USDNL: Uncertainty-Based Single Dropout in Noisy Label Learning

Yuanzhuo Xu[1], Xiaoguang Niu[1,4*], Jie Yang[1], Steve Drew[2], Jiayu Zhou[3], Ruizhi Chen[4]

[1]School of Computer Science, Wuhan University, China
[2] Department of Electrical and Software Engineering, University of Calgary, Canada
[3] Department of Computer Science and Engineering, Michigan State University, USA
[4] LIESMARS, Wuhan University, China
xyzxyz, xgniu, csyangjie, Ruizhi.chen@whu.edu.cn, steve.drew@ucalgary.ca, jiayuz@msu.edu

Abstract

Deep Neural Networks (DNNs) possess powerful prediction capability thanks to their over-parameterization design, although the large model complexity makes it suffer from noisy supervision. Recent approaches seek to eliminate impacts from noisy labels by excluding data points with large loss values and showing promising performance. However, these approaches usually associate with significant computation overhead and lack of theoretical analysis. In this paper, we adopt a perspective to connect label noise with epistemic uncertainty. We design a simple, efficient, and theoretically provable robust algorithm named *USDNL* for DNNs with uncertainty-based dropout. Specifically, we estimate the epistemic uncertainty of the network prediction after early training through single dropout. The epistemic uncertainty is then combined with cross-entropy loss to select the clean samples during training. Finally, we theoretically show the equivalence of replacing selection loss with single cross-entropy loss. Compared to existing small-loss selection methods, USDNL features its simplicity for practical scenarios by only applying dropout to a standard network, while still achieving high model accuracy. Extensive empirical results on both synthetic and real-world datasets show that USDNL outperforms other methods. Our code is available at https://github.com/kovelxyz/USDNL.

Introduction

Deep neural networks (DNNs) have shown impressive capability (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016) to capture complicated patterns between instances and labels when trained with large-scale datasets. The sheer amount of data collected from Web-based image crawling, crowdsourcing, and other possible sources in scale, inevitably brings label noise and severely impacts model performance due to the memorization of noisy labels by DNNs (Arpit et al. 2017; Yao et al. 2020).

General regularization methods (Zhang et al. 2021) fail in noisy label learning due to the strong fitting ability of the model. To alleviate such problems, existing works (Patrini et al. 2017; Xia et al. 2019; Natarajan et al. 2013) focused on correcting the loss by estimating the latent noisy transition matrix, with the requirement of prior knowledge or a subset of data to be correctly labeled for reaching high accuracy. Prior art with theoretically supported algorithms (Liu and Guo 2020; Cheng et al. 2020; Englesson and Azizpour 2021; Wu et al. 2021) guaranteed the robustness of noisy labels by constructing a loss function or a regularization scheme while still failing at high noise rates. Another large family of methods (Cheng et al. 2020; Zhou, Wang, and Bilmes 2020; Nishi et al. 2021; Berthon et al. 2021; Qu, Mo, and Niu 2021) focused on distilling a subset of clean samples from the training data by leveraging the gradient flow or the statistical error of the network on false positive samples. The state-of-the-art stream of methods (Han et al. 2018; Yu et al. 2019; Wei et al. 2020; Yao et al. 2021b; Xia et al. 2022) generated two models from the same sample to select a small subset of samples with small losses as clean labels. While achieving promising performance gains with noisy labels, the small-loss sample selection strategy still faces challenges. Performance-wise, training two models double the parameter quantity and computational overhead. In resource-constrained or time-sensitive environments, the additional time and resource costs are unacceptable due to the resource and latency constraints. Additionally, these selection strategies are often heuristic and lack rigorous theoretical support.

To address the challenges introduced by the small-loss sample selection strategy, in this paper, we adopt a perspective on label noise learning with epistemic uncertainty estimation. Inspired by this connection, we propose USDNL, a simple yet effective algorithm based on a single network with a single dropout (i.e., forward once with dropout) to combat label noise. Specifically, we model the network prediction confidence for noisy and clean samples after early-stage training with epistemic uncertainty estimation. This model is then implemented in DNNs via multiple Monte Carlo dropout samplings. Further, we combine the epistemic uncertainty estimation with prediction cross-entropy to select clean samples. Finally, we theoretically demonstrate the count of dropout required for epistemic uncertainty estimation and cross-entropy computation can be reduced to just one in the clean samples selection task.

Our proposed noisy label learning framework USDNL is easy to deploy with no change to existing network structures. USDNL incurs substantially lower computational overhead and latency, and outperforms other state-of-the-art methods

*The corresponding author.

after experiments on MNIST, CIFAR-10, CIFAR-100, and Clothing1M datasets with various noise settings.

Related Literature

Noise transition matrix estimation. Earlier research focused on estimating the noise rate to construct the noise transition matrix (Patrini et al. 2017; Xia et al. 2019; Natarajan et al. 2013; Zhang, Niu, and Sugiyama 2021). T-Revision (Xia et al. 2019) learned the noise transition matrix utilizing transition-revision. Further, the authors in (Zhu, Song, and Liu 2021) proposed a method for estimating anchor points using cluster representations. The transition matrices were identified in (Li et al. 2021) when the clean class posteriors were sufficiently dispersed. Zhang, Niu, and Sugiyama (2021) proposed a method to estimate the transition matrix with the total variation regularization. Nevertheless, these noise rate estimation methods have constraints and usually fail at high noise rates.

