Supplementary Material for "Progressive Identification of True Labels for Partial-Label Learning"

Jiaqi Lv [† 1] **Miao Xu** [2 3] **Lei Feng** [4] **Gang Niu** [2] **Xin Geng** [1] **Masashi Sugiyama** [2 5]

A. Proof of Lemma 2

Cross-Entropy loss According to (Masnadi-Shirazi & Vasconcelos, 2009), since the ℓ_{CE} is non-negative, minimizing the conditional risk $\mathbb{E}_{p(y|x)}[\ell_{\mathrm{CE}}(\boldsymbol{g}(X), Y)|X], \forall X \in \mathcal{X}$ is an alternative of minimizing $\mathcal{R}(\boldsymbol{g})$. The conditional risk can be written as

$$\mathcal{C}(\boldsymbol{g}) = -\sum_{i=1}^{c} p(Y=i|X)\log(g_i(X)), \quad \text{s.t. } \sum_{i=1}^{c} g_i(X) = 1.$$

By the Lagrange Multiplier method (Bertsekas, 1997), we have

$$\mathcal{L} = -\sum_{i=1}^{c} p(Y=i|X)\log(g_i(X)) + \lambda(\sum_{i=1}^{c} g_i(X) - 1).$$

To minimize $\mathcal{L}$, we take the partial derivative of $\mathcal{L}$ with respect to g_i and set it be 0:

$$g_i^*(X) = \frac{1}{\lambda} p(Y=i|X).$$

Because $\sum_{i=1}^{c} g_i^*(X) = 1$ and $\sum_{i=1}^{c} g_i^*(X) = 1$, we have

$$\sum_{i=1}^{c} g_i^*(X) = \frac{1}{\lambda}\sum_{i=1}^{c} p(Y=i|X) = 1.$$

Therefore, we can obtain $\lambda = 1$ that ensures $g_i^*(X) = p(Y=i|X), \forall i \in [c], \forall X \in \mathcal{X}$, which concludes the proof.

Mean squared error loss Analogously, if the mean squared error loss is used, we can write the optimization problem as

$$\mathcal{C}(\boldsymbol{g}) = \sum_{i=1}^{c} (p(Y=i|X) - g_i(X))^2, \quad \text{s.t. } \sum_{i=1}^{c} g_i(X) = 1.$$

By the Lagrange Multiplier method, we have

$$\mathcal{L} = \sum_{i=1}^{c} (p(Y=i|X) - g_i(X))^2 - \lambda'(\sum_{i=1}^{c} g_i(X) - 1).$$

By setting the derivative to 0, we obtain

$$g_i^*(X) = \frac{\lambda'}{2} + p(Y=i|X).$$

[†]Preliminary work was done during an internship at RIKEN AIP. [1]School of Computer Science and Engineering, Southeast University, Nanjing, China [2]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan [3]The University of Queensland, Australia [4]School of Computer Science and Engineering, Nanyang Technological University, Singapore [5]University of Tokyo, Tokyo, Japan. Correspondence to: Xin Geng <xgeng@seu.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Because $\sum_{i=1}^{c} g_i^*(X) = 1$ and $\sum_{i=1}^{c} g_i^*(X) = 1$, we have

$$\sum_{i=1}^{c} g_i^*(X) = \frac{\lambda' c}{2} + \sum_{i=1}^{c} p(Y = i|X).$$

Since $c \neq 0$, we can obtain $\lambda' = 0$. In this way, $g_i^*(X) = p(Y = i|X), \forall i \in [c], \forall X \in \mathcal{X}$, which concludes the proof. $\square$

B. Proof of Theorem 1

First we prove $\boldsymbol{g}^*$ is the optimal classifier for PLL by substituting the $\boldsymbol{g}^*$ into the PLL risk estimator Eq. (5):

$$\begin{aligned}
\mathcal{R}_{\mathrm{PLL}}(\boldsymbol{g}^*) = \mathbb{E}_{(X,S)\sim p(x,s)}[\min_{i\in S} \ell(\boldsymbol{g}^*(X), \boldsymbol{e}^i)] &= \int \sum_{S\in\mathcal{S}} \min_{i\in S} \ell(\boldsymbol{g}^*(X), \boldsymbol{e}^i) p(s|x)p(x)dX \\
&= \int \sum_{S\in\mathcal{S}} \min_{i\in S} \ell(\boldsymbol{g}^*(X), \boldsymbol{e}^i) \sum_{Y\in\mathcal{Y}} p(s,y|x)p(x)dX \\
&= \int \sum_{Y\in\mathcal{Y}} \sum_{S\in\mathcal{S}} \min_{i\in S} \ell(\boldsymbol{g}^*(X), \boldsymbol{e}^i) p(s|x,y)p(y|x)p(x)dX \\
&= \int \sum_{Y\in\mathcal{Y}} \sum_{S\in\mathcal{S}} \ell(\boldsymbol{g}^*(X), \boldsymbol{e}^{Y_X}) p(s|x,y)p(y|x)p(x)dX \\
&= \int \sum_{Y\in\mathcal{Y}} \ell(\boldsymbol{g}^*(X), \boldsymbol{e}^{Y_X}) \sum_{S\in\mathcal{S}} p(s|x,y)p(y|x)p(x)dX \\
&= \int \sum_{Y\in\mathcal{Y}} \ell(\boldsymbol{g}^*(X), \boldsymbol{e}^{Y_X}) p(x,y)dX = \mathcal{R}(\boldsymbol{g}^*) = 0.
\end{aligned}$$

