Mitigating Label Noise through Data Ambiguation

Julian Lienen[1], Eyke Hüllermeier[2,3]

[1] Paderborn University, Germany [2] LMU Munich, Germany [3] Munich Center for Machine Learning, Germany

PROBLEM SETTING

Setting: Probabilistic classification given instances $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ with discrete space $\mathcal{Y} := \{y_1, \dots, y_K\}$

- Instances $\boldsymbol{x} \in \mathcal{X}$ associated with underlying ground-truth class-conditional probability $p^*(\cdot \mid \boldsymbol{x}) \in \mathbb{P}(\mathcal{Y})$

Goal: Learn probabilistic classifier $\hat{p} : \mathcal{X} \rightarrow \mathbb{P}(\mathcal{Y})$

Problem: Dealing with ***label noise***

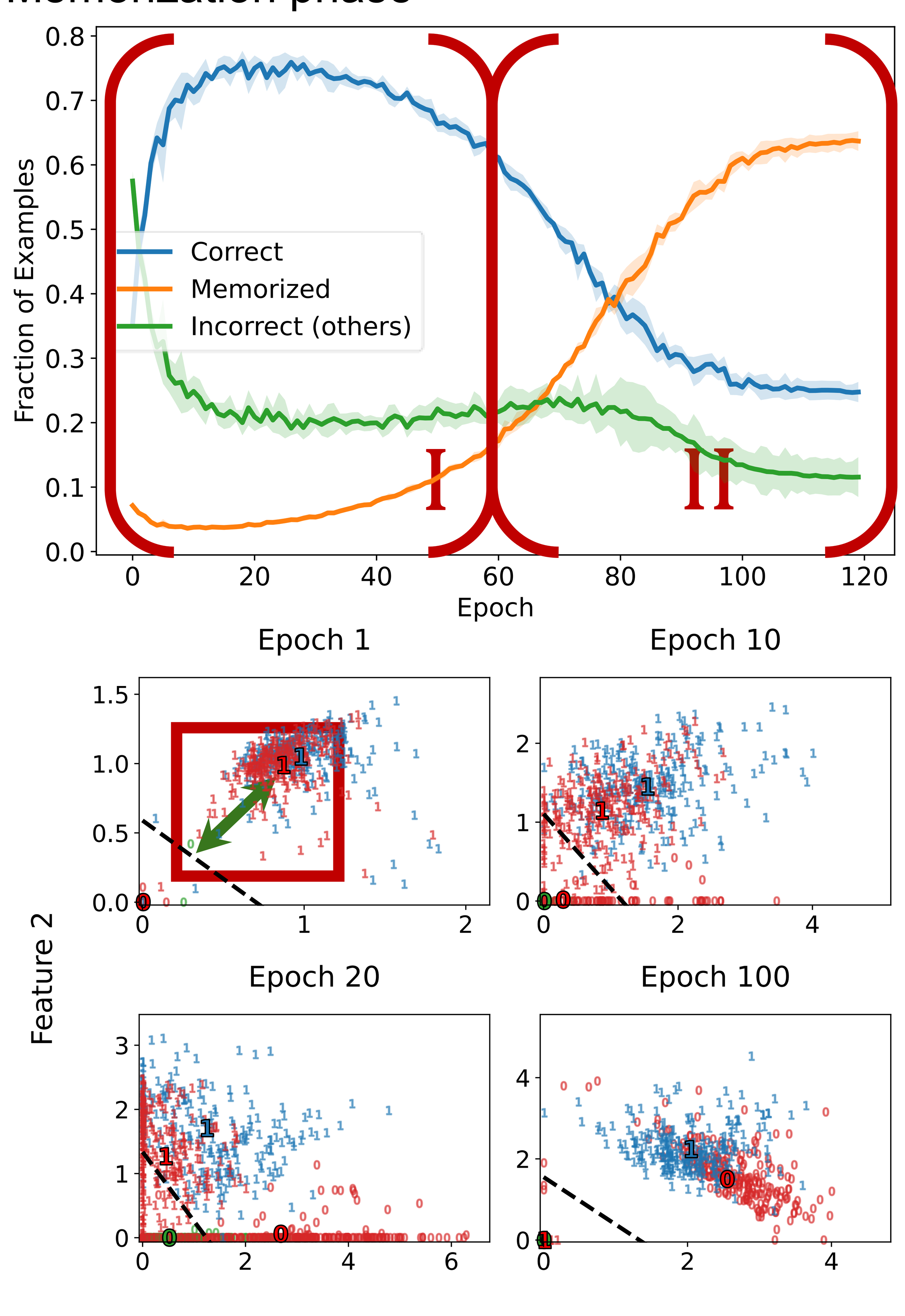
- Observing some instances with corrupted training labels $\tilde{y} \neq y$