Deep learning methods. There have been numerous articles leveraging deep learning methods (Goldberger and Ben-Reuven 2017; Yi and Wu 2019; Yao et al. 2021a; Zhang, Lee, and Agarwal 2021; Xia et al. 2021) to deal with the label noise. For composite multi-module networks, PENCIL (Yi and Wu 2019) supervised network training by modeling noisy labels and updating the noise distribution during back-propagation. Yao et al. (2021a) introduced a structural causal model to instance-dependent label noise modeling through VAE to train robust classifiers. During the early training of a network, the side effects of noisy labels could be reduced by dividing the critical and non-critical parameters (Xia et al. 2021). Li et al. (2022) proposed Sel-CL to select confidence pairs from the noisy datasets for supervised contrastive learning. For network predictions, a simple method was presented (Zhang, Lee, and Agarwal 2021) without any changes to the training but only performed class probability estimation (CPE) on the noisy examples. More lightweight, rectified loss functions were used in (Wu et al. 2021; Hendrycks et al. 2018; Liu and Guo 2020; Zhou et al. 2021; Liu et al. 2021) to robustify the classifiers with little overhead.

Clean samples selection. Another class of methods for label noise learning is to select clean samples for the classifier to learn. Previous work (Arpit et al. 2017) showed that the networks preferentially fit clean samples, making it possible for small-loss selection. Among these methods, the most representative category is called co-training strategy (Wei et al. 2020; Yao et al. 2021b; Han et al. 2018; Yu et al. 2019). JoCoR (Wei et al. 2020) used two networks to make predictions on the data in the same mini-batch. It applied co-regularization constraints and computed a joint loss for each training example to filter out small loss samples. CNLCU (Xia et al. 2022) adopted interval estimation combined with loss uncertainty for sample selection. Other selection methods did not employ dual network training. RCL (Sun et al. 2020) exploited the divergence and agreement between multiple network learning to combat label noise. Kim et al. (2021) focused on latent representation dynamics and filtered instances by using eigendecomposition to measure the distance between the latent distribution and each representation. UniCon (Karim et al. 2022) introduced the unified selection mechanism of Jensen-Shannon divergence to avoid the class imbalance problem of selected clean sets. Compared with these selection methods, USDNL, derived from uncertainty estimation, uses a single network with a single dropout sampling to select clean samples, which is simple in practice and has little overhead.

Preliminaries

Problem setting Suppose $(x, \tilde{y})$ is a sample pair with label noise drawn from $\tilde{\mathcal{H}}$, denoted by $\tilde{\mathcal{H}} := \{(x_1, \tilde{y}_1), \cdots, (x_n, \tilde{y}_n)\} \in (X \times \tilde{Y})$, where $x_n \in X$ denotes the target data space and $\tilde{y} \in \tilde{Y}$ is the corresponding noise label. Each noise example $(x_n, \tilde{y}_n) \in \tilde{\mathcal{H}}$ has an unobservable clean pair $(x_n, y_n) \in \mathcal{H}$. This paper targets to propose a progressive approach capable of distilling and learning clean samples, resulting in a classifier robust to label noise.

Dropout regularization Dropout (Wager, Wang, and Liang 2013) is a common regularization technique used in deep learning to prevent overfitting and improve the generalization ability of the model. We denote $\mathcal{P}^w(y|x)$ the network with parameters w. With dropout, we sample variables in Bernoulli distribution for every unit of the network with the possible rate Φ (i.e., dropout rate) and obtain the sub-model $\mathcal{P}^{\hat{w}}(y|x)$. A unit is dropped once its corresponding variable takes 0. Given a sample set $\{\mathbf{x_i}, \mathbf{y_i}\}_1^M$, we apply L_2 regularization in NN training, resulting in the minimization objective with dropout as follows:

$$\tilde{\mathcal{L}} = \frac{1}{M} \sum\nolimits_{i=1}^{M} -\log \mathcal{P}^{\hat{w}_i}(\mathbf{y_i}|\mathbf{x_i}) + \lambda \left\|\hat{w}_i\right\|_2^2 \qquad (1)$$

A Perspective from Epistemic Uncertainty to Noisy Label Learning

Uncertainty of Neural Network

The uncertainties of neural networks in deep learning include aleatoric uncertainty and epistemic uncertainty (Kendall and Gal 2017). For a fully trained model, given a test sample outside the training set, the aleatoric uncertainty is determined by the inherent noise of the observation. In contrast, the epistemic uncertainty captures the ignorance of the samples beyond the training distribution. For epistemic uncertainty, we have the following property.

Property 1 *The epistemic uncertainty demonstrates the confidence of the network in a prediction. Given a sample* ***x*** *and trained network* ***f*** *on learned distribution* ***S****, perform the epistemic uncertainty estimation H(x) of x. Then H(x) is lower (i.e., more confident) while* ***x*** *is highly correlated with* ***S*** *and higher (i.e., less confident) on the contrary.*

Property 1 reveals that the relevance between an input sample and the learned distribution can be estimated from the epistemic uncertainty of the network prediction, i.e., whether or not the features of inputs have been learned in past training.

Uncertainty in Noisy Label Learning

In noisy label learning tasks, aleatoric uncertainty is not our main consideration because the network always traverses the same training set. In early learning stages, the network preferentially fits clean labels (Arpit et al. 2017). After the early learning phase, we can divide the training set distribution S into two subspaces: learned distributions S_c, and unlearned distributions S_d. When the network traverses the training set again, it will exhibit higher confidence on samples from the learned distributions S_c, while lower confidence on those from unlearned distributions S_d. Therefore, we can select samples with higher confidence (i.e., more likely to be clean) by estimating their epistemic uncertainty after early training stages based on Property 1. We demonstrate this process through a toy example.

Intuition. Given a noisy dataset $D = \{(x, y_c), (x, y_n), (x_{ood}, y)\}$ and a network f, where (x, y_c) denotes a clean sample and (x, y_n) represents its noisy version. Define (x_{ood}, y) as a sample outside the clean distribution. In the early training stage, f preferentially fits (x, y_c). We implement the epistemic uncertainty estimation of f on D. According to Property. 1, the uncertainty score of (x, y_c) and (x, y_n) is low as the x is learned, and $H(x_{ood})$ is high as x_{ood} is not learned in the early training. Thus, we can select the $(x, y_c), (x, y_n)$ by the epistemic uncertainty estimation.