where we have used $\min_{i\in S} \ell(\boldsymbol{g}^*(X), \boldsymbol{e}^i) = \ell(\boldsymbol{g}^*(X), \boldsymbol{e}^{Y_X})$ because ℓ is a proper loss and the derministic assumption is made. This indicates that the PLL risk has been minimized by $\boldsymbol{g}^*$.

On the other hand, we prove $\boldsymbol{g}^*$ is the only solution to Eq. (5) by contradiction, namely, there is at least one other solution $\boldsymbol{h}$ enables $\mathcal{R}_{\mathrm{PLL}}(\boldsymbol{h}) = 0$, and predicts different label $Y^{\boldsymbol{h}} \neq Y_X$ for at least one instance X. Hence for any $S \ni Y_X$ we have

$$\min_{i\in S} \ell(\boldsymbol{h}(X), \boldsymbol{e}^i) = \ell(\boldsymbol{h}(X), \boldsymbol{e}^{Y^{\boldsymbol{h}}}) = 0.$$

Nevertheless, the above equality is always true unless $Y^{\boldsymbol{h}}$ is invariably included in the candidate label set of X, i.e., $\Pr_{S\sim p(s|x,y)}(Y^{\boldsymbol{h}} \in S) = 1$. Obviously, this contradicts the small ambiguity degree condition. Therefore, there is one, and only one minimizer of the PLL risk estimator, which is the same as the minimizer learned from ordinarily labeled data. The proof is complete. $\square$

C. Proof of Theorem 2

First, we show the uniform deviation bound, which is useful to derive the estimation error bound.

Lemma 3. *For any $\delta > 0$, we have with probability at least $1 - \delta$,*

$$\sup\nolimits_{\boldsymbol{g}\in\mathcal{G}} \left| \mathcal{R}_{\mathrm{PLL}}(\boldsymbol{g}) - \widehat{\mathcal{R}}_{\mathrm{PLL}}(\boldsymbol{g}) \right| \leq 2\mathfrak{R}_n(\ell_{\mathrm{PLL}} \circ \mathcal{G}) + M\sqrt{\frac{\log(2/\delta)}{2n}}$$

Proof. Consider the one-side uniform deviation $\sup_{\boldsymbol{g}\in\mathcal{G}} \mathcal{R}_{\mathrm{PLL}}(\boldsymbol{g}) - \widehat{\mathcal{R}}_{\mathrm{PLL}}(\boldsymbol{g})$. Since the loss function ℓ is upper-bounded by M, the change of it will be no more than M/n after replacing some x. Then, by *McDiarmid's inequality* (McDiarmid, 1989), for any $\delta > 0$, with probability at least $1 - \delta/2$, the following holds:

$$\sup\nolimits_{\boldsymbol{g}\in\mathcal{G}} \mathcal{R}_{\mathrm{PLL}}(\boldsymbol{g}) - \widehat{\mathcal{R}}_{\mathrm{PLL}}(\boldsymbol{g}) \leq \mathbb{E}\Big[\sup\nolimits_{\boldsymbol{g}\in\mathcal{G}} \mathcal{R}_{\mathrm{PLL}}(\boldsymbol{g}) - \widehat{\mathcal{R}}_{\mathrm{PLL}}(\boldsymbol{g})\Big] + M\sqrt{\frac{\log(2/\delta)}{2n}}.$$

By *symmetrization* (Vapnik, 1998), it is a routine work to show that

$$\mathbb{E}\Big[\sup_{\boldsymbol{g}\in\mathcal{G}} \mathcal{R}_{\mathrm{PLL}}(\boldsymbol{g}) - \widehat{\mathcal{R}}_{\mathrm{PLL}}(\boldsymbol{g})\Big] \leq 2\mathfrak{R}_n(\ell_{\mathrm{PLL}} \circ \mathcal{G}).$$

The one-side uniform deviation $\sup_{\boldsymbol{g}\in\mathcal{G}} \widehat{\mathcal{R}}_{\mathrm{PLL}}(\boldsymbol{g}) - \mathcal{R}_{\mathrm{PLL}}(\boldsymbol{g})$ can be bounded similarly. □

Then we upper bound $\mathfrak{R}_n(\ell_{\mathrm{PLL}} \circ \mathcal{G})$.

