Deep Learning from Noisy Labels via Robust Nonnegative Matrix Factorization-Based Design

Daniel Grey Wolnick
School of EECS
Oregon State University
Corvallis, OR, USA
wolnickd@oregonstate.edu

Shahana Ibrahim
School of EECS
Oregon State University
Corvallis, OR, USA
ibrahish@oregonstate.edu

Tim Marrinan
Pacific Northwest National Lab
Seattle, WA, USA
timothy.marrinan@pnnl.gov

Xiao Fu
School of EECS
Oregon State University
Corvallis, OR, USA
xiao.fu@oregonstate.edu

***Abstract*—Deep neural networks (DNN) heavily rely on labeled data for supervised training. However, acquiring accurate labels is often a challenging task. Moreover, DNNs easily overfit to noisy labels, hindering the generalization ability. Modeling the label noise using a "confusion matrix" is a widely adopted strategy under such circumstances. A recent work dealt with this problem using a regularizer that is reminiscent of minimum-volume enclosing simplex (MVES)-based matrix factorization. MVES is known for its identifiability of the latent factors, which in turn helps accurately estimate the confusion matrix and rectify its negative effects when training DNNs. However, MVES is highly sensitive to outliers due to its geometric nature. To overcome this limitation, we take insight from the robustification of MVES in the literature to come up with an outlier-resilient noisy label learning criterion. Consequently, when some data samples deviate from the model assumptions, the proposed criterion automatically downweights such outlying data, thereby steering DNN towards identifying the correct model parameters. Our experiment results provide support for the effectiveness of the proposed criterion.**

***Index Terms*—noisy labels, deep learning, volume minimization, nonnegative matrix factorization**

I. Introduction

The success of deep learning can be largely attributed to the use of massive amounts of labeled data. However, collecting accurate labels is a highly nontrivial task. Label noise often arises, especially when the annotators lack expertise [1]. Training deep neural networks (DNNs) using noisy labels often leads to poor generalization performance, as DNNs easily overfit to wrong labels [2], [3]. Hence, it is important to take label noise into consideration when training DNNs.

Many approaches were proposed to learn from noisy labels; see, e.g., [4]–[13]. Among them, an effective and widely used strategy is to explicitly model the label noise generation process using a label transition matrix, namely, the confusion matrix; see [14]–[17]. This leads to a nonnegative matrix factorization (NMF) model [18], [19], with the confusion matrix and the desired (possibly DNN-represented) classifier as the underlying latent factors. Under this model, rectifying the negative impacts brought by the label noise boils down to identifying the latent factors of the NMF model. Several methods were proposed for this purpose, which mostly relied on the so-called *anchor point assumption* [10], [20], which assumes the existence of data items that belong to a specific class with a probability of one. Nonetheless, anchor points are not always available [11].

In the context of NMF, the existence of anchor points is equivalent to the separability condition [18], [21]. It is well-known that the latent factors of the NMF model can be identified without using separability, e.g., via finding a minimum-volume enclosing simplex (MVES) of the data points [22]–[27]. A recent work [12] took the insight of MVES to design a DNN training loss, which was shown to be more effective than the anchor point-based methods. However, MVES is sensitive to outliers due to its geometric nature; see [24], [25], [28], [29]. As a result, when some data deviate from the model assumptions (e.g., when some outlying samples have confusion matrices different from that of the majority of samples), the MVES-based learning criterion may produce poor classification performance.

In this work, we propose to robustify the MVES-based DNN learning criterion. Our approach is inspired by the outlier-robust MVES concept that is often adopted in the hyperspectral unmixing community; see, e.g., [24], [25]. Unlike the hyperspectral imaging works that often use (quasi-)norm based loss functions, the proposed criterion utilizes a robust cross-entropy loss [9], which is better suited for data classification—as the data labels are integers rather than continuous values. The newly designed loss function is differentiable, and thus the involved DNN can be easily learned by back-propagation based algorithms. We train DNN classifiers using the CIFAR-100 [30] and the CIFAR-10 [31] datasets in the presence of outlying samples (whose confusion matrices are not the same as that of the vast majority) and use the test performance to validate our idea.