Estimate Epistemic Uncertainty in DNNs

Uncertainty estimation generally needs to be carried out under the framework of Bayesian networks. Fortunately, the work by Kendall and Gal (2017) allows us to implement the epistemic uncertainty estimation in DNNs training via multiple Monte Carlo (MC) dropout samplings.

Suppose there are T sampled masked model weights $\hat{w}_t \in q_\theta(w)$, where $q_\theta(w)$ is the dropout distribution (i.e., Bernoulli distribution). The prediction uncertainty can be approximated using Monte Carlo integration with the following conditional probability:

$$p(y = k \mid \mathbf{x}, \mathbf{X}, \mathbf{Y}, \mathbf{T}) \approx \frac{1}{T} \sum_{t=1}^{T} \mathrm{Softmax}\left(\mathcal{P}^{\hat{w}_t}(y \mid \mathbf{x})\right) \tag{2}$$

Further, the epistemic uncertainty can be summarized by calculating the average entropy value of the probability vector $\mathbf{p}$:

$$H(\mathbf{p}) = -\sum_{k=1}^{K} p_k \log p_k \tag{3}$$

As shown in the toy example, the epistemic uncertainty $H(\mathbf{p})$ can help us weed out samples outside the clean distribution (i.e., x_{ood}). However, the low epistemic uncertainty (or high prediction confidence) does not indicate that the sample $(x_i, \tilde{y}_i)$ has a clean label because $\tilde{y}_i$ may not match the prediction.

Single Dropout to Combat Label Noisy

Selection Loss with the Epistemic Uncertainty

Samples with lower epistemic uncertainty do not automatically get a pass to have clean labels, as their observed labels may not match the network predictions. We measure the matching gap between a prediction and its observed label by checking the cross-entropy loss. We subsequently mark a sample as clean once it has both low epistemic uncertainty and low matching loss, illustrated below:

$$\mathcal{L}_{sl} = \mathcal{L}_{nll} + \alpha H(\mathbf{p}), \tag{4}$$

where $\mathcal{L}_{nll} = \frac{1}{n}\sum_{i=1}^{n} -\log \mathcal{P}^{\hat{w}_i}(\tilde{y} \mid x)$ is the negative log-likelihood loss function between the observed labels $\tilde{y}$ and network predictions. Additionally, α is the coefficient weight to control $H(\mathbf{p})$.

Single Dropout to Estimate Selection Loss

According to Equation (4), we need to perform multiple dropout to evaluate epistemic uncertainty during the training process. However, multiple dropout are computationally expensive and time-consuming with large datasets. In this section, we first theoretically prove that we only need single dropout while still maintaining its efficiency for both epistemic uncertainty estimation and cross-entropy computation on a clean distribution. We also simplify the selection loss as a single cross-entropy based on the positive correlation of uncertainty and the cross-entropy in a task. Finally, we analyze the efficiency of single dropout on samples outside the clean distribution. Then we give the progressive selection strategy.

Single dropout on clean distribution We first reveal the consistency of a dropout's submodels on a clean distribution (a.k.a learned distribution) in the early learning stages.

Lemma 1 *For a closed dataset training, once the linear model well-fit the clean samples in training set, i.e., $E_t\left(\log \mathcal{P}^{\hat{w}_t}(y_c \mid x_c)\right) \leq \epsilon$, for the fixed learned clean sets $\{x_c, y_c\}$, we have:*

$$E_{t_1,t_2}\left(\|\mathcal{P}^{\hat{w}_{t_1}}(y_c \mid x_c) - \mathcal{P}^{\hat{w}_{t_2}}(y_c \mid x_c)\|\right) \leq c_1\epsilon, \tag{5}$$

where c_1 is the Lipschitz constant.

Lemma 1 indicates the existence of an upper bound on the difference in prediction of the sub-models on the learned distribution (detailed proof in the appendix). This allows us to measure the gap between finite dropout and multiple dropout of epistemic uncertainty estimation.

Theorem 1 ***(Finite dropout on epistemic uncertainty)*** *Define a linear model $\mathcal{P}^{w}(y|x)$ and its sub-model $\mathcal{P}^{\hat{w}}(y|x)$ with w and Bernoulli sampling $\hat{w}$, respectively. We denote $\mathcal{H}$ as the truncated entropy loss function of the sub-model with $\hat{w}$. For the fixed learned clean sets $\{x_c, y_c\}$, we have:*

$$\mathbb{E}_c\left(\left|\mathcal{H}\left(\frac{1}{N}\sum\nolimits_{i=1}^{N}\left(\mathcal{P}^{\hat{w}_i}(y_c \mid x_c)\right)\right) - \mathcal{H}\left(\mathbb{E}\left(\mathcal{P}^{\hat{w}}(y_c \mid x_c)\right)\right)\right|\right) \leq c_2\epsilon, \tag{6}$$

where N is a finite integer (at least 1) and c_2 is the Lipschitz constant.

Theorem. 1 shows that we can represent the epistemic uncertainty of a sub-model by observing a finite number of sub-models (detailed proof in the appendix), This fact makes it possible to evaluate the epistemic uncertainty through a finite number (at least one) of dropout.

The following theorem shows that under the same constraints of Lemma. 1, the gap of the empirical loss between the average of finite sub-models and the expectation of sub-models is also bounded (detailed proof in the appendix).

Theorem 2 ***(Finite dropout on empirical loss)*** *Define a linear model $\mathcal{P}^{w}(y|x)$ and its sub-model $\mathcal{P}^{\hat{w}}(y|x)$ with w and Bernoulli sampling $\hat{w}$, respectively. We denote $\mathcal{L}(\mathcal{P}(\tilde{y}|x)) = -\log(\mathcal{P}(\tilde{y}|x))$ as the cross-entropy loss function of the sub-model with dropout applied on w. For the fixed learned clean sets x_c, y_c, we have*

$$\mathbb{E}_c\left(\left|\frac{1}{N}\sum\nolimits_{i=1}^{N}\left(\mathcal{L}\left(\mathcal{P}^{\hat{w}_i}(\tilde{y}_c \mid x_c)\right)\right) - \mathbb{E}\left(\mathcal{L}\left(\mathcal{P}^{\hat{w}}(\tilde{y}_c \mid x_c)\right)\right)\right|\right) \leq c_3\epsilon, \tag{7}$$

where N is a finite integer (at least 1) and c_3 is the Lipschitz constant satisfying the empirical loss function.