Lemma 4. *Suppose ℓ_{PLL} is defined as Eq. (4), it holds that*

$$\mathfrak{R}_n(\ell_{\mathrm{PLL}} \circ \mathcal{G}) \leq c\mathfrak{R}_n(\ell \circ \mathcal{G}) \leq \sqrt{2}cL_\ell \sum_{y=1}^{c} \mathfrak{R}_n(\mathcal{G}_y).$$

Proof. By definition of ℓ_{PLL}, $\ell_{\mathrm{PLL}} \circ \mathcal{G}(x_i, S_i) = \min_{y\in S} \ell \circ \mathcal{G}(x_i, y) = \min_{y\in[c]} \ell \circ \mathcal{G}(x_i, y)$. Given sample sized n, we first prove the result in the case $c = 2$. The min operator can be written as

$$\min\{z_1, z_2\} = \frac{1}{2}\big[z_1 + z_2 - |z_1 - z_2|\big].$$

In this way, we can write

$$\begin{aligned}
\mathfrak{R}_n(\ell_{\mathrm{PLL}} \circ \mathcal{G}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{g}\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(\boldsymbol{g}(x_i), s_i))\right] \qquad (11)\\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{g}\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \min\{\ell(\boldsymbol{g}(x_i), y_1), \ell(\boldsymbol{g}(x_i), y_2)\}\right]\\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{g}\in\mathcal{G}} \frac{1}{2n}\sum_{i=1}^{n} \sigma_i \Big[\ell(\boldsymbol{g}(x_i), y_1) + \ell(\boldsymbol{g}(x_i), y_2) - |\ell(\boldsymbol{g}(x_i), y_1) - \ell(\boldsymbol{g}(x_i), y_2)|\Big]\right]\\
&\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{g}\in\mathcal{G}} \frac{1}{2n}\sum_{i=1}^{n} \sigma_i \ell(\boldsymbol{g}(x_i), y_1)\right] + \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{g}\in\mathcal{G}} \frac{1}{2n}\sum_{i=1}^{n} \sigma_i \ell(\boldsymbol{g}(x_i), y_2)\right]\\
&+ \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{g}\in\mathcal{G}} \frac{1}{2n}\sum_{i=1}^{n} \sigma_i \Big|\ell(\boldsymbol{g}(x_i), y_1) - \ell(\boldsymbol{g}(x_i), y_2)\Big|\right]\\
&= \frac{1}{2}\Big(\mathfrak{R}_n(\ell \circ \mathcal{G}) + \mathfrak{R}_n(\ell \circ \mathcal{G})\Big) + \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{g}\in\mathcal{G}} \frac{1}{2n}\sum_{i=1}^{n} \sigma_i \Big|\ell(\boldsymbol{g}(x_i), y_1) - \ell(\boldsymbol{g}(x_i), y_2))\Big|\right].
\end{aligned}$$

Since $x \mapsto |x|$ is a 1-Lipschitz function, by *Talagrand's contraction lemma* (Ledoux & Talagrand, 2013), the last term can be bounded:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{g}\in\mathcal{G}} \frac{1}{2n}\sum_{i=1}^{n} \sigma_i \Big|\ell(\boldsymbol{g}(x_i), y_1) - \ell(\boldsymbol{g}(x_i), y_2))\Big|\right] \qquad (12)\\
\leq &\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{g}\in\mathcal{G}} \frac{1}{2n}\sum_{i=1}^{n} \sigma_i \Big(\ell(\boldsymbol{g}(x_i), y_1) - \ell(\boldsymbol{g}(x_i), y_2))\Big)\right] \leq \frac{1}{2}\Big(\mathfrak{R}_n(\ell \circ \mathcal{G}) + \mathfrak{R}_n(\ell \circ \mathcal{G})\Big).
\end{aligned}$$

Combining Eq. (11) and Eq. (12) yields $\mathfrak{R}_n(\ell_{\mathrm{PLL}} \circ \mathcal{G}) \leq \mathfrak{R}_n(\ell \circ \mathcal{G}) + \mathfrak{R}_n(\ell \circ \mathcal{G})$. The general case can be derived from the case $c = 2$ using $\min\{z_1, \ldots, z_c\} = \min\{z_1, \min\{z_2, \ldots, z_c\}\}$ and an immediate recurrence.

Then we apply the Rademacher vector contraction inequality (Maurer, 2016),

$$\mathfrak{R}_n(\ell \circ \mathcal{G}) \leq \sqrt{2}L_\ell \sum_{y=1}^{c} \mathfrak{R}_n(\mathcal{G}_y).$$

The proof is completed. □

Table 4. Summary of benchmark datasets and models.

Dataset	# Train	# Test	# Feature	# Class	Model $\boldsymbol{g}(x;\Theta)$
MNIST	60,000	10,000	784	10	Linear model, MLP (depth 5)
Fashion-MNIST	60,000	10,000	784	10	Linear model, MLP (depth 5)
Kuzushiji-MNIST	60,000	10,000	784	10	Linear model, MLP (depth 5)
CIFAR-10	50,000	10,000	3,072	10	ConvNet (Laine & Aila, 2017), ResNet (He et al., 2016)