TRAINING DYNAMICS WHEN FACING LABEL NOISE

Training dynamics of (overparameterized) models show two distinct phases [1,2]:

I) "Correct concept learning phase"

II) Memorization phase



Idea: Deliberately *ambiguate* labels if the model suggests a different label than the observed training label.

MODELING AMBIGUOUS PROBABILISTIC LABELS

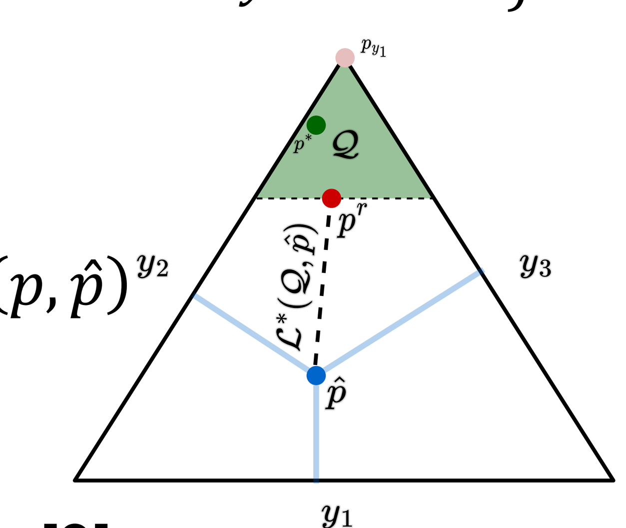
Ambiguation of probabilistic labels by ***credal sets*** $\mathcal{Q}$:

- Modeling beliefs about p^* as *upper probabilities* $\pi : \mathcal{Y} \rightarrow [0,1]$
  - $\pi(y')$ represents upper bound on $p^*(y')$
  - $\pi(y) = 1$ for observed training label y

$$\mathcal{Q}_\pi := \left\{ p \in \mathbb{P}(\mathcal{Y}) \,\middle|\, \forall Y \subseteq \mathcal{Y} \colon \sum_{y' \in Y} p(y') \leq \max_{y' \in Y} \pi(y') \right\}$$