II. Background

A. Problem Statement

Assume that we have a dataset of size N, denoted as $\{\boldsymbol{x}_n\}_{n=1}^N$, where $\boldsymbol{x}_n \in \mathbb{R}^D$ is the D-dimensional feature vector of the nth data item. Let $y_n \in [K] = \{1, \ldots, K\}$ denote the ground-truth label of $\boldsymbol{x}_n$; i.e., each data item $\boldsymbol{x}_n$ belongs to one of the K classes. Assume that y_n is not available to us.

The work of S. Ibrahim and X. Fu was supported in part by the National Science Foundation (NSF) under Project IIS-2007836. The work of D. G. Wolnick was supported in part by the research experiences for undergraduates (REU) program under NSF IIS-2007836.

Instead, an annotator-produced estimation $\widehat{y}_n$ is given as the label (and it is possible that $\widehat{y}_n \neq y_n$). The goal is to train a classifier $\boldsymbol{f}(\cdot)$ using the dataset $\{\boldsymbol{x}_n\}_{n=1}^N$ and the noisy labels $\{\widehat{y}\}_n^N$ such that $\boldsymbol{f}(\boldsymbol{x}_n) = y_n$ and $\boldsymbol{f}(\boldsymbol{x}_{\rm unseen}) = y_{\rm unseen}$—i.e., a predictor that recognizes the ground-truth class label of both the training and unseen (testing) data samples.

Consider the following generative model for noisy labels [10]–[13]:

$$\begin{aligned}&\mathsf{Pr}(\widehat{y}_n = k|\boldsymbol{x}_n)\\ &\quad= \sum_{k'=1}^K \mathsf{Pr}(\widehat{y}_n = k|y_n = k', \boldsymbol{x}_n)\mathsf{Pr}(y_n = k'|\boldsymbol{x}_n). \quad (1)\end{aligned}$$

The conditional probabilities $\mathsf{Pr}(y_n = k|\boldsymbol{x}_n)$ represents the prediction of the ground-truth labels given the data features $\boldsymbol{x}_n$. This probability distribution is what we aim to learn and is represented by $\boldsymbol{f}(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}^K$ such that

$$[\boldsymbol{f}(\boldsymbol{x}_n)]_k := \mathsf{Pr}(y_n = k|\boldsymbol{x}_n). \quad (2)$$

A commonly used assumption in noisy label learning is that the label noise is independent of individual data samples' features; see [10]–[12], [32]. Under this assumption, we have

$$\mathsf{Pr}(\widehat{y}_n = k|y_n = k', \boldsymbol{x}_n) = \mathsf{Pr}(\widehat{y}_n = k|y_n = k'). \quad (3)$$

The right-hand side can be represented using a $K \times K$-sized *label transition matrix* $\boldsymbol{T}$ (which is also often called the confusion matrix), i.e.,

$$[\boldsymbol{T}]_{k,k'} = \mathsf{Pr}(\widehat{y}_n = k|y_n = k').$$

Note that (3) is an important simplification for reducing the complexity of the model—otherwise the amount of model parameters is too large to learn. We can also define a probability vector $\boldsymbol{p}_n$ such that $[\boldsymbol{p}_n]_k = \mathsf{Pr}(\widehat{y}_n = k|\boldsymbol{x}_n)$. Using these notations, we obtain the following model from (1):

$$\boldsymbol{p}_n = \boldsymbol{T}\boldsymbol{f}(\boldsymbol{x}_n), \forall n. \quad (4)$$

The noisy labels $\widehat{y}_n$'s are categorical realizations of the probability distribution represented by $\boldsymbol{p}_n$'s, i.e., $\widehat{y}_n \sim \mathsf{categorical}(\boldsymbol{p}_n)$. In practice, $\boldsymbol{f}(\cdot)$ is represented by a certain machine learning model, oftentimes DNNs in recent years. Hence, learning $\boldsymbol{f}(\cdot)$ from $\{(\boldsymbol{x}_n, \widehat{y}_n)\}_{n=1}^N$ is a DNN training problem using noisy labels.

