Robustness of Accuracy Metric and its Inspirations in Learning with Noisy Labels

Pengfei Chen,[1] **Junjie Ye,**[2*] **Guangyong Chen,**[3*] **Jingwei Zhao,**[2] **Pheng-Ann Heng**[1,3]

[1] The Chinese University of Hong Kong

[2] VIVO AI Lab

[3] Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

{pfchen, pheng}@cse.cuhk.edu.hk, {junjie.ye, jingwei.zhao}@vivo.com, gy.chen@siat.ac.cn

Abstract

For multi-class classification under class-conditional label noise, we prove that the accuracy metric itself can be robust. We concretize this finding's inspiration in two essential aspects: training and validation, with which we address critical issues in learning with noisy labels. For training, we show that maximizing training accuracy on sufficiently many noisy samples yields an approximately optimal classifier. For validation, we prove that a noisy validation set is reliable, addressing the critical demand of model selection in scenarios like hyperparameter-tuning and early stopping. Previously, model selection using noisy validation samples has not been theoretically justified. We verify our theoretical results and additional claims with extensive experiments. We show characterizations of models trained with noisy labels, motivated by our theoretical results, and verify the utility of a noisy validation set by showing the impressive performance of a framework termed noisy best teacher and student (NTS). Our code is released[1].

1   Introduction

In real-world classification tasks, annotation methods such as crowdsourcing systems (Yan et al. 2014) and online queries (Schroff, Criminisi, and Zisserman 2010) inevitably introduce noisy labels. In learning with noisy labels, many works study the robustness, providing theoretical guarantees for some robust loss functions (Ghosh, Kumar, and Sastry 2017; Zhang and Sabuncu 2018; Charoenphakdee, Lee, and Sugiyama 2019; Xu et al. 2019; Ma et al. 2020): a classifier minimizing the loss on noisy distribution is guaranteed to minimize the loss on clean distribution. These works commonly assume that the label noise is conditional on the true class, i.e., the class-conditional noise (Natarajan et al. 2013).

In this paper, we first prove that the accuracy metric itself is robust for common diagonally-dominant class-conditional noise, i.e., *a classifier maximizing its accuracy on the noisy distribution is guaranteed to maximize the accuracy on clean distribution*. It seems counterintuitive since we previously believe maximizing accuracy on noise results in overfitting.

---

*Corresponding authors.

[1]https://github.com/chenpf1025/RobustnessAccuracy

In fact, Deep Neural Networks (DNNs) can achieve 100% training accuracy on finite samples (Zhang et al. 2017) but not the noisy distribution. Theorem 1 shows that for any classifier, the maximized accuracy on noisy distribution is $1 - \varepsilon$, and we obtain an optimal classifier if the maximum is attained, where ε is the noise rate. The definition of robustness used in Theorem 1 is consistent with existing works on robust loss functions. On the one hand, such robustness is not sufficient to guarantee 'good' generalization performance when trained with finite noisy samples because by sampling, we are not directly optimizing the classifier w.r.t. a distribution. On the other hand, we can still show inspirations of Theorem 1 by analyzing the gaps between training/validation accuracy on finite samples and the accuracy on noisy distribution.

For training, even if without any robust loss functions or advanced training strategies, *we can obtain an approximately optimal classifier by maximizing training accuracy on sufficiently many noisy samples*. This claim is justified by Theorem 1 and an additional Theorem 2, which presents a counterpart of the generalization bound derived from Vapnik–Chervonenkis (VC) dimension (Vapnik 1999; Abu-Mostafa, Magdon-Ismail, and Lin 2012). For validation, the accuracy on hold-out noisy samples is an unbiased estimator of the accuracy on noisy distribution. Therefore, *a noisy validation set is reliable*. Together with Theorem 1, the validation bound presented in Theorem 3 formally justifies the utility of a noisy validation set. Some previous works (Zhang and Sabuncu 2018; Nguyen et al. 2020; Xia et al. 2019, 2020) empirically use a noisy validation set to tune hyperparameters, while our contribution is the theoretical justification.

In experiments, we focus on verifying our theoretical results and additional claims. We verify characterizations of models derived from Theorem 1&2, including a dummy classifier given few samples, predicting exactly the noise transition process given more samples, and approaching a global optimal classifier with sufficiently many noisy samples. To show the utility of a noisy validation, we present a framework termed Noisy best Teacher and Student (NTS), which selects a teacher with the highest accuracy on a noisy validation set and retrains a student. It improves many baselines on CIFAR-10/100 with synthetic noise and Clothing1M (Xiao et al. 2015) with real-world noise.
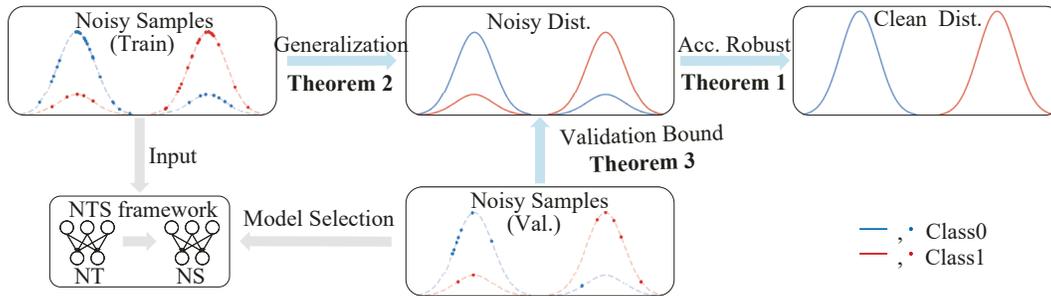
Figure 1: A systematical illustration of learning with noisy labels.

2 Backgrounds

For a K-classes classification problem, we denote the feature space by $\mathcal{X}$ and the label space by $\mathcal{Y} = \{1, 2, \cdots, K\}$. Let $D_{X,Y}$ be the clean distribution of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$. With label noise, we have only access to samples from a noisy distribution $\tilde{D}_{X,\tilde{Y}}$ of $(X, \tilde{Y}) \in \mathcal{X} \times \mathcal{Y}$. The random variable of noisy labels $\tilde{Y}$ is corrupted from Y via a noise transition matrix $T \in [0, 1]^{K \times K}$, s.t.,

$$\Pr[\tilde{Y} = j | Y = i] = T_{i,j}. \tag{1}$$

Then the noise rate is $\varepsilon = 1 - \sum_{i \in \mathcal{Y}} \Pr[Y = i] T_{i,i}$. Such a noise model is common in the literature (Ghosh, Kumar, and Sastry 2017; Zhang and Sabuncu 2018; Han et al. 2018a; Charoenphakdee, Lee, and Sugiyama 2019; Yao et al. 2020b), known as class-conditional noise (Natarajan et al. 2013), s.t.,

Assumption 1. *X is independent of $\tilde{Y}$ conditioning on Y.*

A classifier is denoted by $h \in \mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$, where $\mathcal{H}$ is the hypothesis space. The performance of a classifier is evaluated by its accuracy on clean distribution

$$A_D(h) := \mathbb{E}_{(x,y) \sim D}[\mathbb{1}(h(x) = y)] = \Pr[h(X) = Y], \tag{2}$$

where $\mathbb{1}(\cdot)$ is the indicator function. By training a classifier on a collection of noisy samples $\tilde{S} = \{(x_i, \tilde{y}_i)\}_{i=1}^m$ drawn i.i.d. from $\tilde{D}_{X,\tilde{Y}}$, we can obtain the training accuracy as following,

$$A_{\tilde{S}}(h) := \frac{1}{m} \sum_{i=1}^m \mathbb{1}(h(x_i) = \tilde{y}_i). \tag{3}$$

Thus, an essential problem is to bound the gap between $A_{\tilde{S}}(h)$ and $A_D(h)$. To this end, we introduce the accuracy on noisy distribution,

$$A_{\tilde{D}}(h) := \mathbb{E}_{(x,\tilde{y}) \sim \tilde{D}}[\mathbb{1}(h(x) = \tilde{y})] = \Pr[h(X) = \tilde{Y}]. \tag{4}$$

In this paper, Theorem 1 bridges the gap between $A_{\tilde{D}}(h)$ and $A_D(h)$ by showing the robustness of the accuracy metric. With Theorem 1, the remaining issue is the gap between $A_{\tilde{S}}(h)$ and $A_{\tilde{D}}(h)$, as if without noise.