With Theorem 1 and 2, we can limit the count of dropout samplings in Equation (4) to a finite number:

$$\begin{aligned}\mathcal{L}_{sl} =& \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}\left(\mathcal{P}^{\hat{w}_i}(\tilde{y} \mid x)\right) \\ &+ \alpha\mathcal{H}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{P}^{\hat{w}_i}(y \mid x)\right)\right)\end{aligned} \tag{8}$$

where N is a finite integer (at least 1). We set $N = 1$, and the selection loss becomes the weighted sum of epistemic uncertainty and empirical loss with a single dropout. We further show the agreement between the empirical loss term and the epistemic uncertainty estimation with single dropout in the noisy label learning task (detailed proof in the appendix).

Proposition 1 *In a selection task of clean samples $(x_c, \hat{y}_c)$, for a well-trained linear model $\mathcal{P}^{w}(y_c|x_c)$, the empirical loss $\mathcal{L}\left(\mathcal{P}^{\hat{w}}(\tilde{y_c} \mid x_c)\right)$ and the epistemic uncertainty $H(\mathbf{p})$ with single dropout sampling are positively correlated on clean label samples in the learned distribution space.*

The proposition is intuitively obvious because the empirical loss and epistemic uncertainty exhibit consistency for samples already learned. Therefore, we can replace the whole selection loss with a single empirical loss and eventually get the final selection loss with a single dropout:

$$\mathcal{L}_{sl} = \mathcal{L}\left(\mathcal{P}^{\hat{w}}(\tilde{y} \mid x)\right) \tag{9}$$

Analysis of single dropout on samples outside the clean distribution We have demonstrated the effectiveness of single dropout on clean distributions, i.e., the noise corrupts the label, but the instance is uncontaminated. Most of the existing works focused on clean distributions. However, there is another class of datasets whose noisy samples are outside the clean distribution (a.k.a OOD datasets), such as when CIFAR-10 samples are added to MNIST dataset in the classification task. The possible situations lead us to analyze the performance of single dropout on samples outside the clean distribution.

The network still preferentially fits clean-label samples in the early training stage. In the loss selection stage, for OOD samples, the predictions of the sub-model have a certain degree of randomness due to the dropout. This indicates the sub-model may have low uncertainty with a single dropout. However, based on Equation (9), the OOD sample must also achieve small empirical before it is finally selected. Such condition is difficult to achieve. Even in the worst case, i.e., the OOD sample is selected in this forward pass, it is not overfitted by the model in a single backward.

In general, before an OOD sample is overfitted by the network, it needs to be selected in consecutive forward-pass. This is equivalent to performing multiple dropout while showing low selection loss in every training iteration, which is extremely unlikely. Therefore, the effectiveness of a single dropout is also guaranteed on the OOD dataset. We conduct experiments on OOD datasets to corroborate our analysis.

Algorithm 1: The training pipeline of USDNL

Input: Noisy training set $\tilde{D}$, Network f with w, learning rate η, fixed τ, warm-up epoch T_k and total training epoch T_{max}, total iteration of each epoch I_{max};

for $t = 1, 2, \cdots, T_{max}$ **do**
  Shuffle the training set $\tilde{D}$;
  for $n = 1, 2, \cdots, I_{max}$ **do**
    Fetch mini-batch $\tilde{D}_n$ from $\tilde{D}$ and a sub-model with $\hat{w}$;
    $\hat{y} = f^{\hat{w}}(x), \forall x \in \hat{D}_n$;
    Calculate the selection loss $\mathcal{L}_{sl}$ by Equation (9) using $\hat{y}$ and $\tilde{y}$;
    Obtain the small-loss set $\hat{D}_n$ by Equation (10) from $\tilde{D}_n$;
    Calculate the final loss $\hat{\mathcal{L}}$ on $\hat{D}_n$ by Equation(11);
    Update $w = w - \eta\nabla\hat{\mathcal{L}}$;
  end for
  Update $R(t) = 1 - min\left\{\frac{t}{T_k}\tau, \tau\right\}$
end for

Output: w.

Progressive Clean Sample Selection

We employ a progressive decay learning schedule similar to JoCoR (Wei et al. 2020) for small error sample selection w.r.t. Equation (9). Applying a single dropout to the network during training, we reserve more small error samples for warm-up training and gradually reduce the number of reserved samples to avoid the network fitting noisy samples:

$$\hat{\mathcal{D}}_n = \arg\min_{D'_n:|D'_n|\geq R(t)|D|} \mathcal{L}_{sl}(D'_n) \tag{10}$$

where $\mathcal{D}$ is the whole dataset, and $R(t)$ is the selected ratio decay schedule. After getting a small error sample set $\hat{\mathcal{D}}_n$, we compute its error and feed it into the network for

