
Provably End-to-end Label-noise Learning without Anchor Points

Xuefeng Li^{1,2} Tongliang Liu² Bo Han³ Gang Niu⁴ Masashi Sugiyama^{4,5}

Abstract

In label-noise learning, the *transition matrix* plays a key role in building *statistically consistent* classifiers. Existing consistent estimators for the transition matrix have been developed by exploiting *anchor points*. However, the anchor-point assumption is not always satisfied in real scenarios. In this paper, we propose an end-to-end framework for solving label-noise learning without anchor points, in which we simultaneously optimize two objectives: the cross entropy loss between the noisy label and the predicted probability by the neural network, and the volume of the simplex formed by the columns of the transition matrix. Our proposed framework can identify the transition matrix if the clean class-posterior probabilities are *sufficiently scattered*. This is by far the mildest assumption under which the transition matrix is provably *identifiable* and the learned classifier is statistically consistent. Experimental results on benchmark datasets demonstrate the effectiveness and robustness of the proposed method.

1. Introduction

The success of modern deep learning algorithms heavily relies on large-scale accurately annotated data (Daniely & Granot, 2019; Han et al., 2020b; Xia et al., 2020; Berthon et al., 2021). However, it is often expensive or even infeasible to annotate large datasets. Therefore, cheap but less accurate annotating methods have been widely used (Xiao et al., 2015; Li et al., 2017; Han et al., 2020a; Yu et al., 2020; Zhu et al., 2021a). As a consequence, these alternatives inevitably introduce label noise. Training deep learning models on noisy data can significantly degenerate the test performance due to overfitting to the noisy labels

(Arpit et al., 2017; Zhang et al., 2017; Xia et al., 2021; Wu et al., 2021).

To mitigate the negative impacts of label noise, many methods have been developed and some of them are based on a loss correction procedure. In general, these methods are *statistically consistent*, i.e., these methods guarantee that the classifier learned from the noisy data approaches to the optimal classifier defined on the clean risk as the size of the noisy training set increases (Liu & Tao, 2016; Scott, 2015; Natarajan et al., 2013; Goldberger & Ben-Reuven, 2017; Patrini et al., 2017; Thekumparampil et al., 2018). The idea is that the clean class-posterior $P(\mathbf{Y}|\mathbf{X} = \mathbf{x}) := [P(Y = 1|\mathbf{X} = \mathbf{x}), \dots, P(Y = C|\mathbf{X} = \mathbf{x})]^\top$ can be inferred by utilizing the noisy class-posterior $P(\tilde{\mathbf{Y}}|\mathbf{X} = \mathbf{x})$ and the transition matrix $\mathbf{T}(\mathbf{x})$ where $T_{ij}(\mathbf{x}) = P(\tilde{Y} = j|Y = i, \mathbf{X} = \mathbf{x})$, i.e., $P(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = [\mathbf{T}(\mathbf{x})]^{-1}P(\tilde{\mathbf{Y}}|\mathbf{X} = \mathbf{x})$. While those methods theoretically guarantee the statistical consistency, they all heavily rely on the success of estimating transition matrices.

Generally, the transition matrix is unidentifiable without additional assumptions (Xia et al., 2019). In the literature, methods have been developed to estimate the transition matrices under the so-called *anchor-point* assumption: it assumes the existence of anchor points, i.e., instances belonging to a specific class with probability one (Liu & Tao, 2016). The assumption is reasonable in certain applications (Liu & Tao, 2016; Patrini et al., 2017). However, the violation of the assumption in some cases could lead to a poorly learned transition matrix and a degenerated classifier (Xia et al., 2019). This motivates the development of algorithms without exploiting anchor points (Xia et al., 2019; Liu & Guo, 2020; Xu et al., 2019; Zhu et al., 2021b). However, the performance is not theoretically guaranteed in these works.

Motivation. In this work, our interest lies in designing a consistent algorithm without anchor points, subject to class-dependent label noise, i.e., $\mathbf{T}(\mathbf{x}) = \mathbf{T}$ for any \mathbf{x} in the feature space. Our algorithm is based on a geometric property of the label corruption process. Given an instance \mathbf{x} , the noisy class-posterior probability $P(\tilde{\mathbf{Y}}|\mathbf{X} = \mathbf{x}) := [P(Y = 1|\mathbf{X} = \mathbf{x}), \dots, P(Y = C|\mathbf{X} = \mathbf{x})]^\top$ can be thought of as a point in the C -dimensional space where C is the number of classes. Since we have $P(\tilde{\mathbf{Y}}|\mathbf{X} =$

¹University of New South Wales ²Trustworthy Machine Learning Lab, University of Sydney ³Hong Kong Baptist University ⁴RIKEN AIP ⁵University of Tokyo. Correspondence to: Tongliang Liu <tongliang.liu@sydney.edu.au>.

$\mathbf{x}) = \mathbf{T}P(\mathbf{Y}|\mathbf{X} = \mathbf{x})$ and $\sum_{i=1}^C P(Y = i|\mathbf{X} = \mathbf{x}) = 1$, $P(\tilde{\mathbf{Y}}|\mathbf{X} = \mathbf{x})$ is then a convex combination of the columns of \mathbf{T} . This means that the simplex $\text{Sim}\{\mathbf{T}\}$ formed by the columns of \mathbf{T} encloses $P(\tilde{\mathbf{Y}}|\mathbf{X} = \mathbf{x})$ for any \mathbf{x} (Boyd et al., 2004). Thus, the problem of identifying the transition matrix can be treated as the problem of recovering $\text{Sim}\{\mathbf{T}\}$. However, when no assumption has been made, the problem is ill-defined as $\text{Sim}\{\mathbf{T}\}$ is not identifiable, i.e., there exists an infinite number of simplexes enclosing $P(\tilde{\mathbf{Y}}|\mathbf{X})$, and any of them can be regarded as the true simplex $\text{Sim}\{\mathbf{T}\}$. It is apparent that under the anchor-point assumption, $\text{Sim}\{\mathbf{T}\}$ can be uniquely determined by exploiting anchor points whose noisy class-posterior probabilities are the vertices of $\text{Sim}\{\mathbf{T}\}$. The goal is thus to identify the points which have the largest noisy class-posterior probabilities for each class (Liu & Tao, 2016; Patrini et al., 2017). However, if there are no anchor points, the identified points would not be the vertices of $\text{Sim}\{\mathbf{T}\}$. In this case, existing methods cannot consistently estimate the transition matrices. To recover \mathbf{T} without anchor points, a key observation is that, among all simplexes enclosing $P(\tilde{\mathbf{Y}}|\mathbf{X})$, $\text{Sim}\{\mathbf{T}\}$ is the one with *minimum volume*. See Figure 1 for a geometric illustration. This observation motivates the development of our method which incorporates the minimum volume constraint of $\text{Sim}\{\mathbf{T}\}$ into label-noise learning.

To this end, we propose *Volume Minimization Network* (VolMinNet) to consistently estimate the transition matrix and build a statistically consistent classifier. Specifically, VolMinNet consists of a classification network \mathbf{h}_θ and a trainable transition matrix $\hat{\mathbf{T}}$. We simultaneously optimize $\hat{\mathbf{T}}$ and \mathbf{h}_θ with two objectives: i) the discrepancy between $\hat{\mathbf{T}}\mathbf{h}_\theta(\mathbf{x})$ and the noisy class-posterior distribution $P(\tilde{\mathbf{Y}}|\mathbf{X} = \mathbf{x})$, ii) The volume of the simplex formed by the columns of $\hat{\mathbf{T}}$. The proposed framework is end-to-end, and there is no need for identifying anchor points or *pseudo anchor points* (i.e., instances belonging to a specific class with probability close to one) (Xia et al., 2019). Since our proposed method does not rely on any specific data points, it yields better noise robustness compared with existing methods. With a so-called *sufficiently scattered assumption* where the clean class-posterior distribution is far from uniform, we theoretically prove that $\hat{\mathbf{T}}$ will converge to the true transition matrix \mathbf{T} while $\mathbf{h}_\theta(\mathbf{x})$ converges to the clean class-posterior $P(\mathbf{Y}|\mathbf{X} = \mathbf{x})$. We also prove that the anchor-point assumption is a special case of the sufficiently scattered assumption.

The rest of this paper is organized as follows. In Section 2, we set up the notations and review the background of label-noise learning with anchor points. In Section 3, we introduce our proposed VolMinNet. In Section 4, we present the main theoretical results. In Section 5, we briefly introduce the related works in the literature. Experimental results on both synthetic and real-world datasets are provided in Sec-

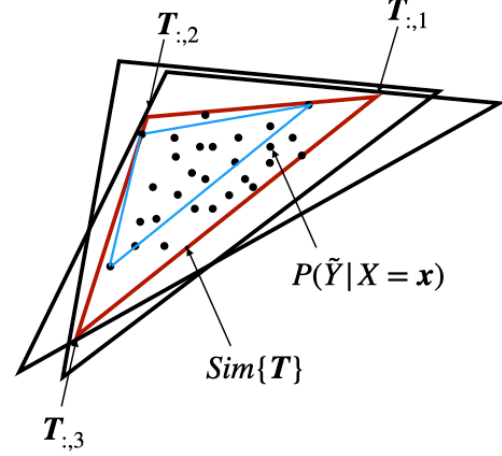


Figure 1. Geometric illustration of the problem of estimating the transition matrix without anchor points. The red triangle is the simplex of \mathbf{T} with vertices denoted by $\mathbf{T}_{:,i}$. When there are no anchor points, the simplex found with existing methods (blue triangle) by using extreme-valued noisy class-posterior probabilities (see Eq. (4)) is not the true simplex (red triangle). It is obvious that among possible enclosing simplexes (black and red triangles), the true simplex $\text{Sim}\{\mathbf{T}\}$ has the minimum volume.

tion 6. Finally, we conclude the paper in Section 7.

