then for symmetric losses, $\ell(f(\boldsymbol{x}),j)=C/(k-1),\forall j\neq i$.

4.2 Data Generation Processes

To provide the main insights, we have to propose some assumptions on the data generation processes. In the following, we formally establish two general problem settings for the generation processes of noise-free PLs and noisy PLs, and for each of them we further propose several particular implementations motivated by realistic scenarios. We follow the assumption in prior reliable PLL works [18], [36], [61] and the classical feature-independent model of label noise [24], [45], [50], [53], [64] that the observation is conditionally independent of input given the true label, as a result there are $s\perp\boldsymbol{x}\mid y, p(s|\boldsymbol{x},y)=p(s|y)$. Then the density of corrupted distribution is formulated as $p(\boldsymbol{x},s)=\sum_{y\in\mathcal{Y}}p(s|y)p(\boldsymbol{x},y)$.

4.2.1 Filtered sampling process for noise-free PLs

In a pioneering study involving PLL generation processes, the PL is assumed to be independently and uniformly sampled given a specific true label [18]. It is inspired by a real-world cost-saving application of labeling: without any prior knowledge, the labeling system generates a random PL for each sample and asks the human annotators whether the set contains the true label. Resampling PLs for samples for which the annotators answered "NO". While it would be easy to make the labeling system have some rudimentary knowledge, for example, "salmon" and "spacecraft" do not usually appear in the same set. Then we generalize the uniformly sampling assumption and propose the *Filtered sampling process*. Formally speaking, given a specific true label, a noise-free PL is assumed to be sampled as a whole:

$$p(s|y)=\begin{cases}\eta_s^y & \text{if } y\in s,\\ 0 & \text{if } y\notin s,\end{cases} \tag{4}$$

where $0\le\eta_s^y\le 1$ is the *sampling probability* of the label set s given the true label y, and $\sum_{y\in s}\eta_s^y=1$.

This generation process can be written in the form of a transition matrix [24]. We enumerate all the label sets s in the PL space $\mathcal{S}$ and specify an index l for each set, i.e., $l_i\in\mathcal{S}$ $(i\in[2^k-2])$. By this notation, we summarize all the probabilities into a *PL transition matrix* $\boldsymbol{Q}^{k\times(2^k-2)}$, where $Q_{ij}=p(s=l_j|y=i)$. Further taking into account the assumed data distribution in Equation (4), we can instantiate $\boldsymbol{Q}$ as $Q_{ij}=\eta_{l_j}^i$ if $i\in l_j$, otherwise $Q_{ij}=0$. Then for all $j\in[2^k-2]$, there is $p(s=l_j|\boldsymbol{x})=\sum_{i\in l_j}p(s=l_j|y=$

1. Notice that if the marginal density $p(\boldsymbol{x})$ is compactly supported, given $\sup_{f\in\mathcal{F}}||f||_\infty\le C_f$ where $C_f>0$ and $\mathcal{F}$ is a chosen function class, $\ell(f(\boldsymbol{x}),y)$ with any surrogate loss function ℓ is bounded. In this paper, "bounded" refers to the property of the loss function itself without regard to specific data distributions, function classes, or regularizations.

$i)p(y=i|\boldsymbol{x})$. Thus we have

$$p(\boldsymbol{x},s)=\boldsymbol{Q}^{\top}p(\boldsymbol{x},y), \tag{5}$$

where $^{\top}$ denotes the transpose.

Flipping model. In the real world, however, there are complex and varying correlations between categories, making it common to get some similar categories mixed up, so that some combinations of labels appear more frequently than others. The probability of an incorrect label appearing in the PL of a sample depends on how similar it is to its true label. We therefore propose the *Flipping model* as a means of representing this more realistic situation, where a noise-free PL is supposed to be generated by adding each label to the candidate-label set independently:

$$p(s|y)=M\prod\nolimits_{i\in s}\eta_i^y\prod\nolimits_{i\notin s}(1-\eta_i^y), \tag{6}$$

where

$$\eta_i^y=p(i\in s|y),\forall i\in\mathcal{Y},\quad M=1/\big(1-\prod\nolimits_{i\neq y}\eta_i^y\big).$$

η_i^y is the *flipping probability* that depicts the probability of i-label being included into the candidate-label set given the specific class label y, and $\eta_y^y=1$. For $i\neq y$, it satisfies $0\leq\eta_i^y<1$. M excludes the set whose cardinality equals k by re-sampling.

Similarly, the PL transition matrix can be formulated as $\boldsymbol{Q}^{k\times k}$ where $Q_{ij}=p(j\in s|y=i)=\eta_j^i$ and the diagonal elements of $\boldsymbol{Q}$ are all 1. If $q(\boldsymbol{x})$ is a k-dimension vector where the j-th element $q_j(\boldsymbol{x})$ is the probability $p(j\in s|\boldsymbol{x})$, then

$$\begin{aligned}&q(\boldsymbol{x})=\boldsymbol{Q}^{\top}p(y|\boldsymbol{x}),\\&p(\boldsymbol{x},s)=M\prod\nolimits_{i\in s}q_i(\boldsymbol{x})\prod\nolimits_{j\notin s}(1-q_j(\boldsymbol{x}))p(\boldsymbol{x}).\end{aligned} \tag{7}$$

The Flipping model is a well-established approach for characterizing the data generation process in weakly supervised learning problems represented by NLL. This model emphasizes the commonalities among categories and quantifies the likelihood that an error category is indistinguishable from the true category, parameterized by a class-level transition matrix. In contrast, in PLL, the Filtered sampling process serves as the "rell/complete[2]" Flipping model, as annotations in PLL form a transition matrix at the collection (i.e., PL) level. However, the Filtered sampling process typically fails to decouple the flipping probabilities of individual labels, making it challenging to discern similarities among categories. Consequently, special consideration of the Flipping model remains necessary.

4.2.2 Global sampling process for noisy PLs

As a reminder, the sampling process entails manual filtering of non-conforming label sets and subsequent resampling, leaving room for potential noisy PLs in the event of errors by human annotators. Consequently, all elements within the PL space possess a non-zero probability of being sampled:

$$p(s|y)=\eta_s^y,\ \forall s\in\mathcal{S}, \tag{8}$$

2. Given a sample, the probability of the generation of its corresponding PLs can necessarily be formulated by the Filtered sampling process, but not necessarily by the Flipping model.

where $0\leq\eta_s^y\leq 1$. If the sampling probability for all noise-free PLs is $(1-\gamma)/(2^{k-1}-1)$, and that for all noisy PLs equals $\gamma/(2^{k-1}-1)$, then the sampling probability under the Global sampling process is said to be uniform. In addition, the density $p(\boldsymbol{x},s)$ takes the same form as Equation (5) while even if $i\notin s^j$, Q_{ij} may be larger than 0.

Next we discuss two specific generation patterns of the Global sampling process in terms of how noise-free PLs are contaminated, i.e., the true class is obfuscated by another similar class, or noise-free PL is deliberately corrupted, respectively.

Confusing model. In this type of setting, the true class was (accidentally) confused with other (similar) classes, leading to the misuse of an incorrect label as the original true label in the PL generation process. Therefore, the *Confusing model* consists of two steps.

