Paper ID: 1148

Clusterability as an Alternative to Anchor Points When Learning with Noisy Labels

Zhaowei Zhu, Yiwen Song, and Yang Liu

{**zwzhu**, yangliu}@ucsc.edu

Code & Dataset

REsponsible & Accountable Learning (REAL)
@ University of California, Santa Cruz

https://github.com/UCSC-REAL

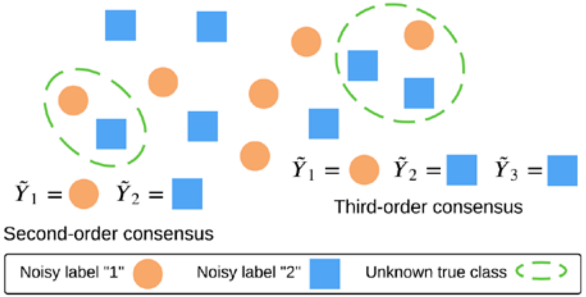


Figure: Illustration of high-order consensuses.

UNIVERSITY OF CALIFORNIA SANTA CRUZ

Noise Transition Matrix **T**

- Each element of **T**: $T_{ij} := \mathbb{P}(\widetilde{Y} = j | Y = i)$ Clean label *i* ➜ Noisy label *j*
- Example: Our self-collected CIFAR-10 human annotations:

(cat➜1,
dog➜2,
ship➜3)

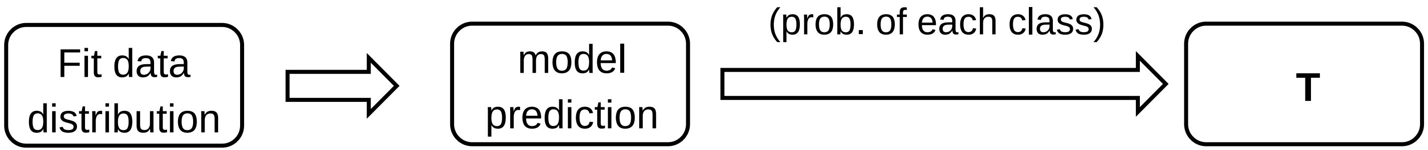
$$T_{11} = \mathbb{P}(\widetilde{Y} = 1 | Y = 1) = 0.6, \ T_{12} = \mathbb{P}(\widetilde{Y} = 2 | Y = 1) = 0.2$$

Why we need T?

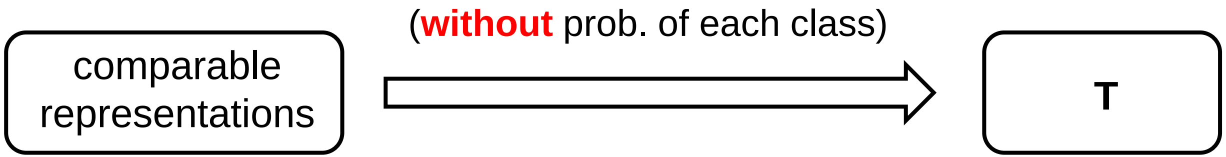
- Knowing T helps build *noise-resistant* classifier

BUT…

- Current methods [1-3] relies on models:



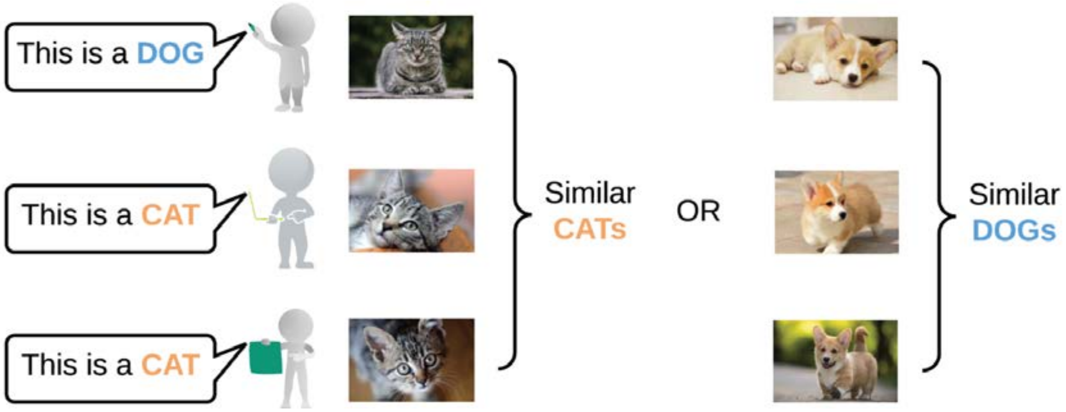
Model-free?



