Candidate Label Set Pruning: A Data-centric Perspective for Deep Partial-label Learning

Shuo He[1] Chaojie Wang[2] Guowu Yang[1] Lei Feng[2]

[1]University of Electronic Science and Technology of China
[2]Nanyang Technological University

Email: shuohe123@gmail.com

What is deep partial label learning?

Using partially label samples to train a deep neural network

Partially labeled samples have a set of candidate labels

Image example

Candidate Label Set

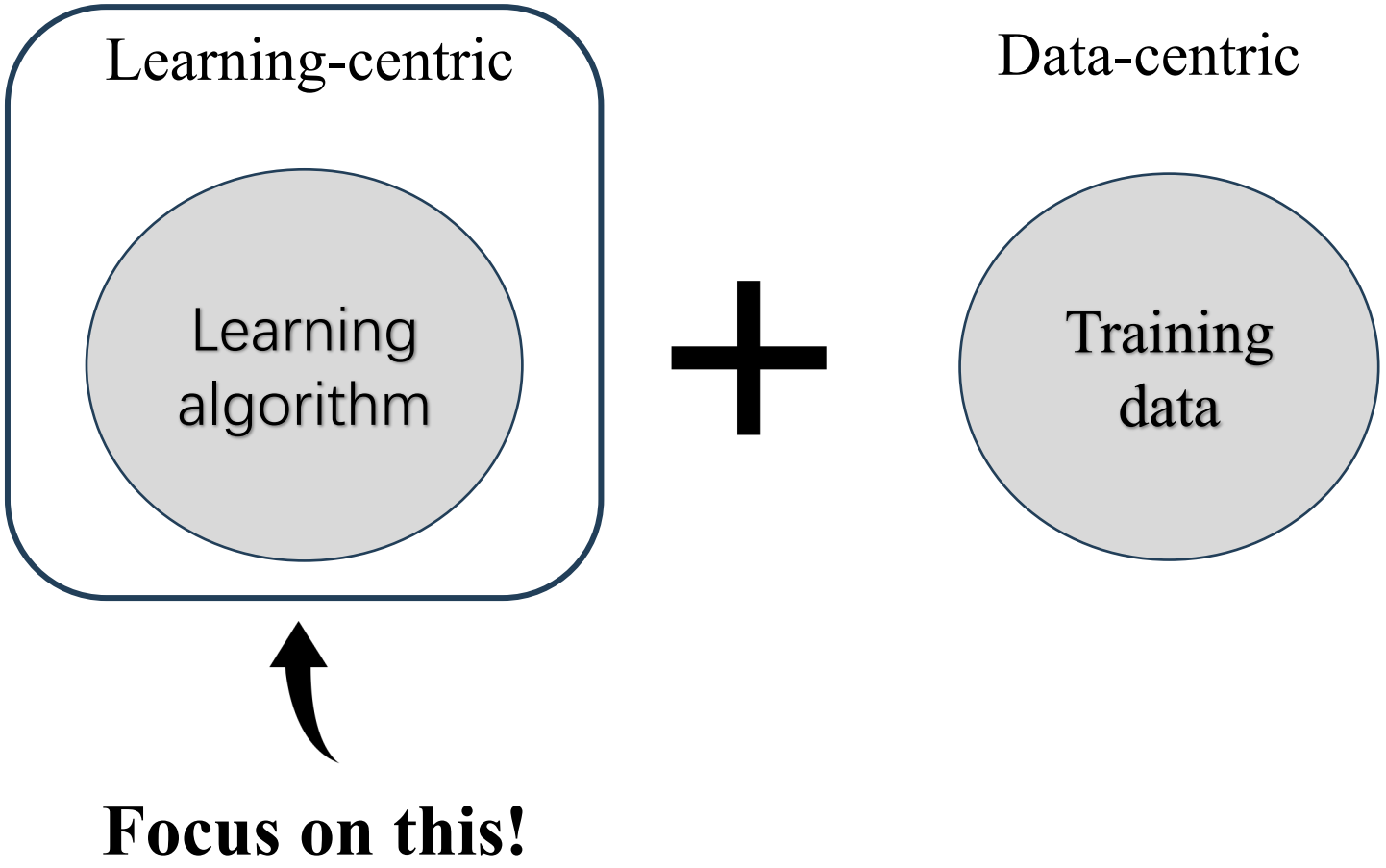
African lion (True label) ✓

Buffalo (False label) ✗

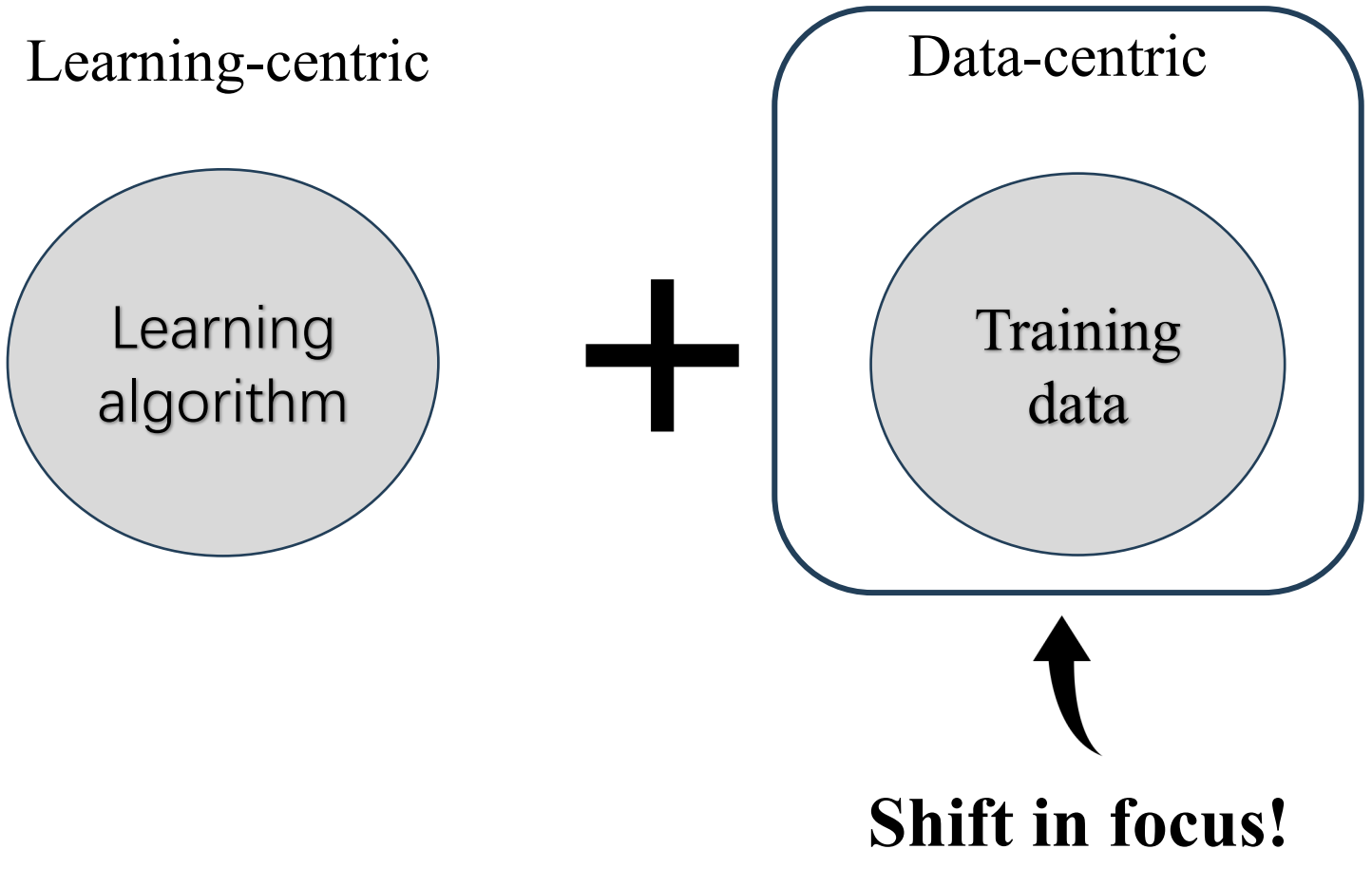
Antelope (False label) ✗

Zebra (False label) ✗

Existing research on deep partial label learning

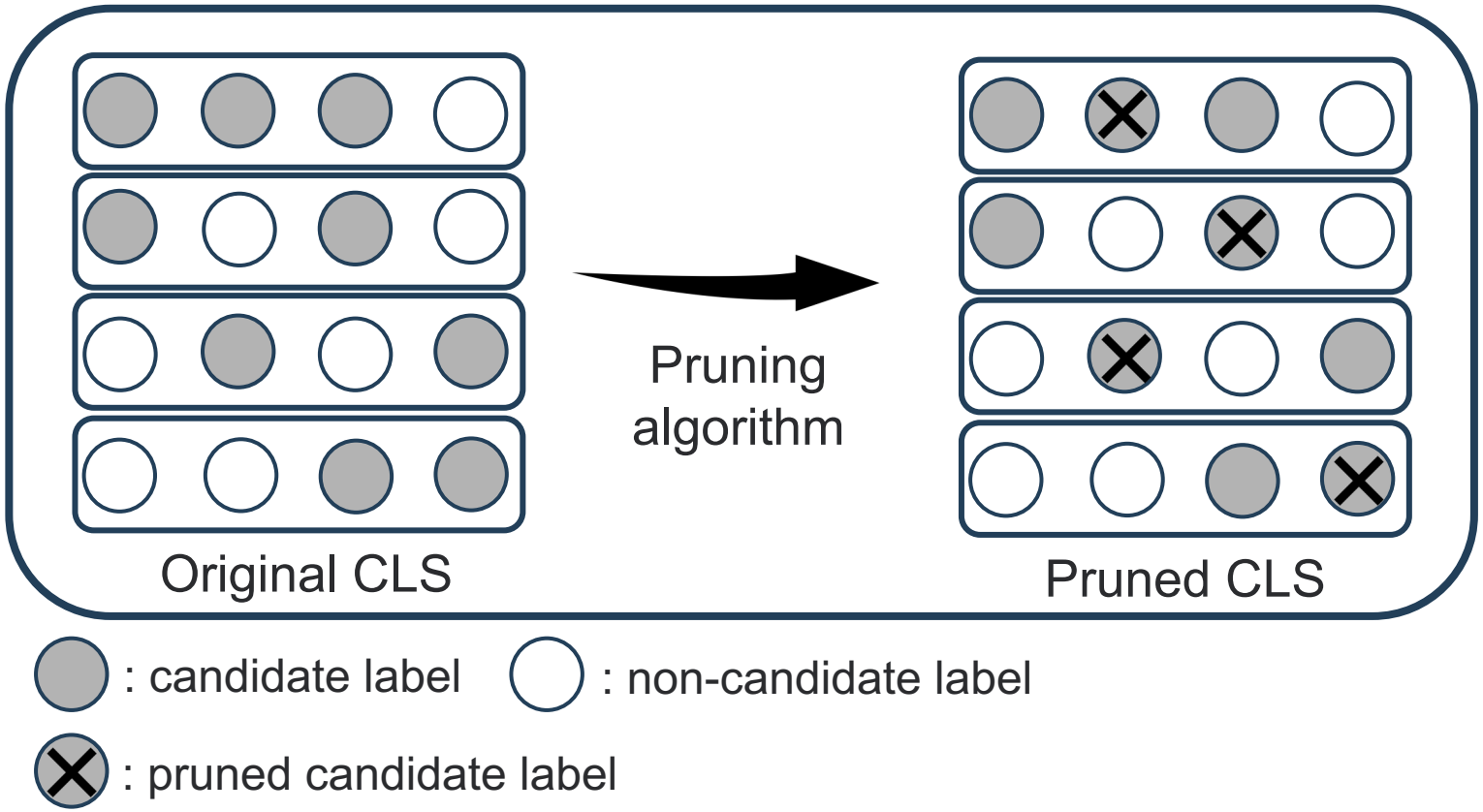


Data-centric perspective for deep partial label learning

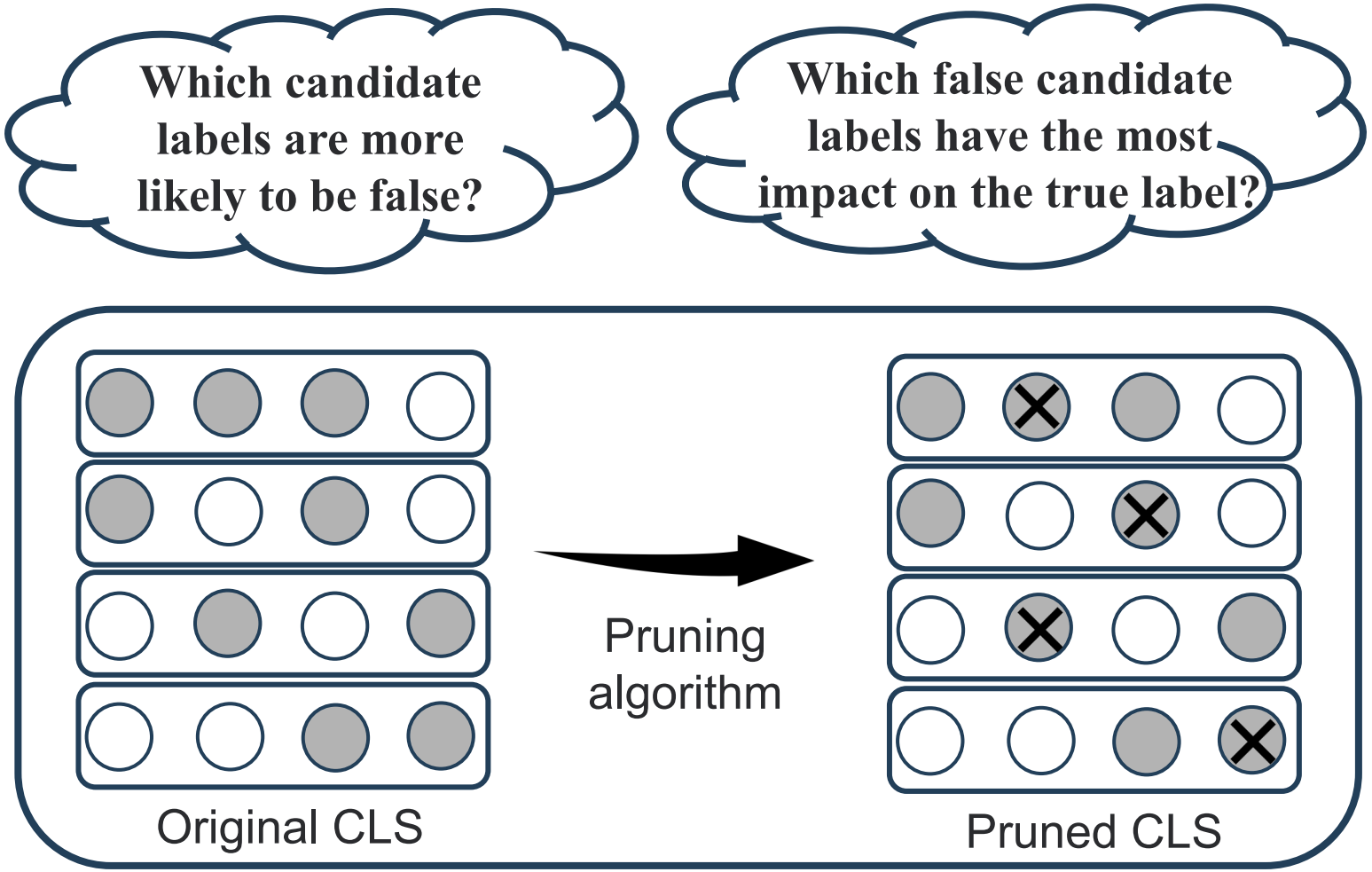


A new task: candidate label set pruning

Eliminating potential false candidate labels before the training process



Key issues of candidate label set pruning