Type	Dataset	Method	Sym-20%	Sym-50%	Sym-80%	Asym-20%	Asym-45%
CCN	MNIST	Standard	82.66 ± 0.38	60.35 ± 0.40	26.33 ± 0.43	90.25 ± 0.18	75.76 ± 0.04
		Co-teaching	97.86 ± 0.01	94.04 ± 0.10	86.84 ± 0.36	99.25 ± 0.05	85.68 ± 4.66
		Co-teaching+	97.86 ± 0.01	96.43 ± 0.06	71.81 ± 0.16	97.88 ± 0.03	82.41 ± 2.88
		JoCoR	98.23 ± 0.04	96.54 ± 0.12	84.58 ± 4.34	98.29 ± 0.47	92.19 ± 1.14
		CDR	98.95 ± 0.07	98.44 ± 0.03	84.44 ± 0.25	99.17 ± 0.02	94.39 ± 0.15
		USDNL	**99.20 ± 0.03**	**98.76 ± 0.03**	**96.61 ± 0.15**	**99.28 ± 0.03**	**98.99 ± 0.01**
	CIFAR-10	Standard	69.54 ± 0.14	42.53 ± 0.54	16.07 ± 0.32	78.87 ± 0.17	67.36 ± 0.21
		Co-teaching	82.26 ± 0.10	73.18 ± 0.19	25.45 ± 2.20	86.63 ± 0.07	62.49 ± 2.61
		Co-teaching+	57.35 ± 0.83	48.77 ± 0.08	10.61 ± 0.47	57.91 ± 0.38	46.00 ± 0.41
		JoCoR	85.69 ± 0.04	79.26 ± 0.45	28.09 ± 2.08	83.50 ± 0.10	70.02 ± 0.67
		CDR	84.10 ± 0.12	71.62 ± 0.32	40.41 ± 0.32	88.01 ± 0.27	75.93 ± 0.26
		CNLCU-H	90.21 ± 0.10	66.57 ± 0.08	27.29 ± 0.12	89.09 ± 0.12	79.10 ± 0.19
		USDNL	**91.91 ± 0.02**	**87.37 ± 0.12**	**44.63 ± 0.88**	**91.85 ± 0.10**	**82.90 ± 0.24**
	CIFAR-100	Standard	35.64 ± 0.12	17.53 ± 0.07	4.30 ± 0.17	39.54 ± 0.30	24.49 ± 0.35
		Co-teaching	50.62 ± 0.14	37.26 ± 0.29	12.54 ± 0.39	48.80 ± 0.37	29.22 ± 0.25
		Co-teaching+	49.31 ± 0.27	39.95 ± 0.33	12.27 ± 1.43	48.83 ± 0.19	28.90 ± 0.28
		JoCoR	53.36 ± 0.13	43.38 ± 0.14	12.96 ± 0.57	47.04 ± 0.20	28.29 ± 0.29
		CDR	60.12 ± 0.14	47.17 ± 0.11	23.40 ± 0.11	61.01 ± 0.22	39.60 ± 0.18
		CNLCU-H	64.05 ± 0.13	38.92 ± 0.07	8.11 ± 0.03	61.30 ± 0.20	33.92 ± 0.02
		USDNL	**70.69 ± 0.07**	**63.42 ± 0.33**	**27.44 ± 0.32**	**66.35 ± 0.32**	**41.24 ± 0.25**
OOD	CIFAR80N-O	Standard	42.57 ± 0.28	27.06 ± 0.04	9.27 ± 0.07	44.03 ± 0.17	29.81 ± 0.30
		Co-teaching	58.97 ± 0.21	45.61 ± 0.48	12.98 ± 0.84	56.76 ± 0.06	37.01 ± 0.26
		Co-teaching+	60.67 ± 0.37	52.23 ± 0.07	3.70 ± 0.84	60.38 ± 0.22	**46.58 ± 1.15**
		JoCoR	58.37 ± 0.31	51.56 ± 0.34	12.83 ± 0.45	55.17 ± 0.14	34.04 ± 0.28
		CDR	59.78 ± 0.02	46.96 ± 0.17	22.32 ± 0.20	59.72 ± 0.06	39.14 ± 0.28
		CNLCU-H	61.30 ± 0.12	34.10 ± 0.46	6.99 ± 0.47	57.50 ± 0.70	34.89 ± 0.42
		USDNL	**71.54 ± 0.35**	**63.98 ± 0.21**	**26.07 ± 0.36**	**69.51 ± 0.17**	43.74 ± 0.50

Table 1: Average test accuracy (%) on CCN and OOD dataset over the last 10 epochs.

backpropagation w.r.t. Equation (1). USDNL is shown in Algorithm 1.

$$\hat{\mathcal{L}} = \frac{1}{|\hat{\mathcal{D}}_n|} \sum_{\boldsymbol{x} \in \hat{\mathcal{D}}_n} \mathcal{L}_{sl}(\boldsymbol{x}) \tag{11}$$

Experiment

In this section, we first describe the settings of the experiment, then show and analyze the results compared to other state-of-art methods. We finally perform the algorithm complexity analysis and the ablation study of the single dropout.

Experiment Setup

Datasets We verify the effectiveness of USDNL on four manually corrupted datasets, i.e., *MNIST* (LeCun 1998), *CIFAR-10*, *CIFAR-100* (Krizhevsky 2009) with artificial corruption, along with a real-world noisy dataset *Clothing1M* (Xiao et al. 2015). The *CIFAR-10* and *CIFAR-100* datasets are artificially corrupted into three categories: class conditional noise (CCN), out-of-distribution (OOD) noise, and instance-dependent noise (IDN). Specifically, we construct symmetric and asymmetric noise of the CCN dataset, and then introduce out-of-distribution samples to build an OOD dataset. We corrupt the labels according to instance features to construct an IDN dataset. No need to construct manually, *Clothing1M* has one million images with real-world noisy labels and ten thousand images with clean labels. (The dataset construction details can be found in the appendix)

Baseline and evaluation metrics We compare USDNL with several state-of-the-art algorithms: Co-teaching (Han et al. 2018), Co-teaching+ (Yu et al. 2019), JoCoR (Wei et al. 2020), CDR (Xia et al. 2021), CNLCU-H (Huang et al. 2022) and Standard. We use the hyperparameters specified in the code provided by the authors. When it comes to the adjustment of the classifier network, we select the hyperparameters according to the requirements of the paper. For metrics, follow (Wei et al. 2020; Han et al. 2018; Yu et al. 2019; Xia et al. 2019; Yao et al. 2021a,b), we evaluate USDNL and other state-of-art methods by comparing the test accuracy on clean sets and the label precision under all noise settings.

