Partial Label Clustering
-Appendix-

A Evaluation Metrics

We use Average Clustering Accuracy (ACC) and Normalized Mutual Information (NMI) metrics, both of which are widely used criteria in the field of clustering. ACC discovers the one-to-one relationship between clusters and classes. Denote c_i as the clustering result of sample $\boldsymbol{x}_i$ and g_i as the ground-truth label of sample $\boldsymbol{x}_i$, ACC is defined as

$$ACC = \frac{1}{n}\sum_{i=1}^{n}\delta(c_i, map(g_i)), \tag{S1}$$

where $\delta(p,q) = 1$, if $p = q$ and $\delta(p,q) = 0$ otherwise. n is the total number of examples and $map(g_i)$ is the mapping function that permutes the clusters to match the ground-truth labels. NMI measures the mutual information entropy between the clusters and the ground-truth labels. Given the ground-truth labels $\mathbf{Y}$ and the clustering results $\mathbf{C}$, NMI is defined as

$$NMI = \frac{\sum_{y\in\mathbf{Y},c\in\mathbf{C}} p(y,c)log(\frac{p(y,c)}{p(y)p(c)})}{\sqrt{\sum_{y\in\mathbf{Y}} p(y)\log p(y)\sum_{c\in\mathbf{C}} p(c)\log p(c)}}, \tag{S2}$$

where $p(y)$ and $p(c)$ represent the marginal probability distribution functions of $\mathbf{Y}$ and $\mathbf{C}$ respectively, and $p(y,c)$ is the joint distribution.

B Proof of Theorem 1

We first give a lemma as follows.

Lemma 1.

$$\mathrm{Tr}(\boldsymbol{AB}) \leq ||\boldsymbol{A}||_F||\boldsymbol{B}||_F. \tag{S3}$$

Proof. By Cauchy-Schwarz, we have

$$\begin{aligned}\mathrm{Tr}(\mathbf{AB}) &= \sum_{i,j} a_{ij}b_{ji} \\ &\leq \sum_i(\sum_j |a_{ij}|^2)^{1/2}(\sum_j |b_{ji}|^2)^{1/2} \\ &\leq (\sum_{i,j} |a_{ij}|^2)^{1/2}(\sum_{i,j} |b_{ji}|^2)^{1/2} \\ &= ||\mathbf{A}||_F||\mathbf{B}||_F.\end{aligned} \tag{S4}$$

This concludes the proof. □

Now we begin the proof of Theorem 1.

Proof. Denote $\mathbf{F} \in [0,1]^{n\times q}$ and $\mathbf{W} \in [0,1]^{n\times n}$ the label confidence matrix and the weight matrix to be optimized. For the convenience of explanation, the terms related to $\mathbf{F}$ in the objective function Eq. (2) can be rewritten as

$$\begin{aligned}&\min_{\mathbf{W}} ||\mathbf{F} - \mathbf{WF}||_F^2 \\ &\text{s.t.} \quad w_{ij} = 0 \text{ if } (\boldsymbol{x}_i, \boldsymbol{x}_j) \notin \mathcal{E}, \\ &\qquad \mathbf{W}^\top\mathbf{1}_n = \mathbf{1}_n, \mathbf{0}_{n\times n} \leq \mathbf{W} \leq \mathbf{N}.\end{aligned} \tag{S5}$$

Let $\mathbf{F}_G$ and $\mathbf{W}_G$ be the ground-truth label matrix and the optimal weight matrix under the ground-truth labels. We assume that $\mathbf{W}_G$ is constructed on the premise that the ground-truth labels of neighboring examples are the same, which can improve clustering performance. Due to the constraint of $\mathbf{W}_G^\top\mathbf{1}_n = \mathbf{1}_n$, we have $||\mathbf{F}_G - \mathbf{W}_G\mathbf{F}_G||_F^2 = 0$. Denote $\Delta_\mathbf{W} = \mathbf{W}_G - \mathbf{W}$, the following inequality holds

$$||\mathbf{F}_G - (\Delta_\mathbf{W} + \mathbf{W})\mathbf{F}_G||_F^2 \leq ||\mathbf{F} - \mathbf{WF}||_F^2. \tag{S6}$$