First, the true label is corrupted. Suppose the class-conditional label noise (CCN) model [24], [45], [50], [53], [64]—the most widely-used model for noisy label classification—is applied, where each instance from class y has a fixed probability of being assigned to label i, that is

$$\bar{y}=\begin{cases}y & \text{with probability } 1-\gamma^y,\\ i,i\in\mathcal{Y},i\neq y & \text{with probability } \bar{\gamma}_i^y,\end{cases} \tag{9}$$

where $0\leq\gamma^y\leq 1$ is the *label noise rate*. The noise is said to be uniform if $\gamma_y=\gamma$ and $\bar{\gamma}_i^y=\gamma/(k-1)$, otherwise it is said to be asymmetric. The corrupting step can be formalized by the *noise transition matrix* $\boldsymbol{T}$ [53], where $T_{ij}=p(\bar{y}=j|y=i)$. Second, the corrupted label $\bar{y}$ serves as the true label y to generate candidate labels, which signifies we require noisy PLs to contain $\bar{y}$, in the same way that noise-free PLs must contain y. Accordingly, the following equation holds:

$$\begin{aligned}p(\boldsymbol{x},s)&=\sum\nolimits_{y,\bar{y}\in\mathcal{Y}}p(s|\bar{y})p(\bar{y}|y)p(y|\boldsymbol{x})p(\boldsymbol{x})\\&=\sum\nolimits_{j=1}^{k}p(s|\bar{y}=j)\underbrace{\sum\nolimits_{i=1}^{k}T_{ij}^{\top}p(y=i|\boldsymbol{x})}_{p(\bar{y}=j|\boldsymbol{x})}p(\boldsymbol{x}),\end{aligned} \tag{10}$$

where $p(s|y)$ can be expanded into the form of Equation (4) or Equation (6).

Destructing model. We believe that noise-free PLs can also be (intentionally) destructed. For example, there may exist spammers who deliberately choose label sets that are totally irrelevant to the tasks. Hence we propose the *Destructing model* that also contains two steps.

Noise-free PLs are first generated by the Filtered sampling process, followed by taking its complement with the *set flipping rate* γ_s:

$$s=\begin{cases}s & \text{with probability } 1-\gamma_s,\\ \bar{s} & \text{with probability } \gamma_s,\end{cases} \tag{11}$$

where $0\leq\gamma_s\leq 1$. Constructing the *complement transition matrix* $\boldsymbol{T}^{(2^k-2)\times(2^k-2)}$, each row (column) of which represents a fixed label set in the PL space. $T_{ij}=\gamma_s$ if the i-th and the j-th label set are complementary to each other, $T_{ii}=1-\gamma_s$ for all $i\in[2^k-2]$, and 0 otherwise. Then the density function of the PL distribution $p(\boldsymbol{x},s)$ is multiplied by $\boldsymbol{T}$ on the basis of Equation (5) or Equation (7).

Similar to the noise-free scenario, the Global sampling process provides a thorough description of the sampling possibilities for all PLs, whereas the Confusing and Destructing models, which are particular instances of the Global sampling process, focus primarily on elucidating the techniques and approaches used to corrupt true labels and generate PLs.

5 Theoretical Results

The studied generation processes of PLs allow us to theoretically understand the properties of the APL losses. In this section, we detail sufficient conditions under which multi-class loss functions make PLL with the APL losses robust in various scenarios, and conclude some instructive findings.

5.1 Robustness to Noise-Free PLs

Lemma 1. *Any symmetric loss satisfies $\tilde{f}^* = f^*$, and any bounded loss satisfies $0 \leq \mathcal{R}(\tilde{f}^*) - \mathcal{R}^* \leq \frac{A'(C_2-C_1)}{A-A'}$, where $A = \frac{1}{2^{k-1}-1}\sum_{j=1}^{k-1}\frac{1}{j}\binom{k-1}{j-1}$ and $A' = \frac{1}{2^{k-1}-1}\sum_{j=2}^{k-1}\frac{1}{j}\binom{k-2}{j-2}$ under the Filtered sampling process with uniform sampling probability.*

Theorem 1. *With the Filtered sampling process, suppose $\mathcal{R}^* = 0$ and $\forall i \neq y$, $\sum_{s:i\in s}\eta_s^y < 1$, then*

(1) for any symmetric loss, $\tilde{f}^* = f^*$;

(2) for any bounded loss, $0 \leq \tilde{\mathcal{R}}(f^*) - \tilde{\mathcal{R}}^* \leq A(C_2 - C_1)$, where $A = \mathbb{E}_{p(\boldsymbol{x},y)}\sum_{i=1}^{k-1}\frac{1}{i}\sum_{s:|s|=i}\eta_s^y$.

We introduce two quantities, A is the expectation of a *weighted sum* of $p(y \in s|y)$ over $p(\boldsymbol{x}, y)$, and A' is the weighted sum of $p(i \in s, i \neq y|y)$. Theorem 1 establishes that under certain conditions, the risk of the Bayes classifier on PL data approaches the minimal PLL risk w.r.t. bounded losses. The tighter the bound of the bounded loss, the more robust the APL loss, and the extreme case is achieved by the symmetric loss which leads to statistical consistency (refer to Section 3.2). The crucial condition for PL-robustness concerns the sampling probability. Since $\sum_{y\in s}\eta_s^y = 1$, the condition $\sum_{s:i\in s}\eta_s^y < 1$, $\forall i \neq y$ signifies that any label other than the true one may not necessarily included in the candidate-label set, i.e., *the domination of true labels*. Another constraint, $\mathcal{R}^* = 0$, implies that classes are separable in fully-supervised classification if the multi-class loss ℓ is classification-calibrated. Note that experimental results later demonstrate that even if this constraint is not satisfied, bounded losses still exhibit good empirical PL-robustness.

In Lemma 1, the uniform sampling probability has already ensured the domination of true labels, and the separability constraint $\mathcal{R}^* = 0$ is eliminated, which implies that even in the stochastic scenario, learning with the APL losses can be PL-robust. Therefore, in this case, the robustness is satisfied without any constraint. In addition, the excess 0-1 risk is bounded, which signifies that the optimal PLL classifier approaches the Bayes classifier, so that the PL-robustness is better guaranteed.

Moreover, if we take into account the shared information among categories the utilize the Flipping model to formalize the generation process, we can obtain the subsequent PL-robustness conditions.

Corollary 2. *With the Flipping model, suppose $\mathcal{R}^* = 0$, then*

(1) for any symmetric loss, $\tilde{f}^* = f^*$;

(2) for any bounded loss, $0 \leq \tilde{\mathcal{R}}(f^*) - \tilde{\mathcal{R}}^* \leq MA(C_2 - C_1)$, where $A = \mathbb{E}_{p(\boldsymbol{x},y)}[\prod_{i\neq y}(1-\eta_i^y) + \frac{1}{2}\sum_{i=1,i\neq y}^{k}\eta_i^y\prod_{j\neq y,i}(1-\eta_j^y) + \frac{1}{3}\sum_{i=1,i\neq y}^{k-1}\sum_{j=i+1,j\neq y}^{k}\eta_i^y\eta_j^y\prod_{m\neq y,i,j}(1-\eta_m^y) + \ldots + \frac{1}{k-1}\sum_{i=1,i\neq y}^{k}(1-\eta_i^y)\prod_{j\neq y,i}\eta_j^y]$.