A formal definition of candidate label set pruning

Definition 1 (α-error and β-coverage pruning). *Given a PLL dataset* $\mathcal{D} = \{\boldsymbol{x}_i, Y_i\}_{i=1}^n$, *for each candidate label set* Y_i, *let* $\tilde{Y}_i$ *denote the set of eliminated candidate labels from* Y_i *and* $\bar{Y}_i$ *denote the pruned candidate label set of* Y_i *(i.e.,* $\bar{Y}_i = Y_i \backslash \tilde{Y}_i$*). The pruning method is* α*-error where* $\alpha = \frac{\sum_{i=1}^n \mathbb{I}[y_i \in \tilde{Y}_i]}{n}$ *and* β*-coverage where* $\beta = \frac{\sum_{i=1}^n |\tilde{Y}_i|}{\sum_{i=1}^n (|Y_i| - 1)}$.

α-error → **How many true labels are pruned incorrectly?**

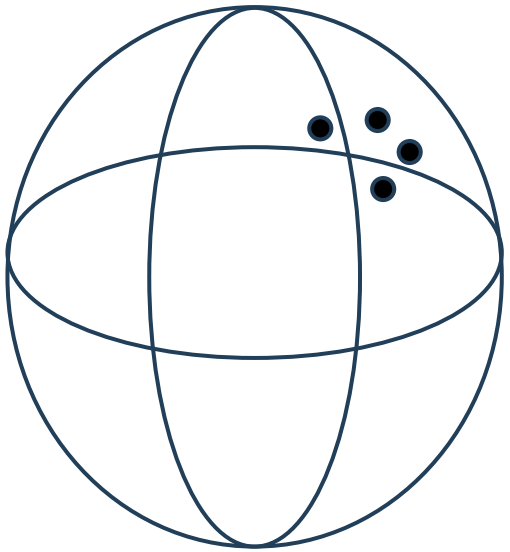
β-coverage → **How many candidate labels are pruned?**

Ideal situation: $\alpha = 0$ and $\beta = 1$ Perfect Pruning

Motivation of the proposed algorithm

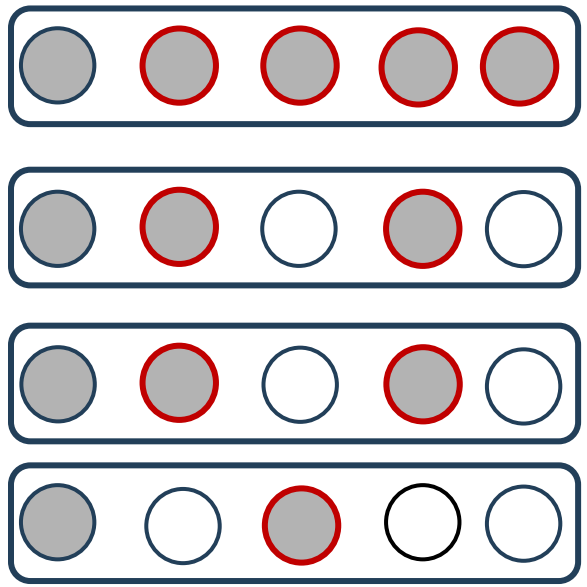
The candidate label that **rarely appears** in its nearby samples' candidate label sets has a high probability of being a false label.