Expand Eq. (S6), we have

$$\begin{aligned}&||\Delta_\mathbf{W}\mathbf{F}_G||_F^2 \\ &\leq ||\mathbf{F}||_F^2 + \mathrm{Tr}(\mathbf{W}^\top\mathbf{W}(\mathbf{F}^\top\mathbf{F} - \mathbf{F}_G^\top\mathbf{F}_G)) \\ &\quad + \mathrm{Tr}((\mathbf{W} + \mathbf{W}^\top)(\mathbf{F}_G^\top\mathbf{F}_G - \mathbf{F}^\top\mathbf{F})) - ||\mathbf{F}_G||_F^2 \\ &\quad + \mathrm{Tr}(\mathbf{F}_G^\top\mathbf{F}_G((\mathbf{I} - \mathbf{W})^\top\Delta_\mathbf{W} + (\mathbf{I} - \mathbf{W})\Delta_\mathbf{W}^\top)).\end{aligned} \tag{S7}$$

According to Lemma 1 and the fact that the Frobenius norm is submultiplicative, we have

$$\begin{aligned}&||\Delta_\mathbf{W}\mathbf{F}_G||_F^2 \\ &\leq ||\mathbf{F}||_F^2 - ||\mathbf{F}_G||_F^2 + 2||\mathbf{F}_G||_F^2||\mathbf{I} - \mathbf{W}||_F||\Delta_\mathbf{W}||_F \\ &\quad (||\mathbf{W}||_F^2 + 2||\mathbf{W}||_F)||\mathbf{F}^\top\mathbf{F} - \mathbf{F}_G^\top\mathbf{F}_G||_F.\end{aligned} \tag{S8}$$

Since $\mathbf{W}$ is upper bounded by the number of samples n, we have $||\mathbf{W}||_F^2 \leq n^2$ and $||\mathbf{I} - \mathbf{W}||_F^2 \leq n^2$. Due to $\mathbf{F}_G$ is the ground-truth label matrix, we have $||\mathbf{F}_G||_F^2 = n$. Furthermore, $\mathbf{F}$ is upper bounded by the number of samples n and the number of classes q, i.e., $||\mathbf{F}||_F^2 \leq nq$. Assume that $||\Delta_\mathbf{W}||_F \geq 1$, which is a reasonable assumption when n is large enough. We have

$$\begin{aligned}||\Delta_\mathbf{W}\mathbf{F}_G||_F^2 \leq &(n^2 + 2n)||\mathbf{F}^\top\mathbf{F} - \mathbf{F}_G^\top\mathbf{F}_G||_F||\Delta_\mathbf{W}||_F \\ &+ (2n^2 + nq - n)||\Delta_\mathbf{W}||_F.\end{aligned} \tag{S9}$$

Controlled UCI Datasets				
Dataset	**# Examples**	**# Features**	**# Class Labels**	**# False Positive Labels (r)**
Ecoli	336	7	8	$r = 1, 2, 3$
Vehicle	846	18	4	$r = 1, 2$
Coil20	1440	1024	20	$r = 1, 2, 3$

Real-World Datasets					
Dataset	**# Examples**	**# Features**	**# Class Labels**	**Avg.# CLs**	**Task Domain**
Lost	1122	108	16	2.23	automatic face naming
MSRCv2	1758	48	23	3.16	object classification
Mirflickr	2780	1536	14	2.76	web image classification
BirdSong	4998	38	13	2.18	bird song classification
LYN10	16526	163	10	1.84	automatic face naming

Table S1: Characteristics of controlled UCI datasets and real-world datasets.

Compared Method	LYN10	
	$\rho = 0.01$	$\rho = 0.02$
PLC (Ours)	$\mathbf{0.525 \pm 0.024}$	$\mathbf{0.556 \pm 0.010}$
DPCLS	0.485 ± 0.016	0.523 ± 0.020
AGGD	0.493 ± 0.016	0.538 ± 0.019
IPAL	0.468 ± 0.015	0.522 ± 0.013
PL-kNN	0.394 ± 0.025	0.426 ± 0.016
PL-SVM	0.485 ± 0.038	0.542 ± 0.016
PARM	$\underline{0.497 \pm 0.010}$	$\underline{0.550 \pm 0.018}$
SSPL	0.445 ± 0.041	0.482 ± 0.021

Table S2: ACCs when compared with PLL and semi-supervised PLL methods on large-scale dataset, where bold and underlined indicate the best and second best results respectively.