[1] G. Patrini et al. “Making deep neural networks robust to label noise: A loss correction approach." *CVPR'17.*
[2] X. Xia et al. “Are anchor points really indispensable in label-noise learning?" *NeurIPS'19.*
[3] C. Northcutt et al. “Condent learning: Estimating uncertainty in dataset labels." *JAIR'21.*

Motivation

- Check *label consensuses* of similar features



Intuition:

➥ Pattern (**DOG**, **CAT**, **CAT**) encodes **T**

Questions:

➥ Find similar features

➥ # similar features

➥ Decode **T**

Find similar features

- k-NN Label Clusterability:
  - Each representation and its k-NN belong to the same true class

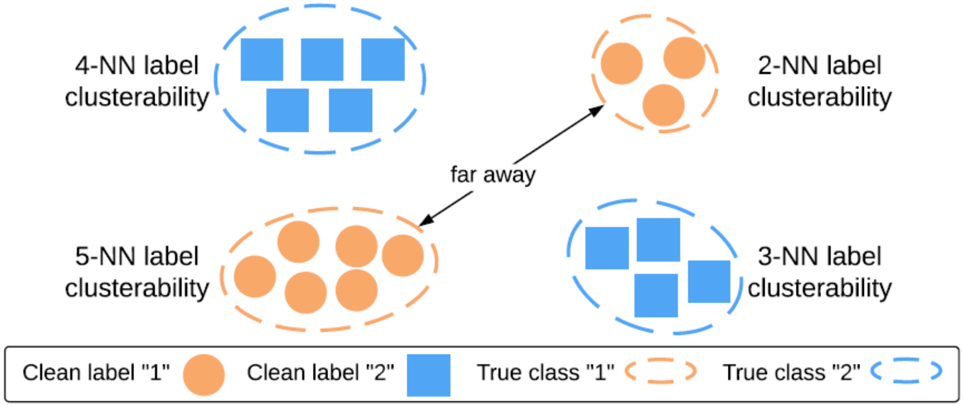


Figure: Illustration of k-NN label clusterability.

Properties:

➥ Larger k is harder
  - 2-NN is sufficient

➥ Local small clusters
  - different breeds of "CATs" may be far away

➥ **NOT** specifying the true class
  - "CAT" or "DOG"? Unknown!

similar features

- 2-NN label clusterability is feasible
  - Feature Extractors: Output of convolutional layers (when DNN overfits a dataset)
  - |E|: Sample size

Table: Ratio of feasible 2-NN tuples. (%)

Feature Extractor	*CIFAR-10* $\|E\| = 5k$	*CIFAR-10* $\|E\| = 50k$	*CIFAR-100* $\|E\| = 5k$	*CIFAR-100* $\|E\| = 50k$
Clean	99.99	99.99	99.88	99.90
Inst. $\eta = 0.2$	87.88	89.06	82.82	84.33
Inst. $\eta = 0.4$	78.15	79.85	64.88	68.31

- 2-NN label clusterability is sufficient to ***uniquely*** get the true **T** (Theorem 1)

UC SANTA CRUZ

Decode **T**

- Check High-Order Consensuses (HOC)