Based on Lemma 3 and 4, the estimation error bound Eq. (7) is proven through

$$\begin{aligned}
\mathcal{R}_{\mathrm{PLL}}(\widehat{\boldsymbol{g}}_{\mathrm{PLL}}) - \mathcal{R}_{\mathrm{PLL}}(\boldsymbol{g}^*_{\mathrm{PLL}}) &= \Big(\mathcal{R}_{\mathrm{PLL}}(\widehat{\boldsymbol{g}}_{\mathrm{PLL}}) - \widehat{\mathcal{R}}_{\mathrm{PLL}}(\widehat{\boldsymbol{g}}_{\mathrm{PLL}})\Big) + \Big(\widehat{\mathcal{R}}_{\mathrm{PLL}}(\widehat{\boldsymbol{g}}_{\mathrm{PLL}}) - \widehat{\mathcal{R}}_{\mathrm{PLL}}(\boldsymbol{g}^*_{\mathrm{PLL}})\Big) \\
&\quad + \Big(\widehat{\mathcal{R}}_{\mathrm{PLL}}(\boldsymbol{g}^*_{\mathrm{PLL}}) - \mathcal{R}_{\mathrm{PLL}}(\boldsymbol{g}^*)\Big) \\
&\le \Big(\mathcal{R}_{\mathrm{PLL}}(\widehat{\boldsymbol{g}}_{\mathrm{PLL}}) - \widehat{\mathcal{R}}_{\mathrm{PLL}}(\widehat{\boldsymbol{g}}_{\mathrm{PLL}})\Big) + \Big(\widehat{\mathcal{R}}_{\mathrm{PLL}}(\boldsymbol{g}^*_{\mathrm{PLL}}) - \mathcal{R}_{\mathrm{PLL}}(\boldsymbol{g}^*_{\mathrm{PLL}})\Big) \\
&\le 2 \sup_{\boldsymbol{g}\in\mathcal{G}} \Big|\mathcal{R}_{\mathrm{PLL}}(\boldsymbol{g}) - \widehat{\mathcal{R}}_{\mathrm{PLL}}(\boldsymbol{g})\Big| \\
&\le 4\sqrt{2}cL_\ell \sum_{y=1}^{c} \mathfrak{R}_n(\mathcal{G}_y) + 2M\sqrt{\frac{\log(2/\delta)}{2n}}.
\end{aligned}$$

□

D. Supplementary Theorem on Section 4

Theorem 3. *The learning objective in Jin & Ghahramani (2003) is a special case of Eq. (8).*

Proof. Recall the learning objective in Jin & Ghahramani (2003) is formulated as:

$$\widehat{\mathcal{R}}_{\mathrm{PLL}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{KL}\big[\boldsymbol{z}_i || \boldsymbol{g}(x_i)\big] = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \log \frac{\boldsymbol{z}_i}{\boldsymbol{g}(x_i)}, \tag{13}$$

where KL divergence is used and $\boldsymbol{z}_i$ represents the prior probability of x_i.

Then in Eq. (8), the loss function can be specified as the cross-entropy loss: $\ell_{\mathrm{CE}}(g_j(x_i), e_j^{s_i}) = -e_j^{s_i}\log(g_j(x_i))$, which is linear in the second term, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} \mathrm{w}_{ij}\ell_{\mathrm{CE}}(g_j(x_i), e_j^{s_i}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} \ell_{\mathrm{CE}}(g_j(x_i), \mathrm{w}_{ij}e_j^{s_i}).$$

Thus, the weights $\mathbf{w}$ can be moved into the loss function and yields:

$$\widehat{\mathcal{R}}_{\mathrm{PLL}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} \ell_{\mathrm{CE}}(g_j(x_i), z_{ij}) = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \log \boldsymbol{g}(x_i), \tag{14}$$

where $z_{ij} = \mathrm{w}_{ij}\mathbb{I}(j \in s_i)$. The optimal $\boldsymbol{g}^*$ of Eq. (14) is essentially equivalent to $\boldsymbol{g}^*$ learned from Eq. (13). Therefore, our method is a strict extension of (Jin & Ghahramani, 2003). □

E. Benchmark Datasets

E.1. Setup

Tabel 4 describes the benchmark datasets and the corresponding models of them.

MNIST This is a grayscale image dataset of handwritten digits from 0 to 9 where the size of the images is 28×28.

The linear model is a linear-in-input model: d-10, and MLP refers to a 5-layer FC with ReLU as the activation function: d-300-300-300-300-10. Batch normalization (Ioffe & Szegedy, 2015) was applied before hidden layers. For both models, the softmax function was applied to the output layer, and ℓ_2-regularization was added. The two models were trained by SGD with the default momentum parameter ($\beta = 0.9$), and the batch size was set to 256.

Fashion-MNIST This is a grayscale image dataset similarly to MNIST. In Fashion-MNIST, each instance is a 28×28 grayscale image and associated with a label from 10 fashion item classes. The models and optimizer were the same as MNIST.

Kuzushiji-MNIST This is another grayscale image dataset similarly to MNIST. In Kuzushiji-MNIST, each instance is a 28×28 grayscale image and associated with a label from 10 cursive Japanese (Kuzushiji) characters. The models and optimizer were the same as MNIST.

CIFAR-10 This dataset consists of 60,000 $32 \times 32 \times 3$ colored image in RGB format in 10 classes.

The detailed architecture of ConvNet (Laine & Aila, 2017) is as follows.