Following (Chen et al. 2019b), we use a confusion matrix $C(h) \in [0, 1]^{K \times K}$ to characterize the performance of h in detail,

$$C_{i,j}(h) := \Pr[h(X) = j | Y = i]. \tag{5}$$

3 Related Works

There have been many robust loss functions (Ghosh, Kumar, and Sastry 2017; Zhang and Sabuncu 2018; Charoenphakdee, Lee, and Sugiyama 2019; Xu et al. 2019; Ma et al. 2020), whereas the $0 - 1$ loss (Bartlett, Jordan, and McAuliffe 2006), justified to be robust under binary classification (Manwani and Sastry 2013; Ghosh, Manwani, and Sastry 2015), has not attracted enough attention. We previously believe that to combat noisy labels, a loss function should be not only robust, but also easy to optimize in training. In this paper, we first show that accuracy metric itself is robust for multi-class classification under common diagonally-dominant class-conditional noise. We then present several related findings useful in training and validation, where the optimization is not a concern since we can easily maximize training accuracy with a surrogate loss such as the cross-entropy loss.

For the generalization bound, the Vapnik-Chervonenkis (VC) dimension (Vapnik 1999; Abu-Mostafa, Magdon-Ismail, and Lin 2012) and the Natarajan dimension (Natarajan 1989; Daniely and Shalev-Shwartz 2014) are classical techniques that have been used in analysing label noise (Natarajan et al. 2013; Khetan, Lipton, and Anandkumar 2018). In this paper, thanks to Theorem 1, a simple counterpart of the generalization bound derived from VC dimension is sufficient to justify our claim that we can obtain an approximately optimal classifier by maximizing training accuracy on sufficiently many noisy samples. It further motivates us to present a thorough characterizations of models trained with noisy labels, including a dummy classifier given few samples, predicting exactly the noise transition process given more samples, and approaching a global optimal classifier with sufficiently many noisy samples.

For robust training methods, apart from using a robust loss function, one can adopt sample-selection/weighting (Malach and Shalev-Shwartz 2017; Han et al. 2018b; Jiang et al. 2018; Yu et al. 2019; Fang et al. 2020), loss-correction (Patrini et al. 2017; Hendrycks et al. 2018), label-correction (Reed et al. 2015; Tanaka et al. 2018; Zheng et al. 2020) and other refined training strategies (Nguyen et al. 2020; Li, Socher, and Hoi 2020). Specifically, the idea of using teacher predictions is common in loss-correction (Arazo et al. 2019), label-correction (Tanaka et al. 2018) and the classical distillation (Hinton, Vinyals, and Dean 2015). It is known that early-stopped teachers (Arpit et al. 2017; Cho and Hariharan 2019) may be better. Our NTS's critical point is a noisy

validation set, theoretically justified by Theorem 1&3, with which we can easily select a good teacher. Thus, we can obtain impressive performance using merely cross-entropy loss, without explicitly preventing overfitting using advanced training techniques. In contrast, existing methods such as Tanaka et al. (2018) uses regularization to avoid memorizing noise when training the teacher network. Our performance gain is not simply due to distilling a teacher, but due to selecting a good teacher using a noisy validation set. Otherwise, distilling a converged model, which memorizes noise, can not improve the generalization much. Moreover, our aim is not merely presenting the algorithm NTS and comparing it with baselines, but to verify the utility of a noisy validation set with NTS's impressive performance. Justifying the utility of a noisy validation set both theoretically and empirically is a critical contribution, considering that almost all methods need to tune hyperparameters, whereas the usage of a noisy validation set has been empirically demonstrated (Zhang and Sabuncu 2018; Nguyen et al. 2020; Xia et al. 2019, 2020) but has not been theoretically justified.

4 Main Results

In a classification task, our ultimate goal is obtaining a global optimal classifier that shall generalize well on clean distribution.

Definition 1. *(Global optimal classifier). The global optimal classifier, denoted by $h^\star$, is a classifier such that $h^\star(X) = Y$ almost everywhere, i.e., the accuracy on clean distribution is $A_D(h^\star) = 1$. Equivalently, the confusion matrix $C(h^\star) = I$, where I is a $K \times K$ identity matrix.*

The definition holds except for a trivial case where there exists uncertainty for Y conditioning on fully knowing X.

4.1 The Accuracy Metric is Robust

We first show that *the accuracy metric itself is robust to diagonally-dominant noise*, where the diagonally-dominant condition is a mild assumption commonly adopted in literature (Ghosh, Kumar, and Sastry 2017; Zhang and Sabuncu 2018; Chen et al. 2019b).

Assumption 2. *The noise transition matrix T is diagonally-dominant, i.e., $\forall i, T_{i,i} > \max_{j \in \mathcal{Y}, j \neq i} T_{i,j}$.*

Theorem 1. *With Assumption 1&2, if there exists a global optimal classifier $h^\star \in \mathcal{H}$, then*
(i) $h^\star$ is the classifier that maximizes the accuracy on the noisy distribution, i.e.,

$$\max_{h \in \mathcal{H}} A_{\tilde{D}}(h) = A_{\tilde{D}}(h^\star) = 1 - \varepsilon, \qquad (6)$$

where $\varepsilon = 1 - \sum_{i \in \mathcal{Y}} \Pr[Y = i] T_{i,i}$ is the noise rate;
(ii) when the accuracy of a classifier on noisy distribution approaches its maximum, the classifier will approach the global optimal classifier, i.e.,

$$A_{\tilde{D}}(h) \to \max_{h \in \mathcal{H}} A_{\tilde{D}}(h) \implies A_D(h) \to 1. \qquad (7)$$

In Appendix A, we show a convergence rate $1 - A_D(h) \leq \frac{1}{\min_{i,j \in \mathcal{Y}, j \neq i}(T_{i,i} - T_{i,j})} \cdot (\max_{h \in \mathcal{H}} A_{\tilde{D}}(h) - A_{\tilde{D}}(h))$. For any classifier, the maximum of accuracy on noisy distribution is $1 - \varepsilon$

rather than 1. It is not contradictory to the observation that we can achieve 100% training accuracy on finite samples with sufficiently powerful DNNs (Zhang et al. 2017) because by sampling, we are maximizing the accuracy on finite samples rather than the accuracy on noisy distribution. On the one hand, we must be aware that the robustness, as in its widely used form in many existing works, is not sufficient to guarantee 'good' generalization performance when trained with finite noisy samples. On the other hand, we can still show the inspirations of Theorem 1 from two essential aspects: training and validation, by analyzing the gaps between accuracy on finite samples and the accuracy on noisy distribution.