If $\eta_i^y = \eta$ then, for any symmetric loss, $\tilde{f}^* = f^*$; for any bounded loss, $0 \leq \mathcal{R}(\tilde{f}^*) - \mathcal{R}^* \leq \frac{A'(C_2-C_1)}{A-A'}$, where $A = \sum_{j=1}^{k-1}\frac{1}{j}\binom{k-1}{j-1}\eta^{j-1}(1-\eta)^{k-j}$ and $A' = \sum_{j=2}^{k-1}\frac{1}{j}\binom{k-2}{j-2}\eta^{j-1}(1-\eta)^{k-j}$.

Comparing this corollary versus Theorem 1, we observe that their sufficient conditions and upper bounds on the risk difference differ in notation, but are conceptually similar. Notably, the flipping probability is not constrained due to the diagonal-dominance of the PL transition matrix, which ensures the prevalence of true labels.

5.2 Robustness to Noisy PLs

In this section, we present formal statements for the Global sampling process, and provide intuitive statements for the specific models of noisy PLs, as articulated in Corollary 4, with more formal mathematical details deferred to Appendix A.

Lemma 2. *Any symmetric loss satisfies $\tilde{f}^* = f^*$, and any bounded loss satisfies $0 \leq \mathcal{R}(\tilde{f}^*) - \mathcal{R}^* \leq \frac{A'(C_2-C_1)}{A-A'}$, where $A = \frac{1-\gamma}{2^{k-1}-1}\sum_{j=1}^{k-1}\frac{1}{j}\binom{k-1}{j-1}$ and $A' = \frac{1-\gamma}{2^{k-1}-1}\sum_{j=2}^{k-1}\frac{1}{j}\binom{k-2}{j-2} + \frac{\gamma}{2^{k-1}-1}\sum_{j=1}^{k-1}\frac{1}{j}\binom{k-2}{j-1}$ under the Global sampling process with uniform sampling probability, if label noise rate $\gamma < 1/2$.*

Theorem 3. *With the Global sampling process, suppose $\mathcal{R}^* = 0$ and the domination relations hold: $d(y) > d(j)\ \forall j \neq y$ where $d(\cdot)$ is defined as $d(i) = \mathbb{E}_{p(\boldsymbol{x},y)}\sum_{j=1}^{k-1}\frac{1}{j}\sum_{s:|s|=j,i\in s}\eta_s^y, \forall i \in \mathcal{Y}$, then*

(1) for any symmetric loss, $\tilde{f}^* = f^*$;

(2) for any bounded loss, $0 \leq \tilde{\mathcal{R}}(f^*) - \tilde{\mathcal{R}}^* \leq d(y)(C_2 - C_1)$.

In Theorem 3, the domination relations mean that the expectation of the weighted sum of $p(y \in s|y)$ is larger than that of $p(i \in s|y)$ for any incorrect label $i \neq y$ over $p(\boldsymbol{x}, y)$. The condition for robustness to noisy PLs tightens the constraint on the dominance relationship compared to noise-free PLs, from the sum of probabilities to the weighted sum of probabilities. Whereas, with the uniform sampling probability, the constraint is relatively loose: the probability of outputting noise-free PLs exceeds fifty percent, requiring only slightly more domain knowledge than a completely random labeling system.

To encompass a variety of cases, we summarize the Confusing and Destructing models in the following corollary and then discuss each in detail.

Corollary 4. *With the Confusing model or the Destructing model, suppose $\mathcal{R}^* = 0$ and the domination relations hold: the expectation of the weighted sum of $p(y \in s|y)$ is always larger than that of $p(i \in s|y), \forall i \neq y$ over $p(\boldsymbol{x}, y)$, then*

(1) for any symmetric loss, $\tilde{f}^* = f^*$;

(2) for any bounded loss, $0 \leq \tilde{\mathcal{R}}(f^*) - \tilde{\mathcal{R}}^* \leq A(C_2 - C_1)$, where A is a constant associated with $p(y \in s|\boldsymbol{x})$.

For the Confusing model, we assume uniform cases in the generation of candidate labels to simplify the process. This

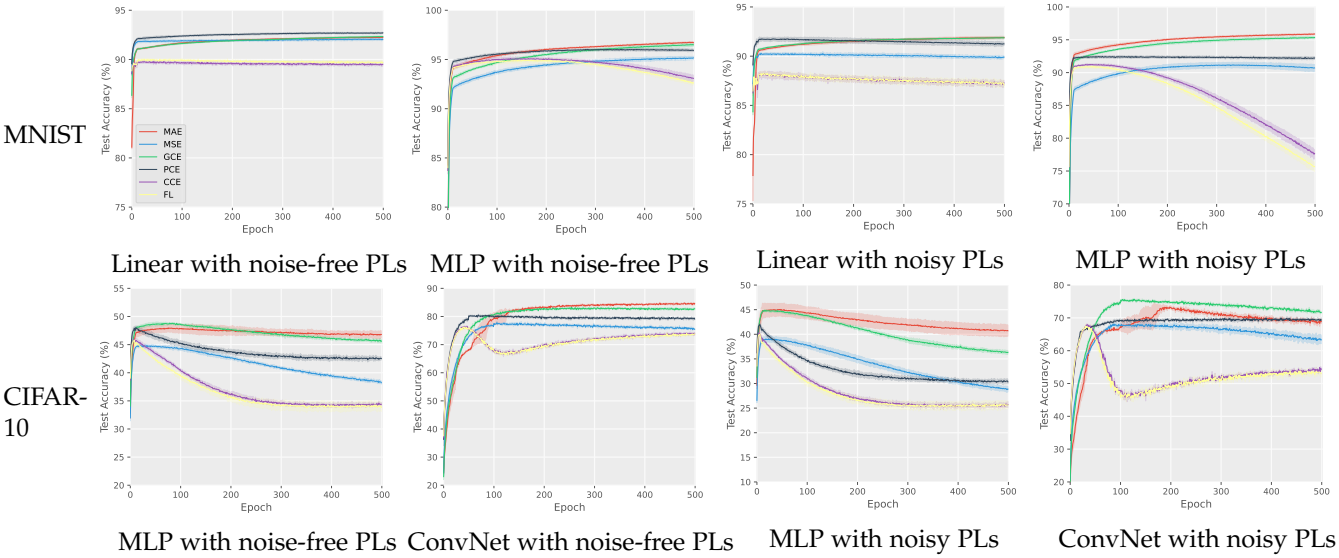

Fig. 2. Test accuracy of our ABS method with bounded versus unbounded losses on benchmarks.

simplification degenerates the domination relations to $\bar{\gamma}_i^y < 1 - \gamma^y,\ \forall i \neq y$, which we were a little surprised to find that it becomes identical to the condition for asymmetric-noise-tolerance (Theorem 3 in [20]). Further supposing that the true labels are corrupted uniformly, we have $\gamma < (k-1)/k$. This is the same as the noise-tolerant condition under uniform noise (Theorem 1 in [20]). These findings indicate that the robustness condition to noisy PLs is *exclusively determined* by the label noise rate as long as the candidate labels are generated uniformly. Moreover, the constraint $\mathcal{R}^* = 0$ is removed and $\mathcal{R}(\tilde{f}^*) - \mathcal{R}^*$ is bounded in this case.

Next we probe into the Destructing model through similar renderings. Assuming uniform generation of candidate labels, the domination relations are reduced to $\gamma_s < 1/k,\ \forall s$. If the probability of destructing every candidate-label set is also uniform, the set flipping rate is also the unreliability rate, namely $\gamma_s = \gamma,\ \forall s$. Then we again show that the PL-robustness condition is solely dependent on the level of unreliability in the candidate-label set.

