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

1. Proofs

1.1. 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 $\boldsymbol{H} = [P(Y|X = x_1), \dots, P(Y|X = x_m)]$ satisfies that $\text{cone}\{\boldsymbol{H}\} \not\subseteq \text{cone}\{\boldsymbol{Q}\}$, for any unitary matrix $\boldsymbol{Q} \in \mathbb{R}^{C\times C}$ that is not a permutation matrix.

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

$$\boldsymbol{H} = [P(Y|X = \boldsymbol{x}^1), ..., P(Y|X = \boldsymbol{x}^C)] = \boldsymbol{I}, \tag{1}$$

where $\boldsymbol{I}$ is the identity matrix. By the definition of the identity matrix $\boldsymbol{I}$, it is clear that $\text{cone}\{\boldsymbol{H}\} = \text{cone}\{\boldsymbol{I}\} \not\subseteq \text{cone}\{\boldsymbol{Q}\}$, for any unitary matrix $\boldsymbol{Q} \in \mathbb{R}^{C\times C}$ 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 $\mathcal{R} \subseteq \text{cone}\{\boldsymbol{H}\}$, where $\mathcal{R} = \{\boldsymbol{v} \in \mathbb{R}^C | \boldsymbol{v}^\top \boldsymbol{1} \geq \sqrt{C-1}\|\boldsymbol{v}\|_2\}$. By Eq. (1), condition (1) of Theorem 1 is equivalent to

$$\mathcal{R} \subseteq \text{cone}\{\boldsymbol{I}\} = \{\boldsymbol{u} | \boldsymbol{u} = \sum_{j=1}^{C} \boldsymbol{e}_j \alpha_j,\ \alpha_j \geq 0,\ \forall j\}. \tag{2}$$

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

$$\boldsymbol{v}^\top \boldsymbol{1} \geq \sqrt{C-1}\|\boldsymbol{v}\|_2, \tag{3a}$$

$$\frac{\boldsymbol{v}^\top}{\|\boldsymbol{v}\|}\boldsymbol{1} = \hat{\boldsymbol{v}}^\top \boldsymbol{1} \geq \sqrt{C-1}, \tag{3b}$$

$$\sum_{i\in\{1,2,\dots,C\}} \hat{\boldsymbol{v}}_i \geq \sqrt{C-1}. \tag{3c}$$

To show $\boldsymbol{v}$ is non-negative is equivalent to prove that $\hat{\boldsymbol{v}}$ is non-negative, i.e., $\forall k \in \{1, \dots, C\}$, $\hat{\boldsymbol{v}}_k \geq 0$. Let $\boldsymbol{u} \in \mathbb{R}^{C-1}$ be the vector which has same elements with $\hat{\boldsymbol{v}}$ except that the kth element $\hat{\boldsymbol{v}}_k$ is removed. Following Eq. 3, we have:

$$\hat{\boldsymbol{v}}_k \geq \sqrt{C-1} - \sum_{i\in\{1,2,\dots,C\}\setminus\{k\}} \hat{\boldsymbol{v}}_i, \tag{4a}$$

$$\hat{\boldsymbol{v}}_k \geq \sqrt{C-1} - \boldsymbol{u}^\top \boldsymbol{1}. \tag{4b}$$

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

$$|\boldsymbol{u}^\top \boldsymbol{1}| \leq \|\boldsymbol{u}\|\|\boldsymbol{1}\|. \tag{5}$$

Then by the definition of $\boldsymbol{u}$ and $\boldsymbol{1}$, we have $\|\boldsymbol{u}\| \leq 1$ and $\|\boldsymbol{1}\| = \sqrt{C-1}$. Combined this with Eq. 4 and Eq. 5, we get:

$$\hat{\boldsymbol{v}}_k \geq \sqrt{C-1} - \|\boldsymbol{u}\|\|\boldsymbol{1}\| \geq 0. \tag{6}$$

This simply implies that $\boldsymbol{v}_k \geq 0$ 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 $P(\boldsymbol{Y}|X)$ 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 cone$\{\boldsymbol{H}\}$ 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 $\boldsymbol{H}$, there is no other unitary matrix which can cover cone$\{\boldsymbol{H}\}$ except permutation matrices. This shows that condition (2) of the sufficiently scattered assumption is also satisfied and the proof is completed. □

1.2. 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, $dual\{\mathcal{K}_2\} \subseteq dual\{\mathcal{K}_1\}$.*