4.2 Training with Noisy Labels

In this section, we analyze the generalization to noisy distribution given noisy samples drawn from it. Intuitively, the training accuracy is a biased estimator of accuracy on noisy distribution. The bias converges to 0 with sufficiently many training samples. This intuition is formalized by the generalization bound in the following Theorem 2. The bound is a counterpart of the generalization bound derived from VC dimension (Vapnik 1999; Abu-Mostafa, Magdon-Ismail, and Lin 2012). It can be generalized very well to multi-class classification by Natarajan dimension (Natarajan 1989; Daniely and Shalev-Shwartz 2014). We first recall the definition of shattering and VC dimension, which can be interpreted as a measurement of the complexity/expressivity of all possible classifiers in a hypothesis space.

Definition 2. *(Shattering and VC dimension (Vapnik 1999; Abu-Mostafa, Magdon-Ismail, and Lin 2012)). Let $\mathcal{H}$ be the hypothesis space of all possible classifiers (functions). We say a set of m points $\{x_1, x_2, \cdots, x_m\} \subset \mathcal{X}$ is **shattered** by $\mathcal{H}$ if all possible binary labeling of the points can be realized by functions in $\mathcal{H}$. The **VC dimension** of $\mathcal{H}$, denoted by $d_{VC}(\mathcal{H})$, is the cardinality of the largest set of points in $\mathcal{X}$ that can be shattered by $\mathcal{H}$.*

Theorem 2. *Considering training a classifier $h \in \mathcal{H}$ on a collection of noisy samples $\tilde{S} = \{x_i, \tilde{y}_i\}_{i=1}^m$ drawn i.i.d. from $\tilde{D}_{X, \tilde{Y}}$, where the hypothesis space $\mathcal{H}$ has a finite VC dimension. Then with probability at least $1 - \delta$,*

$$A_{\tilde{D}}(h) - A_{\tilde{S}}(h) \geq -\sqrt{\frac{8 \left(d_{VC} \cdot (\ln(2m/d_{VC}) + 1) + \ln(4/\delta)\right)}{m}}, \quad (8)$$

where $A_{\tilde{S}}(h)$ is the training accuracy on noisy samples and $A_{\tilde{D}}(h)$ is the accuracy on noisy distribution.

The proof is in Appendix A. The bound converges to 0 with sufficiently large m. The significance is how the bound helps us understand the essence of learning with noisy labels. Firstly, the bound implies that with sufficiently many noisy samples, the accuracy on noisy distribution is approximately lower bounded by the training accuracy. In this case, if we maximize training accuracy, the accuracy on noisy distribution is approximately maximized. Then Theorem 1 implies that the classifier approaches an optimal classifier. Therefore, we conclude that we can obtain an approximately optimal classifier by maximizing training accuracy on sufficiently many noisy samples. Moreover, the results motivate characterizations of neural networks trained with noisy labels. In

particular, with respect to training size m, models trained with noisy labels can be typically categorized into the following three cases:

1) If m is quite small such that the model can not learn any generalized features, then the generalization accuracy typically has a high error and a high variance.

2) If m is relatively large such that the distribution of X is well approximated yet the model still fits all training samples, then $A_{\tilde{S}}(h) \approx 1$, $C(h) \approx T$ and $A_D(h) \approx 1 - \varepsilon$.

3) If m is sufficiently large, then the bound in Theorem 2 approaches 0, and the highest training accuracy that any classifier can achieve approaches $1 - \varepsilon$. When the maximum is attained, we approximately obtain a global optimal classifier, s.t., $A_{\tilde{S}}(h) \approx 1 - \varepsilon$, $C(h) \approx I$ and $A_D(h) \approx 1$.

Case 1) is obvious if training samples are limited; case 2) has been demonstrated in the literature (Chen et al. 2019b); case 3) is derived from Theorem 1&2. In practice, apart from these typical cases, we can observe transition phase, e.g., 1) → 2) and 2) → 3) as we gradually increase m. We will demonstrate these cases in experiments.

To quickly understand the characterizations, we present a tabular example. As shown in Figure 2 (a), we show a classification problem on a discrete space, s.t., $\mathcal{X} = \{(1, \pm2), (2, \pm2), (1, \pm1), (2, \pm1)\}$, $\mathcal{Y} = \{0, 1\}$, where samples in $\{(1, 2), (2, 2), (1, 1), (2, 1)\}$ belong to class 0 and the rest belong to class 1. The data has a uniform distribution, s.t., $\Pr[X = x] = 1/8, \forall x \in \mathcal{X}$. For the noisy distribution, we use a noise transition matrix $T = [0.75, 0.25; 0.25, 0.75]$, i.e, each sample has a probability of $1/4$ to be mislabeled. Now we sample m $i.i.d.$ samples for training, the example sampling results and predictions of a classifier that maximizes training accuracy are illustrated in Figure 2 (b-d).

1) $m = 4$, Figure 2 (b). The optimal training accuracy is 1, and the resulting classifier can not generalize well. Since the classifier's prediction on data-points without a training sample is unknown, the testing accuracy is unknown.

2) $m = 8$, Figure 2 (c). The optimal training accuracy is 1, and the resulting classifier h has a confusion $C(h) = T$. The testing accuracy on clean distribution is 0.75, which equals to $1 - \varepsilon$.

3) $m = 32$, Figure 2 (d). The optimal training accuracy is 0.75, which equals to our theoretical result $1 - \varepsilon$. The resulting classifier is a global optimal classifier with testing accuracy 1 on the clean distribution.

4.3 Validation with Noisy Labels

Validation is a crucial step in the standard machine learning pipeline. When there is label noise, how to conduct validation reliably without clean samples is rarely discussed in previous works. A noisy validation set is used in some previous works (Zhang and Sabuncu 2018; Nguyen et al. 2020; Xia et al. 2019, 2020) but without theoretical justification. In this paper, our Theorem 1 implies that we can approach the optimal classifier by approximately maximizing the accuracy on noisy distribution. Taking a step further, we see that the accuracy on hold-out noisy samples is an unbiased estimator of
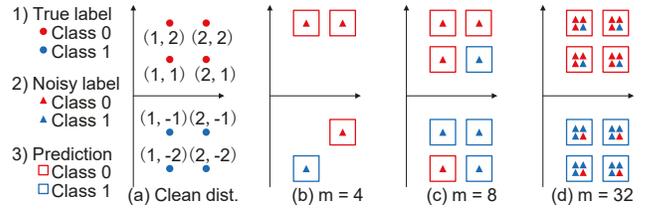
Figure 2: A tabular example of classification under noisy labels. (a): the distribution of data points with true labels. (b-d): examples of sampling m instances with noisy labels for training, where $m = 4, 8, 32$. In (d), we slightly shift the overlapped samples for display.

the accuracy on noisy distribution. Therefore, a noisy validation set is reliable. The reliability is measured by a validation bound presented in the following Theorem 3, which is very tight even with few noisy validation samples.

Theorem 3. *Given a classifier h, considering validating it on a noisy validation set $\tilde{V} = \{(x_i, \tilde{y}_i)\}_{i=1}^n$ drawn i.i.d. from $\tilde{D}_{X,\tilde{Y}}$ and observing the validation accuracy*

$$A_{\tilde{V}}(h) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}(h(x_i) = \tilde{y}_i), \qquad (9)$$

then for any $0 < \delta \leq 1$, we have that with probability at least $1 - \delta$,

$$A_{\tilde{D}}(h) - A_{\tilde{V}}(h) \geq -\sqrt{\frac{\ln(1/\delta)}{2n}}, \qquad (10)$$

where $A_{\tilde{D}}(h)$ is the accuracy on noisy distribution.