2. Label-Noise Learning with Anchor Points

In this section, we review the background of label-noise learning. We follow common notational conventions in the literature of label-noise learning. $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{V} \in \mathbb{R}^{n \times m}$ denote a real-valued n -dimensional vector and a real-valued $n \times m$ matrix, respectively. Elements of a vector are denoted by a subscript (e.g., v_j), while rows and columns of a matrix are denoted by $\mathbf{V}_{i,:}$ and $\mathbf{V}_{:,i}$ respectively. The i th standard basis vector in \mathbb{R}^C is denoted by \mathbf{e}_i . We denote the all-ones vector by $\mathbf{1}$, and $\Delta^{C-1} \subset [0, 1]^C$ is the C -dimensional simplex. In this work, we also make extensive use of convex analysis. Let a set $\mathcal{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_m\}$, and the convex cone of \mathcal{V} is denoted by $\text{cone}(\mathcal{V}) = \{\mathbf{v} | \mathbf{v} = \sum_{j=1}^m \mathbf{v}_j \alpha_j, \alpha_j \geq 0, \forall j\}$. Similarly, the convex hull of \mathcal{V} is defined as $\text{conv}(\mathcal{V}) = \{\mathbf{v} | \mathbf{v} = \sum_{j=1}^m \mathbf{v}_j \alpha_j, \alpha_j \geq 0, \sum_{j=1}^m \alpha_j = 1, \forall j\}$. Specially, when $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ are affinely independent, $\text{conv}(\mathcal{V})$ is also called a simplex which we denote it as $\text{Sim}(\mathcal{V})$.

Let \mathcal{D} be the underlying distribution generating a pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the feature space, $\mathcal{Y} = \{1, 2, \dots, C\}$ is the label space and C is the number of classes. In many real-world applications, samples drawn from \mathcal{D} are unavailable. Before being observed, labels of these samples are contaminated with noise and we obtain a set of corrupted data $\{(\mathbf{x}_i, \tilde{y}_i)\}_{i=1}^n$ where \tilde{y} is the noisy label and we denote by $\tilde{\mathcal{D}}$ the distribution of

the noisy random pair $(X, \tilde{Y}) \in \mathcal{X} \times \mathcal{Y}$.

Given an instance \mathbf{x} sampled from X , \tilde{Y} is derived from the random variable Y through a noise transition matrix $\mathbf{T}(\mathbf{x}) \in [0, 1]^{C \times C}$:

$$P(\tilde{Y}|X = \mathbf{x}) = \mathbf{T}(\mathbf{x})P(Y|X = \mathbf{x}), \quad (1)$$

where $P(\tilde{Y}|X = \mathbf{x}) = [P(Y = 1|X = \mathbf{x}), \dots, P(Y = C|X = \mathbf{x})]^\top$ and $P(Y|X = \mathbf{x}) = [P(Y = 1|X = \mathbf{x}), \dots, P(Y = C|X = \mathbf{x})]^\top$ are the clean class-posterior probability and the noisy class-posterior probability, respectively. The ij -th entry of the transition matrix, i.e., $\mathbf{T}_{ij}(\mathbf{x}) = P(\tilde{Y} = i|Y = j, X = \mathbf{x})$, represents the probability that the instance \mathbf{x} with the clean label $Y = i$ will have a noisy label $\tilde{Y} = j$. Generally, the transition matrix is *non-identifiable* without any additional assumption (Xia et al., 2019). For example, we can decompose the transition matrix with $\mathbf{T}(\mathbf{x}) = \mathbf{T}_1(\mathbf{x})\mathbf{T}_2(\mathbf{x})$. If we define $P'(Y|X = \mathbf{x}) = \mathbf{T}_2(\mathbf{x})P(Y|X = \mathbf{x})$, then $P(\tilde{Y}|X = \mathbf{x}) = \mathbf{T}(\mathbf{x})P(Y|X = \mathbf{x}) = \mathbf{T}_1(\mathbf{x})P'(Y|X = \mathbf{x})$ are both valid. Therefore, in this paper, we study the *class-dependent* and *instance-independent* transition matrix on which the majority of existing methods focus (Han et al., 2018b;a; Patrini et al., 2017; Northcutt et al., 2017; Natarajan et al., 2013). Formally, we have:

$$P(\tilde{Y}|X = \mathbf{x}) = \mathbf{T}P(Y|X = \mathbf{x}), \quad (2)$$

where the transition matrix \mathbf{T} is now independent of the instance \mathbf{x} . In this work, we also assume that the true transition matrix \mathbf{T} is *diagonally dominant*¹. Specifically, the transition matrix \mathbf{T} is diagonally dominant if for every column of \mathbf{T} , the magnitude of the diagonal entry is larger than any non-diagonal entry, i.e., $\mathbf{T}_{ii} > \mathbf{T}_{ij}$ for any $i \neq j$. This assumption has been commonly used in the literature of label-noise learning (Patrini et al., 2017; Xia et al., 2019; Yao et al., 2020b).

As in Eq. (2), the clean class-posterior probability $P(Y|X)$ can be inferred by using the noisy class-posterior probability $P(\tilde{Y}|X)$ and the transition matrix \mathbf{T} as $P(Y|X) = \mathbf{T}^{-\top}P(\tilde{Y}|X)$. For this reason, the transition matrix has been widely exploited to build statistically consistent classifiers, i.e., the learned classifier will converge to the optimal classifier defined with clean risk. Specifically, the transition matrix has been used to modify loss functions to build risk-consistent estimators (Goldberger & Ben-Reuven, 2017; Patrini et al., 2017; Yu et al., 2018; Xia et al., 2019), and has been used to correct hypotheses to build classifier-consistent algorithms (Natarajan et al., 2013; Scott, 2015; Patrini et al., 2017). Thus, the successes of these consistent algorithms rely on an accurately learned transition matrix.

¹The definition of being diagonally dominant is different from the one in matrix analysis, but it has been commonly used in label-noise learning (Xu et al., 2019)

In recent years, considerable efforts have been invested in designing algorithms for estimating the transition matrix. These algorithms rely on a so-called *anchor-point* assumption which requires that there exist anchor points for each class (Liu & Tao, 2016; Xia et al., 2019).

Definition 1 (anchor-point assumption). *For each class $i \in \{1, 2, \dots, C\}$, there exists an instance $\mathbf{x}^i \in \mathcal{X}$ such that $P(Y = i|X = \mathbf{x}^i) = 1$.*

Under the anchor-point assumption, the task of estimating the transition matrix boils down to finding anchor points for each class. For example, given anchor points \mathbf{x}^i , we have

$$P(\tilde{Y} = j | X = \mathbf{x}^i) = \sum_{k=1}^C \mathbf{T}_{kj} P(Y = k | X = \mathbf{x}^i) = \mathbf{T}_{ij}. \quad (3)$$

Namely, the transition matrix can be obtained with the noisy class-posterior probabilities of anchor points. Assuming that we can accurately model the noisy class-posterior $P(\tilde{Y}|X)$ given a sufficient number of noisy data, anchor points can be easily found as follows (Liu & Tao, 2016; Patrini et al., 2017):

$$\mathbf{x}^i = \arg \max_{\mathbf{x}} P(\tilde{Y} = i | X = \mathbf{x}). \quad (4)$$

However, when the anchor-point assumption is not satisfied, points found with Eq. (4) are no longer anchor points. Hence, the above-mentioned method can not consistently estimate the transition matrix with Eq. (3), which will lead to a statistically inconsistent classifier. This motivates us to design a statistically classifier-consistent algorithm which can consistently estimate the transition matrix without anchor points.

3. Volume Minimization Network

In this section, we propose a novel framework for label-noise learning which we call the Volume Minimization Network (VolMinNet). The proposed framework is end-to-end, and there is no need for identifying anchor points or a second stage for loss correction, resulting in better noise robustness than existing methods.

To learn the clean class-posterior $P(Y|X)$, we define a transformation $h_{\theta} : \mathcal{X} \rightarrow \Delta^{C-1}$ where h_{θ} is a differentiable function represented by a neural network with parameters θ . To estimate the transition matrix, we construct a trainable *diagonally dominant column stochastic* matrix $\hat{\mathbf{T}}$, i.e., $\hat{\mathbf{T}} \in [0, 1]^{C \times C}$, $\sum_{j=1}^C \mathbf{T}_{ij} = 1$ and $\mathbf{T}_{ii} > \mathbf{T}_{ij}$ for any $i \neq j$. To learn the noisy class posterior distribution from the noisy data, with some abuse of notation, we define the composition of $\hat{\mathbf{T}}$ and h_{θ} as $\hat{\mathbf{T}}h_{\theta} : \mathcal{X} \rightarrow \Delta^{C-1}$.

