

End-to-End Learning from Crowdsourced Labels: A Signal Processing perspective

*Shahana Ibrahim¹, Panagiotis A. Traganitis²,
Xiao Fu³, Georgios B. Giannakis⁴*

¹Dept. of ECE, University of Central Florida

²Dept. of ECE, Michigan State University

³Dept. of ECE, Oregon State University

⁴Dept. of ECE, University of Minnesota

Acknowledgements: NSF 2128593, 2212318, 2312546, 2007836



UNIVERSITY OF
CENTRAL FLORIDA

MICHIGAN STATE
UNIVERSITY



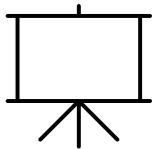
UNIVERSITY OF MINNESOTA
Driven to Discover™

Relevant links



Learning From Crowdsourced Noisy Labels: A Signal Processing Perspective

<https://arxiv.org/abs/2407.06902>



LINK TO SLIDES

https://github.com/shahana-ibrahim/Learning-from-Crowdsourced-Noisy-Labels/blob/main/Crowdsourcing_ICASSP25.pptx



[https://github.com/shahana-ibrahim/
Learning-from-Crowdsourced-Noisy-Labels](https://github.com/shahana-ibrahim/Learning-from-Crowdsourced-Noisy-Labels)

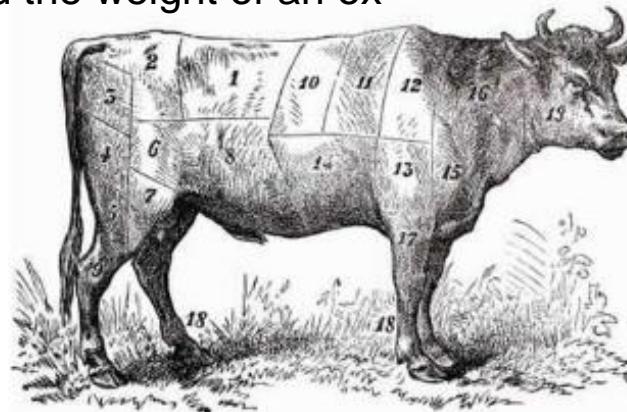
Outline

- Motivation and problem statement
- Part I: Combining crowdsourced labels
- Part II: End-to-end (E2E) learning with crowdsourced labels
- Part III: Other aspects of crowdsourcing
- Conclusions and open issues

The wisdom of crowds

- ❑ **The parable of the ox** (Sir Francis Galton, 1906)

- 787 people guessed the weight of an ox



- Average crowd guess: **1,197 pounds** - True weight: **1,198 pounds!**

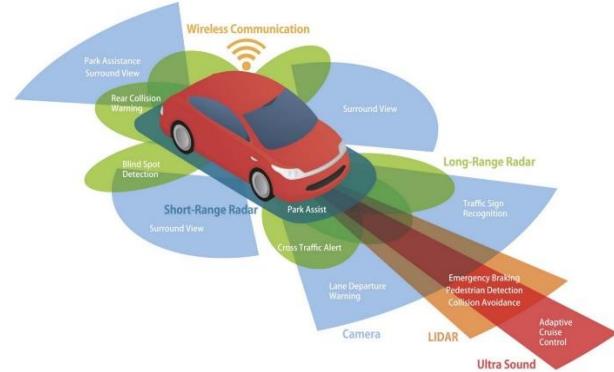
- ❑ **Who wants to be a millionaire** – Ask the audience

- Q.** Can we harness this wisdom in a principled way?