Representation Space



Inconsistency

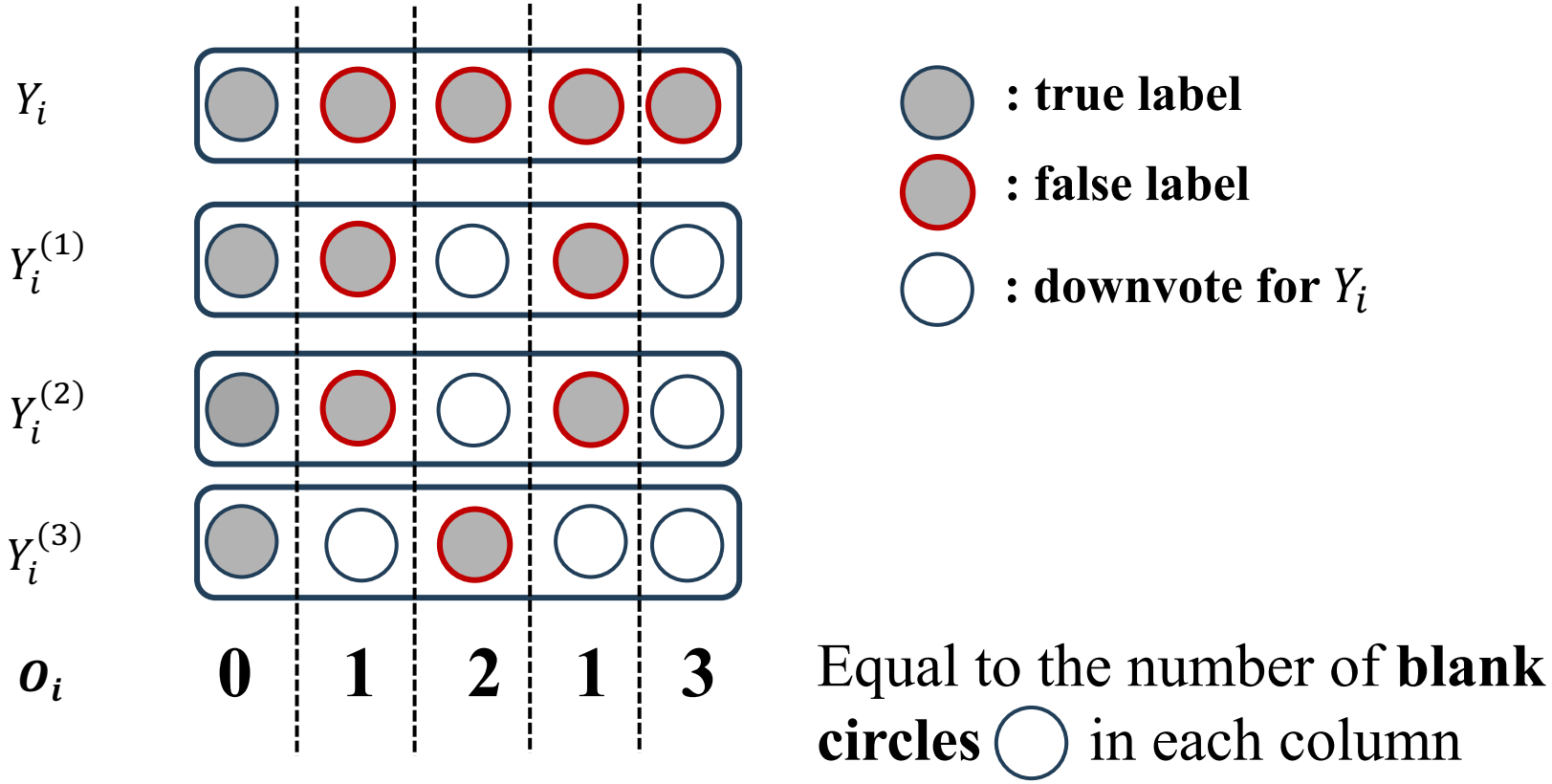
Candidate Label Space



Estimate voting statistic in the algorithm

Each *k*-NN likes a voter to downvote candidate labels

Voting formulation: $O_{ij} = \sum_{v=1}^{k} \mathbb{I}[y_{ij} \neq y_{ij}^{(v)}], \forall j \in Y_i,$

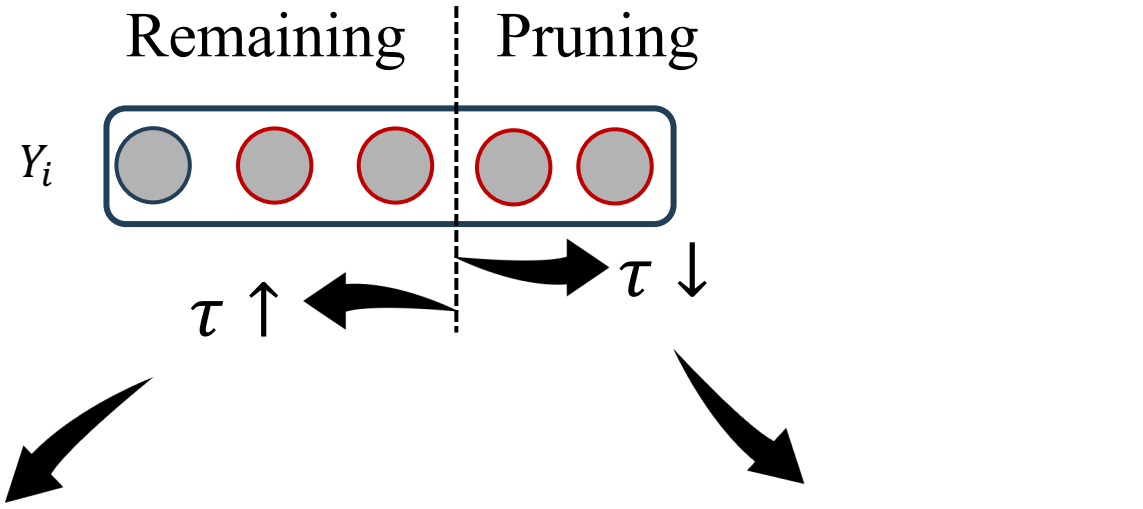


Equal to the number of **blank circles** ◯ in each column

Pruning number in the algorithm

Pruning how many candidate labels?