Lemma 2. *If $\mathbf{A}$ is invertible, then $dual(\mathbf{A}) =$ 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 $P(\boldsymbol{Y}|X)$ and $\boldsymbol{T}$. To this end, let us denote $(\boldsymbol{T}_\star, h_{\boldsymbol{\theta}_\star})$ as a feasible solution of Criterion (5), i.e.,

$$\boldsymbol{T}_\star h_{\boldsymbol{\theta}_\star} = \boldsymbol{T}P(\boldsymbol{Y}|X) = P(\tilde{\boldsymbol{Y}}|X). \tag{7}$$

As defined in sufficient scattered assumption, we have the matrix $\boldsymbol{H} = [P(\boldsymbol{Y}|X = x_1), \dots, P(\boldsymbol{Y}|X = x_m)]$ defined on the set $\mathcal{H} = \{x_1, \dots, x_m\} \subseteq \mathcal{X}$. Let $\boldsymbol{H}_\star = [h_{\boldsymbol{\theta}_\star}(\boldsymbol{x}_1), \dots, h_{\boldsymbol{\theta}_\star}(\boldsymbol{x}_m)]$, it follows that

$$\boldsymbol{T}_\star \boldsymbol{H}_\star = \boldsymbol{T}\boldsymbol{H}. \tag{8}$$

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

$$\boldsymbol{T}_\star = \boldsymbol{T}\boldsymbol{A}^{-\top}, \tag{9}$$

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

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

$$\mathbf{1}^\top\boldsymbol{A}^{-\top} = \mathbf{1}^\top\boldsymbol{H}\boldsymbol{H}_\star^\dagger = \mathbf{1}^\top\boldsymbol{H}_\star^\dagger = \mathbf{1}^\top\boldsymbol{H}_\star\boldsymbol{H}_\star^\dagger = \mathbf{1}^\top. \tag{10}$$

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

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

$$\mathcal{R} \subseteq \text{cone}\{\boldsymbol{H}\} \subseteq \text{cone}(\boldsymbol{A}^{-\top}). \tag{11}$$

By applying Lemmas (1-2) to Eq. (11), we have

$$\text{cone}(\boldsymbol{A}) \subseteq dual\{\mathcal{R}\}, \tag{12}$$

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

$$dual\{\mathcal{R}\} = \{\boldsymbol{v} \in \mathbb{R}^C | \|\boldsymbol{v}\|_2 \leq \mathbf{1}^\top\boldsymbol{v}\}. \tag{13}$$

Then we have the following inequalities:

$$\begin{aligned} |det(\boldsymbol{A})| &\leq \prod_{i=1}^C \|\boldsymbol{A}_{:,i}\|_2 && \text{(14a)} \\ &\leq \prod_{i=1}^C \mathbf{1}^\top\boldsymbol{A}_{:,i} && \text{(14b)} \\ &\leq (\frac{\sum_{i=1}^C \mathbf{1}^\top\boldsymbol{A}_{:,i}}{C})^C && \text{(14c)} \\ &= (\frac{\mathbf{1}^\top\boldsymbol{A}\mathbf{1}}{C})^C = 1, && \text{(14d)} \end{aligned}$$

where (14a) is Hadamard's inequality; (14b) is by Eq. (12); (14c) is by the arithmetic-geometric mean inequality; and (14d) is by Eq. (10).

Note that $|det(\boldsymbol{A})|^{-1} = |det(\boldsymbol{A}^{-\top})|$ and $det(\boldsymbol{T}_\star) = det(\boldsymbol{T}\boldsymbol{A}^{-\top}) = det(\boldsymbol{T})|det(\boldsymbol{A})|^{-1}$ from properties of the determinant, it follows from Eq. (14) that $det(\boldsymbol{T}_\star) \geq det(\boldsymbol{T})$. We also know that $det(\boldsymbol{T}_\star) \leq det(\boldsymbol{T})$ must hold from Criterion (5), hence we have

$$det(\boldsymbol{T}_\star) = det(\boldsymbol{T}) \tag{15}$$

By Hadamard's inequality, the equality in (14a) holds only if $\boldsymbol{A}$ is column-orthogonal, which is equivalent to that $\boldsymbol{A}^{-\top}$ is column-orthogonal. Considering condition (2) in the definition of sufficiently scattered and the property of $\boldsymbol{A}^{-\top}$ that cone$\{\boldsymbol{H}\} \subseteq$ cone$(\boldsymbol{A}^{-\top})$, the only possible choices of column-orthogonal $\boldsymbol{A}^{-\top}$ are