B. Connections to NMF & MVES

The model in (4) can be perceived as an NMF model by stacking $\boldsymbol{p}_n$'s as the columns of a matrix:

$$\begin{aligned}\begin{bmatrix}\boldsymbol{p}_1 & \cdots & \boldsymbol{p}_N\end{bmatrix} &= \boldsymbol{T}\begin{bmatrix}\boldsymbol{f}(\boldsymbol{x}_1) & \cdots & \boldsymbol{f}(\boldsymbol{x}_N)\end{bmatrix}\\ &\iff \boldsymbol{P} = \boldsymbol{T}\boldsymbol{F}. \quad (5)\end{aligned}$$

Note that each column of $\boldsymbol{F}$ resides in the probability simplex, i.e.,

$$\boldsymbol{1}^\top \boldsymbol{f}(\boldsymbol{x}_n) = 1,\ \boldsymbol{f}(\boldsymbol{x}_n) \geq \boldsymbol{0},$$

which is due to its physical meaning [cf. Eq. (2)]. Geometrically, Eq. (5) implies that the $\boldsymbol{p}_n$'s are enclosed by the simplex spanned by the columns of $\boldsymbol{T}$, i.e., $\boldsymbol{p}_n \in \mathrm{conv}\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_K\}$. Hence, identifying the latent factors $\boldsymbol{T}$ and $\boldsymbol{F}$ amounts to identifying this data-enclosing simplex.

It is well-known that such a simplex-identification problem is ill-posed [18], [19], [22], [23], [26] and some additional assumptions are needed to underpin $\boldsymbol{T}$ and $\boldsymbol{F}$. The so-called anchor point assumption is widely adopted for this purpose [10], [20]. Under this assumption, there exists an anchor point $\boldsymbol{x}_{n_k}$ for each class k such that $\mathsf{Pr}(y_{n_k} = k|\boldsymbol{x}_{n_k}) = 1$, i.e., $\boldsymbol{f}(\boldsymbol{x}_{n_k}) = \boldsymbol{e}_k, \forall k$. This is the same as the *separability* condition in NMF [18], [19], [21]. Under this condition, the columns of $\boldsymbol{T}$ can be uniquely identified from $\boldsymbol{P}$. However, the anchor points may not always be available. To deal with this challenge, the work in [12] considered the concept of MVES in NMF [22], [23], [25], [27]. It was shown in the literature [23], [33] that finding the minimum-volume data-enclosing simplex identifies the latent factors of the corresponding NMF model, if the data points are geometrically sufficiently spread—see Fig. 1. MVES stemmed from the hyperspectral unmixing literature in the 1990s [34], and has been widely used to deal with NMF problems when the separability condition does not hold.

The work [12] used the idea of MVES and recast the DNN learning problem as follows:

$$\begin{aligned}\underset{\boldsymbol{T}, \boldsymbol{f}\in\mathcal{F}}{\text{minimize}}\ & \mathsf{vol}(\boldsymbol{T}) && (6a)\\ \text{subject to}\ & \boldsymbol{p}_n = \boldsymbol{T}\boldsymbol{f}(\boldsymbol{x}_n), \forall n, && (6b)\\ & \boldsymbol{1}^\top\boldsymbol{T} = \boldsymbol{1},\ \boldsymbol{T} \geq \boldsymbol{0}, && (6c)\end{aligned}$$

where $\mathsf{vol}(\boldsymbol{T})$ denotes the volume of $\mathrm{conv}\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_K\}$, (6c) is introduced to respect the physical meaning of $\boldsymbol{t}_k$ (conditional probabilities), $\mathcal{F}$ denotes the function class to learn $\boldsymbol{f}$ from, e.g., DNNs. The work [12] also argued for the identifiability of the ground-truth confusion matrix $\boldsymbol{T}$ via solving (6) using the MVES proof from [23]. Note that the probability vectors $\boldsymbol{p}_n$'s are not observed in practice. Instead, we observe their realizations $\widehat{y}_n$'s. Hence, the work [12] employed the cross entropy (CE) loss in order to handle the constraint (6b). Also, a commonly used volume measure $\log|\det(\boldsymbol{T})|$ is chosen for $\mathsf{vol}(\boldsymbol{T})$. Specifically, the following criterion is employed in [12]:

$$\begin{aligned}\underset{\boldsymbol{T}, \boldsymbol{f}\in\mathcal{F}}{\text{minimize}}\ & -\frac{1}{N}\sum_{n=1}^N\sum_{k=1}^K [\widehat{\boldsymbol{y}}_n]_k \log[\boldsymbol{T}\boldsymbol{f}(\boldsymbol{x}_n)]_k + \lambda\log|\det(\boldsymbol{T})| && (7a)\\ \text{subject to}\ & \boldsymbol{1}^\top\boldsymbol{T} = \boldsymbol{1},\ \boldsymbol{T} \geq \boldsymbol{0}, && (7b)\end{aligned}$$

where $\widehat{\boldsymbol{y}}_n$ denotes the one-hot representation of the noisy label $\widehat{y}_n$ and and $\lambda > 0$.

C. Challenges

A notable challenge of MVES is that such a geometric NMF criterion is sensitive to outliers [24], [25], [28]. Even if there exists a single outlying data point, the MVES criterion may produce largely undesired solutions—as the minimum-volume enclosing simplex can be quite different from the ground-truth one; see an illustration in Fig. 1. As a result, if there are

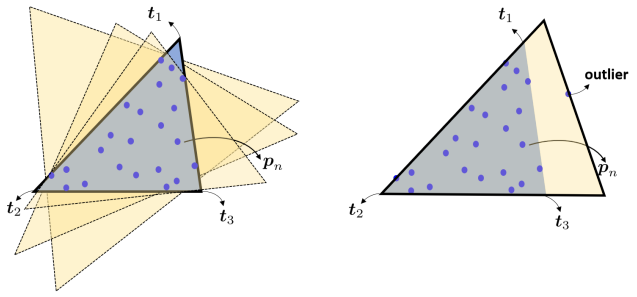

Fig. 1. Illustrating the effect of outlier data points in volume minimization. The dots denote the vectors $\boldsymbol{p}_n$. Blue shaded region denote the ground-truth $\mathsf{conv}(\boldsymbol{t}_1,\ldots,\boldsymbol{t}_K)$ Left) The minimum volume simplex (outlined by the black bold line) is same as $\mathsf{conv}(\boldsymbol{t}_1,\ldots,\boldsymbol{t}_K)$. Right) An outlier impacts volume minimization. Here the minimum volume simplex is different from $\mathsf{conv}(\boldsymbol{t}_1,\ldots,\boldsymbol{t}_K)$.

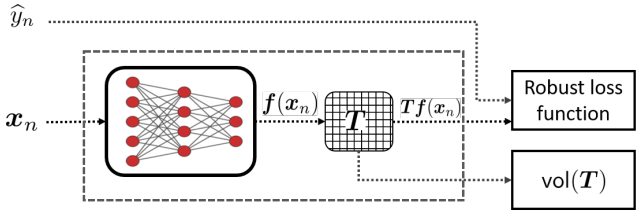

Fig. 2. The proposed `RobVolMinNet` framework.

data points that do not closely obey the model in (4), the criterion in [12] may fail to identify the ground-truth model parameters, $\boldsymbol{T}$ and $\boldsymbol{f}$. Note that for real-world data, the model in (4) can indeed be easily violated, e.g., when some data items suffer from feature-dependent label noise (i.e., when the simplification in (3) is too far from reality). The hyperspectral imaging community has long noticed the challenges arising from outlying data and advocated using robust fitting criteria to enforce $\boldsymbol{P} \approx \boldsymbol{T}\boldsymbol{F}$ [22]–[27]. This suggests that the CE criterion used in (6) may also need be replaced by some outlier-robust surrogates. Indeed, it was widely reported in the machine learning literature that CE is sensitive to outliers and can create overfitting issues [7].