Intuitively, as explained in Section 1, if $\hat{\mathbf{T}}h_{\theta}$ models $P(\tilde{Y}|X)$ perfectly while the simplex of $\hat{\mathbf{T}}$ has the minimum volume, $\hat{\mathbf{T}}$ will converge to the true transition matrix

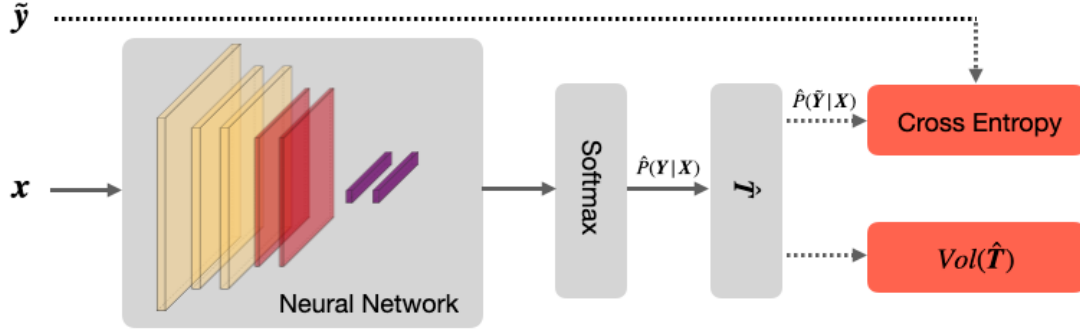


Figure 2. Overview of the proposed VolMinNet. The training in the proposed framework is carried out in an end-to-end manner with two objectives (red blocks) optimized simultaneously.

T and h_θ will converge to $P(Y|X)$. This motivates us to propose the following criterion which corresponds to a constraint optimization problem:

$$\begin{aligned} \min_{\hat{T} \in \mathbb{T}} \text{vol}(\hat{T}) \\ \text{s.t. } \hat{T}h_\theta = P(\tilde{Y}|X), \end{aligned} \quad (5)$$

where $\mathbb{T} = \{\hat{T} \in [0, 1]^{C \times C} \mid \sum_{j=1}^C \hat{T}_{ij} = 1, \hat{T}_{ii} > \hat{T}_{ij}, \forall i \neq j\}$ is the set of diagonally dominant column stochastic matrices. $\text{vol}(\hat{T})$ denotes a measure that is related or proportional to the volume of the simplex formed by the columns of \hat{T} .

To solve criterion (5), we first note that the constraint $\hat{T}h_\theta = P(\tilde{Y}|X)$ can be solved with expected risk minimization (Patrini et al., 2017). The risk is defined as $\hat{R}(h_\theta) = \mathbb{E}_{(x, \tilde{y}) \sim \tilde{D}}[\ell(\hat{T}h_\theta(x), \tilde{y})]$, where ℓ is a loss function and we use the cross-entropy loss throughout this paper. We can then re-write criterion (5) as a Lagrangian under the KKT condition (Karush, 1939; Kuhn & Tucker, 2014) to obtain:

$$\mathcal{L}(\theta, \hat{T}) := \text{vol}(\hat{T}) + \beta \cdot \mathbb{E}_{(x, \tilde{y}) \sim \tilde{D}}[\ell(\hat{T}h_\theta(x), \tilde{y})], \quad (6)$$

where $\beta > 0$ is the KKT multiplier. In the literature, various functions for measuring the volume of the simplex have been investigated (Fu et al., 2015; Li & Bioucas-Dias, 2008; Miao & Qi, 2007). Given \hat{T} is a square matrix, a common choice is $\text{vol}(\hat{T}) = \det(\hat{T})$, where \det denotes the determinant. However, this function is numerically unstable for optimization and computationally hard to deal with. Hence, we adopt another popular alternative $\log \det(\hat{T})$. This function has been widely used in low-rank matrix recovery and non-negative matrix decomposition (Fazel et al., 2003; Liu et al., 2012; Fu et al., 2016). Besides, since we only have access to a set of noisy training examples $\{(x_i, \tilde{y}_i)\}_{i=1}^n$ instead of the distribution \tilde{D} , we employ the empirical risk for training. Formally, we

propose the following objective function:

$$\mathcal{L}(\theta, \hat{T}) := \frac{1}{n} \sum_{i=1}^n \tilde{\ell}(\hat{T}h_\theta(x_i), \tilde{y}_i) + \lambda \cdot \log \det(\hat{T}), \quad (7)$$

where $\lambda > 0$ is a regularization coefficient that balances distribution fidelity versus volume minimization.

The problem remains how to design \hat{T} so that it is differentiable, diagonally dominant and column stochastic. Specifically, we first create a matrix $A \in \mathbb{R}^{C \times C}$ so that diagonal elements $A_{ii} = 1$ for all $i \in \{1, 2, \dots, C\}$, and all other elements $A_{ij} = \sigma(w_{ij})$ for all $i \neq j$ where σ is the sigmoid function $\sigma(w) = \frac{1}{1+e^{-w}}$ and each $w_{ij} \in \mathbb{R}$ is a real-valued variable which will be updated throughout training. Then we do the normalization $\hat{T}_{ij} = \frac{A_{ij}}{\sum_{k=1}^C A_{ik}}$ so that the sum of each column of \hat{T} is equal to one. Since the sigmoid function returns a value in the range 0 to 1, we have $\hat{T}_{ii} > \hat{T}_{ij}$ for all $i, j \in \{1, 2, \dots, C\}$. With this specially designed \hat{T} , we ensure that $\hat{T} \in \mathbb{T}$, i.e., \hat{T} is a diagonally dominant and column stochastic matrix. In addition, \hat{T} is differentiable because the sigmoid function and the normalization operation are differentiable.

With both \hat{T} and h_θ being differentiable, the objective in Eq. (7) can be easily optimized with any standard gradient-based learning rule. This allows us to replace the two-stage loss correction procedure in existing works with an end-to-end learning framework. See Figure 2 for a less formal, more pedagogical explanation of our proposed learning framework.

4. Theoretical Results

In this section, we show that criterion (5) guarantees the consistency of the estimated transition matrix and the learned classifier under the sufficiently scattered assumption. We also show that the anchor-point assumption is a special case of the sufficiently assumption. To explain,

we give a formal definition of the sufficiently scattered assumption:

Definition 2 (Sufficiently Scattered). *The clean class-posterior $P(\mathbf{Y}|X)$ is said to be sufficiently scattered if there exists a set $\mathcal{H} = \{x_1, \dots, x_m\} \subseteq \mathcal{X}$ such that the matrix $\mathbf{H} = [P(\mathbf{Y}|X = x_1), \dots, P(\mathbf{Y}|X = x_m)]$ satisfies (1) $\mathcal{R} \subseteq \text{cone}\{\mathbf{H}\}$, where $\mathcal{R} = \{\mathbf{v} \in \mathbb{R}^C \mid \mathbf{v}^\top \mathbf{1} \geq \sqrt{C-1}\|\mathbf{v}\|_2\}$ and $\text{cone}\{\mathbf{H}\}$ denotes the convex cone formed by columns of \mathbf{H} . (2) $\text{cone}\{\mathbf{H}\} \not\subseteq \text{cone}\{\mathbf{Q}\}$, for any unitary matrix $\mathbf{Q} \in \mathbb{R}^{C \times C}$ that is not a permutation matrix.*

This assumption is evolved from previous works in non-negative matrix decomposition (Fu et al., 2015; 2018) with necessary modification. Intuitively, in the case of $C = 3$, \mathcal{R} corresponds to a “ball” tangential to the triangle formed by a permutation matrix, e.g., $\mathbf{I} = [e_1, e_2, e_3]$. $\text{cone}\{\mathbf{H}\}$ is the polytope inside this triangle. Columns of \mathbf{Q} also form triangles which are rotated versions of the triangle defined by permutation matrices; facets of those triangles are also tangential to \mathcal{R} . Condition (1) of the sufficiently scattered assumption requires that \mathcal{R} is enclosed by $\text{cone}\{\mathbf{H}\}$, i.e., \mathcal{R} is a subset of $\text{cone}\{\mathbf{H}\}$. Condition (2) ensures that given condition (1), $\text{cone}\{\mathbf{H}\}$ is enclosed by the triangle formed by a permutation matrix and not any other unitary matrix.

To understand the sufficiently scattered assumption and its relationship with the anchor-point assumption, we provide several examples in Figure 3. In Figure 3.(a), we show a situation where both the anchor-point assumption and sufficiently scattered assumption are satisfied. Figure 3.(b) shows a situation where the sufficiently scattered assumption is satisfied while the anchor-point assumption is violated. In Figure 3.(c) and 3.(d), both assumptions are violated. However, in Figure 3.(d), only condition (2) of the sufficiently scattered assumption is violated while both conditions of the sufficiently scattered assumption are violated in 3.(c).

The first observation is that if the anchor-point assumption is satisfied, then the sufficiently scattered assumption must hold, but not vice versa. Intuitively, if the anchor-point assumption is satisfied, then there exists a matrix $\mathbf{H} = [P(\mathbf{Y}|X = x^1), \dots, P(\mathbf{Y}|X = x^C)] = \mathbf{I}$ where x^1, \dots, x^C are anchor points for different classes and \mathbf{I} is the identity matrix. From Figure 3.(a), it is clear that $\mathcal{R} \subseteq \text{cone}\{\mathbf{I}\} = \text{cone}\{\mathbf{H}\}$, and $\text{cone}\{\mathbf{H}\}$ can only be enclosed by the convex cone of permutation matrices. This shows that the sufficiently scattered assumption is satisfied. However, from Figure 3.(b), it is clear that the sufficiently scattered assumption is satisfied but not the anchor-point assumption. Formally, we show that:

Proposition 1. *The anchor-point assumption is a sufficient but not necessary condition for the sufficiently scattered assumption when $C > 2$.*

The proof of Proposition 1 is included in the supplementary material. Proposition 1 implies that the anchor-point assumption is a special case of the sufficiently scattered assumption. This means that the proposed framework can also deal with the case where the anchor-point assumption holds. Under the sufficiently scattered assumption, we get our main result:

Theorem 1. *Given sufficiently many noisy data, if $P(\mathbf{Y}|X)$ is sufficiently scattered, then $\hat{\mathbf{T}}_\star = \mathbf{T}$ and $h_{\theta_\star}(x) = P(\mathbf{Y}|X = x)$ must hold, where $(\hat{\mathbf{T}}_\star, \theta_\star)$ are optimal solutions of Eq. (5).*

The proof of Theorem 1 can be found in the supplementary material. Intuitively, if $P(\mathbf{Y}|X)$ is sufficiently scattered, the noisy class-posterior $P(\hat{\mathbf{Y}}|X)$ will be sufficiently spread in the simplex formed by the columns of \mathbf{T} . Then, finding the minimum-volume data-enclosing convex hull of $P(\hat{\mathbf{Y}}|X)$ recovers the ground-truth \mathbf{T} and $P(\mathbf{Y}|X)$.