$$\boldsymbol{A}^{-\top} = \boldsymbol{\Pi}\boldsymbol{\Phi} \tag{16}$$

where $\boldsymbol{\Pi} \in \mathbb{R}^{C\times C}$ is any permutation matrix and $\boldsymbol{\Phi} \in \mathbb{R}^{C\times C}$ is any diagonal matrix with non-zero diagonals. By Eq. (10), we must have $\Phi = \mathrm{I}$. Subsequently, we are left with $\boldsymbol{A}^{-\top} = \boldsymbol{\Pi}$, or equivalently, $\boldsymbol{T}_\star = \boldsymbol{\Pi}\boldsymbol{T}$. Since $\boldsymbol{T}$ and $\boldsymbol{T}_\star$ are both diagonal dominant, the only possible permutation matrix is $\boldsymbol{I}$, which means $\boldsymbol{T}_\star = \boldsymbol{T}$ holds. By Eq. (7), it follows that $h_{\boldsymbol{\theta}_\star} = P(\boldsymbol{Y}|X)$. Hence we conclude that $(\boldsymbol{T}_\star, h_{\boldsymbol{\theta}_\star}) = (\boldsymbol{T}, P(\boldsymbol{Y}|X))$ is the unique optimal solution to criterion (5). □

2. 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 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 1 (estimation error) and Table 1 (classification accuracy). The experimental performance shows that our proposed method outperforms the baseline methods.

References

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

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.

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.

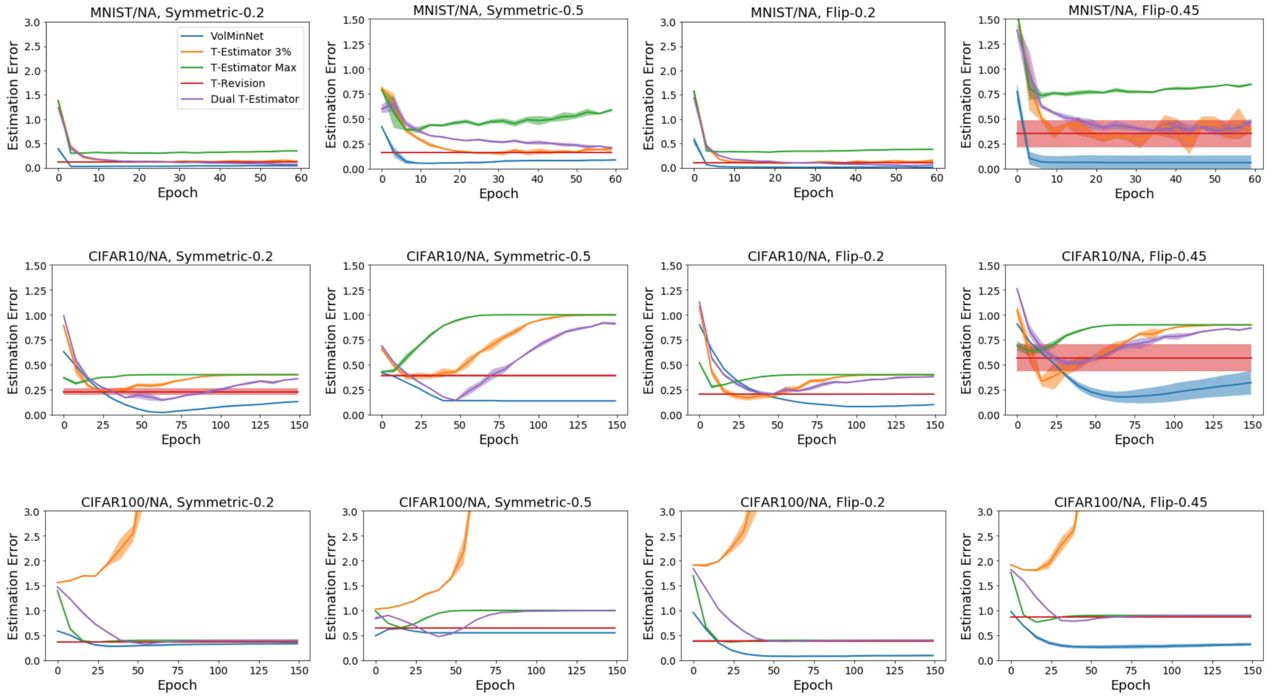

Figure 1. Transition matrix estimation error on MNIST/NA, CIFAR-10/NA, CIFAR-100/NA. Datasets with "/NA" means that possible anchor points are removed. The error bar for the standard deviation in each figure has been shaded. The lower the better.

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

Table 1. 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.