Fusing data versus fusing decisions

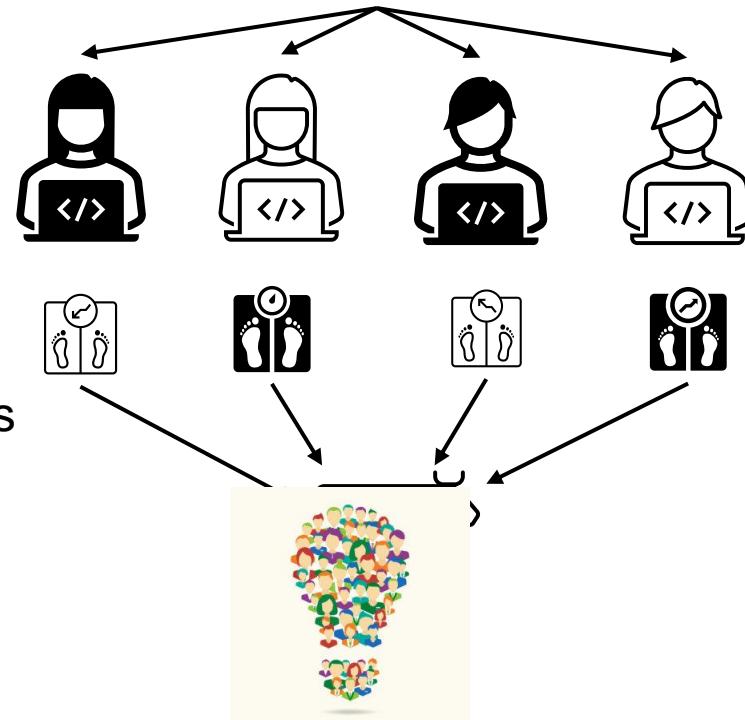
- ❑ Data fusion is costly [Waltz-Llinas'90, Mitchell'07]



- ❑ Distributed (generally noisy) decisions [Tsitsiklis'89]

- ❑ **Crowdsourcing vs hierarchical learning**

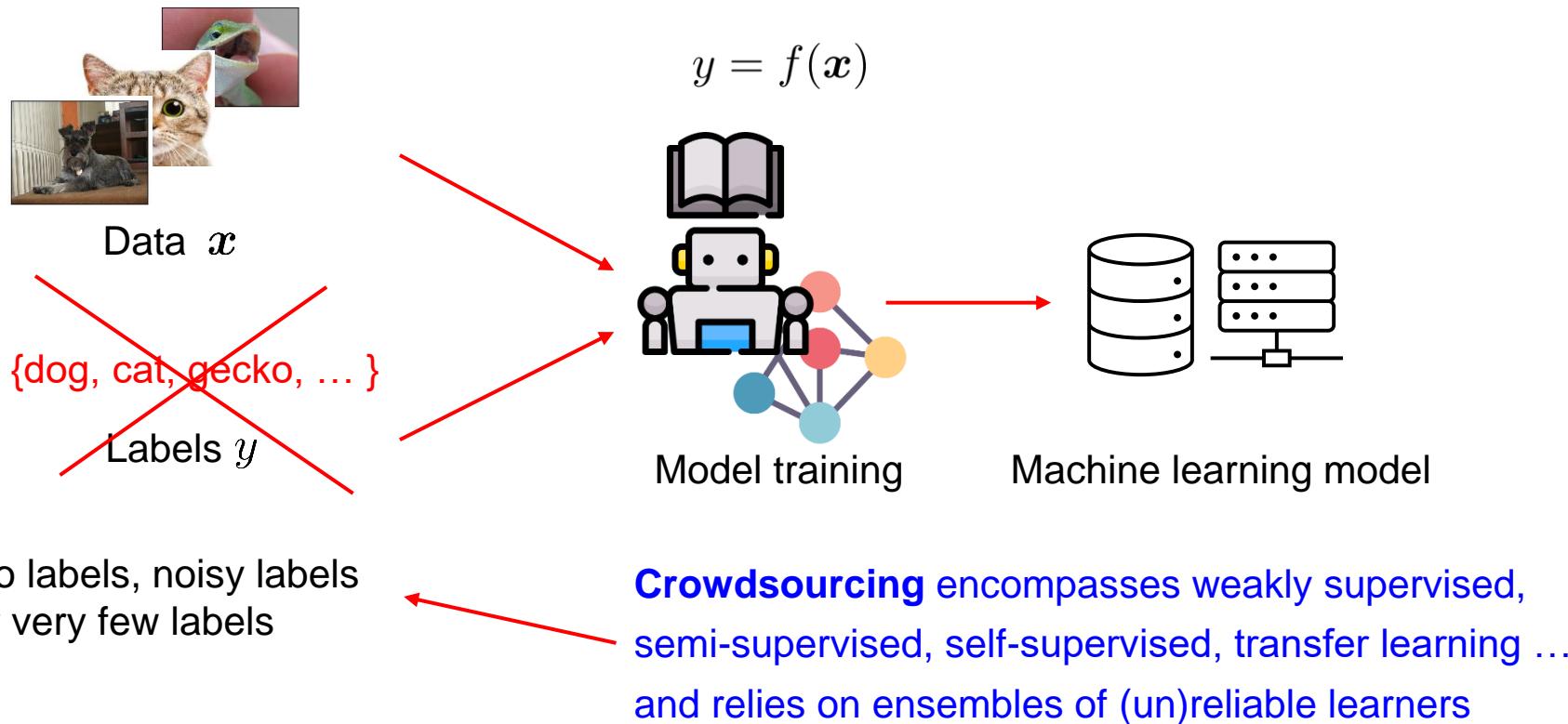
- Can benefit from the wisdom of humans
- Fuses (un) reliable decisions from annotators