III. Proposed Robustification

To enforce $\boldsymbol{P} \approx \boldsymbol{T}\boldsymbol{F}$, the hyperspectral imaging works often use outlier-robust norms or quasi-norms (e.g., the ℓ_1 norm or the ℓ_q quasi-norm where $q \in (0,1)$), see e.g., [24]. However, such surrogates are not suitable for classification, as the data $\widehat{y}_n$ in classification is integer—instead of a continuous-valued pixel as in hyperspectral imaging. Indeed, the ℓ_1 norm-based *mean absolute error* (MAE) loss, i.e.,

$$\sum_{n=1}^{N} \|\widehat{\boldsymbol{y}}_n - \boldsymbol{T}\boldsymbol{f}(\boldsymbol{x}_n)\|_1,$$

is considered computationally "unfriendly" when combined with DNNs under integer $\widehat{y}_n$, as its non-smoothness may cause convergence issues [8]. The CE loss, despite its non-robustness to outliers, encounters much less convergence challenges in the context of classification.

To strike a balance between the difficulty of optimization and the robustness of the learning criterion, we propose to employ the symmetric cross entropy (SCE) loss function proposed in [9]. For a probability vector $\boldsymbol{p} \in \mathbb{R}^K$ and a one hot label $\boldsymbol{y} \in \{0,1\}^K$, the SCE loss is defined as follows:

$$\ell_{\mathsf{sce}}(\boldsymbol{p},\boldsymbol{y}) = -\alpha \sum_{k=1}^{K} [\boldsymbol{y}]_k \log [\boldsymbol{p}]_k - \beta \sum_{k=1}^{K} [\boldsymbol{p}]_k \log [\boldsymbol{y}_\epsilon]_k, \quad (8)$$

where $\alpha, \beta > 0$ are certain scalars and the notation $\boldsymbol{y}_\epsilon$ is introduced to handle the $\log 0$ cases of the one-hot label $\boldsymbol{y}$, in which we have

$$\log[\boldsymbol{y}_\epsilon]_k = \begin{cases} \epsilon, & [\boldsymbol{y}]_k = 0 \\ \log[\boldsymbol{y}]_k, & \text{otherwise.} \end{cases}$$

The first term on the R.H.S of (8) is same as the CE loss. The second term—the reverse cross entropy–is reduced to exactly the MAE loss, when $\epsilon = -2$. To see this, consider the following:

$$\begin{aligned} \ell_{\mathsf{mae}}(\boldsymbol{p},\boldsymbol{y}) &\triangleq \sum_{k=1}^{K} |[\boldsymbol{p}]_k - [\boldsymbol{y}]_k| = (1 - [\boldsymbol{p}]_y) + \sum_{k\neq y} [\boldsymbol{p}]_k \\ &= 2(1 - [\boldsymbol{p}]_y). \\ \ell_{\mathsf{rce}}(\boldsymbol{p},\boldsymbol{y}) &\triangleq -\sum_{k=1}^{K} [\boldsymbol{p}]_k \log [\boldsymbol{y}_\epsilon]_k = -[\boldsymbol{p}]_k \log 1 - \sum_{k\neq y} [\boldsymbol{p}]_k \epsilon \\ &= -\epsilon \sum_{k\neq y} [\boldsymbol{p}]_k = -\epsilon(1 - [\boldsymbol{p}]_y). \end{aligned}$$

Hence, with a proper choice of the hyperparameters, the SCE loss is expected to enjoy a balance between the nice convergence behavior of CE and the outlier robustness of MAE.

Using the SCE loss, we propose the following robustified criterion:

$$\underset{\boldsymbol{T},\boldsymbol{f}\in\mathcal{F}}{\text{minimize}} \ \frac{1}{N}\sum_{n=1}^{N} \ell_{\mathsf{sce}}(\boldsymbol{T}\boldsymbol{f}(\boldsymbol{x}_n), \widehat{\boldsymbol{y}}_n) + \lambda \log|\det(\boldsymbol{T})|, \quad (9a)$$

$$\text{subject to } \mathbf{1}^\top \boldsymbol{T} = \mathbf{1}, \ \boldsymbol{T} \geq \mathbf{0}. \quad (9b)$$

where $\widehat{\boldsymbol{y}}_n$ denotes the one-hot vector for $\widehat{y}_n$ and $\log|\det(\boldsymbol{T})|$ is the MVES regularization with parameter $\lambda > 0$.