0th (input) layer:	(32*32*3)-
1st to 4th layers:	[C(3*3, 128)]*3-Max Pooling-
5th to 8th layers:	[C(3*3, 256)]*3-Max Pooling-
9th to 11th layers:	C(3*3, 512)-C(3*3, 256)-C(3*3, 128)-
12th layers:	Average Pooling-10

where C(3*3, 128) means 128 channels of 3*3 convolutions followed by Leaky-ReLU (LReLU) active function (Maas et al., 2013), [·]*3 means 3 such layers, etc.

The detailed architecture of ResNet (He et al., 2016) was as follows.

0th (input) layer:	(32*32*3)-
1st to 11th layers:	C(3*3, 16)-[C(3*3, 16), C(3*3, 16)]*5-
12th to 21st layers:	[C(3*3, 32), C(3*3, 32)]*5-
22nd to 31st layers:	[C(3*3, 64), C(3*3, 64)]*5-
32nd layer:	Average Pooling-10

where [·, ·] means a building block (He et al., 2016). These two models were trained by SGD with the default momentum parameter and the batch size was 256.

An example of a binomial flipping with $q = 0.1$ and of a pair flipping with $q = 0.5$ used on MNIST are below, respectively:

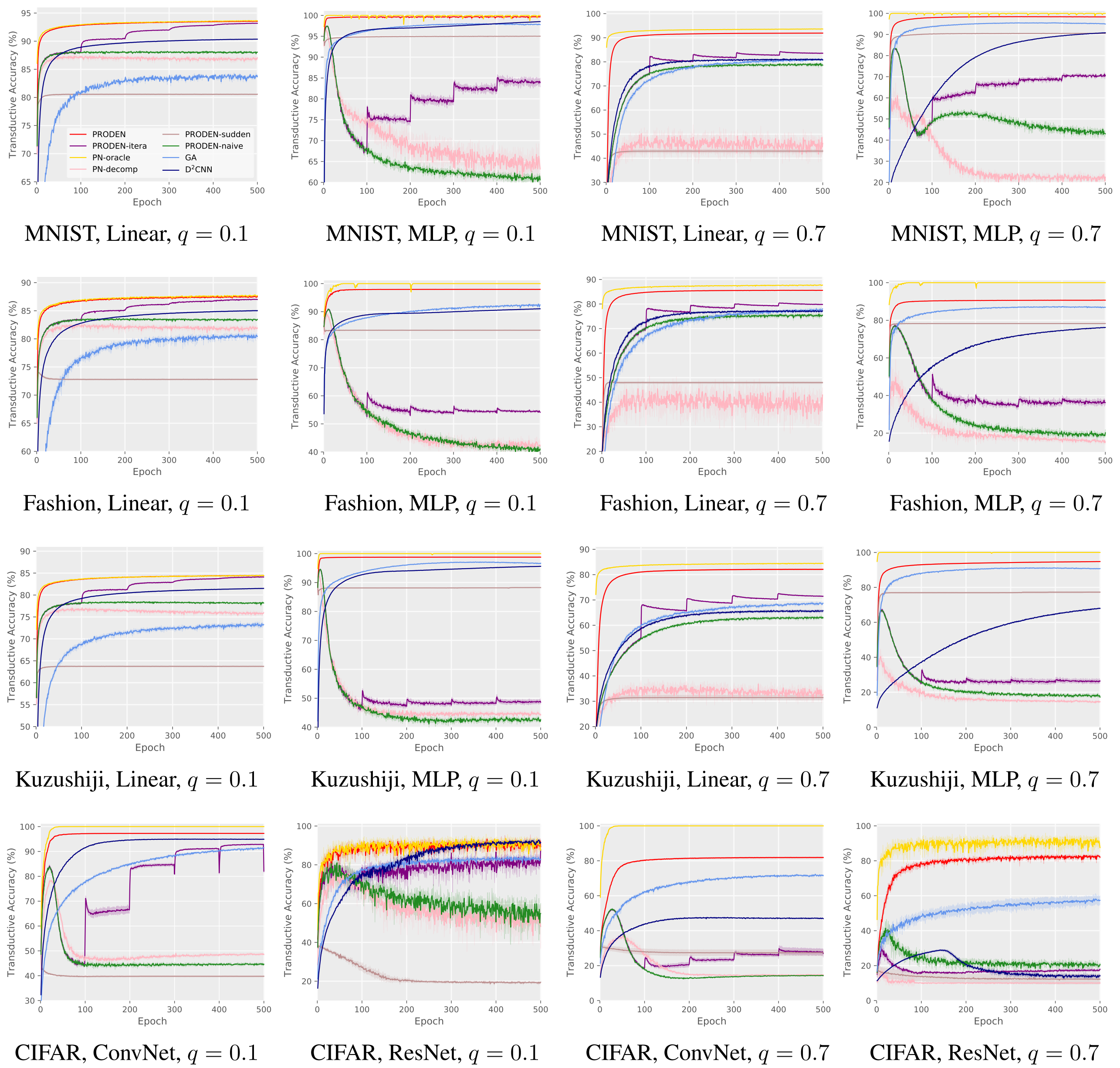
$$
\begin{bmatrix}
1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.1 & 1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.1 & 1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.1 & 1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 1 & 0.1 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 1 & 0.1 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 1 & 0.1 \\
0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 1
\end{bmatrix}
\quad
\begin{bmatrix}
1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0.5 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0.5 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.5 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.5 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.5 \\
0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
$$

E.2. Transductive Results

Figure 3 illustrates the transductive results on the benchmark datasets, i.e., the ability in identifying the true labels in the training set. We can see that PRODEN has a strong ability to find the true labels.