5. Related Works

In this section, we review existing methods in label-noise learning. Based on the statistical consistency of the learned classifier, we divided existing methods for label-noise learning into two categories: heuristic algorithms and statistically consistent algorithms.

Methods in the first category focus on employing heuristics to reduce the side-effect of noisy labels. For example, many methods use a specially designed strategy to select reliable samples (Yu et al., 2019; Han et al., 2018b; Malach & Shalev-Shwartz, 2017; Ren et al., 2018; Jiang et al., 2018; Yao et al., 2020a) or correct labels (Ma et al., 2018; Kremer et al., 2018; Tanaka et al., 2018; Reed et al., 2015). Although those methods empirically work well, there is not any theoretical guarantee on the consistency of the learned classifiers from all these methods.

Statistically consistent algorithms are primarily developed based on a loss correction procedure (Liu & Tao, 2016; Patrini et al., 2017; Zhang & Sabuncu, 2018). For these methods, the noise transition matrix plays a key role in building consistent classifiers. For example, Patrini et al. (2017) leveraged a two-stage training procedure of first estimating the noise transition matrix and then use it to modify the loss to ensure risk consistency. These works rely on anchor points or instances belonging to a specific class with probability one or approximately one. When there are no anchor points in datasets or data distributions, all the aforementioned methods cannot guarantee the statistical consistency. Another approach is to jointly learn the noise transition matrix and classifier. For instance, on top of the softmax layer of the classification network (Goldberger & Ben-Reuven, 2017), a constrained linear layer or a nonlinear softmax layer is added to model the noise tran-

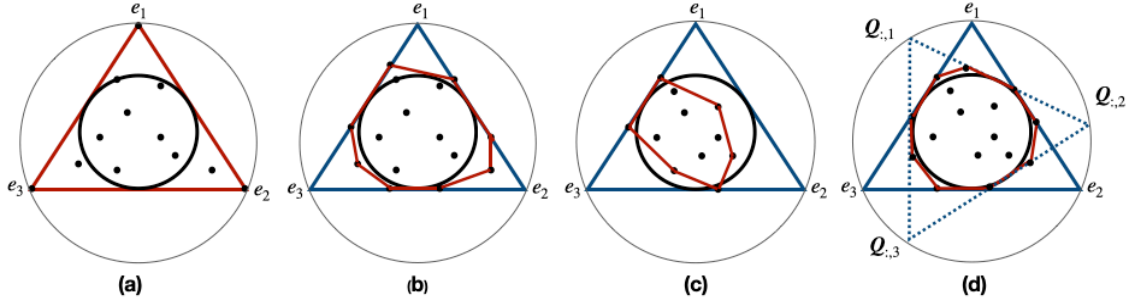


Figure 3. Illustration of the anchor-point assumption and the sufficiently scattered assumption in the case of $C = 3$ by assuming that the viewer are facing the hyperplane $\mathbf{1}^\top \mathbf{x} = 1$ from the positive orthant. The dots are class-posterior probabilities $P(\mathbf{Y}|\mathbf{X} = \mathbf{x})$; the triangle is the non-negative orthant; the inner circle is \mathcal{R} ; the region encompassed by red lines is $\text{cone}\{\mathbf{H}\}$. Clearly, the anchor-point assumption (a) is a special case of the sufficiently scattered assumption (b).

sition matrix (Sukhbaatar et al., 2015). Zhang et al. (2021) concurrently propose a one-step method for the label-noise learning problem and a derivative-free method for estimating the transition matrix. Specifically, their method uses a total variation regularization term to prevent the over-confidence problem of the neural network, which leads to a more accurate noisy class-posterior. However, the anchor-point assumption is still needed for their method. Based on different motivations, assumptions and learning objectives, their method achieves different theoretical results compared with our proposed method. Learning with noisy labels are also closely related to learning with complementary labels where instead of noisy labels, only complementary labels are given for training (Yu et al., 2018; Chou et al., 2020; Feng et al., 2020).

Recently, some methods exploiting semi-supervised learning techniques have been proposed to solve the label-noise learning problem like SELF (Nguyen et al., 2019) and DivideMix (Li et al., 2019). These methods are aggregations of multiple techniques such as augmentations and multiple networks. Noise robustness is significantly improved with these methods. However, these methods are sensitive to the choice of hyperparameters and changes in data and noise types would generate degenerated classifiers. In addition, the computational cost of these methods increases significantly compared with previous methods.

6. Experiments

In this section, we verify the robustness of the proposed volume minimization network (VolMinNet) from two folds: the estimation error of the transition matrix and the classification accuracy.

Datasets We evaluate the proposed method on three synthetic noisy datasets, i.e., MNIST, CIFAR-10 and CIFAR-100 and one real-world noisy dataset, i.e., clothing1M. We

leave out 10% of the training examples as the validation set. The three synthetic datasets contain clean data. We corrupted the training and validation sets manually according to transition matrices. Specifically, we conduct experiments with two commonly used types of noise: (1) Symmetry flipping (Patrini et al., 2017); (2) Pair flipping (Han et al., 2018b). We report both the classification accuracy on the test set and the estimation error between the estimated transition matrix \hat{T} and the true transition matrix T . All experiments are repeated five times on all datasets. Following T-Revision (Xia et al., 2019), we also conducted experiments on datasets where possible anchor points are removed from the datasets. The details and more experimental results can be found in the Supplementary Material.

Clothing1M is a real-world noisy dataset which consists of 1M images with real-world noisy labels. Existing methods like Forward (Patrini et al., 2017) and T-revision (Xia et al., 2019) use the additional 50k clean training data to help initialize the transition matrix and validate on 14k clean validation data. Here we use another setting which is also commonly used in the literature (Xia et al., 2020). We only exploit the 1M data for both transition matrix estimation and classification training. Specifically, we leave out 10% of the noisy training examples as a noisy validation set for model selection. We think this setting is more natural considering that it does not require any clean data. All results of baseline methods are quoted from PTD (Xia et al., 2020) as we have the same setting.

Network structure and optimization For a fair comparison, we implement all methods with default parameters by PyTorch on Tesla V100-SXM2. For MNIST, we use a LeNet-5 network. SGD is used to train the classification network h_θ with batch size 128, momentum 0.9, weight decay 10^3 and a learning rate 10^{-2} . Adam with default parameters is used to train the transition matrix \hat{T} . The algorithm is run for 60 epoch. For CIFAR10, we use a

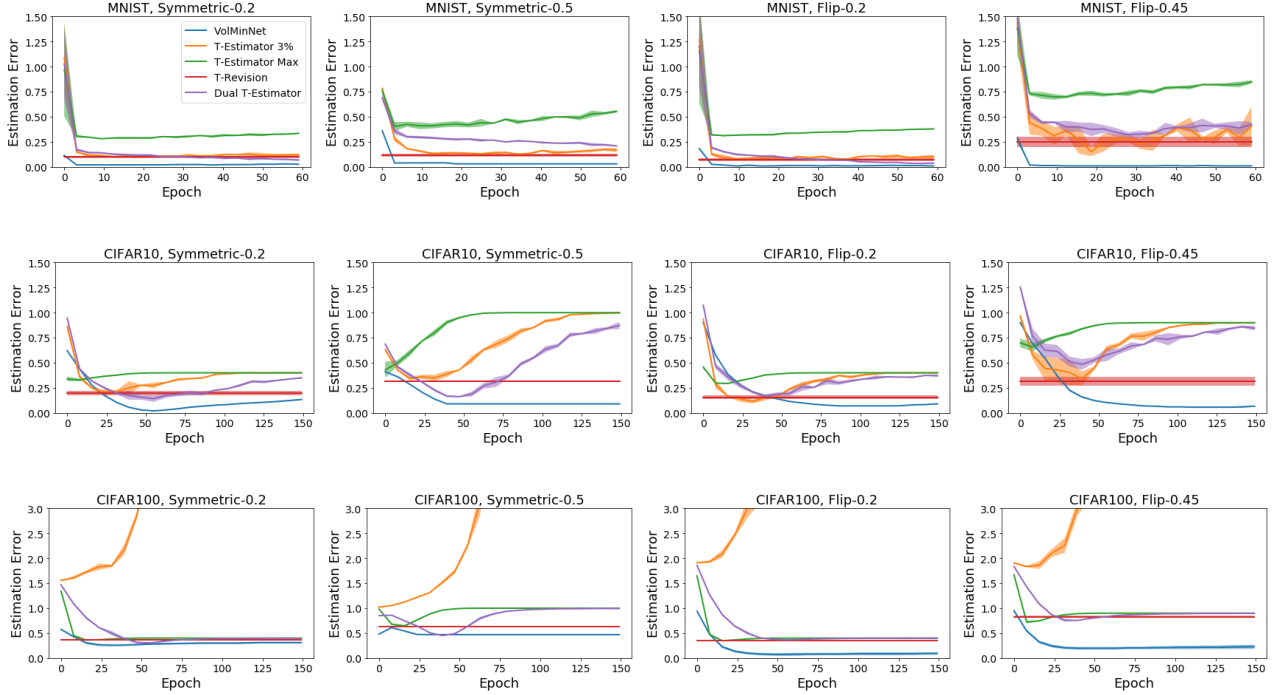


Figure 4. Transition matrix estimation error on MNIST, CIFAR-10 and CIFAR-100. The error bar for the standard deviation in each figure has been shaded. The lower the better.