To implement the criterion, one can choose an appropriate DNN function class $\mathcal{F}$, e.g., ResNet. Any off-the shelf optimizers (e.g., Adam, SGD) can be employed to update the parameters of the proposed network. To prevent $\boldsymbol{T}$ from being rank deficient, we adopt the diagonally dominant parameterization trick from [12]. Specifically, the matrix $\boldsymbol{T}$ is assigned as $\boldsymbol{T}(k,k) = 1, \forall k$ and $\boldsymbol{T}(k,j) = \mathsf{sigmoid}(w_{kj}), \forall k \neq j$. Then, the parameter w_{kj} is updated in every iteration during the training phase. In this way, the nonnegativity constraints on $\boldsymbol{T}$ will also be satisfied, since the sigmoid function output ranges between 0 and 1. To account for the sum-to-one constraints on the columns of $\boldsymbol{T}$, column-wise normalization is performed in each iteration.

TABLE I
AVERAGE TEST ACCURACY OF THE PROPOSED METHODS AND THE BASELINES ON THE CIFAR-100 DATASET ($K = 100$)

Methods	Noise Rate	$\eta = 0.1$	$\eta = 0.2$
RobVolMinNet	$\tau = 0.3$	**63.84 ± 1.47**	**64.34 ± 0.59**
RobVolMinNet($\lambda = 0$)	$\tau = 0.3$	60.77 ± 0.47	60.92 ± 0.50
VolMinNet	$\tau = 0.3$	**62.31 ± 0.47**	**62.56 ± 0.98**
GCE	$\tau = 0.3$	59.20 ± 0.78	59.36 ± 0.64
CE	$\tau = 0.3$	56.12 ± 0.44	56.18 ± 0.42
RobVolMinNet	$\tau = 0.5$	**56.41 ± 0.91**	**56.61 ± 0.53**
RobVolMinNet($\lambda = 0$)	$\tau = 0.5$	51.09 ± 0.94	51.81 ± 0.71
VolMinNet	$\tau = 0.5$	**53.97 ± 0.55**	**53.57 ± 0.75**
GCE	$\tau = 0.5$	51.78 ± 0.32	51.83 ± 0.40
CE	$\tau = 0.5$	44.52 ± 0.68	44.88 ± 0.84

TABLE II
AVERAGE TEST ACCURACY OF THE PROPOSED METHODS AND THE BASELINES ON THE CIFAR-10 DATASET ($K = 10$)

Methods	Noise Rate	$\eta = 0.1$	$\eta = 0.2$
RobVolMinNet	$\tau = 0.3$	**87.85 ± 0.14**	**86.74 ± 0.29**
RobVolMinNet($\lambda = 0$)	$\tau = 0.3$	**87.93 ± 0.08**	86.16 ± 0.17
VolMinNet	$\tau = 0.3$	87.70 ± 0.13	86.01 ± 0.34
GCE	$\tau = 0.3$	87.53 ± 0.16	**86.63 ± 0.15**
CE	$\tau = 0.3$	85.22 ± 0.15	83.04 ± 0.20
RobVolMinNet	$\tau = 0.5$	**82.80 ± 0.21**	**81.49 ± 0.41**
RobVolMinNet($\lambda = 0$)	$\tau = 0.5$	80.45 ± 0.51	**79.55 ± 0.27**
VolMinNet	$\tau = 0.5$	**80.75 ± 0.15**	79.32 ± 0.31
GCE	$\tau = 0.5$	79.78 ± 0.54	77.37 ± 0.63
CE	$\tau = 0.5$	76.93 ± 0.46	75.46 ± 0.67

We name our framework as Robust Volume-Minimization-based Deep Neural Network (RobVolMinNet). The illustration of the learning system is shown in Fig. 2.

IV. EXPERIMENTS

In this section, we present experiment results to showcase the effectiveness of the proposed approach.

Baselines. We consider the following baselines: VolMinNet [12], which employs the CE loss function along with MVES regularization (cf. (7)); GCE [8], which utilizes another variant of CE named as generalized cross entropy; and the plain vanilla cross entropy loss-based DNN training, denoted as CE. We also compare the proposed method with the regularization parameter λ set to zero,denoted as RobVolMinNet ($\lambda = 0$).