E.3. Test Results in the Pair Case

Figure 4 illustrates the test results on the benchmark datasets in the pair case. They show a similar phenomenon to Figure 2 that PRODEN is affected slightly and CCN is affected severely when the ambiguity degree goes large.



MNIST, Linear, $q = 0.1$ | MNIST, MLP, $q = 0.1$ | MNIST, Linear, $q = 0.7$ | MNIST, MLP, $q = 0.7$

Fashion, Linear, $q = 0.1$ | Fashion, MLP, $q = 0.1$ | Fashion, Linear, $q = 0.7$ | Fashion, MLP, $q = 0.7$

Kuzushiji, Linear, $q = 0.1$ | Kuzushiji, MLP, $q = 0.1$ | Kuzushiji, Linear, $q = 0.7$ | Kuzushiji, MLP, $q = 0.7$

CIFAR, ConvNet, $q = 0.1$ | CIFAR, ResNet, $q = 0.1$ | CIFAR, ConvNet, $q = 0.7$ | CIFAR, ResNet, $q = 0.7$

Figure 3. Transductive accuracy for various models and datasets. Dark colors show the mean accuracy of 5 trials and light colors show standard deviation. Fashion is short for Fashion-MNIST, Kuzushiji is short of Kuzushiji-MNIST, CIFAR is short of CIFAR-10.

F. UCI datasets

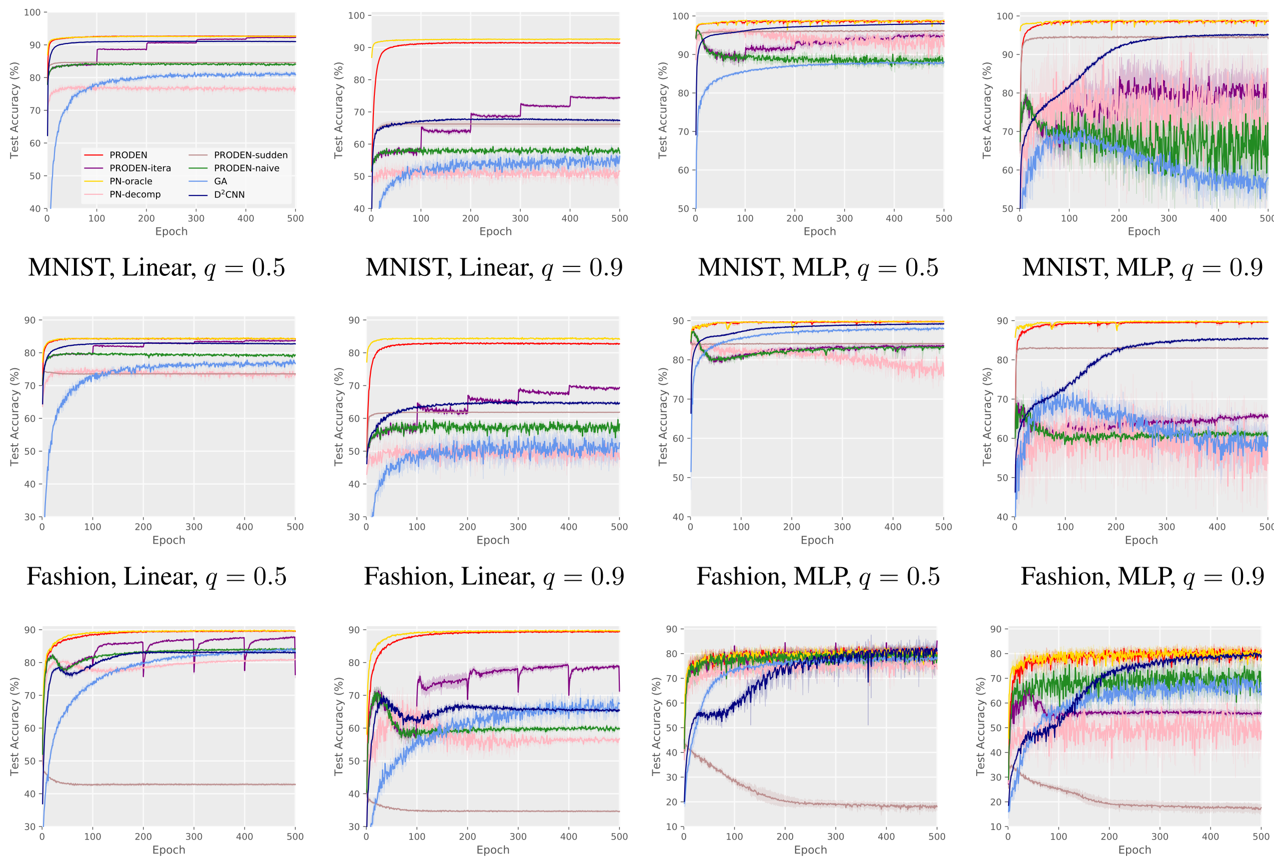
F.1. Characteristic of the UCI Datasets and Setup

Table 5 summaries the characteristic of the UCI datasets. We normalized these dataset by the Z-scores by convention and use the linear model trained by SGD with momentum 0.9.