- ❑ **Redundancy** is key to discover truth and assess annotator reliability!

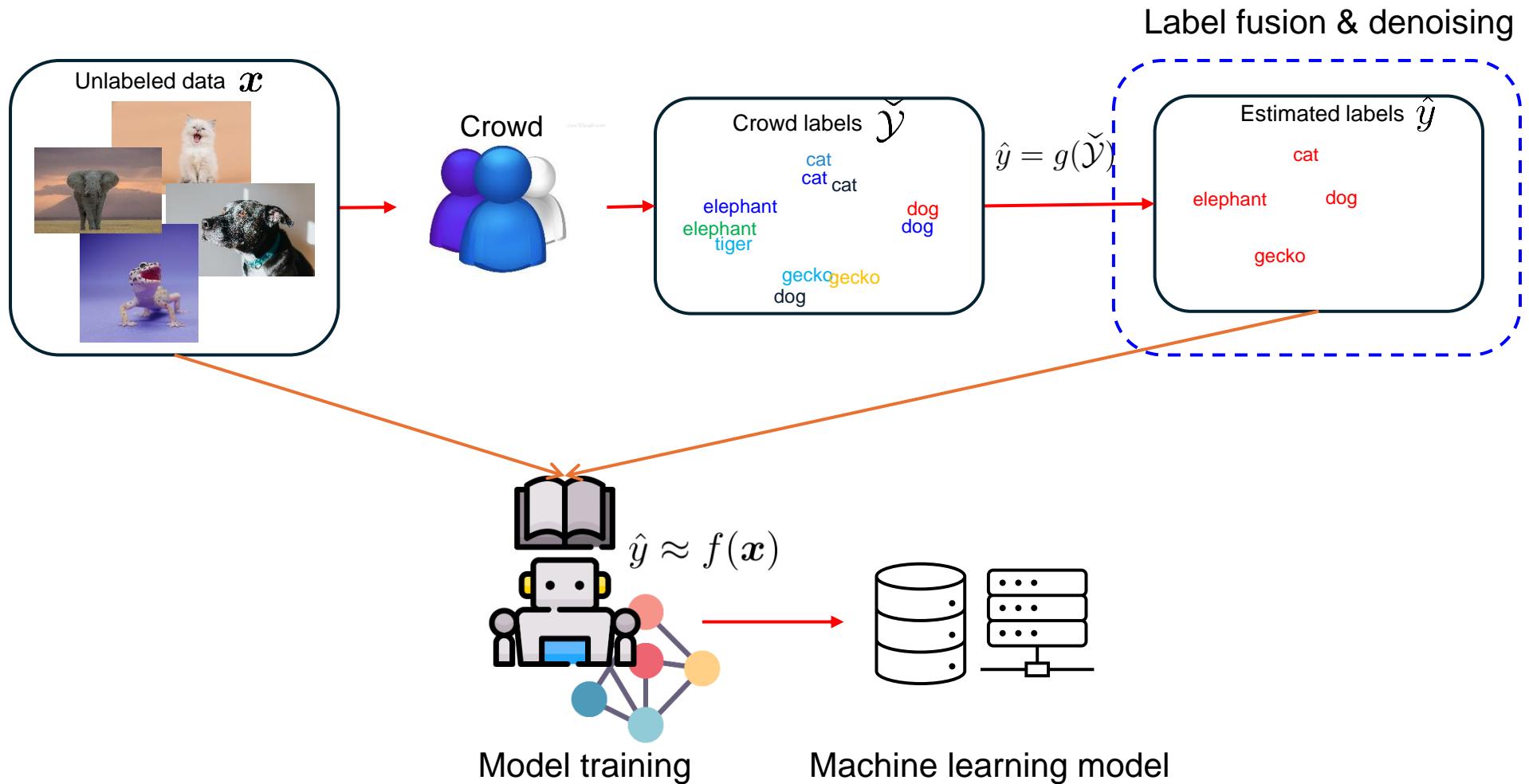
Crowdsourcing in the learning context

□ Machine learning pipeline

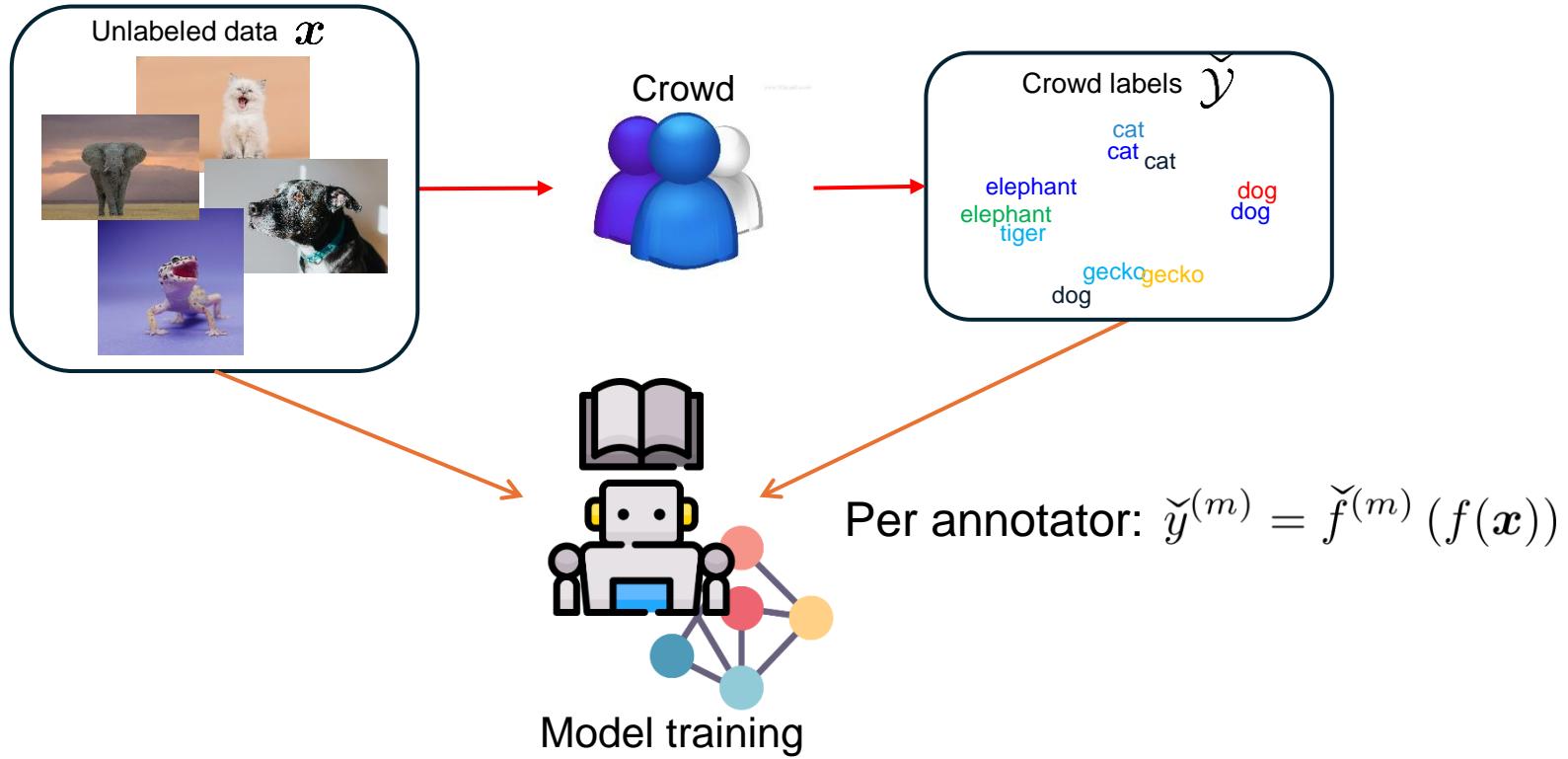


□ Two major paradigms to **learn from crowds**

Two-stage: Label denoising followed by learning

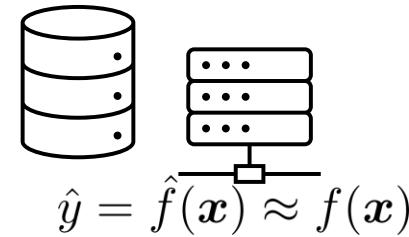


Learning with noisy labels – End-to-end learning