Note that $\mathbf{F}_G^\top \mathbf{F}_G$ is a positive semidefinite matrix, thus its eigenvalues are non-negative. Taking λ as the smallest eigenvalue of $\mathbf{F}_G^\top \mathbf{F}_G$, we have $\lambda ||\Delta_\mathbf{W}||_F^2 \leq ||\Delta_\mathbf{W} \mathbf{F}_G||_F^2$. Thus, Eq. (S9) can be further relaxed as

$$\begin{aligned}\lambda ||\Delta_\mathbf{W}||_F^2 \leq & (n^2 + 2n) ||\mathbf{F}^\top \mathbf{F} - \mathbf{F}_G^\top \mathbf{F}_G||_F ||\Delta_\mathbf{W}||_F \\ & + (2n^2 + nq - n) ||\Delta_\mathbf{W}||_F.\end{aligned} \tag{S10}$$

Let $||\overline{\Delta}_\mathbf{W}||_F$ be the average distance of each corresponding position between $\mathbf{W}_G$ and $\mathbf{W}$, i.e., $||\overline{\Delta}_\mathbf{W}||_F = \frac{1}{n^2} ||\mathbf{W}_G - \mathbf{W}||_F$. Dividing n^2 on both sides of Eq. (S10), we finally have

$$||\overline{\Delta}_\mathbf{W}||_F \leq \frac{n+2}{\lambda n} ||\mathbf{F}^\top \mathbf{F} - \mathbf{F}_G^\top \mathbf{F}_G||_F + \frac{2n + q - 1}{\lambda n}. \tag{S11}$$

This concludes the proof of Theorem 1. □

C Complexity Analysis

The computational complexity of our algorithm is dominated by steps 7-11. Before alternative optimization, we use the KD-Tree method to find the k nearest neighbors for each sample in the dataset with the complexity of $\mathcal{O}(kn \log n)$. In steps 7-9, we use interior point method [Ye and Tse, 1989] to solve a series of QP problems with the complexity of $\mathcal{O}(nk^3)$. Similarly, step 10 solves a QP problem with the complexity of $\mathcal{O}(n^3 q^3)$. When dealing with large datasets, we can transform the original problem into a series of subproblems as Eq. (9) with the complexity of $\mathcal{O}(nq^3)$. Step 11 solves the problem by KKT conditions with the complexity of $\mathcal{O}(n^3)$. In summary, the overall complexity of our algorithm in each iteration is $\mathcal{O}(kn \log n + nk^3 + n^3 q^3 + n^3)$ and $\mathcal{O}(kn \log n + nk^3 + nq^3 + n^3)$ for large datasets.

D Details of Compared Datasets

Table S1 summarizes the characteristics of controlled UCI datasets and real-world datasets. Following the widely-used partial label data generation protocol [Cour *et al.*, 2011], we generate artificial partial label datasets under the parameter r which controls the number of false-positive labels[1].

The real-world datasets are collected from various domains including Lost [Cour *et al.*, 2009] for automatic face naming, MSRCv2 [Liu and Dietterich, 2014] for object classification, Mirflickr [Huiskes and Lew, 2008] for web image classification and BirdSong [Raich, 2012] for bird song classification.

E Supplementary Experimental Results

E.1 More Comparison to Constrained Clustering

Fig. S1 illustrates the ACCs and NMIs of our PLC method compared to the constrained clustering methods on the datasets Ecoli $r = 3$ and Coil20 $r = 3$. Our PLC method ranks first in 87.5% (21/24) cases which further proves the effectiveness of our PLC method.

E.2 More Comparison to PLL & Semi-supervised PLL

Fig. S2 illustrates the ACCs of our PLC method compared to the PLL and semi-supervised PLL methods on synthetic UCI datasets. Our PLC method achieves superior or competitive performance against the comparing methods with a lower proportion of partial labeled samples. According to Fig. S2, PLC method achieves superior performance against PL-KNN and PL-SVM in 100% (32/32) cases, against IPAL, DP-CLS, PARM and SSPL in 96.88% (31/32) cases, and against

[1]For Vehicle, the setting $r = 3$ is not considered as there are only four class labels in the label space