Remarks. Our theoretical analysis on the APL losses for learning with PL data reveals several critical conditions for achieving PL-robustness. We now make several observations about all the theorems.

- The key factor that makes the APL losses PL-robust is consistent across all scenarios: the weighted sum of the probabilities of the true labels associated with the instances dominates;
- For noisy PL data, if the candidate labels are generated uniformly, the robustness condition is exclusively determined by the degree of unreliability rate;
- In the uniform case, the excess 0-1 risk w.r.t. the clean distribution of the optimal PLL classifier can be upper bounded in proportion to the difference between the upper and lower bounds of the loss function $C_2 - C_1$. A smaller difference implies a smaller upper bound on the excess risk, implying a more robust APL loss, and then the most desirable situation (statistical consistency) can be achieved by the symmetric losses. Its constant of proportionality increases with the probability of incorrect labels, accord with the intuition that higher stochasticity makes learning difficult;
- In the more general cases, when learning with a bounded loss, we can only bound the excess $\tilde{\ell}$-risk under the corrupted distribution, whose upper bound, up to a constant less than 1, equals to $C_2 - C_1$.

The above theoretical findings provide guidance to the design of losses of ABS. Note that IBS is heuristic and not really ERM-based (refer to Section. 6.3), and on this account, the robustness of IBS could hardly be proven.

5.3 Estimation Error Bound

Let us review the relation between robustness and consistency again. Now we have proved the condition that bounds RHS2 in Equation (1) (assuming $\mathcal{F}$ is instantiated to be a DNN). Then we establish the estimation error bound and show that as the number of training data approaches infinity, RHS1 is also bounded.

Suppose $\mathcal{G}_y$ be a class of real functions, and $\mathcal{F} = \oplus_{y\in[k]}\mathcal{G}_y$ be a k-valued function class. Assume there are $C_f > 0$ and $C_\ell > 0$ such that $\sup_{f\in\mathcal{F}} ||f||_\infty \leq C_f$ and $\sup_{x\in\mathcal{X}, f\in\mathcal{F}, y\in\mathcal{Y}} \ell(f(x), y) \leq C_\ell$, and assume $\ell(f(x), y)$ is ρ-Lipschitz continuous for all $||f||_\infty \leq C_f$. The *Rademacher complexity* of $\mathcal{G}_y$ over $p(x)$ with sample size n is defined as $\mathfrak{R}_n(\mathcal{G}_y)$ [5], [48]. Then we have the following estimation error bound.

Theorem 5. *For any $\delta > 0$, we have with probability at least $1 - \delta$,*

$$\tilde{\mathcal{R}}(\hat{\tilde{f}}) - \tilde{\mathcal{R}}(\tilde{f}^\bigstar) \leq 4\sqrt{2}k\rho \sum_{y=1}^{k} \mathfrak{R}_n(\mathcal{G}_y) + 2C_\ell \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{12}$$

As $n \to \infty$, $\mathfrak{R}_n(\mathcal{G}_y) \to 0$ for all parametric models with a bounded norm such as DNNs trained with weight decay [39], which signifies $\tilde{\mathcal{R}}(\hat{\tilde{f}}) \to \tilde{\mathcal{R}}(\tilde{f}^\bigstar)$.

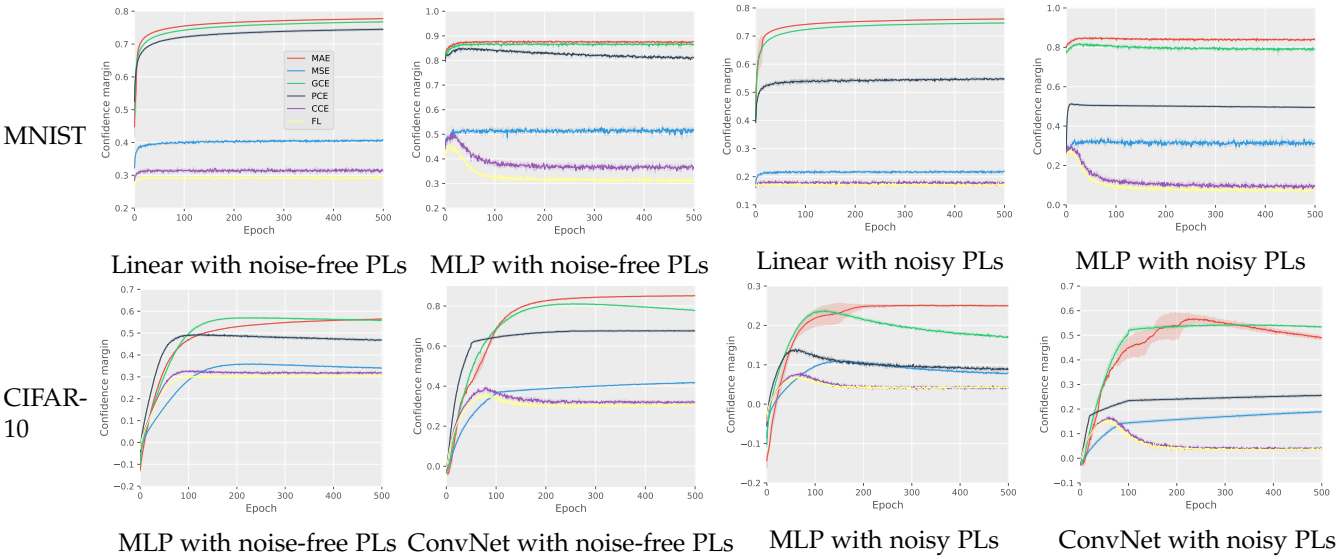

Fig. 3. Confidence margin of our ABS method with bounded versus unbounded losses on benchmarks.

6 Experimental Findings

In this section, we provide some empirical understandings of our ABS method, experimentally validating our theoretical findings on benchmark datasets, which then enlighten an improvement of IBS methods. The implementation is based on PyTorch [52] and experiments were carried out with NVIDIA Tesla V100 GPU.

6.1 Empirical Understanding of APL Losses

Our ABS method with bounded loss functions are robust. We first run a set of experiments on MNIST [34] and CIFAR-10 [32] to verify whether our ABS method with bounded losses is robust to both noise-free PLs and noisy PLs, whereas they are not with unbounded losses. We generate noise-free PLs by the Flipping model with a uniform flipping probability of 0.1, and then noisy PLs by the Confusing model with $\gamma = 0.3$. On each dataset, we train two networks using the APL losses with different multi-class losses, e.g., bounded versus unbounded ones in Table 1. The RCE loss is, up to a constant of proportionality, equivalent to the MAE loss and omitted. We set the focusing parameter 0.5 for the FL loss. Detailed settings are in Section 6.2.

The test accuracy with different losses is presented in Figure 2. As we have theoretically proved, learning with bounded losses is robust: after reaching a peak in test accuracy, their test accuracy is relatively flat throughout the training process. However, unbounded losses exhibit significant overfitting in most cases. Specifically, the symmetric loss MAE has the smoothest curve, while other bounded losses would be slightly overfitting in difficult learning scenarios. The same results are shown across different datasets, under different data settings, with different models. In general, the more difficult the learning scenario is (e.g., harder datasets and weaker supervised information), the larger the gap between bounded loss and unbounded loss is, because unbounded losses overfit more severely.