ResNet-18 network. SGD is used to train both the classification network h_θ and the transition matrix \hat{T} with batch size 128, momentum 0.9, weight decay 10^3 and an initial learning rate 10^{-2} . The algorithm is run for 150 epoch and the learning rate is divided by 10 after the 30th and 60th epoch. For CIFAR100, we use a ResNet-32 network. SGD is used to train the classification network h_θ with batch size 128, momentum 0.9, weight decay 10^3 and an initial learning rate 10^{-2} . Adam with default parameters is used to train the transition matrix \hat{T} . The algorithm is run for 150 epoch and the learning rate is divided by 10 after the 30th and 60th epoch. For CIFAR-10 and CIFAR-100, we perform data augmentation by horizontal random flips and 32×32 random crops after padding 4 pixels on each side. For clothing1M, we use a ResNet-50 pre-trained on ImageNet. We only use the 1M noisy data to train and validate the network. For the optimization, SGD is used to train both the classification network h_θ and the transition matrix \hat{T} with momentum 0.9, weight decay 10^3 , batch size 32, and run with learning rates 2×10^3 and 2×10^5 for 5 epochs each. For each epoch, we ensure the noisy labels for each class are balanced with undersampling. Throughout all experiments, we fixed $\lambda = 10^{-4}$ and the trainable weights w of \hat{T} are initialized with $\ln \frac{1}{C-2}$ (roughly -2 for MNIST and CIFAR10, -4.5 for CIFAR100 and -2.5 for clothing1M).

6.1. Transition Matrix Estimation

For evaluating the effectiveness of estimating the transition matrix, we compare the proposed method with the following methods: (1) T-estimator max (Patrini et al., 2017), which identify the extreme-valued noisy class-posterior probabilities from given samples to estimate the transition matrix. (2) T-estimator 3% (Patrini et al., 2017), which takes a α -percentile in place of the argmax of Equation 4. (3) T-Revision (Xia et al., 2019), which introduces a slack variable to revise the noise transition matrix after initializing the transition matrix with T-estimator. (4) Dual T-estimator (Yao et al., 2020b), which introduces an intermediate class to avoid directly estimating the noisy class-posterior and factorizes the transition matrix into the product of two easy-to-estimate transition matrices.

To show that the proposed method is more robust in estimating the transition matrix, we plot the estimation error for the transition matrix, i.e., $\|T - \hat{T} - \Delta\hat{T}\|_1 / \|T\|_1$. Figure 5 depicts estimation errors of transition matrices estimated by the proposed VolMinNet and other baseline methods. For all different settings of noise on three different datasets (original intact datasets), VolMinNet consistently gives better results compared to the baselines, which shows its superior robustness against label noise. For example, on CIFAR100 (Flip-0.45), our method achieves estimation er-

Provably End-to-end Label-noise Learning without Anchor Points

	MNIST		CIFAR-10		CIFAR-100	
	Sym-20%	Sym-50%	Sym-20%	Sym-50%	Sym-20%	Sym-50%
Decoupling	97.04 \pm 0.06	94.58 \pm 0.08	77.32 \pm 0.35	54.07 \pm 0.46	41.92 \pm 0.49	22.63 \pm 0.44
MentorNet	97.21 \pm 0.06	95.56 \pm 0.15	81.35 \pm 0.23	73.47 \pm 0.15	42.88 \pm 0.41	32.66 \pm 0.40
Co-teaching	97.07 \pm 0.10	95.20 \pm 0.23	82.27 \pm 0.07	75.55 \pm 0.07	48.48 \pm 0.66	36.77 \pm 0.52
Forward	98.60 \pm 0.19	97.77 \pm 0.16	85.20 \pm 0.80	74.82 \pm 0.78	54.90 \pm 0.74	41.85 \pm 0.71
T-Revision	98.72 \pm 0.10	98.23 \pm 0.10	87.95 \pm 0.36	80.01 \pm 0.62	62.72 \pm 0.69	49.12 \pm 0.22
DMI	98.70 \pm 0.02	98.12 \pm 0.21	87.54 \pm 0.20	82.68 \pm 0.21	62.65 \pm 0.39	52.42 \pm 0.64
Dual T	98.43 \pm 0.05	98.15 \pm 0.12	88.35 \pm 0.33	82.54 \pm 0.19	62.16 \pm 0.58	52.49 \pm 0.37
VolMinNet	98.74 \pm 0.08	98.23 \pm 0.16	89.58 \pm 0.26	83.37 \pm 0.25	64.94 \pm 0.40	53.89 \pm 1.26

	MNIST		CIFAR-10		CIFAR-100	
	Pair-20%	Pair-45%	Pair-20%	Pair-45%	Pair-20%	Pair-45%
Decoupling	96.93 \pm 0.07	94.34 \pm 0.54	77.12 \pm 0.30	53.71 \pm 0.99	40.12 \pm 0.26	27.97 \pm 0.12
MentorNet	96.89 \pm 0.04	91.98 \pm 0.46	77.42 \pm 0.00	61.03 \pm 0.20	39.22 \pm 0.47	26.48 \pm 0.37
Co-teaching	97.00 \pm 0.06	96.25 \pm 0.01	80.65 \pm 0.20	73.02 \pm 0.23	42.79 \pm 0.79	27.97 \pm 0.20
Forward	98.84 \pm 0.10	75.06 \pm 12.61	88.21 \pm 0.48	77.44 \pm 6.89	56.12 \pm 0.54	36.88 \pm 2.32
T-Revision	98.89 \pm 0.08	84.56 \pm 8.18	90.33 \pm 0.52	78.94 \pm 2.58	64.33 \pm 0.49	41.55 \pm 0.95
DMI	98.84 \pm 0.09	97.92 \pm 0.76	89.89 \pm 0.45	73.15 \pm 7.31	59.56 \pm 0.73	38.17 \pm 2.02
Dual T	98.86 \pm 0.04	96.71 \pm 0.12	89.77 \pm 0.25	76.53 \pm 2.51	67.21 \pm 0.43	47.60 \pm 0.43
VolMinNet	99.01 \pm 0.07	99.00 \pm 0.07	90.37 \pm 0.30	88.54 \pm 0.21	68.45 \pm 0.69	58.90 \pm 0.89

Table 1. Classification accuracy (percentage) on MNIST, CIFAR-10 and CIFAR-100.

ror around 0.25, while baseline methods can only reach at around 0.75. These results show that our method establishes the new state of the art in estimating transition matrices.

6.2. Classification accuracy Evaluation

We compare the classification accuracy of the proposed method with the following methods: (1) Decoupling (Malach & Shalev-Shwartz, 2017). (2) MentorNet (Jiang et al., 2018). (3) Co-teaching (Han et al., 2018b). (4) Forward (Patrini et al., 2017). (5) T-Revision (Xia et al., 2019). (7) DMI (Xu et al., 2019). (8) Dual T (Yao et al., 2020b). Note that we did not compare the proposed method with some methods like SELF (Nguyen et al., 2019) and DivideMix (Li et al., 2019). This is because these methods are aggregations of semi-supervised learning techniques which have high computational complexity and are sensitive to the choice of hyperparameters. In this work, we are more focusing on solving the label noise learning without anchor points theoretically.

In Table 3, we present the classification accuracy by the proposed VolMinNet and baseline methods on synthetic noisy datasets. VolMinNet outperforms baseline methods on almost all settings of noise. This result is natural after we have shown that VolMinNet leads to smaller estimation error of the transition matrix compared with baseline methods. While the differences of accuracy among different methods are marginal for symmetric noise, VolMinNet outperforms baselines by over 10% with Pair-45% noise

Decoupling	MentorNet	Co-teaching	Forward
54.53	56.79	60.15	69.91
T-Revision	DMI	PTD	VolMinNet
70.97	70.12	71.67	72.42

Table 2. Classification accuracy (percentage) on Clothing1M. Only noisy data are exploited for training and validation.

and has much smaller standard deviations. These results show the clear advantage of the proposed VolMinNet. It has better robustness to different settings of noise and datasets compared to baseline methods.

Finally, we show the results on Clothing1M in Table 2. As explained in the previous section, Forward and T-Revision exploited the 50k clean data and their noisy versions in 1M noisy data to help initialize the noise transition matrix, which is not practical in real-world settings. For a fair comparison, we report results by only using the 1M noisy data to train and validate the network. As shown in Table 2, our method outperforms previous transition matrix based methods and heuristic methods on the Clothing1M dataset. In addition, the performance on the Clothing1M dataset shows that the proposed method has certain robustness against instance-dependent noise as well.

7. Discussion and Conclusion

In this paper, we considered the problem of label-noise learning without anchor points. We relax the anchor-point assumption with our proposed VolMinNet. The consis-

tency of the estimated transition matrix and the learned classifier are theoretically proved under the sufficiently scattered assumption. Experimental results have demonstrated the robustness of the proposed VolMinNet. Future work should focus on improving the estimation of the noisy class posterior which we believe is the bottleneck of our method.