Compared Method	LYN10					
	$\rho = 0.05$	$\rho = 0.10$	$\rho = 0.15$	$\rho = 0.20$	$\rho = 0.30$	$\rho = 0.40$
PLC (Ours)	$\mathbf{0.595 \pm 0.007}$	$\mathbf{0.624 \pm 0.008}$	$\mathbf{0.643 \pm 0.011}$	$\mathbf{0.659 \pm 0.007}$	$\mathbf{0.670 \pm 0.010}$	$\mathbf{0.675 \pm 0.009}$
K-means	0.372 ± 0.033	0.366 ± 0.022	0.366 ± 0.023	0.371 ± 0.019	0.366 ± 0.031	0.361 ± 0.018
SC	0.301 ± 0.005	0.299 ± 0.007	0.304 ± 0.008	0.297 ± 0.006	0.307 ± 0.008	0.302 ± 0.010
SSC-TLRR	$\underline{0.379 \pm 0.007}$	$\underline{0.415 \pm 0.010}$	$\underline{0.467 \pm 0.006}$	$\underline{0.392 \pm 0.024}$	$\underline{0.382 \pm 0.012}$	$\underline{0.423 \pm 0.008}$
DP-GLPCA	0.278 ± 0.011	0.284 ± 0.009	0.287 ± 0.009	0.287 ± 0.008	0.292 ± 0.013	0.299 ± 0.012
SSSC	0.340 ± 0.011	0.346 ± 0.009	0.362 ± 0.010	0.358 ± 0.011	0.351 ± 0.019	0.332 ± 0.011

Table S3: Experimental results on ACCs when compared with constrained clustering methods on large-scale datasets, where bold and underlined indicate the best and second best results respectively.

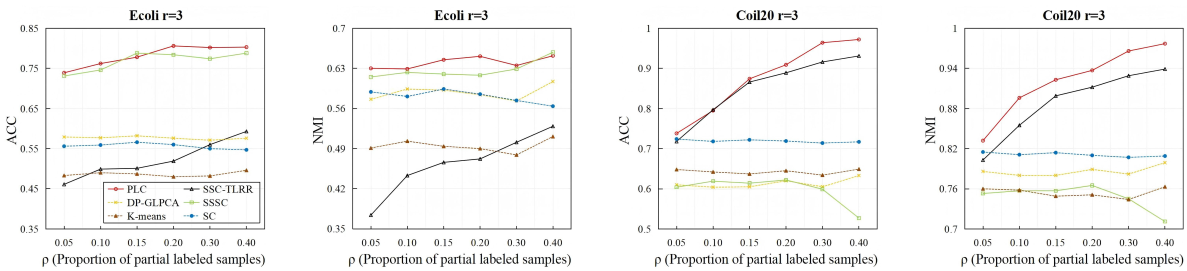

Figure S1: ACCs and NMIs when compared with constrained clustering methods under different proportions of partial label training samples on the datasets Ecoli $r = 3$ and Coil20 $r = 3$.