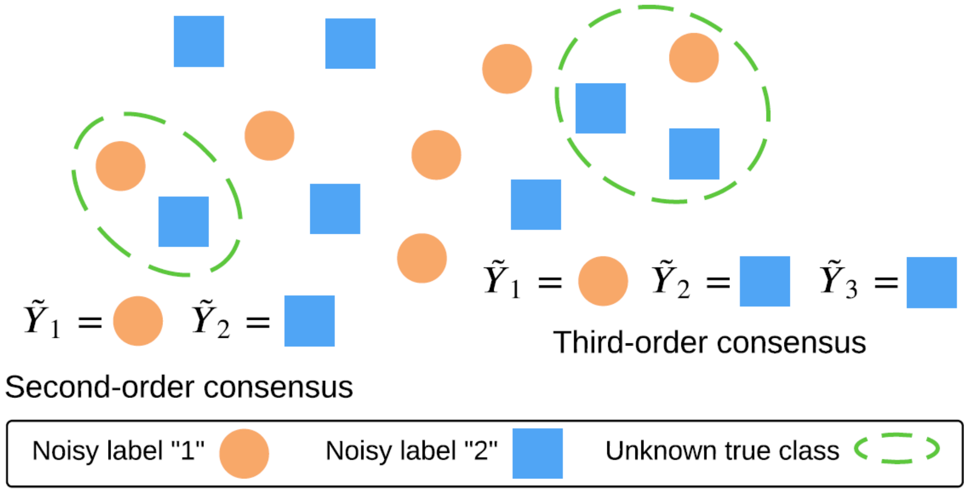


Figure: Illustration of high-order consensuses.

For each consensus pattern, we can:

➥ Count the frequency → ***estimates***

➥ Calculate the probability → ***functions***

Then:

➥ Solve equations:

(numerical) **estimates** **=** *(analytical)* **functions**

➥ Get:

- Noise transition ***T***
- Clean prior ***p***

UC SANTA CRUZ

Calculate the probability (Binary example)

$$\boldsymbol{T} = \begin{bmatrix} 1-e_1 & e_1 \\ e_2 & 1-e_2 \end{bmatrix}$$

$$e_1 = \mathbb{P}(\widetilde{Y} = 2 | Y = 1)$$

$$e_2 = \mathbb{P}(\widetilde{Y} = 1 | Y = 2)$$

$$\boldsymbol{p} = [p; 1-p]$$

- 1st-order (2 patterns)

Pattern "**CAT**"

Probability:

True: CAT. Report: CAT (Correct)

$$p \cdot (1 - e_1)$$

True: DOG. Report: CAT (Wrong)

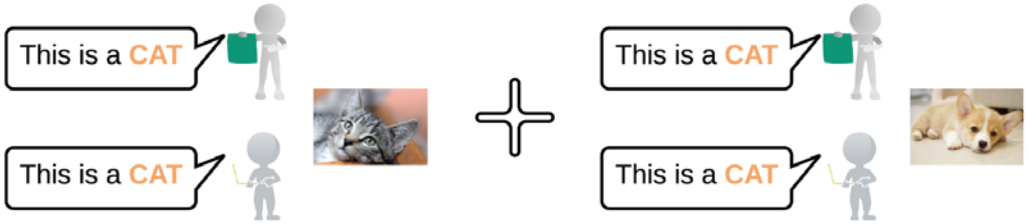
$$(1 - p) \cdot e_2$$

Calculate the probability (Binary example)

- 2nd-order (4 patterns)

Pattern “(CAT,CAT)”

Probability:



True: CAT. Report: CAT, CAT (Correct) + True: DOG. Report: CAT, CAT (Wrong)

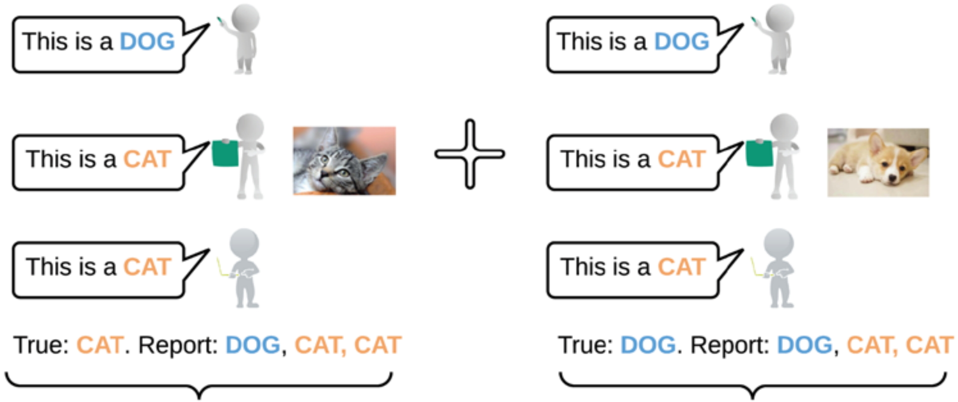
$$p \cdot (1 - e_1)^2 + (1 - p) \cdot e_2^2$$