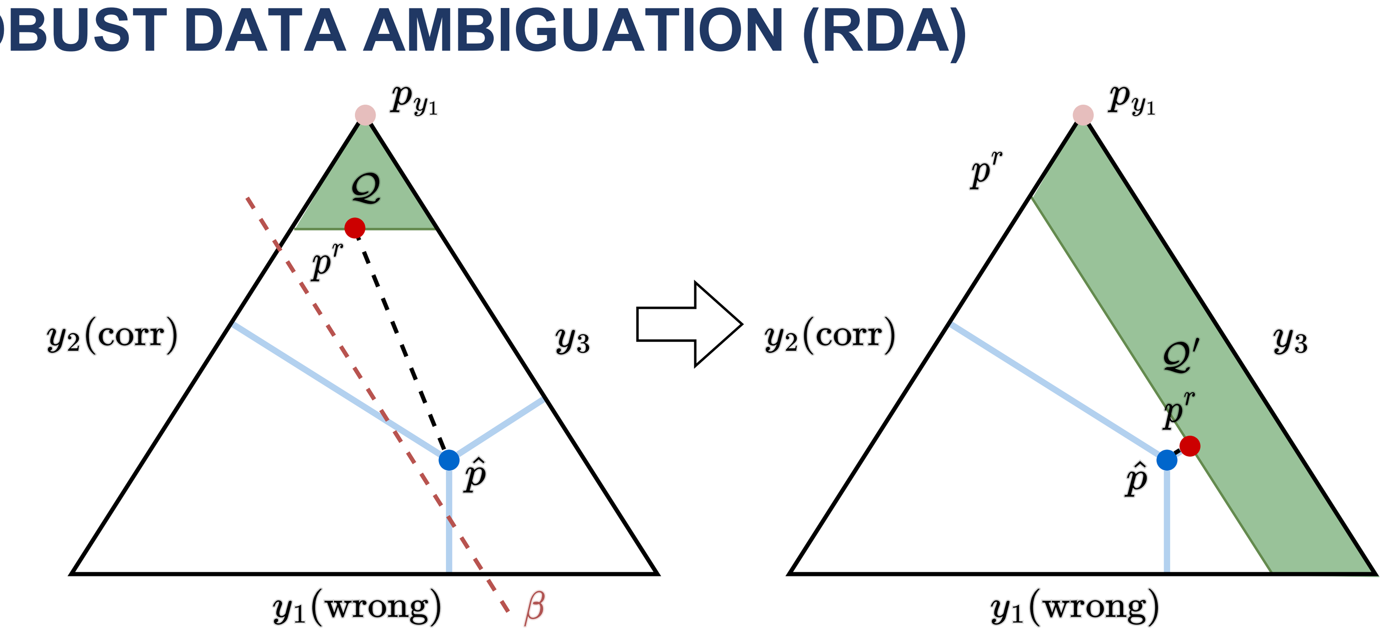
Learning from credal sets by ***label relaxation*** [4]:

$$\mathcal{L}^*(\mathcal{Q}_\pi, \hat{p}) := \min_{p \in \mathcal{Q}_\pi} \mathcal{L}(p, \hat{p})$$



- Probabilistic loss $\mathcal{L}$, efficient analytical solution
- Features data disambiguation [3]

ROBUST DATA AMBIGUATION (RDA)



Algorithm 1 Robust Data Ambiguation (RDA) Loss

Require: Training instance $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$, model prediction $\widehat{p}(\boldsymbol{x}) \in \mathbb{P}(\mathcal{Y})$, confidence threshold $\beta \in [0, 1]$, relaxation parameter $\alpha \in [0, 1)$