F.2. Comparing Methods

The comparing PLL methods are listed as follows.

- *SURE* (Feng & An, 2019): an iterative EM-based method [suggested configuration: $\lambda, \beta \in$



MNIST, Linear, $q = 0.5$ — MNIST, Linear, $q = 0.9$ — MNIST, MLP, $q = 0.5$ — MNIST, MLP, $q = 0.9$

Fashion, Linear, $q = 0.5$ — Fashion, Linear, $q = 0.9$ — Fashion, MLP, $q = 0.5$ — Fashion, MLP, $q = 0.9$

CIFAR, ConvNet, $q = 0.5$ — CIFAR, ConvNet, $q = 0.9$ — CIFAR, ResNet, $q = 0.5$ — CIFAR, ResNet, $q = 0.9$

Figure 4. Test accuracy on MNIST, Fashion-MNIST, and CIFAR-10 in the pair case.

$\{0.001, 0.01, 0.05, 0.1, 0.3, 0.5, 1\}$].

- *CLPL* (Cour et al., 2011): a parametric method that transforms the PLL problem to the binary learning problem [suggested configuration: SVM with squared hinge loss].

- *ECOC* (Zhang et al., 2017): a disambiguation-free method that adapts the binary decomposition strategy to PLL [suggested configuration: $L = log_2(l)$].

- *PLSVM* (Nguyen & Caruana, 2008): a SVM-based method that differentiates candidate labels from non-candidate labels by maximizing the margin between them [suggested configuration: $\lambda \in \{10^{-3}, \ldots, 10^3\}$].

- *PLkNN* (Hullermeier & Beringer, 2006): a non-parametric approach that adapts k-nearest neighbors method to handle partially labeled data [suggested configuration: $k \in \{5, 6, \ldots, 10\}$].

- *IPAL* (Zhang & Yu, 2015): a non-parametric method that applies the label propagation strategy to iteratively update the weight of each candidate label [suggested configuration: $\alpha = 0.95, k = 10, T = 100$].

F.3. Results

Tabel 6 provides additional experiments to investigate the performances of each comparing methods on the UCI datasets with the pair flipping strategy. It shows that PRODEN generally achieves superior performance against other parametric comparing methods. Our advantage is a less obvious compared with the non-parametric method IPAL, whereas the performance of PRODEN could be easily increased by employing a deeper network.

Table 5. Summary of UCI datasets and models.

Dataset	# Train	# Test	# Feature	# Class	Model $g(x; \Theta)$
Yeast	1,335	149	8	10	Linear model
Texture	4,950	550	40	11	Linear model
Dermatology	329	37	34	6	Linear model
Synthetic-Control	540	60	60	6	Linear model
20Newsgroups	16,961	1,885	300	20	Linear model

Table 6. Test accuracy (mean±std) on the UCI datasets in the pair case.

	q	Yeast	Texture	Dermatology	Synthetic Control	20Newsgroups
PRODEN	0.5	**56.38±4.71%**	**99.71±0.12%**	**96.16±3.27%**	98.05±1.58%	78.05±0.97%
	0.9	**44.03±4.12%**	**99.35±0.31%**	93.68±6.80%	96.33±1.12%	**70.71±0.72%**
PRODEN-itera	0.5	56.26±4.74%	99.13±0.40%●	93.15±2.56%	87.40±7.25%●	76.90±0.92%
	0.9	42.62±5.52%	78.74±4.09%●	68.14±6.89%●	57.90±5.08%●	57.15±0.71%●
GA	0.5	24.39±4.40%●	94.25±0.48%●	73.81±6.18%●	64.68±3.08%●	63.68±0.67%●
	0.9	16.50±3.43%●	61.85±2.29%●	48.03±11.42%●	37.15±6.91%●	45.86±0.95%●
D^{2}CNN	0.5	56.38±4.71%	98.71±0.28%●	95.89±3.75%	78.87±11.94%●	74.38±0.90%●
	0.9	41.52±7.03%	86.45±4.87%●	87.84±6.58%	62.92±12.36%●	64.16±0.36%●
SURE	0.5	51.69±3.81%	98.18±0.17%●	95.71±2.49%	78.67±5.26%●	70.21±0.88%●
	0.9	37.61±3.40%●	98.00±0.42%●	93.42±6.23%	52.33±6.49%●	61.01±0.93%●
CLPL	0.5	55.06±4.74%	98.80±0.22%●	94.52±3.36%	75.83±4.29%●	77.92±0.76%
	0.9	40.66±4.69%	90.21±4.77%●	87.67±3.06%	52.33±4.65%●	65.03±0.32%●
ECOC	0.5	54.55±3.92%	99.47±0.17%●	94.71±2.29%	96.67±2.08%	**78.67±1.11%**
	0.9	42.51±5.19%	69.69±4.82%●	92.97±6.61%	94.50±1.32%	70.11±1.63%
PLSVM	0.5	52.59±2.04%	93.38±2.22%●	93.97±4.50%	91.83±3.08%●	76.21±2.30%
	0.9	41.89±3.90%	82.24±6.58%●	93.15±4.22%	80.67±9.42%●	70.76±2.16%
PLkNN	0.5	53.60±2.81%	97.11±0.28%●	94.52±3.06%	94.83±2.60%●	43.77±0.72%●
	0.9	43.80±4.50%	92.98±0.64%●	92.05±4.38%	89.83±4.54%●	38.77±0.90%●
IPAL	0.5	51.80±5.08%	99.30±0.37%	95.89±2.74%	**98.50±1.37%**	76.83±0.51%
	0.9	36.60±2.58%●	98.95±0.58%●	**95.34±2.29%**○	**98.50±1.09%**○	69.15±0.73%