The proof is in Appendix A. To see how tight the bound is, considering $n = 1000$ noisy samples and $\delta = 0.01$, the gap in Eq. (10) is no larger than 0.048. It means we shall obtain an error no larger than 0.048 with at least probability 0.99. Theorem 1&3 justify that it is reliable to use a small noisy validation set for model selection. Previously, there has not been a theoretically widely-accepted criterion for validation without clean samples. For example, many previous experiments on CIFAR-10 use the $10k$ clean samples to tune hyperparameters, which may not be practical in learning with noisy labels since we assume no access to many clean samples. Therefore, our theoretical justification for a noisy validation set is critical, especially considering lack of clean samples and the unavoidable needs of model selection in scenarios such as hyperparameters-tuning and early stopping.

5 Experiments
5.1 Training with Noisy Labels

Theorem 1&2 imply that we can obtain an optimal classifier by maximizing training accuracy on sufficiently many noisy samples, even if without any advanced training strategies. Models' characterizations motivated by the theoretical results have been discussed in Section 4.2. Here we conduct experiments to visualize neural networks' typical characterizations when trained with noisy labels. As shown in Figure 3, we adopt two classical synthetic datasets 'circles'
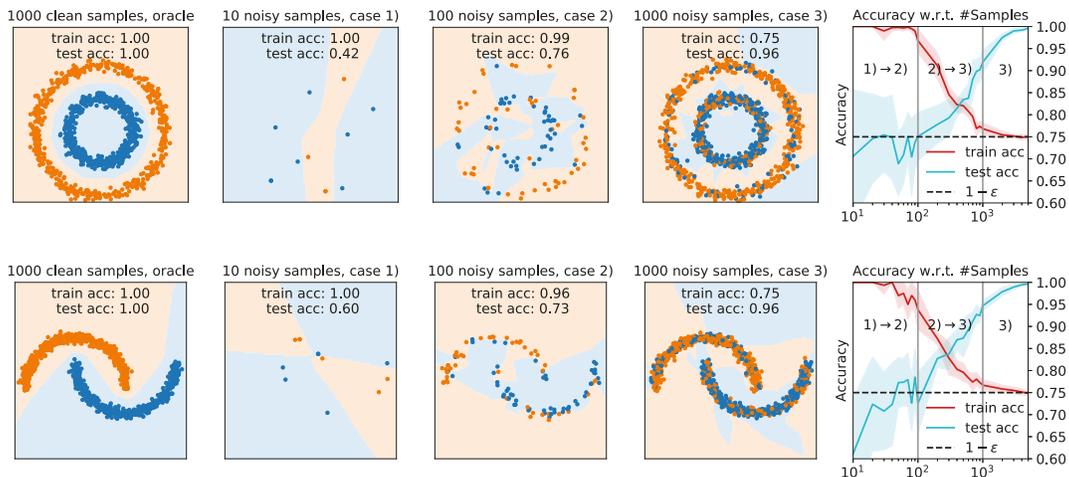
Figure 3: Synthetic training samples and the classifier's decision regions. From left to right: an oracle case trained on clean samples; three typical cases when trained with noisy labels; train/test accuracy (at the last training step) w.r.t. number of training samples. Orange: class 0; Blue: class 1. First raw: 'circles' dataset; Second raw: 'moons' dataset. The ε is noise rate.

and 'moons', and pollute them with a noise transition matrix $T = [0.7, 0.3; 0.2, 0.8]$. We implement an MLP with two hidden layer of size 32. We train the classifier using full-batch gradient descent for at most $20K$ steps to ensure the training accuracy is optimized until convergence, and test on $10k$ clean samples. We vary the training sizes to demonstrate different characterizations and repeat each experiment 10 times. Results in Figure 3 cover the three typical characterizations summarized in Section 4.2. We also present curves of train/test accuracy at the last step w.r.t. number of training samples, where the shadow indicates the standard deviation. It clearly demonstrates transitions 1) $\rightarrow$ 2) $\rightarrow$ 3) as the number of training samples increases. For case 2), the training accuracy is approximately 1 and the testing accuracy is around $1 - \varepsilon = 0.75$; as the number of samples increase, the trained network finally approaches case 3), with the training accuracy converging to $1 - \varepsilon$ and the testing accuracy approaching 1. For deep networks and high dimensional space, we may not have sufficient samples to approximately obtain an optimal classifier. Still, Theorem 1&2 and numerical experiments here justify the utility of data augmentation techniques (Cubuk et al. 2019; He et al. 2016) in learning with noisy labels: data augmentation narrows the gap between noisy samples and noisy distribution. Previously, it is intuitively believed but not explained that why augmented noisy samples, which have the same noise, improves generalization.

5.2 A Noisy Validation Set is Reliable

To verify the utility of a noisy validation set, which has been theoretically justified by Theorem 1&3, we present the NTS framework. The implementation consists of two steps, (1) noisy best teacher (NT): training on given noisy labels and selecting a teacher model with the highest accuracy on the noisy validation set, (2) noisy best student (NS): training on labels predicted by NT and selecting a student model similarly. We empirically validate NTS on extensive benchmark datasets, including CIFAR-10 and CIFAR-100 with uniform

noise and realistic asymmetric noise flipped between similar classes such as DEER→HORSE, as well as the large-scale benchmark Clothing1M (Xiao et al. 2015), which contains 1 million real-world noisy training samples.

CIFAR-10 and CIFAR-100. We use Wide ResNet-28-10 (WRN-28-10) (Zagoruyko and Komodakis 2016) as the classifier on CIFAR-10 and CIFAR-100. We corrupt the training set which has 50000 samples and randomly split 5000 noisy samples for validation. We conduct experiments on *uniform noise* (symmetric noise) and *asymmetric noise*, following previous settings (Han et al. 2018b; Patrini et al. 2017; Ren et al. 2018; Nguyen et al. 2020; Chen et al. 2019a). Uniform noise is generated by uniformly flipping labels to other classes (Han et al. 2018b; Ren et al. 2018; Han et al. 2020), which is the most studied scenario in literature. For asymmetric noise, following (Patrini et al. 2017; Zhang and Sabuncu 2018; Nguyen et al. 2020), on CIFAR-10, it is generated by flipping labels between similar classes, i.e., TRUCK→AUTOMOBILE, BIRD→AIRPLANE, DEER→HORSE, and CAT↔DOG, with a given probability; on CIFAR-100, it is generated by flipping each class into the next circularly with a given probability, which is also known as pair noise (Han et al. 2018b). Apart from implementing NTS on the most common Cross Entropy (CE) loss, we show NTS is also applicable to other advanced training methods, including Generalized Cross Entropy (GCE) (Zhang and Sabuncu 2018) loss, which is proved to be noise-robust; Co-teaching (Co-T) (Han et al. 2018b), which is an effective method that uses two networks to select small-loss training samples for each other; Determinant based Mutual Information (DMI), which is an information-theoretic robust loss. In NTS, all student networks share the same training schedule and hyperparameters with their teachers. Strong data augmentation (Cubuk et al. 2019) is implemented, while we will show in the ablation that the improvement obtained by NTS is consistent with/without data augmentation. More training details can be found in Appendix B.

Table 1 and Figure 4 with surrounding body text.