Acknowledgements

XL was supported by an Australian Government RTP Scholarship. TL was supported by Australian Research Council Project DE-190101473. BH was supported by the RGC Early Career Scheme No. 22200720, NSFC Young Scientists Fund No. 62006202 and HKBU CSD Departmental Incentive Grant. GN and MS were supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, Japan. MS was also supported by Institute for AI and Beyond, UTokyo. Authors also thank for the help from Dr. Alan Blair, Kevin Lam, Yivan Zhang and members of the Trustworthy Machine Learning Lab at the University of Sydney.

References

- Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. A closer look at memorization in deep networks. In *ICML*, pp. 233–242, 2017.
- Berthon, A., Han, B., Niu, G., Liu, T., and Sugiyama, M. Confidence scores make instance-dependent label-noise learning possible. In *ICML*, 2021.
- Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.
- Chou, Y.-T., Niu, G., Lin, H.-T., and Sugiyama, M. Unbiased risk estimators can mislead: A case study of learning with complementary labels. In *ICML*, pp. 1929–1938. PMLR, 2020.
- Daniely, A. and Granot, E. Generalization bounds for neural networks via approximate description length. In *NeurIPS*, pp. 13008–13016, 2019.
- Fazel, M., Hindi, H., and Boyd, S. P. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *ACC*, pp. 2156–2162, 2003.
- Feng, L., Kaneko, T., Han, B., Niu, G., An, B., and Sugiyama, M. Learning with multiple complementary labels. In *ICML*, pp. 3072–3081. PMLR, 2020.
- Fu, X., Ma, W.-K., Huang, K., and Sidiropoulos, N. D. Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain. *IEEE Transactions on Signal Processing*, 63(9):2306–2320, 2015.
- Fu, X., Huang, K., Yang, B., Ma, W.-K., and Sidiropoulos, N. D. Robust volume minimization-based matrix factorization for remote sensing and document clustering. *IEEE Transactions on Signal Processing*, 64(23):6254–6268, 2016.
- Fu, X., Huang, K., and Sidiropoulos, N. D. On identifiability of nonnegative matrix factorization. *IEEE Signal Processing Letters*, 25(3):328–332, 2018.
- Goldberger, J. and Ben-Reuven, E. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.
- Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I., Zhang, Y., and Sugiyama, M. Masking: A new perspective of noisy supervision. In *NeurIPS*, pp. 5836–5846, 2018a.
- Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pp. 8527–8537, 2018b.
- Han, B., Niu, G., Yu, X., Yao, Q., Xu, M., Tsang, I., and Sugiyama, M. Sigua: Forgetting may make learning with noisy labels more robust. In *ICML*, pp. 4006–4016. PMLR, 2020a.
- Han, B., Yao, Q., Liu, T., Niu, G., Tsang, I. W., Kwok, J. T., and Sugiyama, M. A survey of label-noise representation learning: Past, present and future, 2020b.
- Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pp. 2309–2318, 2018.
- Karush, W. Minima of functions of several variables with inequalities as side constraints. *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*, 1939.
- Kremer, J., Sha, F., and Igel, C. Robust active label correction. In *AISTATS*, pp. 308–316, 2018.
- Kuhn, H. W. and Tucker, A. W. Nonlinear programming. In *Traces and emergence of nonlinear programming*, pp. 247–258. Springer, 2014.
- Li, J. and Bioucas-Dias, J. M. Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data. In *IGARSS*, pp. III–250, 2008.
- Li, J., Socher, R., and Hoi, S. C. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2019.
- Li, W., Wang, L., Li, W., Agustsson, E., and Van Gool, L. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

- Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012.
- Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.
- Liu, Y. and Guo, H. Peer loss functions: Learning from noisy labels without knowing noise rates. In *ICML*, pp. 6226–6236. PMLR, 2020.
- Ma, X., Wang, Y., Houle, M. E., Zhou, S., Erfani, S. M., Xia, S.-T., Wijewickrema, S., and Bailey, J. Dimensionality-driven learning with noisy labels. In *ICML*, pp. 3361–3370, 2018.
- Malach, E. and Shalev-Shwartz, S. "Decoupling" when to update" from "how to update". In *NeurIPS*, pp. 960–970, 2017.
- Miao, L. and Qi, H. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 45(3):765–777, 2007.
- Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In *NeurIPS*, pp. 1196–1204, 2013.
- Nguyen, D. T., Mummadi, C. K., Ngo, T. P. N., Nguyen, T. H. P., Beggel, L., and Brox, T. Self: Learning to filter noisy labels with self-ensembling. In *ICLR*, 2019.
- Northcutt, C. G., Wu, T., and Chuang, I. L. Learning with confident examples: Rank pruning for robust classification with noisy labels. In *UAI*, 2017.
- Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pp. 1944–1952, 2017.
- Reed, S. E., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. Training deep neural networks on noisy labels with bootstrapping. In *ICLR, Workshop Track Proceedings*, 2015.
- Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In *ICML*, pp. 4331–4340, 2018.
- Scott, C. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *AISTATS*, pp. 838–846, 2015.
- Sukhbaatar, S., Estrach, J. B., Paluri, M., Bourdev, L., and Fergus, R. Training convolutional networks with noisy labels. In *ICLR*, 2015.
- Tanaka, D., Ikami, D., Yamasaki, T., and Aizawa, K. Joint optimization framework for learning with noisy labels. In *CVPR*, pp. 5552–5560, 2018.
- Thekumparampil, K. K., Khetan, A., Lin, Z., and Oh, S. Robustness of conditional gans to noisy labels. In *NeurIPS*, pp. 10271–10282, 2018.
- Wu, S., Xia, X., Liu, T., Han, B., Gong, M., Wang, N., Liu, H., and Niu, G. Class2simi: A new perspective on learning with label noise. In *ICML*. PMLR, 2021.
- Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., and Sugiyama, M. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, pp. 6835–6846, 2019.
- Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., and Sugiyama, M. Part-dependent label noise: Towards instance-dependent label noise. *NeurIPS*, 2020.
- Xia, X., Liu, T., Han, B., Gong, C., Wang, N., Ge, Z., and Chang, Y. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*, 2021.
- Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. Learning from massive noisy labeled data for image classification. In *CVPR*, pp. 2691–2699, 2015.
- Xu, Y., Cao, P., Kong, Y., and Wang, Y. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*, pp. 6222–6233, 2019.
- Yao, Q., Yang, H., Han, B., Niu, G., and Kwok, J. T.-Y. Searching to exploit memorization effect in learning with noisy labels. In *ICML*, pp. 10789–10798. PMLR, 2020a.
- Yao, Y., Liu, T., Han, B., Gong, M., Deng, J., Niu, G., and Sugiyama, M. Dual t: Reducing estimation error for transition matrix in label-noise learning. In *NeurIPS*, 2020b.
- Yu, X., Liu, T., Gong, M., and Tao, D. Learning with biased complementary labels. In *ECCV*, pp. 68–83, 2018.
- Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., and Sugiyama, M. How does disagreement help generalization against label corruption? In *ICML*, pp. 7164–7173, 2019.
- Yu, X., Liu, T., Gong, M., Zhang, K., Batmanghelich, K., and Tao, D. Label-noise robust domain adaptation. In *ICML*, pp. 10913–10924. PMLR, 2020.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Zhang, Y., Niu, G., and Sugiyama, M. Learning noise transition matrix from only noisy labels via total variation regularization. In *ICML*, 2021.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, pp. 8778–8788, 2018.

Zhu, Z., Liu, T., and Liu, Y. A second-order approach to learning with instance-dependent label noise. 2021a.

Zhu, Z., Song, Y., and Liu, Y. Clusterability as an alternative to anchor points when learning with noisy labels. In *ICML*, 2021b.

A. Appendix

A.1. Proof of Proposition 1

To prove proposition 1, we first show that the anchor-point assumption is a sufficient condition for the sufficiently scattered assumption. In other words, we need to show that if the anchor-point assumption is satisfied, then two conditions of the sufficiently scattered assumption must hold.

We start with condition (2) of the sufficiently scattered assumption. We need to show that if the anchor-point assumption is hold, then there exists a set $\mathcal{H} = \{x_1, \dots, x_m\} \subseteq \mathcal{X}$ such that the matrix $\mathbf{H} = [P(Y|X = x_1), \dots, P(Y|X = x_m)]$ satisfies that $\text{cone}\{\mathbf{H}\} \not\subseteq \text{cone}\{\mathbf{Q}\}$, for any unitary matrix $\mathbf{Q} \in \mathbb{R}^{C \times C}$ that is not a permutation matrix.

Since the anchor-point assumption is satisfied, then there exists a matrix $\mathbf{H} = [P(Y|X = \mathbf{x}^1), \dots, P(Y|X = \mathbf{x}^C)]$ where $\mathbf{x}^1, \dots, \mathbf{x}^C$ are anchor points for each class. From the definition of anchor points, we have $P(Y|X = \mathbf{x}^i) = \mathbf{e}_i$. This implies that

$$\mathbf{H} = [P(Y|X = \mathbf{x}^1), \dots, P(Y|X = \mathbf{x}^C)] = \mathbf{I}, \quad (8)$$

where \mathbf{I} is the identity matrix. By the definition of the identity matrix \mathbf{I} , it is clear that $\text{cone}\{\mathbf{H}\} = \text{cone}\{\mathbf{I}\} \not\subseteq \text{cone}\{\mathbf{Q}\}$, for any unitary matrix $\mathbf{Q} \in \mathbb{R}^{C \times C}$ that is not a permutation matrix. This shows that condition (2) of the sufficiently scattered assumption is satisfied if the anchor-point assumption is hold.