How do the models fit the candidate labels when training with the APL losses? As we discussed before, ABS methods are free from identifying the true labels during training. One may wonder how ABS methods learn from PL data. This raises the question: is the learned model able to identify the true label of the training samples? We investigate this problem by looking at the *confidence margin* between the model's output on the true label and the maximum output on the other labels, i.e., $\text{confidence} - \text{margin}(x_i) = g_{y_i}(x_i) - \max_{j \neq y_i} g_j(x_i)$. The larger the margin is, the greater the likelihood that the model will successfully identify the true label for the training samples is. In Figure 3, we illustrate the mean confidence margin over the training set. We find that the margins trained with bounded losses are generally much higher than those trained with unbounded losses. This means that although our ABS method does not explicitly disambiguate the candidate-label sets during the training phase, our ABS method with bounded losses is still able to robustly fit the true labels against the interference of other candidates, thereby explaining the good prediction performance in the test set.

6.2 Evaluation on Benchmark Datasets

Setup. Experiments were conducted on four widely-used benchmark datasets including MNIST, Fashion-MNIST [65], Kuzushiji-MNIST [11], and CIFAR-10. On each dataset, we generated PLs by (Case 1) the Filtered sampling process with a uniform sampling probability; (Case 2) the Flipping model with a uniform flipping probability of 0.1; (Case 3) the Confusing model where the label noise rate equals 0.3 and the candidate labels were generated according to Case 2; (Case 4) the Destructing model where the candidate labels were generated according to Case 1 and the set flipping rate equals 0.3. We leave out 10% of the corrupted training samples as a validation set, which is for model selection. Further experiments on the generation processes with different parameters can be found in Appendix C.

We employed various base models including a linear-in-input model (Linear), a 5-layer perceptron (MLP), and a 12-layer convolutional neural network (ConvNet) [25]. Linear was trained on MNIST-like datasets, ConvNet was trained

TABLE 2
Means±standard deviations of test accuracy in percentage with different data generations.

Dataset	Model	Case	Bounded				Unbounded	
			MAE	MSE	GCE	PCE	CCE	FL
MNIST	Linear	1	**91.17 ± 0.07**	87.61±0.07	90.40±0.12	86.28±0.40	85.13±0.29	85.10±0.24
		2	92.25±0.08	92.03±0.10	92.33±0.06	**92.71 ± 0.08**	89.51±0.17	89.71±0.13
		3	**91.90 ± 0.13**	89.90±0.10	91.86±0.10	91.25±0.19	87.28±0.27	87.08±0.27
		4	**90.37 ± 0.19**	82.70±0.49	85.87±0.48	80.61±0.41	67.78±1.19	67.63±0.84
	MLP-5	1	**94.44 ± 0.19**	82.28±0.83	92.77±0.21	87.62±0.27	84.80±0.50	84.15±0.73
		2	**96.71 ± 0.05**	95.16±0.19	96.49±0.15	95.92±0.13	93.08±0.30	92.62±0.18
		3	**95.87 ± 0.85**	80.71±0.52	95.34±0.19	91.14±0.25	77.68±0.53	75.66±0.74
		4	**93.18 ± 0.47**	89.12±1.03	89.19±0.47	86.52±0.82	77.57±1.32	76.45±0.91
Fashion-MNIST	Linear	1	81.50±1.82	80.23±0.29	**82.50 ± 0.21**	79.85±0.38	77.92±0.49	77.56±0.58
		2	81.08±0.02	83.59±0.24	**84.72 ± 0.12**	84.35±0.12	81.88±0.36	81.91±0.09
		3	83.55±2.42	81.12±0.30	**84.25 ± 0.09**	81.60±0.28	79.36±0.24	79.37±0.31
		4	76.40±0.50	73.87±0.37	**78.52 ± 0.47**	61.50±0.67	43.83±0.38	40.39±0.93
	MLP	1	**86.37 ± 0.64**	81.26±0.47	84.29±0.10	76.01±0.49	51.29±0.83	49.36±0.48
		2	**87.52 ± 0.09**	84.49±0.27	87.32±0.31	85.47±0.05	74.10±0.64	73.50±0.24
		3	**85.74 ± 0.27**	81.62±0.33	84.42±0.37	80.33±0.51	53.09±0.74	52.75±0.81
		4	**82.41 ± 0.25**	73.49±0.53	66.83±0.87	65.73±0.87	25.76±0.38	22.01±0.53
Kuzushiji-MNIST	Linear	1	63.14±0.14	59.65±0.62	**65.58 ± 0.32**	58.78±0.46	54.90±0.61	55.02±0.48
		2	64.35±0.09	**68.54 ± 0.31**	65.04±0.20	67.40±1.93	63.79±0.34	63.74±0.44
		3	63.55±0.22	65.26±0.21	64.24±0.29	**67.75 ± 0.17**	59.82±0.36	60.16±0.73
		4	**60.34 ± 0.67**	56.16±0.80	51.98±0.44	51.47±0.80	36.48±0.59	35.76±0.31
	MLP	1	**81.72 ± 0.50**	64.60±1.06	75.71±0.15	57.17±0.22	36.07±0.16	26.13±0.46
		2	**86.99 ± 0.25**	74.24±0.13	86.59±0.24	80.90±0.41	66.06±0.51	64.81±0.41
		3	78.86±2.25	69.90±0.61	**79.26 ± 0.54**	72.89±0.18	56.43±0.88	54.96±0.64
		4	**66.61 ± 0.29**	54.93±0.18	43.06±0.44	42.58±0.43	23.93±0.44	23.35±0.63
CIFAR-10	MLP	1	**41.46 ± 0.72**	33.11±0.69	31.98±0.40	31.10±0.42	22.47±0.49	22.75±0.28
		2	**48.69 ± 0.25**	38.99±0.42	46.00±0.30	42.94±0.44	36.81±0.22	36.37±0.35
		3	**40.83 ± 0.48**	29.05±0.49	35.30±0.30	30.90±0.49	28.04±0.34	27.31±0.18
		4	**31.62 ± 0.60**	25.52±0.55	20.75±0.35	21.10±0.37	13.51±0.34	13.70±0.35
	ConvNet	1	66.86±1.63	61.67±0.31	**68.65 ± 0.71**	34.93±0.59	33.67±0.79	34.17±0.89
		2	**86.65 ± 0.13**	75.47±0.11	82.74±0.29	79.11±0.32	72.82±0.61	73.83±0.21
		3	68.88±0.65	63.30±1.16	**71.65 ± 0.27**	69.35±0.65	53.13±0.63	52.56±1.14
		4	**51.96 ± 4.21**	49.58±0.62	43.83±0.36	43.51±0.61	17.29±0.30	17.02±0.14

on CIFAR-10, and MLP was trained on all datasets. The optimizer was stochastic gradient descent with momentum 0.9. We trained each model 500 epochs with the mini-batch size set to 256, and recorded the test accuracy of the hyper-parameters (learning rate and weight decay) with the best validation accuracy. We did not use any manual learning rate decay and early stopping.

Results. Tables 2 shows the test accuracy over 5 trials. The best and comparable methods based on the paired t-test at the significance level 5% were highlighted in boldface. We can see that the bounded loss always outperforms the unbounded loss, especially on the complex models. In difficult scenarios, i.e., unreliable PLL, the accuracy of the complex models trained with bounded losses is almost always better than their linear counterpart, but unbounded losses can make the complex models overfit very badly on some tasks, causing their performance to become worse.