Dataset	Base	Model	Uniform Noise			Asymmetric Noise		
			0.2	0.4	0.6	0.2	0.3	0.4
CIFAR-10	CE	Last	86.30 ± 0.20	74.18 ± 2.27	59.15 ± 1.31	89.95 ± 1.23	86.75 ± 0.64	79.47 ± 1.60
		NT	93.63 ± 0.21	90.09 ± 0.41	82.48 ± 0.47	94.59 ± 0.10	93.33 ± 0.32	89.07 ± 1.44
		NS	$\mathbf{95.73 \pm 0.11}$	$\mathbf{93.76 \pm 0.18}$	$\mathbf{87.88 \pm 0.23}$	$\mathbf{95.09 \pm 0.10}$	$\mathbf{94.80 \pm 0.44}$	$\mathbf{91.49 \pm 1.68}$
	GCE	Last	95.02 ± 0.22	92.72 ± 0.25	86.99 ± 0.24	94.15 ± 0.45	91.55 ± 0.65	82.41 ± 1.61
		NT	94.81 ± 0.07	92.53 ± 0.04	87.11 ± 0.10	94.41 ± 0.17	92.72 ± 0.08	87.60 ± 1.15
		NS	$\mathbf{95.34 \pm 0.24}$	$\mathbf{94.50 \pm 0.30}$	$\mathbf{91.32 \pm 0.12}$	$\mathbf{95.08 \pm 0.09}$	$\mathbf{94.11 \pm 0.08}$	$\mathbf{90.43 \pm 0.73}$
	Co-T	Last	95.09 ± 0.07	92.18 ± 0.36	87.32 ± 0.38	93.56 ± 0.08	91.39 ± 0.29	88.79 ± 0.36
		NT	95.17 ± 0.14	92.42 ± 0.29	87.18 ± 0.29	94.44 ± 0.14	92.15 ± 0.14	89.93 ± 0.10
		NS	$\mathbf{96.26 \pm 0.14}$	$\mathbf{94.73 \pm 0.20}$	$\mathbf{92.11 \pm 0.40}$	$\mathbf{95.47 \pm 0.09}$	$\mathbf{93.73 \pm 0.36}$	$\mathbf{91.32 \pm 0.30}$
	DMI	Last	94.20 ± 0.20	41.60 ± 29.08	11.68 ± 0.72	95.29 ± 0.21	36.29 ± 37.87	10.35 ± 2.76
		NT	$\mathbf{94.25 \pm 0.30}$	91.10 ± 0.46	79.38 ± 5.17	95.14 ± 0.26	94.12 ± 0.52	89.04 ± 3.31
		NS	94.10 ± 0.11	$\mathbf{91.11 \pm 0.53}$	$\mathbf{83.46 \pm 0.46}$	$\mathbf{95.29 \pm 0.08}$	$\mathbf{95.03 \pm 0.08}$	$\mathbf{92.28 \pm 1.33}$
CIFAR-100	CE	Last	70.61 ± 0.20	55.96 ± 0.62	38.25 ± 0.58	70.10 ± 0.01	61.51 ± 0.21	51.11 ± 0.13
		NT	73.38 ± 0.47	66.46 ± 0.26	57.94 ± 0.30	74.20 ± 0.35	68.27 ± 0.44	54.46 ± 1.22
		NS	$\mathbf{76.85 \pm 0.29}$	$\mathbf{71.71 \pm 0.59}$	$\mathbf{63.69 \pm 0.25}$	$\mathbf{77.07 \pm 0.21}$	$\mathbf{70.51 \pm 0.30}$	$\mathbf{55.61 \pm 1.56}$
	GCE	Last	70.95 ± 0.25	55.32 ± 0.35	20.21 ± 0.59	62.59 ± 0.62	52.75 ± 0.50	41.75 ± 0.69
		NT	73.54 ± 0.25	65.48 ± 0.22	$\mathbf{24.22 \pm 0.88}$	71.55 ± 0.17	60.30 ± 0.18	$\mathbf{43.51 \pm 1.69}$
		NS	$\mathbf{75.41 \pm 0.26}$	$\mathbf{67.87 \pm 0.13}$	23.34 ± 1.30	$\mathbf{73.52 \pm 0.41}$	$\mathbf{61.38 \pm 1.04}$	43.07 ± 1.43
	Co-T	Last	78.24 ± 0.28	71.63 ± 0.33	64.93 ± 0.37	74.82 ± 0.07	67.72 ± 0.36	59.05 ± 1.01
		NT	78.51 ± 0.20	72.33 ± 0.43	66.32 ± 0.05	75.45 ± 0.24	69.92 ± 0.39	60.22 ± 0.49
		NS	$\mathbf{80.02 \pm 0.06}$	$\mathbf{76.16 \pm 0.49}$	$\mathbf{72.30 \pm 0.18}$	$\mathbf{77.68 \pm 0.84}$	$\mathbf{74.02 \pm 1.49}$	$\mathbf{62.66 \pm 0.67}$
	DMI	Last	73.80 ± 0.40	66.80 ± 0.22	58.09 ± 0.06	74.54 ± 0.20	68.25 ± 0.24	54.69 ± 0.96
		NT	73.80 ± 0.41	66.82 ± 0.16	58.35 ± 0.10	74.79 ± 0.22	68.65 ± 0.08	54.64 ± 1.05
		NS	$\mathbf{73.89 \pm 0.34}$	$\mathbf{66.95 \pm 0.23}$	$\mathbf{58.35 \pm 0.06}$	$\mathbf{74.82 \pm 0.23}$	$\mathbf{68.91 \pm 0.41}$	$\mathbf{54.74 \pm 1.03}$

Table 1: Results of WRN-28-10 on CIFAR-10 and CIFAR-100 under uniform noise and asymmetric noise. Each experiment is repeated three times. The best accuracy under each base method is in bold.
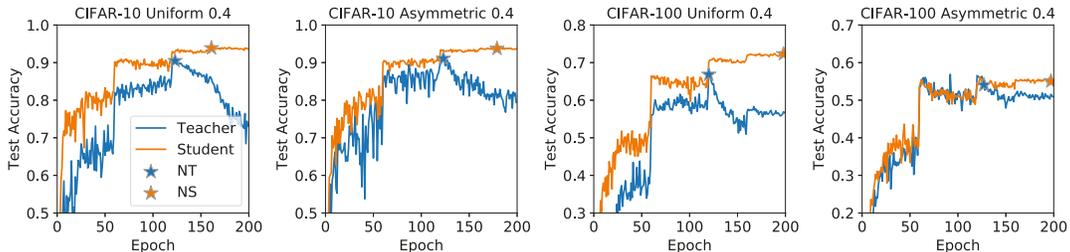
Figure 4: Test accuracy of the teacher and student during training epochs. NT and NS are selected using a noisy validation set, justified by Theorem 1 and Theorem 3.

Table 1 summarizes results on CIFAR-10 and CIFAR-100, where we repeat each experiment three times. The results verify that our NTS framework and the noisy validation set not only improve performance for CE, but are also applicable to many advanced training algorithms. Firstly, in all noise settings, the best accuracy is always achieved by our NS. Secondly, a noisy validation, which is justified by Theorem 1&3, is an enhancement for all methods. When severe overfitting happens at the later stage (e.g., CE), we can get a significantly better NT than the last epoch. Even if existing methods can reduce overfitting, our NT still achieves higher or comparable accuracy than the last epoch. Moreover, we can not tune hyperparameters of existing methods without the noisy validation, whereas conduct validation on clean samples may violate the fair setting in learning with noisy labels. Hence our Theorem 1&3 and the NTS framework are critical, which justify the utility of a noisy validation set theoretically and empirically. Finally, we see that in most cases, we get a better NS by learning knowledge form NT. Figure 4 shows that we benefit from early-stopped teachers (Cho and Hariharan 2019). It is known that early stopping may help due to the memorization effect (Arpit et al. 2017; Yao et al. 2020a), s.t., DNNs learn simple and correct patterns first before memorizing noise. Our critical point is justifying the noisy validation set, with which we can easily select the good teachers.