Trade-off parameter $\gamma_i = \lceil \tau(|Y_i| - 1) \rceil,$



Pruning more labels

High pruning error

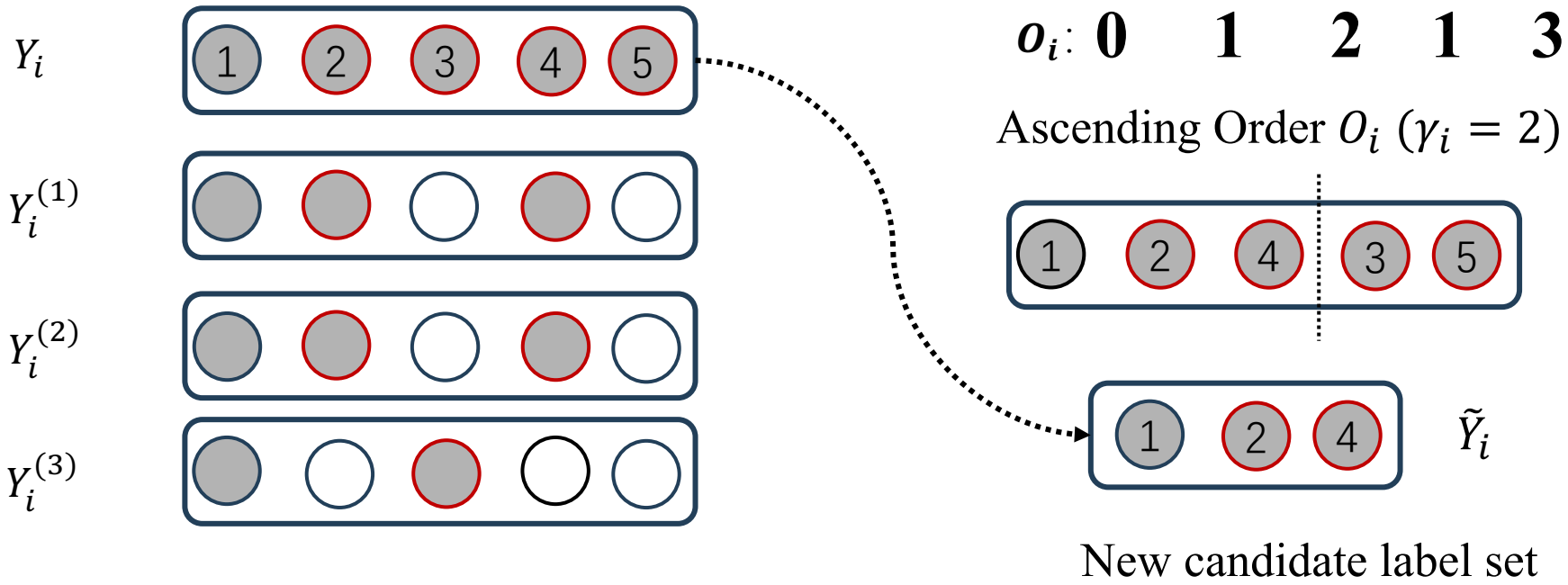
Pruning less labels

Make no sense

Pruning by order in the algorithm

Pruning each sample by the order of O_i

$$\tilde{Y}_i = \text{Top-}\gamma_i\text{-argmax}_{j \in Y_i}(O_{ij}),$$



Insights in the algorithm

Analyze the pruning error of the algorithm is very important!

How do parameters k and τ affect the pruning error?

How do feature quality and label ambiguity affect the pruning error?

Definition of label distinguishability

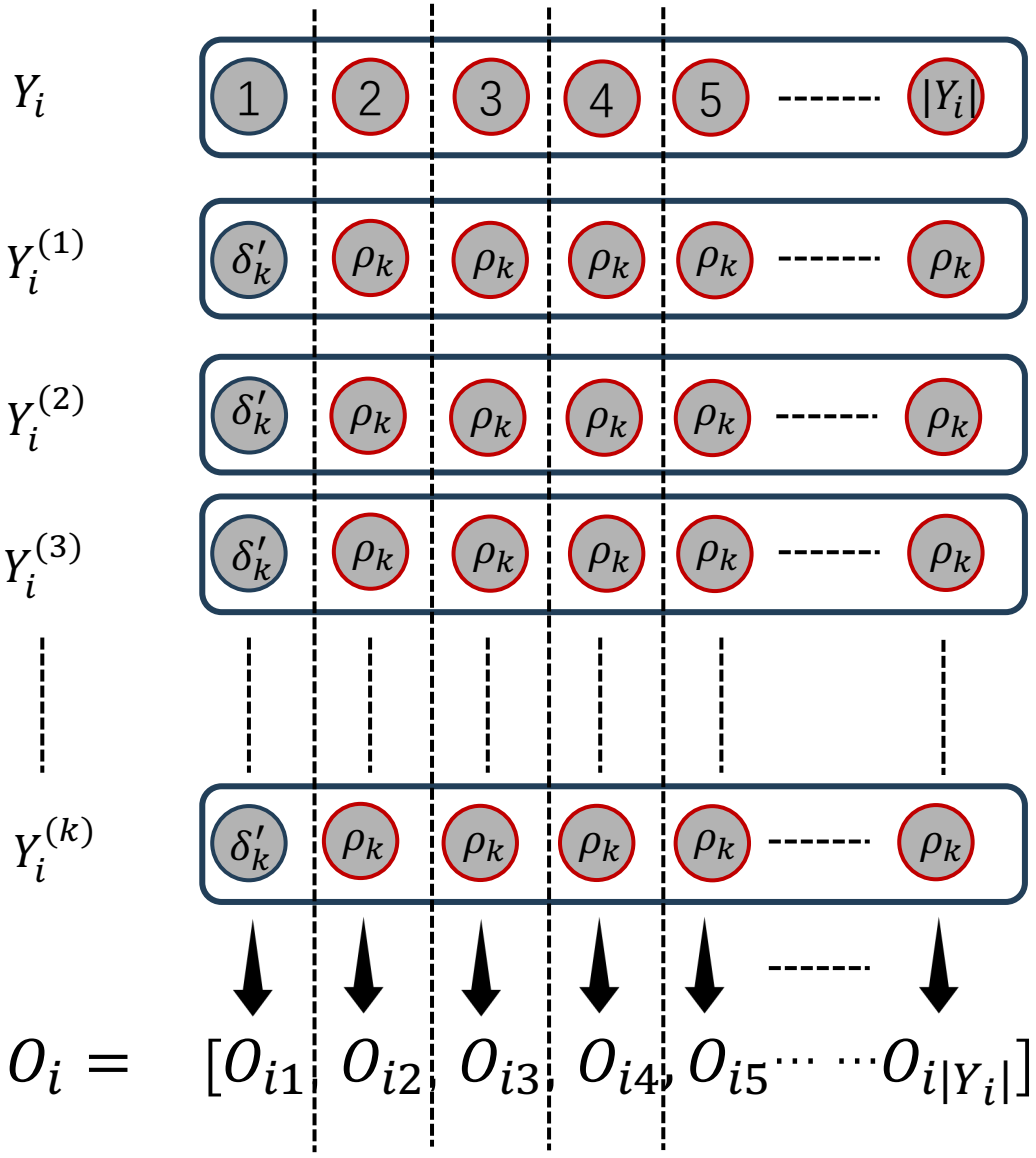
Condition for feature quality and label ambiguity

Definition 2 ((k, δ_k, ρ_k)-label distinguishability). *A PLL dataset* $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i)\}_{i=1}^{n}$ *satisfies* (k, δ_k, ρ_k) *label distinguishability if:* $\forall (\boldsymbol{x}_i, Y_i) \in \mathcal{D}$, *the true label* $y_i \in Y_i$ *is inside the candidate label set* $Y_i^{(j)}$ *of its each k-NN example* $(\boldsymbol{x}_i^{(j)}, Y_i^{(j)})$, *with probability at least* $1 - \delta_k$, *and each false candidate label* $y_i' \in Y_i \backslash \{y_i\}$ *is inside the candidate label set* $Y_i^{(j)}$ *of its each k-NN example* $(\boldsymbol{x}_i^{(j)}, Y_i^{(j)})$ *with probability no more than* ρ_k.

feature quality $\boldsymbol{1 - \delta_k}$

label ambiguity $\boldsymbol{\rho_k}$

Model the algorithm



$O_{i1} \sim B(k, \delta'_k)$ VS. $O_{ij} \sim B(k, \rho_k)$

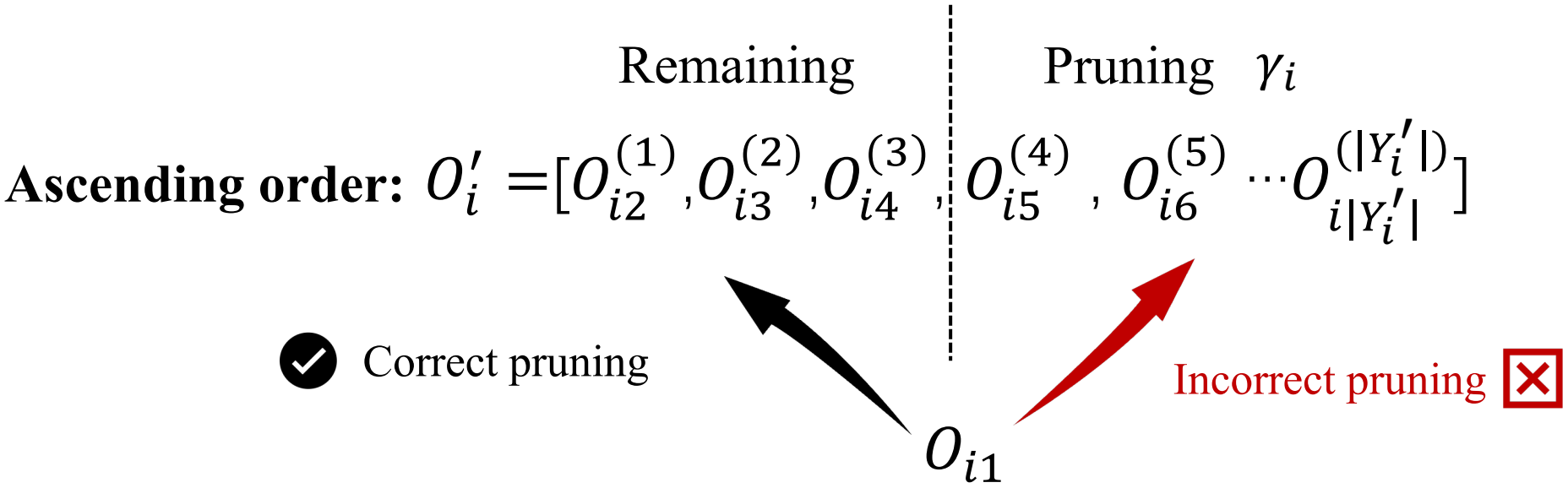
True label False label

Bernoulli experiment

Binomial distribution

$\delta'_k = 1 - \delta_k$

Incorrect pruning event in the algorithm



Incorrect pruning event: $[O_{ij}^{(|Y_i'|-\gamma_i+1)} < O_{i1}]$

Upper bounds of the pruning error

Upper bound of the pruning error

Theorem 1. *Assume that the (k, δ_k, ρ_k)-label distinguishability is satisfied. For each PLL example $(\boldsymbol{x}_i, Y_i)$, let us denote that the y-th label in the candidate label set Y_i is the true label, and the y'-th label in the false candidate label set $Y_i' = Y_i \backslash \{y\}$ is an arbitrary false candidate label, i.e., $y' \neq y$. Given the number of eliminated candidate labels γ_i, then the probability of getting an incorrect pruning can be upper bounded by*