Calculate the probability (Binary example)

- 3rd-order (8 patterns)

Pattern “(DOG,CAT,CAT)”

Probability:



$$p \cdot e_1 \cdot (1 - e_1)^2 + (1 - p) \cdot (1 - e_2) \cdot e_2^2$$

3rd-order is **sufficient!**

[4] Y. Liu et al. “Surrogate scoring rules." *ACM EC'20.*

High-Order Consensuses (HOC)

$$T_{ij} := \mathbb{P}(\widetilde{Y} = j | Y = i)$$

$$p_i = \mathbb{P}(Y = i)$$

Consensus Equations

- 1st-order (K equations): $\boldsymbol{c}^{[1]} := \boldsymbol{T}^\top \boldsymbol{p}$
- 2nd-order (K^2 equations): $\boldsymbol{c}_r^{[2]} := (\boldsymbol{T} \circ \boldsymbol{T}_r)^\top \boldsymbol{p}, \ r \in [K]$
- 3rd-order (K^3 equations): $\boldsymbol{c}_{r,s}^{[3]} := (\boldsymbol{T} \circ \boldsymbol{T}_r \circ \boldsymbol{T}_s)^\top \boldsymbol{p}, \ r, s \in [K]$

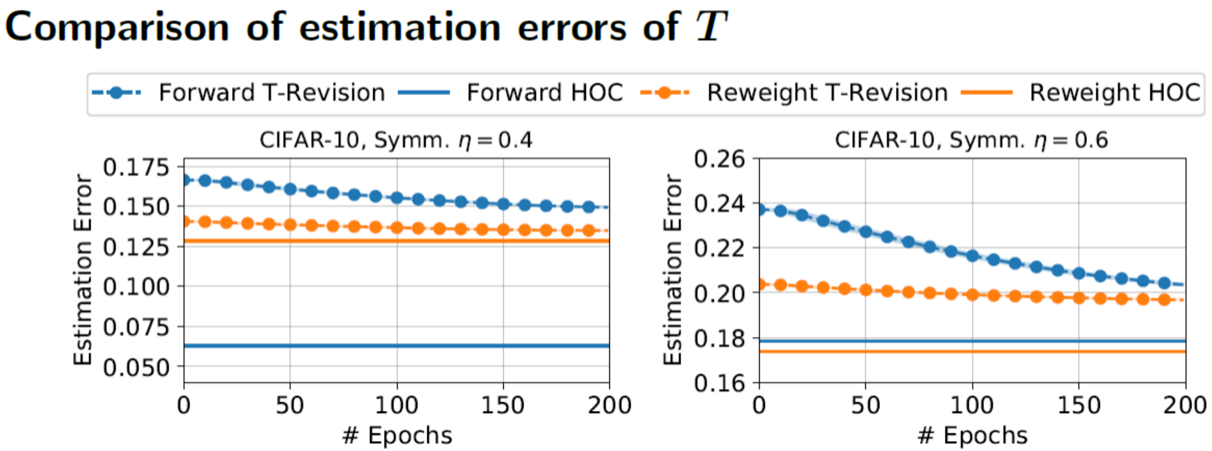
(Numbers = Functions)

Theorem 1: With 2-NN label clusterability, nonsingular and informative T, perfect knowledge of counts, *the consensus equations return the true T* ***uniquely.***

UC SANTA CRUZ

Experiment

- HOC can estimate T accurately



Experiment

- Loss Correction + HOC performs well

Challenging *instance-dependent* label noise:
Estimate ***T*** for each *local group*

Our method is:
1. flexible to extension
2. high sample complexity

✨**Our self-collected CIFAR-10 human annotations:**

- From Amazon Mechanical Turk (MTurk) in February 2020
- Collect each image with a cost of ¢10 per image

Table: Test accuracy (%) with human noise

Method	Clothing1M	**Human CIFAR-10**
Forward [1]	70.83	86.82
T-Revision [2]	71.67	85.92
CORES^2	73.24	89.98
HOC	**73.39**	**90.62**

Thank you !

Code & Dataset

Take a look at our **HOC estimator** and **self-collected** CIFAR-10 human annotations:

REsponsible & Accountable Learning (REAL) La
@ University of California, Santa Cruz

https://github.com/UCSC-REAL