Base	CE			DMI			DivideMix		
Validation	Last	NT	NS	Last	NT	NS	Last	NT	NS
Clean	67.91	68.99	**70.02**	71.82	72.23	**72.99**	73.09/73.48/73.81	73.98/74.31/74.65	**74.48/74.70/74.89**
Noisy	68.05	69.12	**70.21**	72.01	72.18	**72.82**	73.63/73.38/73.84	73.75/73.73/74.18	**73.99/74.09/74.36**

Table 2: Results of ResNet-50 on Clothing1M, using clean/noisy validation. The three results of DivideMix indicate the first model/second model/ensemble model (following Li, Socher, and Hoi (2020)). The utility of NTS and a noisy validation set is verified: 1) results of clean/noisy validation are comparable, 2) NTS obtains consistent performance gain over each baseline.

Dataset	Model	No Augmentation		Standard Augmentation		Strong Augmentation	
		Uniform 0.4	Asymmetric 0.4	Uniform 0.4	Asymmetric 0.4	Uniform 0.4	Asymmetric 0.4
CIFAR-10	Last	60.94	73.82	70.10	77.54	74.18	79.47
	NT	72.15	75.30	86.14	87.66	90.09	89.07
	NS	**75.04**	**76.69**	**89.45**	**90.23**	**93.76**	**91.49**
CIFAR-100	Last	33.83	42.31	53.72	49.26	55.96	51.11
	NT	41.40	41.20	62.96	55.92	66.46	54.46
	NS	**41.72**	**52.61**	**65.92**	**59.27**	**71.71**	**55.61**

Table 3: Ablation results of WRN-28-10 trained with CE. The best accuracy is in bold.

Clothing1M. Clothing1M is a large-scale benchmark containing real-world noise. Strictly following the standard setting (Patrini et al. 2017; Tanaka et al. 2018; Xu et al. 2019; Li, Socher, and Hoi 2020), we first access a ResNet-50 pretrained on ImageNet, then use 1 million noisy samples of Clothing1M for training, $14k$ and $10k$ clean data respectively for validation and test. Moreover, to demonstrate the utility of a noisy validation set, we additionally show a more challenging setting where the clean validation set is assumed unavailable. In this case, we randomly draw $14k$ samples from the noisy training set for validation. More training details can be found in Appendix B. Table 2 summarizes results on Clothing1M. Since DivideMix (Li, Socher, and Hoi 2020) trains two networks, apart from the ensemble result obtained by averaging two networks' predictions following the original paper (Li, Socher, and Hoi 2020), we also report results of each network for a fair comparison with other baselines. For NTS, when we use DivideMix as the teacher, a student is trained for each teacher. Results of both student and their average ensemble are reported similarly. In Table 2, we report results of the last epoch, NT and NS for each baseline method. The utility of NTS and a noisy validation set is verified on this real-world noisy dataset, because 1) results of clean/noisy validation are comparable, 2) NTS obtains consistent performance gain over each baseline. Notably, without our theoretical justification for a noisy validation set, it would be intractable to tune hyperparameters for all baselines.

Ablation. To comprehensively verify the utility of a noisy validation set when there is no augmentation, standard augmentation (He et al. 2016) or strong augmentation (Cubuk et al. 2019), we conduct an ablation study on CIFAR-10 and CIFAR-100. Results are presented in Table 3. It is not surprising that data augmentation can improve test accuracy. Still, it has not been made clear why augmented noisy samples,

which have the same noisy labels, improve generalization. This paper reveals that the reason is the robustness of accuracy, which fills the gap between the noisy and clean distribution. Augmented samples, though have the same noise, can narrow the gap between training samples and the noisy distribution. In rare cases such as CIFAR-100 with 0.4 asymmetric noise, strong augmentation can be harmful since too much randomness may result in underfitting. Importantly, we see that the performance gain of NTS, which uses a noisy validation set, is consistent under different augmentations or no augmentation. The results verify that NTS is practical and a noisy validation set is reliable.

6 Conclusion

In this paper, we target at revealing the essence of learning with noisy labels. We prove that the accuracy metric itself is robust for multi-class classification under common diagonally-dominant class-conditional noise, i.e., a classifier maximizing its accuracy on the noisy distribution is guaranteed to maximize the accuracy on clean distribution. We then show inspirations of this finding in two essential aspects: training and validation. For training, maximizing training accuracy on sufficiently many noisy samples yields an approximately optimal classifier. Characterizations of models trained with noisy samples are derived. For validation, it is reliable to use a noisy validation set. Justifying the noisy validation set is a critical contribution, considering that almost all methods need to tune hyperparameters, whereas conducting model-selection reliably without clean samples has not been theoretically justified in previous works. In experiments, we verify our theoretical results and additional claims by visualizing characterizations of DNNs and demonstrating the impressive performance of NTS.

Acknowledgments

A Proofs

Lemma 1. *With Assumption 1, given any classifier $h : \mathcal{X} \mapsto \mathcal{Y}$, $\tilde{Y}$ is independent of $h(X)$ conditioning on Y.*

The conclusion easily follows since $\tilde{Y}$ is independent of X conditioning on Y (Assumption 1).

Proof. (**Theorem 1**). (i) Without loss of generality, we assume each true class has nonzero probability, i.e., $\Pr[Y = i] > 0, \forall i \in \mathcal{Y}$. Otherwise we can remove the class with probability 0.

Recall that the confusion matrix $C(h)$ for a classifier h is defined as $C_{i,j}(h) := \Pr[h(X) = j | Y = i]$. Thus, $0 \leq C_{i,j}(h) \leq 1$ and for any $i \in \mathcal{Y}$, $\sum_{j \in \mathcal{Y}} C_{i,j}(h) = 1$.

With Assumption 2, for any $i \in \mathcal{Y}$, we have $\sum_{j \in \mathcal{Y}} T_{i,j} \cdot C_{i,j}(h) \leq \sum_{j \in \mathcal{Y}} T_{i,i} \cdot C_{i,j}(h) = T_{i,i}$.

With Lemma 1, we have, $\forall i, j \in \mathcal{Y}$, $\Pr[h(X) = j | Y = i, \tilde{Y} = j] = \Pr[h(X) = j | Y = i] = C_{i,j}(h)$. Therefore,

$$
\begin{aligned}
A_{\tilde{D}}(h) &= \Pr[h(X) = \tilde{Y}] \\
&= \sum_{i \in \mathcal{Y}} \Pr[Y = i] \cdot \Pr[h(X) = \tilde{Y} | Y = i] \\
&= \sum_{i \in \mathcal{Y}} (\Pr[Y = i] \sum_{j \in \mathcal{Y}} T_{i,j} \cdot C_{i,j}(h)) \\
&\leq \sum_{i \in \mathcal{Y}} \Pr[Y = i] \cdot T_{i,i} \\
&= 1 - \varepsilon.
\end{aligned}
\tag{11}
$$

The maximum is attained if and only if $C(h) = I$, i.e., when h is a global optimal classifier $h^\star$.