$$\mathbb{P}(O_{iy'}^{(\xi_i)} < O_{iy}) \leqslant \sum_{j=1}^{k} \sum_{m=\xi_i}^{|Y_i'|} \binom{|Y_i'|}{m} \eta^m (1-\eta)^{(|Y_i'|-m)} b_{\delta_k}(k, j),$$

where $\xi_i = (|Y_i'| - \gamma_i + 1)$, $\binom{n}{r} = \frac{n!}{r!(n-r)!}$ is the combination formula, $b_{\delta_k}(k, j) = \binom{k}{j} \delta_k^j (1-\delta_k)^{k-j}$ denotes the probability mass function of a binomial distribution $B(k, \delta_k)$, and $\eta = I_{\rho_k}(k-j+1, j)$ where $I_{\rho_k}(k, j) = \int_0^{\rho_k} t^{k-1}(1-t)^{j-1} dt$ is the regularized incomplete beta function.

Upper bound of the extra pruning error caused by increasing the pruning number

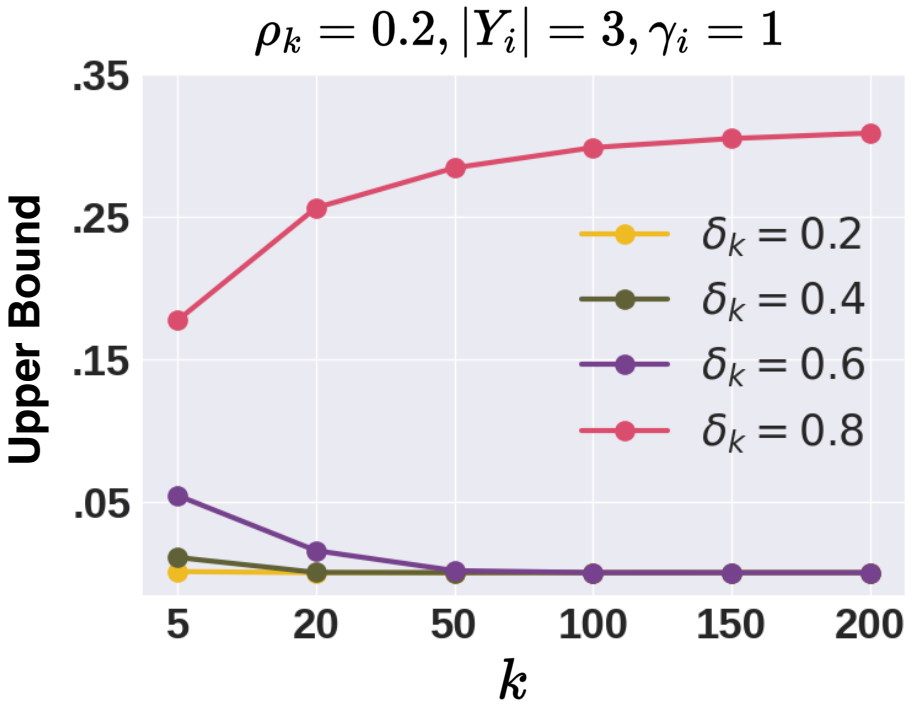
Theorem 2. *Given the same assumption of the (k, δ_k, ρ_k)-label distinguishability and notations in Theorem 1, when increasing the number of eliminated candidate labels (i.e., $\gamma_i^2 > \gamma_i^1$), the extra pruning error can be bounded by*

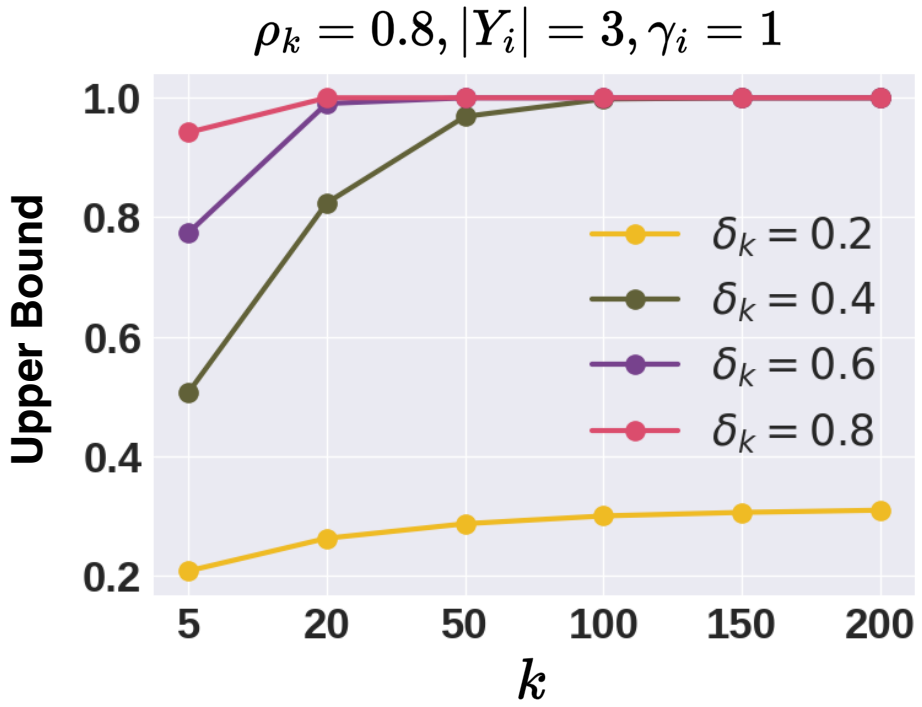
$$\mathbb{P}(O_{iy'}^{(\xi_i^2)} < O_{iy}) - \mathbb{P}(O_{iy'}^{(\xi_i^1)} < O_{iy}) \leqslant \sum_{j=1}^{k} \sum_{m=\xi_i^2}^{\xi_i^1 - 1} \binom{|Y_i'|}{m} \eta^m (1-\eta)^{|Y_i'|-m} b_{\delta_k}(k, j),$$

where $\xi_i^1 = (|Y_i'| - \gamma_i^1 + 1)$, $\xi_i^2 = (|Y_i'| - \gamma_i^2 + 1)$, and other notations are the same as those used in Theorem 1.

How does the parameter k affect the upper bound?

Suppose: $|Y_i| = 3 \quad \gamma_i = 1 \quad \delta_k = [0.2, 0.4, 0.6, 0.8] \quad p_k = [0.2, 0.8]$



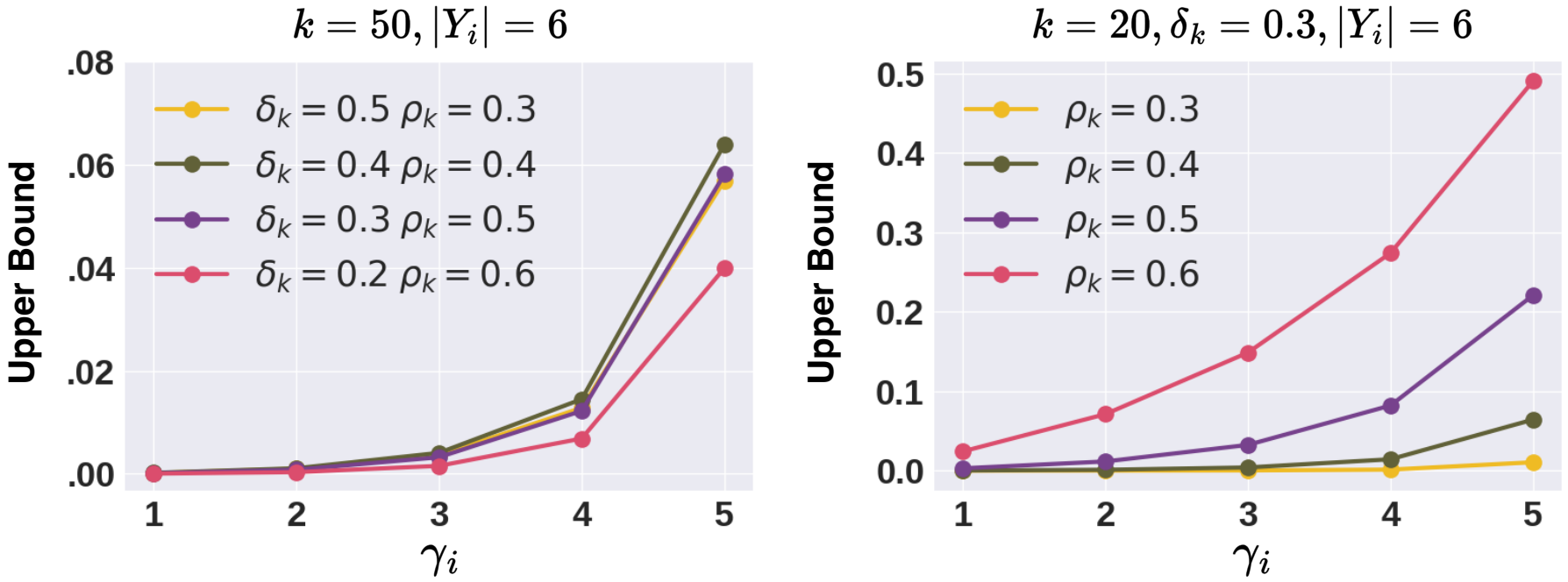


Observation: increasing the value of k is not always good!