TABLE 3
PRODEN & Our ABS method on CIFAR datasets. E and A stand for early stopping and robust warm start, respectively. The best and equivalent based on the paired t-test at the significance level 5% are shown in boldface by comparing the 1st and 3rd columns, the 2nd and 4th columns. "—" means that we skipped the experiments under the reliable PLL setting. The best combination is underline.

Dataset	Model	Case	E:	X	✓	X	✓	Our ABS
			A:	X	X	✓	✓	w/ E
CIFAR-10	MLP	1		48.28±0.62	48.33±0.54	**48.77 ± 0.26**	**48.54 ± 0.36**	—
		2		51.49±0.30	52.17±0.21	**52.31 ± 0.32**	**52.70 ± 0.14**	—
		3		38.52±0.16	44.27±0.13	**39.00 ± 0.38**	**46.67 ± 0.30**	45.86±0.14
		4		29.19±0.63	34.04±0.56	**29.30 ± 0.43**	**35.12 ± 0.42**	34.54±0.36
	ConvNet	1		85.38±0.11	85.52±0.25	**85.93 ± 0.31**	**86.09 ± 0.20**	—
		2		88.62±0.19	88.29±0.33	**89.42 ± 0.17**	**89.05 ± 0.29**	—
		3		64.59±0.70	73.05±0.31	**66.86 ± 0.72**	**75.99 ± 0.39**	76.72±0.12
		4		47.35±0.47	53.47±0.38	**50.77 ± 0.82**	**56.68 ± 0.49**	55.50±0.24
CIFAR-100	ConvNet	2		54.49±0.46	59.90±0.53	**58.31 ± 0.32**	**60.60 ± 0.22**	—
		3		34.78±0.31	42.04±0.22	**36.27 ± 0.43**	**43.67 ± 0.22**	44.75±0.15
		4		37.37±0.32	47.29±0.31	**39.55 ± 0.24**	**47.86 ± 0.20**	46.13±0.24

6.3 Enhancing IBS Methods with ABS

We revisit the SOTA IBS methods PRODEN [41], RC [18], and LW [61]. Their typical learning objective is as follows:

$$\hat{\mathcal{R}}(f;\ell) = \frac{1}{n}\sum_{i=1}^{n}\big[\sum_{j\in s_i} w_j^i \ell(g_j(\boldsymbol{x}_i), j) + \sum_{j\notin s_i} w_j^i \ell(g_j(\boldsymbol{x}_i), j)\big],$$

where w_j is a weight for $j \in [k]$. The weights of the labels that are more likely to be true are progressively increased. Generally speaking, they initialize the weights to be

$$w_j^i = 1/|s_i|,\ \forall j \in s_i.$$

They train a learning model f with the uniform weights for several epochs for a warm start, and then update w and f seamlessly for the remaining epochs. We highlight that uniform weights are necessary to break the circular dependency existing between w and f: f needs to be trained with reasonable w and w needs to be estimated by well-trained f. The success of the algorithms is built on the observation that even if each sample has multiple candidate labels, f will remember the true one first [3]. Thus they adjust w by the output of f. It indicates that an IBS method has to be pretrained a little in an ABS manner.

While we note that they always use the CCE loss in both the ABS-style phase and the subsequent IBS-style phase, which could potentially select incorrect true labels at the very beginning and negatively affect the model training. Therefore, we introduce an enhanced principle to incorporate our theoretical findings into existing IBS methods: training the learning model with *a robust warm start* to avoid overfitting. We replace their loss functions of the first 20 epochs with the APL loss with MAE, and then switch back to their original objective function. The hyper-parameters are tuned according to the original methods.

We considered Case 1, 2, 3, and 4 for CIFAR-10, and Case 2, 3, and 4 where the candidate labels were generated by the Flipping model with a uniform flipping probability of 0.01 for CIFAR-100 [32]. We used the same training/validation setting, models, and optimizer as in Section 6.2. We summarize the results without/with early stopping of PRODEN in Table 3, which means that we report the last epoch or the epoch in which the best validation accuracy was reached during training. The results of RC and LW are put in Appendix C.

From Table 3, we can see that the enhanced method with the robust warm start has significant performance improvements over its original version. The model has better performance when early stopping is not deployed, suggesting that the robust warm start helps not to remember incorrect labels at an early stage. Even after using early stopping, our enhanced version also allows for further performance improvements. Moreover, we presented the results of our ABS method with early stopping in the last column. It is usually comparable with the highest accuracy, meaning that our proposal is a simple yet effective baseline for unreliable PLL. The results on CIFAR-10 are also shown in Figure 1.

7 Conclusion

In this paper, we rethought the forgotten ABS in the era of deep PLL, and improved it theoretically and practically. Theoretically, we proposed two problem settings with different data generation models for noise-free and noisy PLs, and analyzed the conditions that ABS is robust to PLs, which filled the theoretical gap in the robustness analysis of PLL. Practically, we conducted extensive experiments to confirm our theoretical findings, and showed that the IBS methods could be improved from our work, which pushed forward PLL as a whole.

Acknowledgments

BL and XG was supported by the National Key Research and Development Plan of China (No.2018AAA0100104), and

the National Natural Science Foundation of China (62125602, 62076063). LF was supported by the National Natural Science Foundation of China (Grant No. 62106028) and CAAI-Huawei MindSpore Open Fund. NX was supported by China Postdoctoral Science Foundation (2021M700023), Jiangsu Province Science Foundation for Youths (BK20210220). GN and MS were supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, Japan. MS was also supported by the Institute for AI and Beyond, UTokyo.

REFERENCES

[1] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proceedings of 36th International Conference on Machine Learning (ICML'19)*, 2019, pp. 242–252.

[2] M. Anthony and P. Bartlett, *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

[3] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. C. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," in *Proceedings of 34th International Conference on Machine Learning (ICML'17)*, vol. 70, 2017, pp. 233–242.

[4] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, p. 138–156, 2006.

[5] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. 11, pp. 463–482, 2002.

[6] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems 32 (NeurIPS'19)*, 2019, pp. 5050–5060.

[7] D. C. Brabham, "Crowdsourcing as a model for problem solving: An introduction and cases," *Convergence*, vol. 14, no. 1, pp. 75–90, 2008.

[8] V. Cabannes, A. Rudi, and F. R. Bach, "Structured prediction with partial labelling through the infimum loss," in *Proceedings of 37th International Conference on Machine Learning (ICML'20)*, 2020, pp. 1230–1239.

[9] C. Chen, V. M. Patel, and R. Chellappa, "Learning from ambiguously labeled face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 7, pp. 1653–1667, 2017.

[10] Y. Chen, V. M. Patel, J. K. Pillai, R. Chellappa, and P. J. Phillips, "Dictionary learning from ambiguously labeled data," in *Proceedings of 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'13)*, 2013, pp. 353–360.

[11] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical japanese literature," *arXiv preprint arXiv:1812.01718*, 2018.

[12] T. Cour, B. Sapp, and B. Taskar, "Learning from partial labels," *Journal of Machine Learning Research*, vol. 12, no. 5, pp. 1501–1536, 2011.

[13] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp. 303–314, 1989.

[14] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, "Gradient descent finds global minima of deep neural networks," in *Proceedings of 36th International Conference on Machine Learning (ICML'19)*, 2019, pp. 1675–1685.

[15] V. Feldman and C. Zhang, "What neural networks memorize and why: Discovering the long tail via influence estimation," *arXiv preprint arXiv:2008.03703*, 2020.