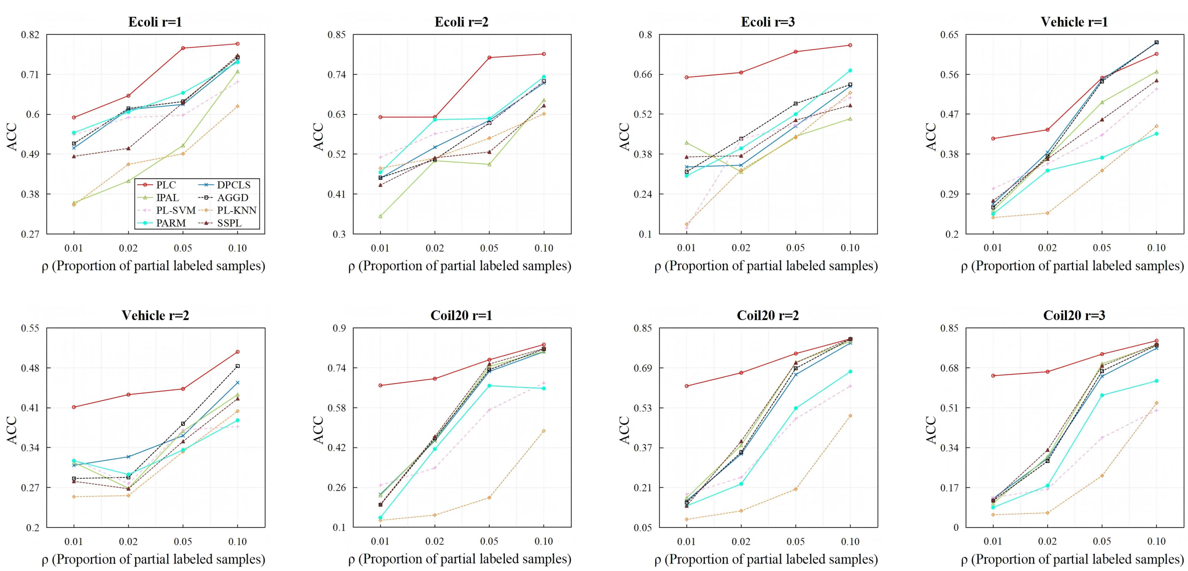

Figure S2: ACCs when compared with PLL and semi-supervised PLL methods under different proportions of partial label training samples on synthetic UCI datasets.

	DPCLS	AGGD	IPAL	PL-SVM	PL-KNN	SSPL	PARM	SSC-TLRR	DP-GLPCA	SSSC
(i)	11/6/1	11/7/1	14/4/0	17/1/0	18/0/0	16/2/0	17/1/0	30/0/0	30/0/0	30/0/0
(ii)	28/4/0	25/6/1	29/3/0	32/0/0	32/0/0	27/5/0	32/0/0	31/17/0	48/0/0	35/13/0
Total	39/10/1	36/13/1	43/7/0	49/1/0	50/0/0	43/7/0	49/1/0	61/17/0	78/0/0	65/13/0

Table S4: Win/tie/loss counts on the classification performance of PLC against the PLL, semi-supervised PLL and constrained clustering methods on all datasets. (i), (ii) indicate the summaries on real-world datasets and synthetic UCI datasets. "Total" denotes the summary on all the datasets.

AGGD in 93.75% (30/32) cases. The experimental results further prove that our PLC method performs well in the case of fewer partial label training samples.

E.3 Experiment on Large-scale Dataset

LYN10 is a large-scale dataset for automatic face naming, consisting of samples from the top 10 classes of the Yahoo!News [Guillaumin *et al.*, 2010] dataset. The characteristics of LYN10 are shown in Table S2. Table S2 reports the ACCs of our PLC method compared to PLL and semi-supervised PLL methods on the large-scale dataset. Table S3 reports the ACCs of our PLC method compared to constrained clustering methods on the large-scale dataset. Our PLC method performs well on the large-scale dataset and ranks first in all experimental settings.

E.4 Significance Analysis

Table S4 reports win/tie/loss counts between our PLC methods and ten comparing methods on the real-world datasets and synthetic UCI datasets according to the pairwise t-test at the significance level of 0.05. We can find that our PLC method statistically outperforms the PLL and semi-supervised PLL methods (the first seven columns) in 88.3% (309/350) cases and statistically outperforms the constrained clustering methods (the last three columns) in 87.2% (204/234) cases.

References

[Cour *et al.*, 2009] Timothee Cour, Benjamin Sapp, Chris Jordan, and Ben Taskar. Learning from ambiguously labeled images. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 919–926, 2009.

[Cour *et al.*, 2011] Timothee Cour, Benjamin Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12(4):1501–1536, 2011.

[Guillaumin *et al.*, 2010] Matthieu Guillaumin, Jakob J. Verbeek, and Cordelia Schmid. Multiple instance metric learning from automatically labeled bags of faces. In *European Conference on Computer Vision*, 2010.

[Huiskes and Lew, 2008] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. *ACM*, page 39, 2008.

[Liu and Dietterich, 2014] L. P. Liu and T. G. Dietterich. A conditional multinomial mixture model for superset label learning. *Advances in Neural Information Processing Systems*, 1:548–556, 2014.

[Raich, 2012] Fern Raviv Raich. Rank-loss support instance machines for miml instance annotation. *SIGKDD explorations*, 2012(CDaROM), 2012.

[Ye and Tse, 1989] Yinyu Ye and Edison T. S. Tse. An extension of karmarkar's projective algorithm for convex quadratic programming. *Mathematical Programming*, 44:157–179, 1989.