Datasets. We use the CIFAR-100 dataset [30] and the CIFAR-10 dataset [31]. Both CIFAR-100 and CIFAR-10 comprise a collection of $60,000$ labeled color images of 32×32 pixels in size. The CIFAR-100 contains 100 different classes and the CIFAR-10 has 10 different classes. For both datasets, we utilize $45,000$ images for training, $5,000$ images for validation, and $10,000$ images for testing.

Noisy Label Generation To generate noisy labels for the data items, we employ the following strategy. We control the overall noise rate of the labels using a parameter $\tau \in (0, 1)$. Using this parameter, we generate the ground-truth confusion matrix $\boldsymbol{T}$ such that the diagonal entries of $\boldsymbol{T}$ are chosen as $[\boldsymbol{T}]_{k,k} = 1 - \tau, \forall k$ and the off-diagonal entries are chosen as $[\boldsymbol{T}]_{k,j} = \frac{\tau}{K-1}, \forall k \neq j$. We select $1 - \eta$ fraction of the data samples uniformly at random and generate their labels using the feature-independent confusion matrix $\boldsymbol{T}$.

The remaining η percent of the samples are considered as outliers. For such outlying data, we adopt a feature-dependent noise generation process. Specifically, we follow the strategy in [35] to generate outliers so that the probabilities associated with observing the noisy label $\widehat{y}_n$ depend on multiple factors including the ground-truth label y_n, the noise rate parameter τ, and also the feature vector $\boldsymbol{x}_n$. Note that, when η is small, this label noise generation process is different from that of the majority of the samples which follows a feature-independent noise generation model using $\boldsymbol{T}$.

Network Structure and Parameters For CIFAR-100 and CIFAR-10, we employ the ResNet-32 and the ResNet-18 network architectures [36], respectively. The SGD optimizer is employed to train the parameters of the network with a batch size of 128, momentum of 0.9, weight decay of 10^{-3}, and an initial learning rate of 10^{-2}. We run the algorithms for a maximum of 80 epochs. We set hyperparameters as $\alpha = 0.9$, $\beta = 0.4$, $\epsilon = 0.0001$, and $\lambda = 0.0001$. For all the methods, we use validation set to select the best model parameters for reporting the classification accuracy on testing set.

Results. We present the average classification accuracy on the testing sets for CIFAR-100 and CIFAR-10 in Table I and II, respectively. The first and second best performances are highlighted in bold letters. All experiments are repeated five times, and the standard deviation is also provided in the tables. Our proposed criterion demonstrates performance advantages over all the baselines considered for different percent of outlying samples (i.e., for different values of η). Notably, our method, RobVolMinNet, consistently outperforms VolMinNet, highlighting the effect of outlier-robustness in our approach. One can also note that both MVES-based approaches, RobVolMinNet and VolMinNet, outperform RobVolMinNet($\lambda = 0$) in most of the cases, underscoring the importance of incorporating the MVES regularization in the learning objective.

V. CONCLUSION

In this work, we revisited the noisy label learning framework that uses the MVES matrix factorization-based regularization. MVES was employed in this framework to guarantee the identifiability of the noise transition matrix. However, from a matrix factor identfiication viewpoint, MVES is known to be sensitive to outliers. We observed that such sensitivity could limit its effectiveness in noisy label deep learning. We proposed an outlier-resilient noisy label learning criterion by combining a robust neural network training loss function with the MVES regularization. Our robustification makes use of a symmetric cross-entropy function that is suited for integer data, which is different from classic robust MVES formulations that were designed for continuous (e.g., image) data. We tested our new criterion using synthetically generated label noise and observed nontrivial improvement relative to the non-robust version in classification accuracy on the CIFAR-100 and CIFAR-10 datasets.

REFERENCES

[1] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Applied statistics*, pp. 20–28, 1979.

[2] D. Arpit, S. Jastrzundefinedbski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," in *Proceedings of International Conference on Machine Learning*, 2017, p. 233–242.

[3] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proceedings of International Conference on Learning Representations*, 2016.

[4] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?" in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, 2019, pp. 7164–7173.

[5] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. W. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 8536–8546.

[6] Q. Yao, H. Yang, B. Han, G. Niu, and J. T.-Y. Kwok, "Searching to exploit memorization effect in learning with noisy labels," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., vol. 119, 2020, pp. 10 789–10 798.