(ii) Note that $\sum_{j \in \mathcal{Y}} C_{i,j}(h) = 1, \forall i \in \mathcal{Y}$. Using Eq. (11), we obtain the gap between $A_{\tilde{D}}(h)$ and its maximum,

$$
\begin{aligned}
&\max_{h \in \mathcal{H}} A_{\tilde{D}}(h) - A_{\tilde{D}}(h) \\
&= \sum_{i \in \mathcal{Y}} \Pr[Y = i] \cdot T_{i,i} - \sum_{i \in \mathcal{Y}} (\Pr[Y = i] \sum_{j \in \mathcal{Y}} T_{i,j} \cdot C_{i,j}(h)) \\
&= \sum_{i \in \mathcal{Y}} (\Pr[Y = i] (T_{i,i} - \sum_{j \in \mathcal{Y}} T_{i,j} \cdot C_{i,j}(h))) \\
&= \sum_{i \in \mathcal{Y}} (\Pr[Y = i] (T_{i,i} \cdot \sum_{j \in \mathcal{Y}} C_{i,j}(h) - \sum_{j \in \mathcal{Y}} T_{i,j} \cdot C_{i,j}(h))) \\
&= \sum_{i,j \in \mathcal{Y}, j \neq i} \Pr[Y = i] \cdot (T_{i,i} - T_{i,j}) \cdot C_{i,j}(h).
\end{aligned}
\tag{12}
$$

Therefore,

$$
\begin{aligned}
&\frac{\max_{h \in \mathcal{H}} A_{\tilde{D}}(h) - A_{\tilde{D}}(h)}{\min_{i,j \in \mathcal{Y}, j \neq i} (T_{i,i} - T_{i,j})} \\
&= \frac{\sum_{i,j \in \mathcal{Y}, j \neq i} \Pr[Y = i] \cdot (T_{i,i} - T_{i,j}) \cdot C_{i,j}(h)}{\min_{i,j \in \mathcal{Y}, j \neq i} (T_{i,i} - T_{i,j})} \\
&\geq \sum_{i,j \in \mathcal{Y}, j \neq i} \Pr[Y = i] \cdot C_{i,j}(h) \\
&= \sum_{i \in \mathcal{Y}} \Pr[Y = i](1 - C_{i,i}(h)) \\
&= 1 - \sum_{i \in \mathcal{Y}} \Pr[Y = i] C_{i,i}(h) \\
&= 1 - \sum_{i \in \mathcal{Y}} \Pr[Y = i] \Pr[h(X) = Y | Y = i] \\
&= 1 - A_D(h).
\end{aligned}
\tag{13}
$$

That is

$$
0 \leq 1 - A_D(h) \leq \frac{\max_{h \in \mathcal{H}} A_{\tilde{D}}(h) - A_{\tilde{D}}(h)}{\min_{i,j \in \mathcal{Y}, j \neq i} (T_{i,i} - T_{i,j})}.
\tag{14}
$$

Then the conclusion follows from the above inequality. $\square$

Proof. (**Theorem 2**). Note that accuracy metric is equivalent to error metric defined by $0 - 1$ loss. Therefore, using the uniform error bound (Vapnik 1999; Abu-Mostafa, Magdon-Ismail, and Lin 2012) for general $h \in \mathcal{H}$, we have, $\forall \epsilon > 0$,

$$
\begin{aligned}
&\Pr\left[\sup_{h \in \mathcal{H}} (A_{\tilde{S}}(h) - A_{\tilde{D}}(h)) \geq \epsilon\right] \\
&\leq 4(2em/d_{VC}(\mathcal{H}))^{d_{VC}(\mathcal{H})} e^{-m\epsilon^2/8}
\end{aligned}
\tag{15}
$$

Let $\delta = 4(2em/d_{VC}(\mathcal{H}))^{d_{VC}(\mathcal{H})} e^{-m\epsilon^2/8}$, we get $\Pr\left[\sup_{h \in \mathcal{H}} (A_{\tilde{S}}(h) - A_{\tilde{D}}(h)) \geq \epsilon\right] \leq \delta$.

Thus, $\Pr\left[\inf_{h \in \mathcal{H}} (A_{\tilde{D}}(h) - A_{\tilde{S}}(h)) \leq -\epsilon\right] \leq \delta$.

Then the conclusion follows by substituting $\epsilon = \sqrt{\frac{8 (d_{VC}(\mathcal{H}) \cdot (\ln(2m/d_{VC}(\mathcal{H})) + 1) + \ln(4/\delta))}{m}}$. $\square$

Proof. (**Theorem 3**). The noisy validation samples $\tilde{V} = \{(x_i, \tilde{y}_i)\}_{i=1}^n$ are drawn i.i.d. from $\tilde{D}_{X,\tilde{Y}}$. Therefore, the indication function on each sample $\mathbb{1}(h(x_i) = \tilde{y}_i), i = 1, 2, \cdots, n$ are i.i.d., such that $\mathbb{E}_{(x,\tilde{y}) \sim \tilde{D}}[\mathbb{1}(h(x) = \tilde{y})] = A_{\tilde{D}}(h)$. Using Hoeffding's inequality (Bentkus et al. 2004) for these i.i.d. Bernoulli random variables with expectation $A_{\tilde{D}}(h)$, we have

$$
\Pr[A_{\tilde{V}}(h) - A_{\tilde{D}}(h) \geq \epsilon] \leq e^{-2n\epsilon^2},
\tag{16}
$$

where $A_{\tilde{V}}(h) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}(h(x_i) = \tilde{y}_i)$ is the observed validation accuracy. Equivalently,

$$
\Pr[A_{\tilde{D}}(h) - A_{\tilde{V}}(h) \leq -\epsilon] \leq e^{-2n\epsilon^2}.
\tag{17}
$$

Let $\delta = e^{-2n\epsilon^2}$, then the conclusion follows by substituting $\epsilon = \sqrt{\frac{\ln(1/\delta)}{2n}}$. $\square$

B   More Details on Experiments

Here we present more experimental details. All models are trained with Tesla V100 GPU. Our code is released [2].

B.1   CIFAR-10 and CIFAR-100

We add label noise to the initial training set with 50000 samples and randomly split 5000 noisy samples for validation. Wide ResNet-28-10 (WRN-28-10) (Zagoruyko and Komodakis 2016) is used as the classifier. Our NTS framework is implemented on four typical methods: Cross Entropy (CE), Generalized Cross Entropy (GCE) (Zhang and Sabuncu 2018), Co-teaching (Co-T) (Han et al. 2018b) and Determinant based Mutual Information (DMI). Though each basic method may require some specific hyperparameters, we ensure that in each method, training of the teacher and the student share exactly the same setting.

CE, GCE and Co-T shares the same batch size of 128 and learning rate schedule, i.e., training using SGD optimizer for 200 epochs, with a initial learning rate of 0.1, which is decreased by a factor of 5 after 60, 120 and 160 epochs. Following its original paper and official implementation, DMI uses a model pretrained by CE as initialization and requires a larger batch size of 256 and a smaller learning rate, which is tuned in $\{10^{-4}, 10^{-5}, 10^{-6}\}$ and fixed to 10^{-6} finally. It is trained using SGD optimizer for 100 epochs without learning rate change. In all methods, the SGD optimizer is implemented with momentum 0.9 and weight decay 5×10^{-4}. For GCE, following its original paper, there is a warm-up epoch, before which no loss pruning is applied. It is set as the epoch of first and second learning rate change for CIFAR-10 and CIFAR-100 respectively (Zhang and Sabuncu 2018). We tune the warm-up epoch in $\{30, 60, 120\}$ and finally use 60, 120 for CIFAR-10 and CIFAR-100 respectively. For Co-T, we also find that a warm-up (tuned in $\{30, 60, 120\}$) where no sample selection is applied can help the training, especially under strong noise, hence we use a warm-up epoch 60 for 0.6 uniform noise and 0.4 asymmetric noise.

We use strong augmentation and dropout 0.2 for all methods by default, except for GCE on CIFAR-100, where we found it works better with standard augmentation and no dropout. Here, standard augmentation (He et al. 2016) includes per-pixel normalization, horizontal random flip and 32×32 random crop after padding with 4 pixels on each side. Strong augmentation (Cubuk et al. 2019) includes operations described in (Cubuk et al. 2019), i.e., ShearX/Y, TranslateX/Y, Rotate, AutoContrast, Invert, Equalize, Solarize, Posterize, Contrast, Color, Brightness, Sharpness and Cutout (DeVries and Taylor 2017). In the ablation study presented in Section 5.2, we conduct experiments without augmentation, with standard augmentation and with strong augmentation, which verifies that the utility of NTS and a noisy validation set is consistent in all these settings.