Network setting For a fair comparison, different schemes use the same network structure on the same dataset. Specifically, we use a LeNet for *MNIST*, a 9-CNN network for *CIFAR-10* and *CIFAR-100*, and non-pretrained Resnet-18 for *Clothing1M*. USDNL adds several dropout layers to the standard network. Without complex parameter tuning on different datasets like other methods, we incorporate a unified dropout rate of 0.25.

Type	Dataset	Method	20%	30%	40%	50%
IDN	CIFAR10	Standard	71.39 ± 0.41	63.24 ± 0.10	53.98 ± 0.24	43.38 ± 0.54
		Co-teaching	85.54 ± 0.04	82.45 ± 0.06	78.70 ± 0.19	64.48 ± 2.90
		Co-teaching+	61.50 ± 0.48	57.63 ± 0.49	52.26 ± 0.17	40.99 ± 5.60
		JoCoR	87.46 ± 0.36	84.93 ± 0.55	78.26 ± 0.15	58.14 ± 6.22
		CDR	77.24 ± 0.28	68.52 ± 0.50	57.81 ± 0.61	46.55 ± 0.62
		CausalNL	81.47 ± 0.32	80.38 ± 0.37	77.53 ± 0.45	**77.39 ± 1.24**
		CNLCU-H	87.79 ± 0.15	83.98 ± 0.03	71.98 ± 0.07	55.24 ± 1.15
		USDNL	**87.90 ± 0.38**	**85.23 ± 0.53**	**80.55 ± 0.65**	68.10 ± 3.59
	CIFAR100	Standard	39.25 ± 0.66	33.31 ± 0.39	27.89 ± 0.61	22.03 ± 0.05
		Co-teaching	55.63 ± 0.11	51.34 ± 0.14	45.37 ± 0.66	36.21 ± 0.32
		Co-teaching+	54.30 ± 0.25	52.85 ± 0.08	49.66 ± 0.15	42.13 ± 0.62
		JoCoR	58.57 ± 0.04	54.14 ± 0.94	47.98 ± 0.55	38.15 ± 0.26
		CDR	54.34 ± 0.59	47.61 ± 0.41	40.21 ± 0.25	31.85 ± 0.25
		CausalNL	41.47 ± 0.32	40.98 ± 0.362	34.02 ± 0.95	32.129 ± 2.23
		CNLCU-H	58.03 ± 0.19	50.02 ± 0.18	40.75 ± 1.25	31.84 ± 0.24
		USDNL	**64.82 ± 0.32**	**61.35 ± 0.29**	**55.82 ± 0.56**	**46.00 ± 0.62**

Table 2: Average test accuracy (%) on IDN-CIFAR10 and IND-CIFAR100 over the last 10 epochs.

Method	Best	Last
Standard	67.22	64.68
co-teaching	69.21	68.51
Decouping	68.48	67.32
JoCoR	70.30	69.70
JoSRC	71.78	70.69
Class2Simi	71.43	71.25
BLTM	71.32	70.89
CNLCU-H	71.08	70.27
USDNL	**72.17**	**71.74**

Table 3: Classification accuracy (%) on *Clothing1M* under the structure of ResNet-18.

Comparison Results

Results on the CCN dataset Table 1 shows the test accuracy compared with other state-of-art methods on CCN-*MNIST*. We achieve the optimal results across all noise settings, especially in the case of Sym-80%, where the test accuracy of USDNL only drops by at most 2%.

Table 1 shows the experimental results on *CIFAR-10* and *CIFAR-100* with different noise rates under two types of noise settings. USDNL achieves the best results in all competitions. It is more beneficial to test accuracy, especially in medium and high noise conditions. Figure 1 illustrates the details of test accuracy and label precision v.s. epochs (see result of *CIFAR-100* in the appendix). The high initial test accuracy proves that the model preferentially fits clean labels in early training stages. Subsequent declines indicate that overfitting has occurred, but robust methods stopped or alleviated this trend. USDNL achieves consistently higher test accuracy with less training time, thanks to the more precise clean sample selection process compared to other methods, shown in Figure 1.

Results on the OOD dataset Table 1 shows the results on *CIFAR-80N-O*. The detailed test accuracy and label precision can be found in the the appendix. USDNL achieves the optimal performance except for pairflip-45% noise rate. The small gap in the noise rate of 45% may be due to the fact that a large proportion of two-class label flips affected the convergence of the network's early learning, as the ratio of correct and incorrect labels is close. The experimental results verify USDNL's excellent ability to resist OOD noise samples, and further justify our previous analysis.

Results on the IDN dataset Tabel. 2 shows the experimental results in the IDN version of CIFAR. Instance-dependent label noise is more challenging to classify. Nevertheless, USDNL achieves excellent performance on almost all settings against the IDN dataset, especially under complex classes and high noise conditions. We noticed that the accuracy of USDNL under the 50% noise rate of CIFAR10 has decreased, and the reason we analyze is that the ability of uncertainty estimation is reduced in extremely high noise case of IDN. The results prove the great robustness of USDNL under instance-dependent noise with no special network design, as well as better capability to select clean samples. The detailed test accuracy and label precision v.s. epochs can be found in the appendix.

Results on the real-world noisy dataset Finally, we test USDNL against other baselines on the real-world dataset *Clothing1M*. We add additional state-of-the-art methods JoSRC (Yao et al. 2021b), Class2Simi (Wu et al. 2021) and BLTM (Yang et al. 2022) to the baseline. The results are shown in Table 3. USDNL outperforms all other baselines with the same classifier *ResNet18*, further demonstrating the outstanding robustness of USDNL to real-world noise.