[16] L. Feng and B. An, "Partial label learning by semantic difference maximization," in *Proceedings of 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*, 2019, pp. 2294–2300.

[17] ——, "Partial label learning with self-guided retraining," in *Proceedings of 33rd AAAI Conference on Artificial Intelligence (AAAI'19)*, 2019, pp. 3542–3549.

[18] L. Feng, J. Lv, B. Han, M. Xu, G. Niu, X. Geng, B. An, and M. Sugiyama, "Provably consistent partial-label learning," in *Advances in Neural Information Processing Systems 33 (NeurIPS'20)*, 2020, pp. 10 948–10 960.

[19] Y. Gao and M. Zhang, "Discriminative complementary-label learning with weighted loss," in *Proceedings of 36th International Conference on Machine Learning (ICML'21)*, vol. 139, 2021, pp. 3587–3597.

[20] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of 31st AAAI Conference on Artificial Intelligence (AAAI'17)*, 2017, pp. 1919–1925.

[21] A. Ghosh, N. Manwani, and P. Sastry, "Making risk minimization tolerant to label noise," *Neurocomputing*, vol. 160, pp. 93–107, 2015.

[22] C. Gong, T. Liu, Y. Tang, J. Yang, J. Yang, and D. Tao, "A regularization approach for instance-based superset label learning," *IEEE Transactions on Cybernetics*, vol. 48, no. 3, pp. 967–978, 2018.

[23] B. Han, Q. Yao, T. Liu, G. Niu, I. W. Tsang, J. T. Kwok, and M. Sugiyama, "A survey of label-noise representation learning: Past, present and future," *arXiv preprint arXiv:2011.04406*, 2020.

[24] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems 31 (NeurIPS'18)*, 2018, pp. 8527–8537.

[25] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017, p. 4700–4708.

[26] E. Hüllermeier and J. Beringer, "Learning from ambiguously labeled examples," *Intelligent Data Analysis*, vol. 10, no. 5, pp. 419–439, 2006.

[27] E. Hüllermeier and W. Cheng, "Superset learning based on generalized loss minimization," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2015, pp. 260–275.

[28] T. Ishida, G. Niu, W. Hu, and M. Sugiyama, "Learning from complementary labels," in *Advances in Neural Information Processing Systems 30 (NeurIPS'17)*, 2017, pp. 5639–5649.

[29] T. Ishida, G. Niu, A. K. Menon, and M. Sugiyama, "Complementary-label learning for arbitrary losses and models," in *Proceedings of 36th International Conference on Machine Learning (ICML'19)*, 2019, pp. 2971–2980.

[30] R. Jin and Z. Ghahramani, "Learning with multiple labels," in *Advances in Neural Information Processing Systems 16 (NeurIPS'03)*, 2003, pp. 921–928.

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

[32] A. Krizhevsky and G. Hinton, *Learning multiple layers of features from tiny images*. Citeseer, 2009.

[33] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proceedings of 5th International Conference on Learning Representations (ICLR'17)*, 2017.

[34] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[35] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of 16th IEEE International Conference on Computer Vision (ICCV'17)*, 2017, pp. 2980–2988.

[36] L. Liu and T. G. Dietterich, "A conditional multinomial mixture model for superset label learning," in *Advances in Neural Information Processing Systems 25 (NIPS'12)*, 2012, pp. 548–556.

[37] ——, "Learnability of the superset label learning problem," in *Proceedings of 31st International Conference on Machine Learning (ICML'14)*, 2014, pp. 1629–1637.

[38] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 447–461, 2015.

[39] N. Lu, G. Niu, A. K. Menon, and M. Sugiyama, "On the minimal supervision for training any binary classifier from only unlabeled data," in *Proceedings of 7th International Conference on Learning Representations (ICLR'19)*, 2019.

[40] J. Luo and F. Orabona, "Learning from candidate labeling sets," in *Advances in Neural Information Processing Systems 23 (NeurIPS'10)*, 2010, pp. 1504–1512.

[41] J. Lv, M. Xu, L. Feng, G. Niu, X. Geng, and M. Sugiyama, "Progressive identification of true labels for partial-label learning," in *Proceedings of 37th International Conference on Machine Learning (ICML'20)*, 2020, pp. 6500–6510.

[42] G. Lyu, S. Feng, T. Wang, C. Lang, and Y. Li, "Gm-pll: Graph matching based partial label learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 2, pp. 521–535, 2021.

[43] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. V. D. Maaten, "Exploring the limits of weakly supervised pretraining," in *Proceedings of 15th European Conference on Computer Vision (ECCV'18)*, 2018, pp. 181–196.

[44] N. Manwani and P. S. Sastry, "Noise tolerance under risk minimization," *IEEE Transactions on Cybernetics*, vol. 43, pp. 1146–1151, 2013.

[45] A. Menon, b. Van Rooyen, C. Ong, and B. Williamson, "Learning from corrupted binary labels via class-probability estimation," in *Proceedings of 32nd International Conference on Machine Learning (ICML'15)*, 2015, pp. 125–134.

[46] A. K. Menon, A. S. Rawat, S. J. Reddi, and S. Kumar, "Can gradient clipping mitigate label noise?" in *Proceedings of 8th International Conference on Learning Representations (ICLR'20)*, 2020.

[47] T. Miyato, S. i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, p. 1979–1993, 2018.

[48] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT Press, 2012.

[49] N. Morgan and H. Bourlard, "Generalization and parameter estimation in feedforward nets: Some experiments," in *Advances in neural information processing systems 2 (NeurIPS'89)*, 1989, pp. 630–637.

[50] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Advances in Neural Information Processing Systems 26 (NeurIPS'13)*, vol. 26, 2013, pp. 1196–1204.

[51] N. Nguyen and R. Caruana, "Classification with partial labels," in *Proceedings of 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'08)*, 2008, pp. 551–559.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32 (NeurIPS'19)*, 2019, pp. 8024–8035.

[53] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, 2017, pp. 1944–1952.

[54] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[55] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, "On the expressive power of deep neural networks," in *Proceedings of 34th International Conference on Machine Learning (ICML'17)*, 2017, pp. 2847–2854.

[56] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[57] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of 16th IEEE International Conference on Computer Vision (ICCV'17)*, 2017, pp. 843–852.

[58] C. Tang and M. Zhang, "Confidence-rated discriminative partial label learning," in *Proceedings of 31st AAAI Conference on Artificial Intelligence (AAAI'17)*, 2017, pp. 2611–2617.

[59] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems 30 (NeurIPS'17)*, 2017, pp. 1195–1204.

[60] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of 17th IEEE International Conference on Computer Vision (ICCV'19)*, 2019, pp. 322–330.

[61] H. Wen, J. Cui, H. Hang, J. Liu, Y. Wang, and Z. Lin, "Leveraged weighted loss for partial label learning," in *Proceedings of 36th International Conference on Machine Learning (ICML'21)*, 2021, pp. 11 091–11 100.

[62] X. Wu and M. Zhang, "Towards enabling binary decomposition for partial label learning," in *Proceedings of 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*, 2018, pp. 2868–2874.

[63] Z. Wu and M. Sugiyama, "Learning with proper partial labels," *arXiv preprint arXiv:2112.12303*, 2021.