B.2   Clothing1M

Strictly following the standard setting (Patrini et al. 2017; Tanaka et al. 2018; Xu et al. 2019; Li, Socher, and Hoi 2020), we first access a ResNet-50 pre-trained on ImageNet, then use 1 million noisy samples of Clothing1M for training, $14k$ and $10k$ clean data respectively for validation and test. For DMI (Xu et al. 2019) and DivideMix (Li, Socher, and Hoi 2020), we reproduce results following their official implementations. In NTS, we find that simply training the student using normal cross-entropy loss works well. We train the ResNet-50 using SGD optimizer for 10 epochs, with a batchsize 256 and an initial learning rate of 0.001, which is decreased by a factor of 10 after 5 epochs. The SGD optimizer is implemented with momentum 0.9 and weight decay 0.001. We use standard data augmentation with per-pixel normalization, horizontal random flip and 224×224 random crop. Note that DivideMix (Li, Socher, and Hoi 2020) trains two networks, hence we train a student for each teacher.

Using exactly the same training hyperparameters, we also demonstrate a more challenging setting, where the $14k$ clean validation samples are assumed unavailable. We randomly sample $14k$ noisy samples (i.e., $1k$ samples per class) from the 1 million noisy training samples for validation. This setting further verifies the utility of a noisy validation set.

References

Abu-Mostafa, Y. S.; Magdon-Ismail, M.; and Lin, H. 2012. Learning from data: a short course. *AMLBook.com* .

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2019. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*.

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. In *International Conference on Machine Learning*.

Bartlett, P. L.; Jordan, M. I.; and McAuliffe, J. D. 2006. Convexity, classification, and risk bounds. *Journal of the American Statistical Association* .

Bentkus, V.; et al. 2004. On Hoeffding's inequalities. *The Annals of Probability* .

Charoenphakdee, N.; Lee, J.; and Sugiyama, M. 2019. On Symmetric Losses for Learning from Corrupted Labels. In *International Conference on Machine Learning*.

Chen, P.; Liao, B.; Chen, G.; and Zhang, S. 2019a. A meta approach to defend noisy labels by the manifold regularizer PSDR. *arXiv preprint arXiv:1906.05509* .

Chen, P.; Liao, B. B.; Chen, G.; and Zhang, S. 2019b. Understanding and Utilizing Deep Neural Networks Trained with Noisy Labels. In *International Conference on Machine Learning*.

Cho, J. H.; and Hariharan, B. 2019. On the efficacy of knowledge distillation. In *IEEE International Conference on Computer Vision*.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. Autoaugment: Learning augmentation strategies from data. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Daniely, A.; and Shalev-Shwartz, S. 2014. Optimal learners for multiclass problems. In *Conference on Learning Theory*.

---

[2] https://github.com/chenpf1025/RobustnessAccuracy

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* .

Fang, T.; Lu, N.; Niu, G.; and Sugiyama, M. 2020. Rethinking Importance Weighting for Deep Learning under Distribution Shift. In *Advances in Neural Information Processing Systems*.

Ghosh, A.; Kumar, H.; and Sastry, P. 2017. Robust loss functions under label noise for deep neural networks. In *AAAI Conference on Artificial Intelligence*.

Ghosh, A.; Manwani, N.; and Sastry, P. 2015. Making risk minimization tolerant to label noise. *Neurocomputing* .

Han, B.; Niu, G.; Yu, X.; Yao, Q.; Xu, M.; Tsang, I.; and Sugiyama, M. 2020. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*.

Han, B.; Yao, J.; Niu, G.; Zhou, M.; Tsang, I.; Zhang, Y.; and Sugiyama, M. 2018a. Masking: A new perspective of noisy supervision. In *Advances in Neural Information Processing Systems*.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Hendrycks, D.; Mazeika, M.; Wilson, D.; and Gimpel, K. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* .

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels. In *International Conference on Machine Learning*.

Khetan, A.; Lipton, Z. C.; and Anandkumar, A. 2018. Learning From Noisy Singly-labeled Data. In *International Conference on Learning Representations*.

Li, J.; Socher, R.; and Hoi, S. C. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *International Conference on Learning Representations*.

Ma, X.; Huang, H.; Wang, Y.; Erfani, S. R. S.; and Bailey, J. 2020. Normalized Loss Functions for Deep Learning with Noisy Labels. In *International Conference on Machine Learning*.

Malach, E.; and Shalev-Shwartz, S. 2017. Decoupling "when to update" from "how to update". In *Advances in Neural Information Processing Systems*.

Manwani, N.; and Sastry, P. 2013. Noise tolerance under risk minimization. *IEEE Transactions on cybernetics* .

Natarajan, B. K. 1989. On learning sets and functions. *Machine Learning* .

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. In *Advances in Neural Information Processing Systems*.

Nguyen, T.; Mummadi, C.; Ngo, T.; Beggel, L.; and Brox, T. 2020. SELF: learning to filter noisy labels with self-ensembling. In *International Conference on Learning Representations*.

Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2015. Training deep neural networks on noisy labels with bootstrapping. In *International Conference on Learning Representations*.

Ren, M.; Zeng, W.; Yang, B.; and Urtasun, R. 2018. Learning to Reweight Examples for Robust Deep Learning. In *International Conference on Machine Learning*.

Schroff, F.; Criminisi, A.; and Zisserman, A. 2010. Harvesting image databases from the web. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint optimization framework for learning with noisy labels. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Vapnik, V. N. 1999. An overview of statistical learning theory. *IEEE Transactions on neural networks* .

Xia, X.; Liu, T.; Han, B.; Wang, N.; Gong, M.; Liu, H.; Niu, G.; Tao, D.; and Sugiyama, M. 2020. Part-dependent label noise: Towards instance-dependent label noise. In *Advances in Neural Information Processing Systems*.

Xia, X.; Liu, T.; Wang, N.; Han, B.; Gong, C.; Niu, G.; and Sugiyama, M. 2019. Are Anchor Points Really Indispensable in Label-Noise Learning? In *Advances in Neural Information Processing Systems*.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Xu, Y.; Cao, P.; Kong, Y.; and Wang, Y. 2019. L_DMI: A Novel Information-theoretic Loss Function for Training Deep Nets Robust to Label Noise. In *Advances in Neural Information Processing Systems*.

Yan, Y.; Rosales, R.; Fung, G.; Subramanian, R.; and Dy, J. 2014. Learning from multiple annotators with varying expertise. *Machine learning* .

Yao, Q.; Yang, H.; Han, B.; Niu, G.; and Kwok, J. T.-Y. 2020a. Searching to exploit memorization effect in learning with noisy labels. In *International Conference on Machine Learning*.

Yao, Y.; Liu, T.; Han, B.; Gong, M.; Deng, J.; Niu, G.; and Sugiyama, M. 2020b. Dual T: Reducing estimation error for transition matrix in label-noise learning. In *Advances in Neural Information Processing Systems*.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does Disagreement Help Generalization against Label Corruption? In *International Conference on Machine Learning*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. In *Proceedings of the British Machine Vision Conference*.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

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

Zheng, S.; Wu, P.; Goswami, A.; Goswami, M.; Metaxas, D.; and Chen, C. 2020. Error-Bounded Correction of Noisy Labels. In *International Conference on Machine Learning*.