[7] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, p. 1919–1925.

[8] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, p. 8792–8802.

[9] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, nov 2019, pp. 322–330.

[10] G. Patrini, A. Rozza, A. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: a loss correction approach," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. United States of America: IEEE, Institute of Electrical and Electronics Engineers, 2017, pp. 2233–2241.

[11] X. Xia, T. Liu, N. Wang, B. Han, C. Gong, G. Niu, and M. Sugiyama, "Are anchor points really indispensable in label-noise learning?" in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[12] X. Li, T. Liu, B. Han, G. Niu, and M. Sugiyama, "Provably end-to-end label-noise learning without anchor points," in *Proceedings of International Conference on Machine Learning*, 2021, pp. 6403–6413.

[13] Z. Zhu, Y. Song, and Y. Liu, "Clusterability as an alternative to anchor points when learning with noisy labels," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., vol. 139, 2021, pp. 12 912–12 923.

[14] S. Ibrahim, X. Fu, N. Kargas, and K. Huang, "Crowdsourcing via pairwise co-occurrences: Identifiability and algorithms," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 7847–7857.

[15] S. Ibrahim and X. Fu, "Crowdsourcing via annotator co-occurrence imputation and provable symmetric nonnegative matrix factorization," in *Proceedings of International Conference on Machine Learning*, vol. 139, 2021, pp. 4544–4554.

[16] S. Ibrahim, T. Nguyen, and X. Fu, "Deep learning from crowdsourced labels: Coupled cross-entropy minimization, identifiability, and regularization," in *The Eleventh International Conference on Learning Representations*, 2022.

[17] T. Nguyen, S. Ibrahim, and X. Fu, "Deep clustering with incomplete noisy pairwise annotations: A geometric regularization approach," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202. PMLR, 23–29 Jul 2023, pp. 25 980–26 007.

[18] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications." *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, 2019.

[19] N. Gillis, *Nonnegative Matrix Factorization*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2020.

[20] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 447–461, 2016.

[21] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in neural information processing systems*, vol. 16, 2003.

[22] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4418 –4432, Nov. 2009.

[23] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2306–2320, 2015.

[24] X. Fu, K. Huang, B. Yang, W.-K. Ma, and N. D. Sidiropoulos, "Robust volume minimization-based matrix factorization for remote sensing and document clustering," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6254–6268, 2016.

[25] J. M. Bioucas-Dias, "A variable splitting augmented lagrangian approach to linear spectral unmixing," in *Proc. IEEE WHISPERS'09*, 2009, pp. 1–4.

[26] X. Fu, K. Huang, and N. D. Sidiropoulos, "On identifiability of nonnegative matrix factorization," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 328–332, 2018.

[27] J. Li and J. M. Bioucas-Dias, "Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data," in *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, vol. 3, 2008, pp. III – 250–III – 253.

[28] W.-K. Ma, J. Bioucas-Dias, T.-H. Chan, N. Gillis, P. Gader, A. Plaza, A. Ambikapathi, and C.-Y. Chi, "A signal processing perspective on hyperspectral unmixing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 67–81, Jan 2014.

[29] A. Ambikapathi, T.-H. Chan, W.-K. Ma, and C.-Y. Chi, "Chance-constrained robust minimum-volume enclosing simplex algorithm for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4194–4209, 2011.

[30] A. Krizhevsky, "Learning multiple layers of features from tiny images," in *Technical Report*, 2009.

[31] ——, "Learning multiple layers of features from tiny images," *Technical Report, University of Toronto*, 2009.

[32] B. Han, J. Yao, G. Niu, M. Zhou, I. Tsang, Y. Zhang, and M. Sugiyama, "Masking: A new perspective of noisy supervision," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[33] C.-H. Lin, W.-K. Ma, W.-C. Li, C.-Y. Chi, and A. Ambikapathi, "Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 10, pp. 5530–5546, Oct 2015.

[34] M. D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 3, pp. 542–552, 1994.

[35] X. Xia, T. Liu, B. Han, N. Wang, M. Gong, H. Liu, G. Niu, D. Tao, and M. Sugiyama, "Part-dependent label noise: Towards instance-dependent label noise," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 7597–7610.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.