[64] X. Xia, T. Liu, N. Wang, B. Han, C. Gong, G. Niu, and M. Sugiyama, "Are anchor points really indispensable in label-noise learning?" in *Advances in Neural Information Processing Systems 32 (NeurIPS'19)*, 2019, pp. 6838–6849.

[65] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[66] N. Xu, J. Lv, and X. Geng, "Partial label learning via label enhancement," in *Proceedings of 33rd AAAI Conference on Artificial Intelligence (AAAI'19)*, 2019, pp. 5557–5564.

[67] Y. Yao, C. Gong, J. Deng, X. Chen, J. Wu, and J. Yang, "Deep discriminative cnn with temporal ensembling for ambiguously-labeled image classification," in *Proceedings of 34th AAAI Conference on Artificial Intelligence (AAAI'20)*, 2020, pp. 12 669–12 676.

[68] Y. Yao, C. Gong, J. Deng, and J. Yang, "Network cooperation with progressive disambiguation for partial label learning," in *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2020, pp. 471–488.

[69] X. Yu, T. Liu, M. Gong, and D. Tao, "Learning with biased complementary labels," in *Proceedings of 15th European Conference on Computer Vision (ECCV'18)*, 2018, pp. 68–83.

[70] Z. Zeng, S. Xiao, K. Jia, T. Chan, S. Gao, D. Xu, and Y. Ma, "Learning by associating ambiguously labeled images," in *Proceedings of 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'13)*, 2013, pp. 708–715.

[71] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proceedings of 5th International Conference on Learning Representations (ICLR'17)*, 2017.

[72] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, ""mixup: Beyond empirical risk minimization," in *Proceedings of 6th International Conference on Learning Representations (ICLR'18)*, 2018.

[73] M. Zhang and F. Yu, "Solving the partial label learning problem: An instance-based approach," in *Proceedings of 24th International Joint Conference on Artificial Intelligence (IJCAI'15)*, 2015, pp. 4048–4054.

[74] M. Zhang, F. Yu, and C. Tang, "Disambiguation-free partial label learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2155–2167, 2017.

[75] M. Zhang, B. Zhou, and X. Liu, "Partial label learning via feature-aware disambiguation," in *Proceedings of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, 2016, pp. 1335–1344.

[76] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in Neural Information Processing Systems 31 (NeurIPS'18)*, 2018, pp. 8778–8788.

[77] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[78] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Lecture Notes in Computer Science*, vol. 3, no. 1, p. 1–130, 2009.

Jiaqi Lv received the B.E. degree and Ph.D. degree in software engineering from the Southeast University, China in 2015 and 2021, respectively. She is currently a postdoctoral researcher at the RIKEN Center for Advanced Intelligence Project, Japan. Her research interests include theory, algorithms, and applications of machine learning. She has also served as a Conference Program Committee Member for ICML, NeurIPS, ICLR, IJCAI and AAAI.

Biao Liu received the BSc degree in computer science from Nanjing University of Science and Technology, China in 2021. Currently, he is a master student at Southeast University. His main research interests include machine learning and data mining.

Bo An is a President's Council Chair Associate Professor in Computer Science and Engineering, and Co-Director of Artificial Intelligence Research Institute (AI.R) at Nanyang Technological University, Singapore. His current research interests include artificial intelligence, multiagent systems, computational game theory, reinforcement learning, and optimization. He has published over 100 referred papers at AAMAS, IJCAI, AAAI, ICAPS, KDD, UAI, EC, WWW, ICLR, NeurIPS, ICML, JAAMAS, AIJ and ACM/IEEE Transactions. He was named to IEEE Intelligent Systems' "AI's 10 to Watch" list for 2018. He was invited to be an Advisory Committee member of IJCAI'18. He is PC Co-Chair of AAMAS'20. He is a member of the editorial board of JAIR and is the Associate Editor of AIJ, JAAMAS, IEEE Intelligent Systems, ACM TAAS, and ACM TIST. He was elected to the board of directors of IFAAMAS, senior member of AAAI, and Distinguished member of ACM.

Lei Feng is a machine learning researcher with research interests in reliable machine learning and weakly supervised learning. He is currently a professor at College of Computer Science, Chongqing University, China and a visiting scientist at Imperfect Information Learning Team, RIKEN Center for Advanced Intelligence Project, Japan. He received the Ph.D. degree from School of Computer Science and Engineering, Nanyang Technological University, Singapore, in 2021, supervised by Prof. Bo An. He was a research intern at Imperfect Information Learning Team, RIKEN Center for Advanced Intelligence Project, supervised by Prof. Masashi Sugiyama and Dr. Gang Niu. Before the Ph.D. study, he received the B.E. degree from College of Computer and Information Science, Southwest University, China, in 2017.

Gang Niu is currently a research scientist (indefinite-term) at RIKEN Center for Advanced Intelligence Project. He received the PhD degree in computer science from Tokyo Institute of Technology in 2013. Before joining RIKEN as a research scientist, he was a senior software engineer at Baidu and then an assistant professor at The University of Tokyo. He has published more than 70 journal articles and conference papers, including 28 ICML, 17 NeurIPS (1 oral and 3 spotlights), and 2 ICLR (1 oral) papers. He has served as an area chair 16 times, including ICML 2019–2022, NeurIPS 2019–2021, and ICLR 2021–2022, and will serve as a publication chair for ICML 2022.

Ning Xu received the B.S. and M.S. degrees from University of Science and Technology of China and Chinese Academy of Sciences China, respectively, and the PhD degree from Southeast University, China. He is now an assistant professor in the School of Computer Science and Engineering at Southeast University, China. His main research interests include machine learning and data mining. He has been selected to the AInet Fellow Program by DAAD in 2020, and won the China Computer Federation Outstanding Doctoral Dissertation Award in 2021.

Xin Geng received the B.S. and M.S. degrees in computer science from Nanjing University, Nanjing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from Deakin University in 2008. He is currently a Professor and the Dean of the School of Computer Science and Engineering, Southeast University, Nanjing. He has authored or coauthored over 80 refereed articles in these areas, including those published in prestigious journals and top international conferences. His research interests include machine learning, pattern recognition, and computer vision. He is a Distinguished Fellow of IETI. He has been an Associate Editor of IEEE T-MM, FCS, and MFC, a Steering Committee Member of PRICAI, a Program Committee Chair for conferences such as PRICAI 18, VALSE 13, and so on, an Area Chair for conferences such as CVPR, ACMMM, PRCV, CCPR, and a Senior Program Committee Member for conferences such as IJCAI, AAAI, ECAI, and so on.

Miao Xu is a lecturer at The University of Queensland and visiting scientist at the RIKEN Center for Advanced Intelligence Project. She is an active researcher in the AI field and has published paper on top conferences and journals such as ICML, NeurIPS, AAAI, IJCAI, PAKDD, and IEEE TKDE. She has served as a program committee member, session chair and reviewer at top-tier AI conferences and journals including area chair for IJCAI2021 and senior program committee member for KDD2022.

Masashi Sugiyama received the PhD degree in Computer Science from Tokyo Institute of Technology, Japan in 2001. He has been Professor at The University of Tokyo since 2014 and concurrently appointed as Director of RIKEN Center for Advanced Intelligence Project in 2016. His research interests include theory, algorithms, and applications of machine learning. He served as a Program co-chair and General co-chair of the Neural Information Processing Systems conference in 2015 and 2016, respectively. Masashi Sugiyama received the Japan Society for the Promotion of Science Award and the Japan Academy Medal in 2017.