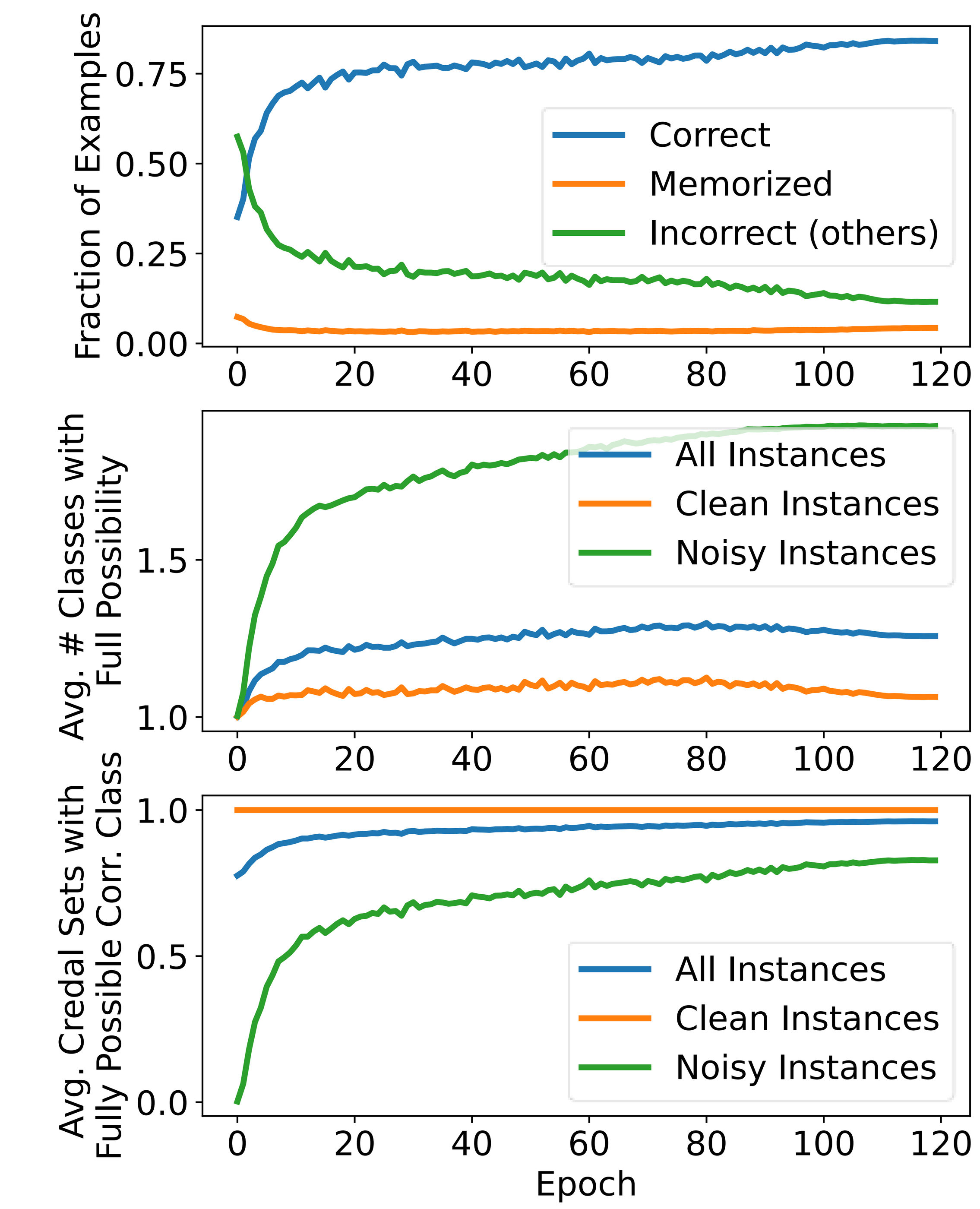
1: Construct π as in Eq. (4) with

$$\pi(y') = \begin{cases} 1 & \text{if } y' = y \vee \widehat{p}(y' \mid \boldsymbol{x}) \geq \beta \\ \alpha & \text{otherwise} \end{cases}$$

2: **return** $\mathcal{L}^*(Q_\pi, \widehat{p}(\boldsymbol{x}))$ as specified in Eq. (4), where Q_π is derived from π

EXPERIMENTS

Empirical results show **suppression of memorization** effects, leading to **improved robustness** against label noise.



Loss	Add. Param.	CIFAR-10 Sym. 25 %	CIFAR-10 Sym. 50 %	CIFAR-10 Sym. 75 %	CIFAR-100 Sym. 25 %	CIFAR-100 Sym. 50 %	CIFAR-100 Sym. 75 %
CE	✗	79.05 ±0.67	55.03 ±1.02	30.03 ±0.74	58.27 ±0.36	37.16 ±0.46	13.66 ±0.45
LS ($\alpha = 0.1$)	✗	76.66 ±0.69	53.95 ±1.47	29.03 ±1.21	59.75 ±0.24	37.61 ±0.61	13.53 ±0.51
LS ($\alpha = 0.25$)	✗	77.48 ±0.32	53.08 ±1.95	28.29 ±0.65	59.84 ±0.57	39.80 ±0.38	14.18 ±0.44
LR ($\alpha = 0.1$)	✗	80.53 ±0.39	57.55 ±0.95	29.83 ±0.87	57.52 ±0.58	36.77 ±0.54	13.23 ±0.14
LR ($\alpha = 0.25$)	✗	80.43 ±0.09	60.18 ±1.01	31.36 ±0.91	57.67 ±0.11	37.15 ±0.14	13.41 ±0.24
GCE	✗	90.82 ±0.10	83.36 ±0.65	54.34 ±0.37	68.06 ±0.31	58.66 ±0.28	**26.85** ±1.28
NCE	✗	79.05 ±0.12	63.94 ±1.74	38.23 ±2.63	19.32 ±0.81	11.09 ±1.03	6.12 ±7.57
NCE+AGCE	✗	87.57 ±0.10	83.05 ±0.81	51.16 ±6.44	64.15 ±0.23	39.64 ±1.66	7.67 ±1.25
NCE+AUL	✗	88.89 ±0.29	84.18 ±0.42	**65.98** ±1.56	69.76 ±0.31	57.41 ±0.41	17.72 ±1.27
CORES	✗	88.60 ±0.28	82.44 ±0.29	47.32 ±17.03	60.36 ±0.67	46.01 ±0.44	18.23 ±0.28
RDA (ours)	✗	**91.48** ±0.22	**86.47** ±0.42	48.11 ±15.41	**70.03** ±0.32	**59.83** ±1.15	26.75 ±8.83