Suggestion: a larger (smaller) value of k for high (low) quality features and low (high) label ambiguity

How does the pruning number γ_i affect the upper bound?

Suppose: $|Y_i| = 6 \quad \delta_k = [0.2, 0.4, 0.6, 0.8] \quad p_k = [0.2, 0.8] \quad k = [20, 50]$



Observation: increasing the value of γ_i could be feasible!

Suggestion: a larger (smaller) value of γ_i for high (low) quality features and low (high) label ambiguity

Experiment

Dataset and setting

Simulated-benchmark

CIFAR-10

CIFAR-100

Tiny-ImageNet

PASCAL VOC

Candidate label generation

Random (Uniform)

Label-dependent (LD)

Instance-dependent (ID)

Long-tailed (LT)

Real-world

Experiment

Deep PLL methods

CC [NeurIPS2020]
PRODEN [ICML2020]
LWS [ICML2021]
CAVL [ICLR2022]
PiCO [ICLR2022]
CRDPLL [ICML2022]

II-PLL methods

ABLE [IJCAI2022]
IDGP [ICLR2023]
POP [ICML2023]

ID-PLL methods

SoLar [NeurIPS2022]
RECORDS [ICLR2023]

LT-PLL methods

Experiment

Evaluation metrics

1. CLSP

How many true labels are pruned incorrectly?

How many false candidate labels are pruned?

α −Error

β − Coverage

2. Performance Improvement

Does the pruning algorithm improve learning methods?

Accuracy Comparison

Experiment

Feature extractors

Representation of different qualities

ResNet-S

Supervised learning

ResNet-SSL

Self-supervised learning

CLIP

ALBEF

BLIP-2

Multi-modal pretrained

ResNet-I

ImageNet1K pretrained

Experiment

Parameter

Values of k and τ used in the algorithm

Setup	Uniform			LD		ID			LT		×
	C-10	C-100	T-I	C-10	C-100	C-10	C-100	T-I	C-10	C-100	VOC
τ	0.6	0.6	0.4	0.6	0.6	0.2	0.2	0.2	0.2	0.2	0.1
k	150	150	150	50	150	5	5	50	50	50	5

Experimental results

CLSP performance

The proposed algorithm effectively reduces the number of candidate labels

Dataset	CIFAR-10				CIFAR-100				Tiny-ImageNet			VOC
	0.4	0.6	LD	ID	0.05	0.1	H-0.5	ID	0.01	0.05	ID	×
O Avg. C	4.6	6.4	2.5	3.3	6.0	10.9	3.0	4.8	3.0	11.0	8.3	2.5
P Avg. C	**2.1**	**2.9**	**1.2**	**2.6**	**2.9**	**5.0**	**1.3**	**4.4**	**2.3**	**7.2**	**7.1**	**1.8**
α-error (%)	.18	.16	.36	.22	.43	.50	2.9	.54	.46	.65	.38	5.2
β-coverage	.69	.64	.89	.32	.62	.59	.82	.08	.49	.37	.16	.47

O (P) Avg. C means that the average number of original (pruned) candidate labels

Experimental results

CLSP performance

Multi-modal feature extractors are better

Dataset	q	ResNet-S	ResNet-SSL	ResNet-I	CLIP	ALBEF	BLIP-2
CIFAR-10	0.4	**91.71**	79.47	89.55	90.82	91.14	91.61
	0.6	**90.05**	76.98	87.87	89.03	89.42	89.97
	LD	**97.74**	78.91	91.94	95.63	96.33	97.48
	ID	70.41	63.73	68.03	69.62	69.85	**70.50**
CIFAR-100	0.05	**89.16**	71.28	87.42	88.07	88.39	89.08
	0.1	**87.87**	67.11	85.55	86.35	86.86	87.69
	H-0.5	**96.00**	69.93	86.43	89.07	90.70	93.89
	ID	**34.35**	16.85	30.39	31.31	31.64	32.53
Tiny-ImageNet	0.01	79.87	70.53	82.40	82.71	82.66	**82.97**
	0.05	71.25	62.04	73.58	74.05	74.00	**74.57**
	ID	48.22	45.72	49.20	49.01	**49.07**	49.05
VOC	×	**86.33**	73.48	78.10	79.93	80.05	78.99

The F1 score of different feature extractors used in the proposed algorithm

Experimental results

Test accuracy comparison

Dataset	q	CC	PRODEN	LWS	CAVL	PiCO	CR	ABLE	IDGP	POP
C-10	0.4	81.67	81.11	84.85	78.14	94.20	96.99	94.53	92.34	95.19
		86.45	**82.07**	**86.68**	**81.25**	**94.44**	**97.24**	**94.97**	**93.38**	**95.64**
	0.6	71.16	79.81	81.67	54.52	92.96	96.47	93.69	89.48	94.57
		84.62	**82.42**	**85.87**	**80.99**	**94.32**	**97.21**	**94.92**	**92.52**	**95.48**
	LD	89.57	81.83	86.18	80.43	94.59	97.24	94.77	92.47	95.63
		90.81	**82.49**	**87.54**	**82.17**	**94.49**	**97.58**	**95.19**	**92.82**	**95.87**
	ID	73.92	78.03	78.70	67.21	91.08	87.89	91.17	84.45	93.63
		77.57	**81.92**	**84.67**	**77.97**	**93.41**	**95.90**	**93.99**	**92.08**	**95.05**
C-100	0.05	64.05	48.68	51.18	41.00	72.31	83.16	74.43	68.39	76.35
		64.36	**49.72**	**52.69**	**49.62**	**72.66**	**83.53**	**75.08**	**68.86**	**76.85**
	0.1	62.31	46.26	45.57	21.34	56.80	82.51	74.80	67.62	74.38
		64.05	**48.38**	**51.62**	**45.48**	**72.51**	**83.39**	74.76	**68.55**	**75.95**
	H-0.5	63.72	29.28	51.29	48.76	72.46	82.93	74.11	68.22	74.90
		65.56	**40.92**	**53.40**	**48.88**	**73.10**	**83.38**	**75.59**	**68.53**	**75.32**
	ID	63.06	49.83	53.18	47.35	71.04	80.76	74.04	66.71	73.36
		63.30	**50.11**	52.74	**48.24**	**71.78**	**80.93**	**74.49**	**67.23**	**74.26**

Dataset	q	CC	PRODEN	LWS	CAVL	CRDPLL
Tiny-ImageNet	0.01	65.04	65.21	66.92	64.97	67.48
		65.35	**65.28**	**66.98**	**65.43**	**67.56**
	0.05	63.06	63.02	64.34	35.53	65.99
		63.42	**63.55**	**65.61**	**52.29**	**66.21**
	ID	61.06	59.12	61.56	53.52	63.70
		62.11	**60.15**	**62.30**	**55.95**	**64.27**

The proposed CLSP algorithm significantly improves deep PLL methods under different datasets and settings.

Specially, the proposed CLSP algorithm has more significant improvement on the instance-dependent case.

Experimental results

Test accuracy comparison

Dataset	q	ϕ	CC	PRODEN	LWS	CAVL	CR	SoLar	RE
C-10-LT	0.3	50	75.31	76.73	77.28	44.18	71.53	84.48	79.95
			77.51	**78.66**	**78.79**	**44.21**	**78.37**	**84.69**	79.80
		100	67.36	65.58	65.52	43.39	82.61	75.53	70.86
			70.48	**71.23**	**72.49**	**43.41**	**84.35**	**76.82**	**71.43**
	0.5	50	59.90	62.95	62.22	42.84	47.92	82.41	75.48
			66.85	**65.30**	**64.13**	**48.36**	**64.94**	**83.57**	**76.48**
		100	55.36	55.37	57.19	44.85	60.43	71.50	65.73
			63.05	**62.19**	**62.13**	42.59	**68.05**	70.85	**67.08**
C-100-LT	0.03	50	43.70	43.93	43.40	32.39	45.06	48.31	39.91
			44.67	**44.69**	**45.24**	**33.33**	**45.80**	**49.27**	**40.30**
		100	39.32	38.71	38.07	30.55	51.24	42.76	36.42
			40.14	**39.52**	**39.24**	**31.75**	**52.52**	**43.81**	**37.44**
	0.05	50	42.37	39.53	38.89	28.43	42.92	46.39	44.82
			42.66	**41.18**	**40.70**	**29.46**	**44.45**	**47.01**	**46.06**
		100	35.59	34.94	34.43	26.16	48.91	40.94	40.03
			37.52	**37.16**	**36.31**	**26.76**	**49.75**	**42.33**	**40.44**

Dataset		CC	PRODEN	LWS	CAVL	CRDPLL	SoLar	RECORDS
VOC	Test	34.21	43.59	17.78	32.25	21.26	64.53	62.52
		50.81	**47.85**	**28.08**	**49.45**	**38.53**	**65.42**	**65.38**
	Trans.	72.15	77.26	64.46	77.43	67.09	76.56	40.33
		77.51	**79.01**	**71.07**	**86.88**	**75.09**	**82.35**	**68.32**

Specially, the proposed CLSP algorithm has more significant improvement on the long-tailed case than the class-balanced case.