Next, we show that condition (1) will also be satisfied, i.e., the convex cone $\mathcal{R} \subseteq \text{cone}\{\mathbf{H}\}$, where $\mathcal{R} = \{\mathbf{v} \in \mathbb{R}^C | \mathbf{v}^\top \mathbf{1} \geq \sqrt{C-1} \|\mathbf{v}\|_2\}$. By Eq. (8), condition (1) of Theorem 1 is equivalent to

$$\mathcal{R} \subseteq \text{cone}\{\mathbf{I}\} = \{\mathbf{u} | \mathbf{u} = \sum_{j=1}^C \mathbf{e}_j \alpha_j, \alpha_j \geq 0, \forall j\}. \quad (9)$$

This means that all elements in \mathcal{R} must be in the non-negative orthant of \mathbb{R}^C , i.e., for all $\mathbf{v} \in \mathcal{R}$, $v_i \geq 0$ for all $i \in \{1, \dots, C\}$. Consider $\mathbf{v} \in \mathcal{R}$ and let $\hat{\mathbf{v}}$ be the normalized vector of \mathbf{v} , by definition of \mathcal{R} we have the following chain:

$$\mathbf{v}^\top \mathbf{1} \geq \sqrt{C-1} \|\mathbf{v}\|_2, \quad (10a)$$

$$\frac{\mathbf{v}^\top \mathbf{1}}{\|\mathbf{v}\|} = \hat{\mathbf{v}}^\top \mathbf{1} \geq \sqrt{C-1}, \quad (10b)$$

$$\sum_{i \in \{1, 2, \dots, C\}} \hat{v}_i \geq \sqrt{C-1}. \quad (10c)$$

To show \mathbf{v} is non-negative is equivalent to prove that $\hat{\mathbf{v}}$ is non-negative, i.e., $\forall k \in \{1, \dots, C\}$, $\hat{v}_k \geq 0$. Let $\mathbf{u} \in \mathbb{R}^{C-1}$ be the vector which has same elements with $\hat{\mathbf{v}}$ except that the k th element \hat{v}_k is removed. Following Eq. 10, we have:

$$\hat{v}_k \geq \sqrt{C-1} - \sum_{i \in \{1, 2, \dots, C\} \setminus \{k\}} \hat{v}_i, \quad (11a)$$

$$\hat{v}_k \geq \sqrt{C-1} - \mathbf{u}^\top \mathbf{1}. \quad (11b)$$

By the Cauchy-Schwarz inequality, we get the following inequality:

$$|\mathbf{u}^\top \mathbf{1}| \leq \|\mathbf{u}\| \|\mathbf{1}\|. \quad (12)$$

Then by the definition of \mathbf{u} and $\mathbf{1}$, we have $\|\mathbf{u}\| \leq 1$ and $\|\mathbf{1}\| = \sqrt{C-1}$. Combined this with Eq. 11 and Eq. 12, we get:

$$\hat{v}_k \geq \sqrt{C-1} - \|\mathbf{u}\| \|\mathbf{1}\| \geq 0. \quad (13)$$

This simply implies that $\mathbf{v}_k \geq 0$ for all $k \in \{1, 2, \dots, C\}$ and we have proved that the anchor-point assumption is a sufficient condition of the sufficiently scattered assumption.

We now prove that the anchor-point assumption is not a necessary condition for the sufficiently scattered assumption. Suppose $P(\mathbf{Y}|\mathbf{X})$ has the property that $x^1, x^2, \dots, x^C \notin \mathcal{X}$ which means that the anchor-point assumption is not satisfied. We also assume that there exist a set $\mathcal{H} = \{x_1, \dots, x_m\} \subseteq \mathcal{X}$ such that $\text{cone}\{\mathbf{H}\}$ covers the whole non-negative orthant except the area along each axis (area formed by noisy class-posterior of anchor points). Since these areas along each axis are not part of \mathcal{R} when $C > 2$, it is clear that condition (1) of the sufficiently scattered assumption is satisfied. Besides, by definition of \mathbf{H} , there is no other unitary matrix which can cover $\text{cone}\{\mathbf{H}\}$ except permutation matrices. This shows that condition (2) of the sufficiently scattered assumption is also satisfied and the proof is completed. \square

A.2. Proof of Theorem 1

The insights of our proof are from previous works in non-negative matrix factorisation (Fu et al., 2015). To proceed, let us first introduce following classic lemmas in convex analysis:

Lemma 1. *If \mathcal{K}_1 and \mathcal{K}_2 are convex cones and $\mathcal{K}_1 \subseteq \mathcal{K}_2$, then, $\text{dual}\{\mathcal{K}_2\} \subseteq \text{dual}\{\mathcal{K}_1\}$.*

Lemma 2. *If \mathbf{A} is invertible, then $\text{dual}(\mathbf{A}) = \text{cone}(\mathbf{A}^{-\top})$.*

Readers are referred to Boyd et al.(2004) for details. Our purpose is to show that criterion (5) has unique solutions which are the ground-truth $P(\mathbf{Y}|X)$ and \mathbf{T} . To this end, let us denote $(\mathbf{T}_*, h_{\theta_*})$ as a feasible solution of Criterion (5), i.e.,

$$\mathbf{T}_* h_{\theta_*} = \mathbf{T} P(\mathbf{Y}|X) = P(\tilde{\mathbf{Y}}|X). \quad (14)$$

As defined in sufficient scattered assumption, we have the matrix $\mathbf{H} = [P(\mathbf{Y}|X = x_1), \dots, P(\mathbf{Y}|X = x_m)]$ defined on the set $\mathcal{H} = \{x_1, \dots, x_m\} \subseteq \mathcal{X}$. Let $\mathbf{H}_* = [h_{\theta_*}(x_1), \dots, h_{\theta_*}(x_m)]$, it follows that

$$\mathbf{T}_* \mathbf{H}_* = \mathbf{T} \mathbf{H}. \quad (15)$$

Note that both \mathbf{T}_* and \mathbf{T} have full rank because they are diagonally dominant square matrices by definition. In addition, since the sufficiently scattered assumption is satisfied, $\text{rank}(\mathbf{H}) = C$ also holds (Fu et al., 2015). Therefore, there exists an invertible matrix $\mathbf{A} \in \mathbb{R}^{C \times C}$ such that

$$\mathbf{T}_* = \mathbf{T} \mathbf{A}^{-\top}, \quad (16)$$

where $\mathbf{A}^{-\top} = \mathbf{H} \mathbf{H}_*^\dagger$ and $\mathbf{H}_*^\dagger = \mathbf{H}_*^\top (\mathbf{H}_* \mathbf{H}_*^\top)^{-1}$ is the pseudo-inverse of \mathbf{H}_* .

Since $\mathbf{1}^\top \mathbf{H} = \mathbf{1}^\top$ and $\mathbf{1}^\top \mathbf{H}_* = \mathbf{1}^\top$ by definition, we get

$$\mathbf{1}^\top \mathbf{A}^{-\top} = \mathbf{1}^\top \mathbf{H} \mathbf{H}_*^\dagger = \mathbf{1}^\top \mathbf{H}_*^\dagger = \mathbf{1}^\top \mathbf{H}_* \mathbf{H}_*^\dagger = \mathbf{1}^\top. \quad (17)$$

Let $\mathbf{v} \in \text{cone}\{\mathbf{H}\}$, which by definition takes the form $\mathbf{v} = \mathbf{H} \mathbf{u}$ for some $\mathbf{u} \geq 0$. Using $\mathbf{H} = \mathbf{A}^{-\top} \mathbf{H}_*$, \mathbf{v} can be expressed as $\mathbf{v} = \mathbf{A}^{-\top} \tilde{\mathbf{u}}$ where $\tilde{\mathbf{u}} = \mathbf{H}_* \mathbf{u} \geq 0$. This implies that \mathbf{v} also lies in $\text{cone}\{\mathbf{A}^{-\top}\}$, i.e. $\text{cone}\{\mathbf{H}\} \subseteq \text{cone}\{\mathbf{A}^{-\top}\}$.

Recall Condition (1) of the sufficiently scattered assumption, i.e., $\mathcal{R} \subseteq \text{cone}\{\mathbf{H}\}$ where $\mathcal{R} = \{\mathbf{v} \in \mathbb{R}^C | \mathbf{1}^\top \mathbf{v} \geq \sqrt{C-1} \|\mathbf{v}\|_2\}$. It implies

$$\mathcal{R} \subseteq \text{cone}\{\mathbf{H}\} \subseteq \text{cone}(\mathbf{A}^{-\top}). \quad (18)$$

By applying Lemmas (1-2) to Eq. (18), we have

$$\text{cone}(\mathbf{A}) \subseteq \text{dual}\{\mathcal{R}\}, \quad (19)$$

where $\text{dual}\{\mathcal{R}\}$ is the dual cone of \mathcal{R} , which can be shown to be

$$\text{dual}\{\mathcal{R}\} = \{\mathbf{v} \in \mathbb{R}^C | \|\mathbf{v}\|_2 \leq \mathbf{1}^\top \mathbf{v}\}. \quad (20)$$

Then we have the following inequalities:

$$|\det(\mathbf{A})| \leq \prod_{i=1}^C \|\mathbf{A}_{:,i}\|_2 \quad (21a)$$

$$\leq \prod_{i=1}^C \mathbf{1}^\top \mathbf{A}_{:,i} \quad (21b)$$

$$\leq \left(\frac{\sum_{i=1}^C \mathbf{1}^\top \mathbf{A}_{:,i}}{C} \right)^C \quad (21c)$$

$$= \left(\frac{\mathbf{1}^\top \mathbf{A} \mathbf{1}}{C} \right)^C = 1, \quad (21d)$$

where (21a) is Hadamard's inequality; (21b) is by Eq. (19); (21c) is by the arithmetic-geometric mean inequality; and (21d) is by Eq. (17).