Single Dropout *vs* Multiple Dropout

In previous sections, we have theoretically proved that the dropout operation can be constrained to once. To verify the theory, we conduct experiments with different dropout times. The network loss under multiple dropout follows Equation (8). Figure 2 shows the experimental results on

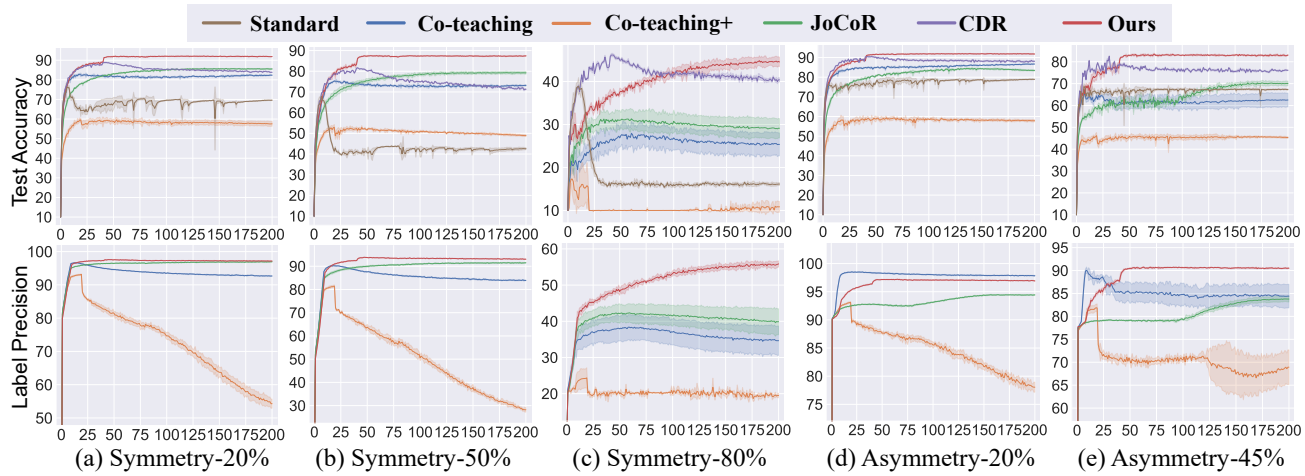

Figure 1: CCN results on CIFAR-10 dataset. Top: test accuracy(%) vs. epochs; bottom: label precision(%) vs. epochs.

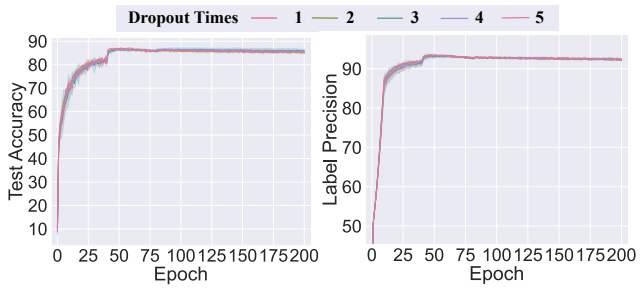

Figure 2: The results of dropout times on CIFAR-10.

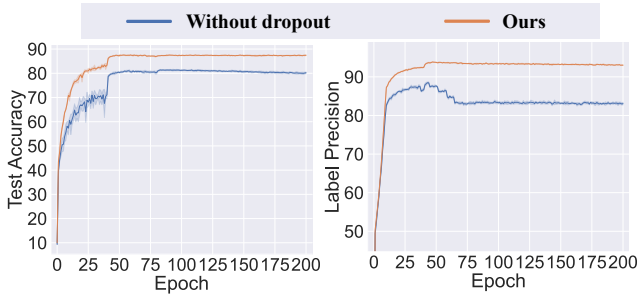

Figure 3: The ablation results w/o dropout on CIFAR-10.

CCN-CIFAR-10-Sym50%, with the number of dropout increasing from 1 to 5. The results show that there is almost no difference in the prediction accuracy between multiple dropout and single dropout, which further validates our theory.

Ablation Study and Complexity Analysis

To verify the effectiveness of the dropout, we conduct ablation experiments on the CIFAR dataset. When the dropout is removed, USDNL becomes the standard method with label selection (Wei et al. 2020). Figure 3 shows the results of the ablation experiments on CCN-CIFAR-10-Sym50%. It is obvious that the models without dropout quickly fit the noisy labels and eventually affect the selection of clean samples. The ablation results demonstrate that simply adding dropout to the general single network will lead to excellent robustness to label noise. To compare algorithm complexity, we run each algorithm for 5 epochs on the RTX-2080Ti platform and count the mean and standard deviation of a single epoch run. Table 4 shows the result on IDN-CIFAR-10-Sym50%, demonstrating a significantly lower computational overhead of USDNL.

Standard	Co-teaching	Co-teaching+	JoCoR
29.1±0.7	49.4±2.2	44.6±3.5	54.7±2.0
CDR	CausalNL	CNLCU-H	USDNL
38.7±1.3	112.4±6.2	66.1±1.2	**32.3±0.1**

Table 4: Training time comparison under 9-CNN network (*second/epoch*).

Conclusion

This paper proposes a simple, effective, and general algorithm named USDNL to combat label noise in DNNs. With dropout applied on general networks, we demonstrate theoretically and experimentally the use of single dropout to estimate uncertainties. We further select the clean samples by single cross-entropy loss with the dropout. We conduct experiments on various artificial datasets (i.e., *CCN, IDN, OOD*) and real-world noisy datasets (i.e., *Clothing1M*). The results demonstrate that USDNL achieves the best performance in almost all settings over other state-of-the-art methods with less computational overhead. The ablation study on *CIFAR* shows the effectiveness of the dropout in USDNL, which again validates our theorem.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China NSFC under Grant 61872431; in part by the Wuhan University—Huawei GeoInformatics Innovation Lab, in part by University of Calgary Start-Up Funding 10032260. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Arpit, D.; Jastrzębski, 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*, 233–242. PMLR.