Loss	Add. Param.	CIFAR-10N Random 1	CIFAR-10N Random 2	CIFAR-10N Random 3	CIFAR-10N Aggregate	CIFAR-10N Worst	CIFAR-100N Noisy
CE	✗	82.96 ±0.23	83.16 ±0.52	83.49 ±0.34	88.74 ±0.13	64.93 ±0.79	52.88 ±0.14
LS ($\alpha = 0.1$)	✗	82.76 ±0.47	82.10 ±0.21	82.12 ±0.37	88.63 ±0.11	63.10 ±0.38	53.48 ±0.45
LS ($\alpha = 0.25$)	✗	82.95 ±1.57	83.86 ±2.05	82.61 ±0.25	87.03 ±2.29	66.14 ±6.89	53.98 ±0.27
LR ($\alpha = 0.1$)	✗	83.00 ±0.36	82.64 ±0.31	82.82 ±0.21	88.41 ±0.29	66.62 ±0.33	52.01 ±0.04
LR ($\alpha = 0.25$)	✗	82.14 ±0.49	81.87 ±0.34	82.46 ±0.11	88.07 ±0.45	66.44 ±0.14	52.22 ±0.29
GCE	✗	88.85 ±0.19	88.96 ±0.32	88.73 ±0.11	90.85 ±0.32	77.24 ±0.47	55.43 ±0.47
NCE	✗	81.88 ±0.27	81.02 ±0.32	81.48 ±0.13	84.62 ±0.49	69.40 ±0.10	21.12 ±0.67
NCE+AGCE	✗	89.48 ±0.28	88.95 ±0.10	89.25 ±0.29	90.65 ±0.44	81.27 ±0.44	51.42 ±0.65
NCE+AUL	✗	89.42 ±0.22	89.36 ±0.15	88.94 ±0.55	90.92 ±0.19	81.28 ±0.47	56.58 ±0.41
CORES	✗	86.09 ±0.57	86.48 ±0.27	86.02 ±0.22	89.23 ±0.10	76.80 ±0.96	53.04 ±0.29
RDA (ours)	✗	**90.43** ±0.03	**90.09** ±0.29	**90.40** ±0.01	**91.71** ±0.38	**82.91** ±0.83	**59.22** ±0.26

- ➢ Robust "off-the-shelf" loss function against label noise without adding complexity
- ➢ On-the-fly loss calculation, no additional parameters

[1] Chang, H., *et al.* Active Bias: Training More Accurate Neural Networks by Emphasizing High Variance Samples. In *NeurIPS*, 2017.

[2] Liu, S., *et al.* Early-Learning Regularization Prevents Memorization of Noisy Labels. In *NeurIPS*, 2020.

[3] Hüllermeier, E., and Cheng, W. Superset Learning Based on Generalized Loss Minimization. In *ECML PKDD*, 2015.

[4] Lienen, J., and Hüllermeier, E. From Label Smoothing to Label Relaxation. In *AAAI*, 2021.