G. Characteristic of the Real-world Datasets and Setup

Tabel 7 summarizes the characteristic of the real-world datasets and the corresponding models. The preprocessing, model and optimizer were the same as UCI datasets.

Table 7. Summary of real-world partial label datasets.

Dataset	# Examples	# Feature	# Class	# Avg. CLs	Task Domain	Model $\boldsymbol{g}(x;\Theta)$
Lost	1122	108	16	2.23	automatic face naming (Panis & Lanitis, 2014)	Linear model
BirdSong	4998	38	13	2.18	bird song classification (Briggs et al., 2012)	Linear model
MSRCv2	1758	48	23	3.16	object classification (Liu & Dietterich, 2012)	Linear model
Soccer Player	17472	279	171	2.09	automatic face naming (Zeng et al., 2013)	Linear model
Yahoo! News	22991	163	219	1.91	automatic face naming (Guillaumin et al., 2010)	Linear model

References

Bertsekas, D. P. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

Briggs, F., Fern, X. Z., and Raich, R. Rank-loss support instance machines for miml instance annotation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*, pp. 534–542, Beijing, China, 2012.

Cour, T., Sapp, B., and Taskar, B. Learning from partial labels. *Journal of Machine Learning Research*, 12(5):1501–1536, 2011.

Feng, L. and An, B. Partial label learning with self-guided retraining. In *Proceedings of 33rd AAAI Conference on Artificial Intelligence (AAAI'19)*, pp. 3542–3549, Honolulu, HI, 2019.

Guillaumin, M., Verbeek, J., and Schmid, C. Multiple instance metric learning from automatically labeled bags of faces. *Lecture Notes in Computer Science*, 63(11):634–647, 2010.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the 29th IEEE conference on Computer Vision and Pattern Recognition (CVPR'16)*, pp. 770–778, Las Vegas, NV, 2016.

Hullermeier, E. and Beringer, J. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5):419–439, 2006.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML'15)*, pp. 448–456, Lille, France, 2015.

Jin, R. and Ghahramani, Z. Learning with multiple labels. In *Advances in Neural Information Processing Systems 16 (NIPS'03)*, pp. 921–928, Vancouver, Canada, 2003.

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. In *Proceedings of 5th International Conference on Learning Representations (ICLR'17)*, Toulon, France, 2017.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, 2013.

Liu, L. and Dietterich, T. G. A conditional multinomial mixture model for superset label learning. In *Advances in Neural Information Processing Systems 25 (NIPS'12)*, pp. 548–556, Lake Tahoe, NV, 2012.

Maas, A. L., Hannun, A. Y., and Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of 30th International Conference on Machine Learning (ICML'13)*, volume 30, pp. 3, Atlanta, GA, 2013.

Masnadi-Shirazi, H. and Vasconcelos, N. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in Neural Information Processing Systems 22 (NIPS'09)*, pp. 1049–1056, Vancouver, Canada, 2009.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory (ALT'16)*, pp. 3–17, 2016.

McDiarmid, C. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Nguyen, N. and Caruana, R. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*, pp. 381–389, Las Vegas, NV, 2008.

Panis, G. and Lanitis, A. An overview of research activities in facial age estimation using the fg-net aging database. In *Proceedings of the 13th European Conference on Computer Vision (ECCV'14)*, pp. 737–750, Zurich, Switzerland, 2014.

Vapnik, V. N. Statistical learning theory. *John Wiley & Sons*, 1998.

Zeng, Z., Xiao, S., Jia, K., Chan, T., Gao, S., Xu, D., and Ma, Y. Learning by associating ambiguously labeled images. In *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'13)*, pp. 708–715, Portland, OR, 2013.

Zhang, M. and Yu, F. Solving the partial label learning problem: An instance-based approach. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI'15)*, pp. 4048–4054, Buenos Aires, Argentina, 2015.

Zhang, M., Yu, F., and Tang, C. Disambiguation-free partial label learning. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2155–2167, 2017.