Note that $|\det(\mathbf{A})|^{-1} = |\det(\mathbf{A}^{-\top})|$ and $\det(\mathbf{T}_*) = \det(\mathbf{T} \mathbf{A}^{-\top}) = \det(\mathbf{T}) |\det(\mathbf{A})|^{-1}$ from properties of the determinant, it follows from Eq. (21) that $\det(\mathbf{T}_*) \geq \det(\mathbf{T})$. We also know that $\det(\mathbf{T}_*) \leq \det(\mathbf{T})$ must hold from Criterion (5), hence we have

$$\det(\mathbf{T}_*) = \det(\mathbf{T}) \quad (22)$$

By Hadamard's inequality, the equality in (21a) holds only if \mathbf{A} is column-orthogonal, which is equivalent to that $\mathbf{A}^{-\top}$ is column-orthogonal. Considering condition (2) in the definition of sufficiently scattered and the property of $\mathbf{A}^{-\top}$ that $\text{cone}\{\mathbf{H}\} \subseteq \text{cone}(\mathbf{A}^{-\top})$, the only possible choices of column-orthogonal $\mathbf{A}^{-\top}$ are

$$\mathbf{A}^{-\top} = \mathbf{\Pi} \mathbf{\Phi} \quad (23)$$

where $\mathbf{\Pi} \in \mathbb{R}^{C \times C}$ is any permutation matrix and $\mathbf{\Phi} \in \mathbb{R}^{C \times C}$ is any diagonal matrix with non-zero diagonals. By Eq. (17), we must have $\mathbf{\Phi} = \mathbf{I}$. Subsequently, we are left with $\mathbf{A}^{-\top} = \mathbf{\Pi}$, or equivalently, $\mathbf{T}_* = \mathbf{\Pi} \mathbf{T}$. Since \mathbf{T} and \mathbf{T}_* are both diagonal dominant, the only possible permutation matrix is \mathbf{I} , which means $\mathbf{T}_* = \mathbf{T}$ holds. By Eq. (14), it follows that $h_{\theta_*} = P(\mathbf{Y}|X)$. Hence we conclude that $(\mathbf{T}_*, h_{\theta_*}) = (\mathbf{T}, P(\mathbf{Y}|X))$ is the unique optimal solution to criterion (5). \square

B. Experiments on datasets where possible anchor points are manually removed.

Following Xia et al.(2019), to show the importance of anchor points, we remove possible anchor points from the datasets, i.e., instances with large estimated class-posterior probability $P(Y|X)$, before corrupting the training and validation sets. For MNIST we removed 40% of the instances with the largest estimated class posterior probabilities in each class. For CIFAR-10 and CIFAR-100, we removed 10% of the instances with the largest estimated class

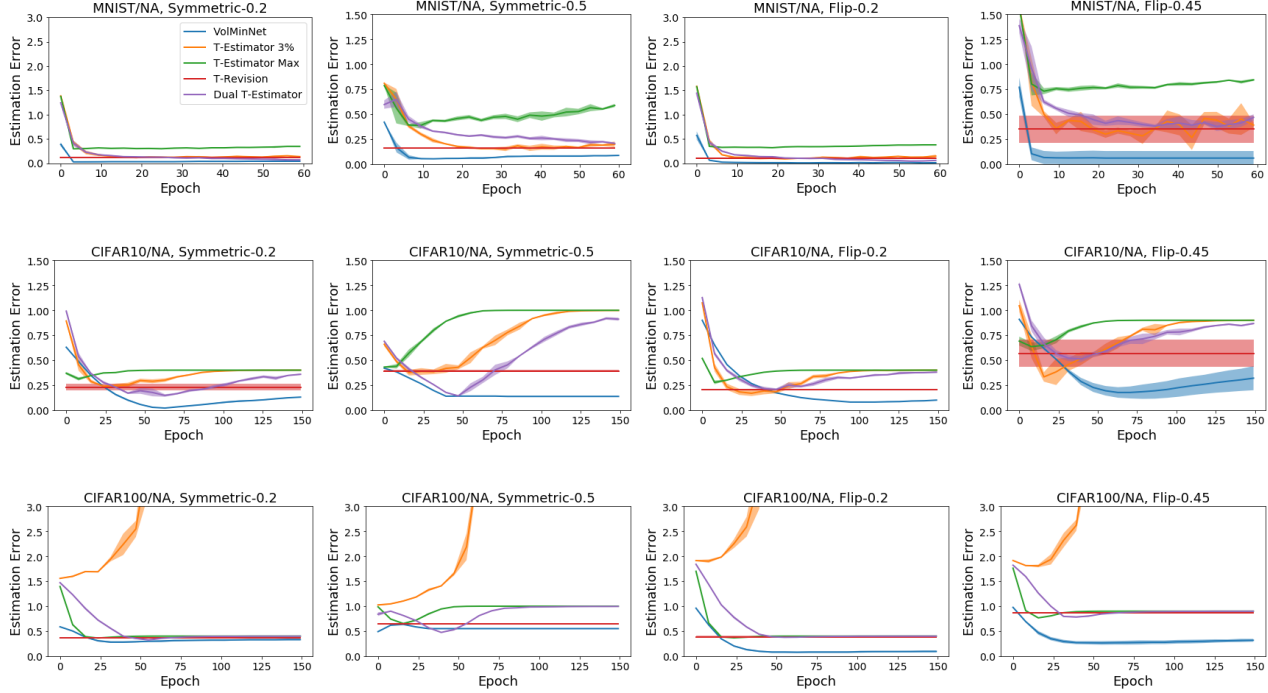


Figure 5. Transition matrix estimation error on MNIST/NA, CIFAR-10/NA, CIFAR-100/NA. Datasets with “/NA” means that possible anchor points are removed. The error bar for the standard deviation in each figure has been shaded. The lower the better.

	MNIST/NA		CIFAR-10/NA		CIFAR-100/NA	
	Sym-20%	Sym-50%	Sym-20%	Sym-50%	Sym-20%	Sym-50%
Decoupling	96.72 \pm 0.16	92.72 \pm 0.33	75.51 \pm 0.38	49.96 \pm 0.51	38.83 \pm 0.37	20.42 \pm 0.53
MentorNet	97.10 \pm 0.10	95.10 \pm 0.14	80.25 \pm 0.52	71.65 \pm 0.28	39.72 \pm 0.35	29.39 \pm 0.35
Co-teaching	97.06 \pm 0.12	94.89 \pm 0.10	81.74 \pm 0.32	73.38 \pm 0.45	44.92 \pm 0.11	33.13 \pm 0.88
Forward	98.46 \pm 0.07	97.59 \pm 0.05	84.25 \pm 0.22	70.00 \pm 3.07	50.58 \pm 0.68	36.79 \pm 1.86
T-Revision	98.72 \pm 0.13	97.86 \pm 0.11	86.81 \pm 0.19	74.10 \pm 2.34	59.57 \pm 1.13	43.75 \pm 0.84
DMI	98.42 \pm 0.03	97.87 \pm 0.18	83.42 \pm 0.54	77.82 \pm 0.45	56.29 \pm 0.28	41.81 \pm 0.70
Dual T	98.61 \pm 0.12	97.91 \pm 0.12	86.70 \pm 0.06	78.92 \pm 0.42	56.99 \pm 1.00	42.04 \pm 1.96
VolMinNet	98.72 \pm 0.06	97.94 \pm 0.07	88.72 \pm 0.03	82.38 \pm 0.65	63.40 \pm 1.25	51.04 \pm 1.23
	MNIST/NA		CIFAR-10/NA		CIFAR-100/NA	
	Pair-20%	Pair-45%	Pair-20%	Pair-45%	Pair-20%	Pair-45%
Decoupling	96.92 \pm 0.06	93.29 \pm 0.57	77.06 \pm 0.26	50.81 \pm 0.73	40.42 \pm 0.47	26.21 \pm 0.67
MentorNet	96.88 \pm 0.04	88.17 \pm 0.70	77.62 \pm 0.28	57.60 \pm 0.35	39.11 \pm 0.41	25.17 \pm 0.36
Co-teaching	96.96 \pm 0.07	95.34 \pm 0.09	80.70 \pm 0.18	69.15 \pm 0.89	43.04 \pm 0.73	26.67 \pm 0.29
Forward	98.61 \pm 0.33	78.51 \pm 17.48	85.87 \pm 0.82	53.92 \pm 11.39	51.37 \pm 0.99	34.69 \pm 1.37
T-Revision	98.71 \pm 0.31	82.65 \pm 14.61	87.52 \pm 0.58	53.96 \pm 14.67	59.70 \pm 1.43	38.35 \pm 0.60
DMI	98.78 \pm 0.11	97.46 \pm 1.38	86.14 \pm 1.52	70.01 \pm 5.63	54.05 \pm 1.09	35.03 \pm 2.91
Dual T	98.76 \pm 0.13	85.77 \pm 7.85	89.02 \pm 0.40	65.17 \pm 0.72	59.07 \pm 3.79	36.95 \pm 3.19
VolMinNet	98.87 \pm 0.11	97.80 \pm 2.43	89.26 \pm 0.22	84.48 \pm 3.85	64.88 \pm 1.87	56.07 \pm 3.35

Table 3. Classification accuracy (percentage) on MNIST, CIFAR-10, CIFAR-100 and MNIST/NA, CIFAR-10/NA, CIFAR-100/NA. Datasets with “/NA” means that possible anchor points are removed.

posterior probabilities in each class. We add "/NA" following the dataset's name denote those datasets which are modified by removing possible anchor points. The detailed experimental results are shown in Figure 5 (estimation error) and Table 3 (classification accuracy). The experimental performance shows that our proposed method outperforms the baseline methods.