Experimental results

Train accuracy comparison

Dataset	q	ϕ	CC	PRODEN	LWS	CAVL	CR	SoLar	RE
C-10	0.3	50	94.57	95.16	95.10	86.37	96.82	98.65	85.18
			95.57	**96.13**	**96.02**	**86.88**	**98.26**	**98.87**	**87.89**
		100	94.86	95.05	94.95	89.47	97.09	97.63	78.26
			96.36	**96.65**	**96.58**	**90.20**	**98.04**	**98.12**	**82.23**
	0.5	50	89.33	90.85	90.57	83.74	91.47	96.22	79.51
			91.53	**92.30**	**92.18**	**86.46**	**95.35**	**97.45**	78.33
		100	90.68	91.68	91.72	87.03	91.62	95.43	73.05
			93.22	**93.95**	**93.58**	**88.15**	**94.23**	**96.65**	71.12
C-100	0.03	50	92.62	91.85	91.84	83.20	95.45	94.72	87.42
			93.81	**93.58**	**93.43**	**85.12**	**96.41**	**95.90**	**88.48**
		100	92.54	91.76	91.46	86.17	95.61	93.56	87.97
			93.97	**93.41**	**93.20**	**87.04**	**96.38**	**94.82**	**88.57**
	0.05	50	89.10	86.63	86.57	77.38	92.50	91.23	90.30
			90.29	**88.63**	**88.54**	**79.61**	**93.72**	**91.86**	**91.33**
		100	88.63	86.98	87.28	79.15	92.28	90.62	90.18
			90.29	**89.19**	**89.14**	**79.61**	**93.72**	**92.21**	**91.51**

Dataset	q	CC	PRODEN	LWS	CAVL	CRDPLL
Tiny-ImageNet	0.01	96.93	96.83	97.08	96.57	97.49
		97.66	**97.62**	**97.75**	**97.45**	**98.03**
	0.05	89.56	89.21	89.96	55.14	90.70
		91.36	**91.04**	**90.84**	**77.09**	**92.31**
	ID	75.81	74.41	75.86	67.39	77.19
		77.50	**75.93**	**77.36**	**70.65**	**78.43**

The proposed CLSP algorithm indeed boosts label disambiguation in deep PLL methods.

More training samples are correctly identified after pruning a part of candidate labels.

Experimental results

Train accuracy comparison

Dataset	q	ϕ	CC	PRODEN	LWS	CAVL	CR	SoLar	RE
C-10-LT	0.3	50	94.57	95.16	95.10	86.37	96.82	98.65	85.18
			95.57	**96.13**	**96.02**	**86.88**	**98.26**	**98.87**	**87.89**
		100	94.86	95.05	94.95	89.47	97.09	97.63	78.26
			96.36	**96.65**	**96.58**	**90.20**	**98.04**	**98.12**	**82.23**
	0.5	50	89.33	90.85	90.57	83.74	91.47	96.22	79.51
			91.53	**92.30**	**92.18**	**86.46**	**95.35**	**97.45**	78.33
		100	90.68	91.68	91.72	87.03	91.62	95.43	73.05
			93.22	**93.95**	**93.58**	**88.15**	**94.23**	**96.65**	71.12
C-100-LT	0.03	50	92.62	91.85	91.84	83.20	95.45	94.72	87.42
			93.81	**93.58**	**93.43**	**85.12**	**96.41**	**95.90**	**88.48**
		100	92.54	91.76	91.46	86.17	95.61	93.56	87.97
			93.97	**93.41**	**93.20**	**87.04**	**96.38**	**94.82**	**88.57**
	0.05	50	89.10	86.63	86.57	77.38	92.50	91.23	90.30
			90.29	**88.63**	**88.54**	**79.61**	**93.72**	**91.86**	**91.33**
		100	88.63	86.98	87.28	79.15	92.28	90.62	90.18
			90.29	**89.19**	**89.14**	**79.61**	**93.72**	**92.21**	**91.51**

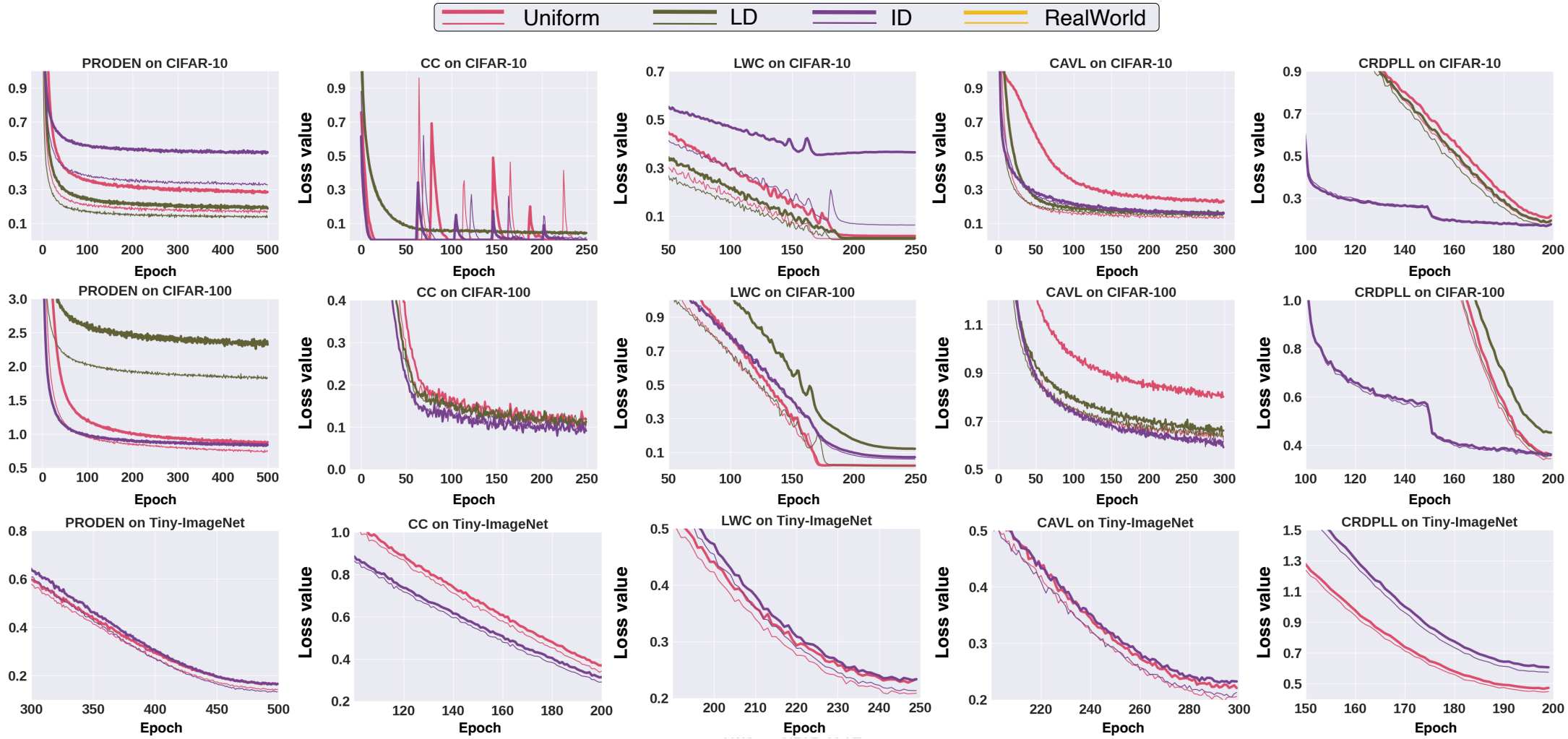
Dataset		CC	PRODEN	LWS	CAVL	CRDPLL	SoLar	RECORDS
VOC	Test	34.21	43.59	17.78	32.25	21.26	64.53	62.52
		50.81	**47.85**	**28.08**	**49.45**	**38.53**	**65.42**	**65.38**
	Trans.	72.15	77.26	64.46	77.43	67.09	76.56	40.33
		77.51	**79.01**	**71.07**	**86.88**	**75.09**	**82.35**	**68.32**

The same situation can also be observed in the long-tailed case.

Specially, on the real-world dataset PASCAL VOC, the performance is also significant.