Berthon, A.; Han, B.; Niu, G.; Liu, T.; and Sugiyama, M. 2021. Confidence scores make instance-dependent label-noise learning possible. In *International Conference on Machine Learning*, 825–836. PMLR.

Cheng, H.; Zhu, Z.; Li, X.; Gong, Y.; Sun, X.; and Liu, Y. 2020. Learning with instance-dependent label noise: A sample sieve approach. *arXiv preprint arXiv:2010.02347*.

Englesson, E.; and Azizpour, H. 2021. Generalized Jensen-Shannon Divergence Loss for Learning with Noisy Labels. *Advances in Neural Information Processing Systems*, 34.

Goldberger, J.; and Ben-Reuven, E. 2017. Training deep neural-networks using a noise adaptation layer. *International Conference on Machine Learning*.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hendrycks, D.; Mazeika, M.; Wilson, D.; and Gimpel, K. 2018. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems*, 31.

Huang, Y.; Bai, B.; Zhao, S.; Bai, K.; and Wang, F. 2022. Uncertainty-aware Learning Against Label Noise on Imbalanced Datasets. In *Association for the Advancement of Artifcial Intelligence*.

Karim, N.; Rizve, M. N.; Rahnavard, N.; Mian, A.; and Shah, M. 2022. UNICON: Combating Label Noise Through Uniform Selection and Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9676–9686.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Kim, T.; Ko, J.; Cho, S.; Choi, J.; and Yun, S.-Y. 2021. FINE Samples for Learning with Noisy Labels. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Tront*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

LeCun, Y. 1998. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Li, J.; Zhang, M.; Xu, K.; Dickerson, J. P.; and Ba, J. 2021. How does a Neural Network's Architecture Impact its Robustness to Noisy Labels? In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Li, S.; Xia, X.; Ge, S.; and Liu, T. 2022. Selective-Supervised Contrastive Learning with Noisy Labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, B.; Sun, M.; Wang, D.; Tan, P.-N.; and Zhou, J. 2021. Learning Deep Neural Networks under Agnostic Corrupted Supervision. In *International Conference on Machine Learning*, 6957–6967. PMLR.

Liu, Y.; and Guo, H. 2020. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, 6226–6236. PMLR.

Natarajan, N.; Dhillon, I. S.; Ravikumar, P. K.; and Tewari, A. 2013. Learning with noisy labels. *Advances in neural information processing systems*, 26.

Nishi, K.; Ding, Y.; Rich, A.; and Hollerer, T. 2021. Augmentation Strategies for Learning With Noisy Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8022–8031.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1944–1952.

Qu, Y.; Mo, S.; and Niu, J. 2021. DAT: Training Deep Networks Robust to Label-Noise by Matching the Feature Distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6821–6829.

Sun, M.; Xing, J.; Chen, B.; and Zhou, J. 2020. Robust collaborative learning with noisy labels. In *2020 IEEE International Conference on Data Mining (ICDM)*, 1274–1279. IEEE.

Wager, S.; Wang, S.; and Liang, P. S. 2013. Dropout training as adaptive regularization. *Advances in neural information processing systems*, 26.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13726–13735.

Wu, S.; Xia, X.; Liu, T.; Han, B.; Gong, M.; Wang, N.; Liu, H.; and Niu, G. 2021. Class2simi: A noise reduction perspective on learning with noisy labels. In *International Conference on Machine Learning*, 11285–11295. PMLR.

Xia, X.; Liu, T.; Han, B.; Gong, C.; Wang, N.; Ge, Z.; and Chang, Y. 2021. Robust early-learning: Hindering the memorization of noisy labels. In *International Conference on Learning Representations*.

Xia, X.; Liu, T.; Han, B.; Gong, M.; Yu, J.; Niu, G.; and Sugiyama, M. 2022. Sample Selection with Uncertainty of Losses for Learning with Noisy Labels. In *International Conference on Learning Representations*.

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? *Advances in Neural Information Processing Systems*, 32.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2691–2699.

Yang, S.; Yang, E.; Han, B.; Liu, Y.; Xu, M.; Niu, G.; and Liu, T. 2022. Estimating instance-dependent bayes-label transition matrix using a deep neural network. In *International Conference on Machine Learning*, 25302–25312. PMLR.

Yao, Q.; Yang, H.; Han, B.; Niu, G.; and Kwok, J. T.-Y. 2020. Searching to exploit memorization effect in learning with noisy labels. In *International Conference on Machine Learning*, 10789–10798. PMLR.

Yao, Y.; Liu, T.; Gong, M.; Han, B.; Niu, G.; and Zhang, K. 2021a. Instance-dependent Label-noise Learning under a Structural Causal Model. *Advances in Neural Information Processing Systems*, 34.

Yao, Y.; Sun, Z.; Zhang, C.; Shen, F.; Wu, Q.; Zhang, J.; and Tang, Z. 2021b. Jo-src: A contrastive approach for combating noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5192–5201.

Yi, K.; and Wu, J. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7017–7025.

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*, 7164–7173. PMLR.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115.

Zhang, M.; Lee, J.; and Agarwal, S. 2021. Learning from noisy labels with no change to the training process. In *International Conference on Machine Learning*, 12468–12478. PMLR.

Zhang, Y.; Niu, G.; and Sugiyama, M. 2021. Learning noise transition matrix from only noisy labels via total variation regularization. In *International Conference on Machine Learning*, 12501–12512. PMLR.

Zhou, T.; Wang, S.; and Bilmes, J. 2020. Robust curriculum learning: from clean label detection to noisy label self-correction. In *International Conference on Learning Representations*.

Zhou, X.; Liu, X.; Jiang, J.; Gao, X.; and Ji, X. 2021. Asymmetric loss functions for learning with noisy labels. In *International Conference on Machine Learning*, 12846–12856. PMLR.

Zhu, Z.; Song, Y.; and Liu, Y. 2021. Clusterability as an alternative to anchor points when learning with noisy labels. In *International Conference on Machine Learning*, 12912–12923. PMLR.