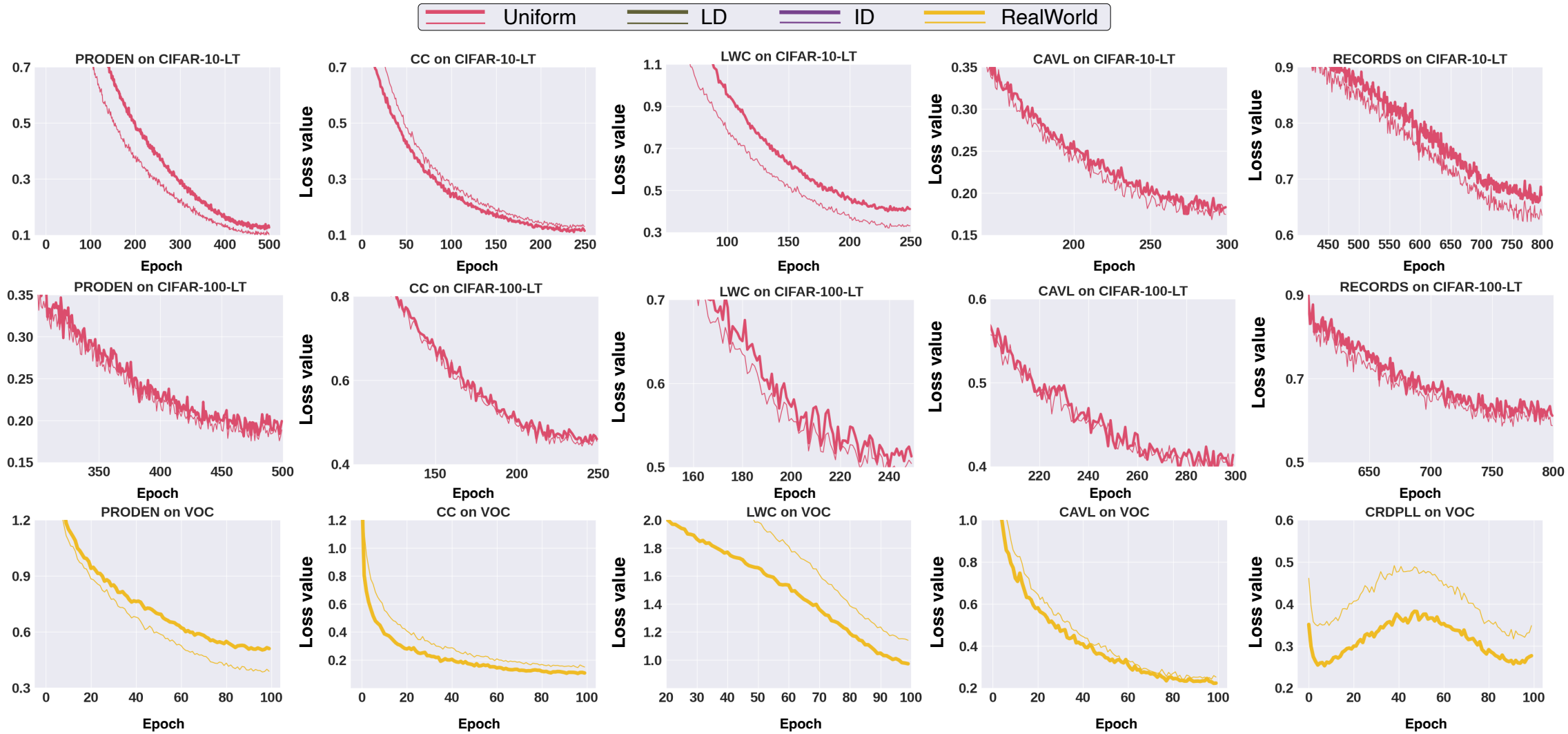
Experimental results

Train loss comparison



Experimental results

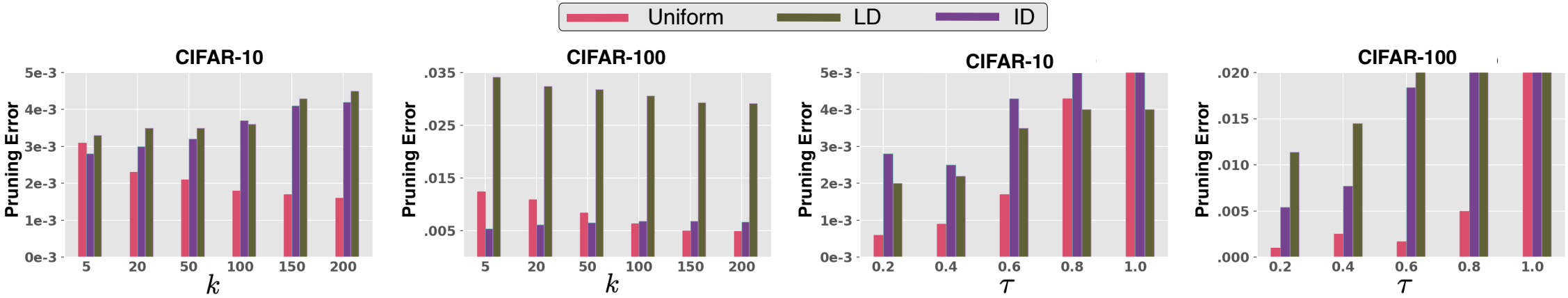
Train loss comparison



Experimental results

Parameter analysis

These results are consistent to the theorem-inspired observations

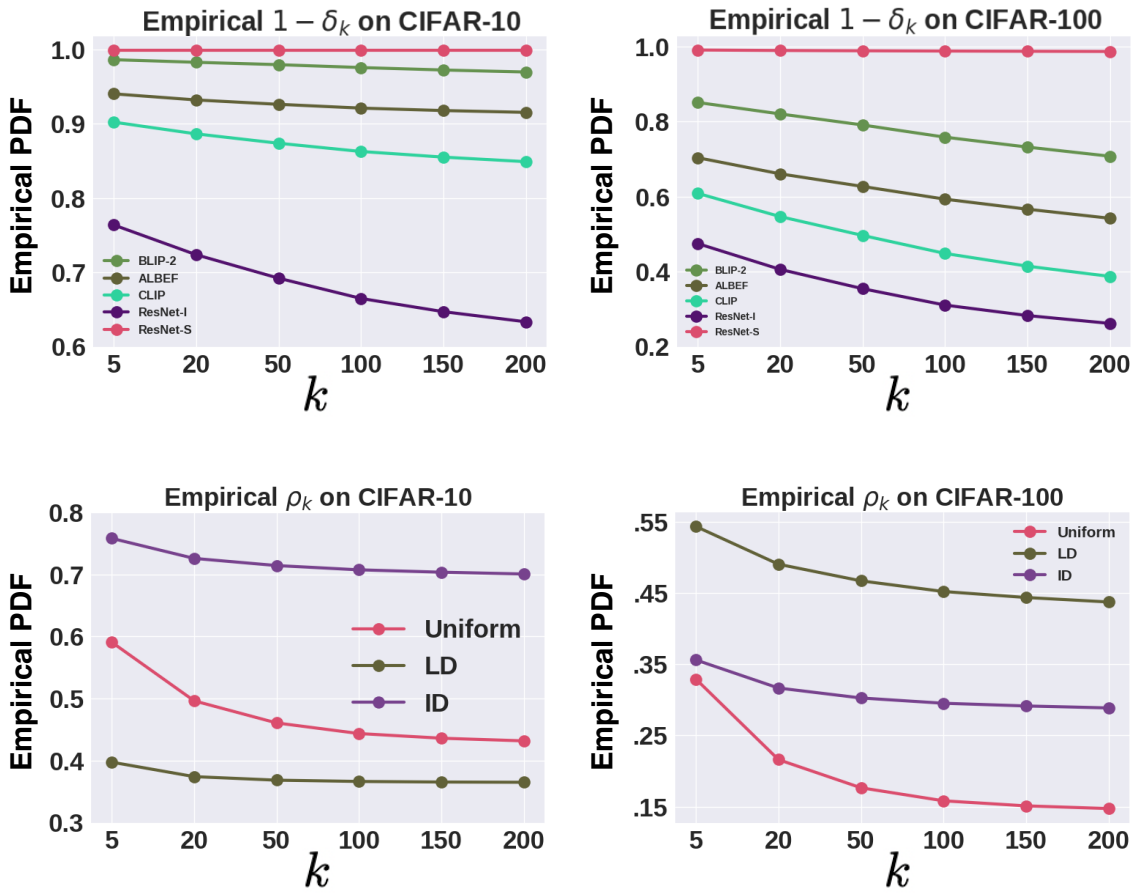


Different values of k and τ

Experimental results

Empirical parameters

Empirically calculated values of δ_k and ρ_k in label distinguishability



Conclusion and future work

- Pioneer data-centric research for deep partial-label learning
- Propose an efficient and effective CLSP algorithm
- Achieve good empirical results

In future, it's also interesting to :

- Design better methods
- Propose better evaluation metrics
- Follow a data-centric perspective for other weakly supervised learning

Thank you!

Email: shuohe123@gmail.com