Testing phase

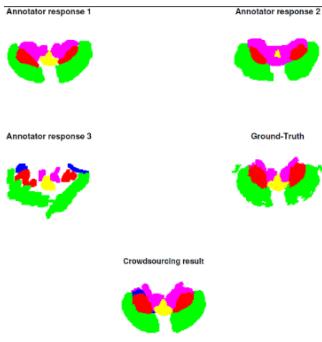
$$\mathbf{x}_{\text{test}} \rightarrow \hat{f}(\mathbf{x}_{\text{test}}) \rightarrow \mathbf{y}_{\text{test}}$$



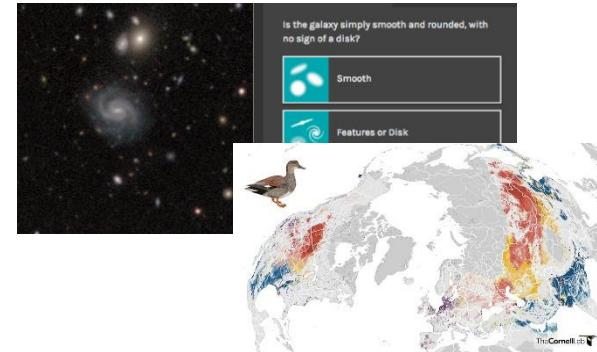
Machine learning model

Crowdsourcing across disciplines

□ Medical image analysis



□ Citizen Science



□ Crowdsensing in 5G and 6G



□ Facilitated via commercial services



Crowdsourcing: “Hidden” catalyst behind AI tools

Send us a Tip! | Shop | Subscribe

We may earn a comm

GIZMODO

The Future Is Here

Search Q

HOME LATEST NEWS GADGETS SCIENCE EARTHER IO9 AI SPACE EN ESPAÑOL VIDEO

ARTIFICIAL INTELLIGENCE

ChatGPT Is Powered by Human Contractors Getting Paid \$15 Per Hour

The well known chatbot is automated, but that automation is guided by low-paid human workers labelling data.

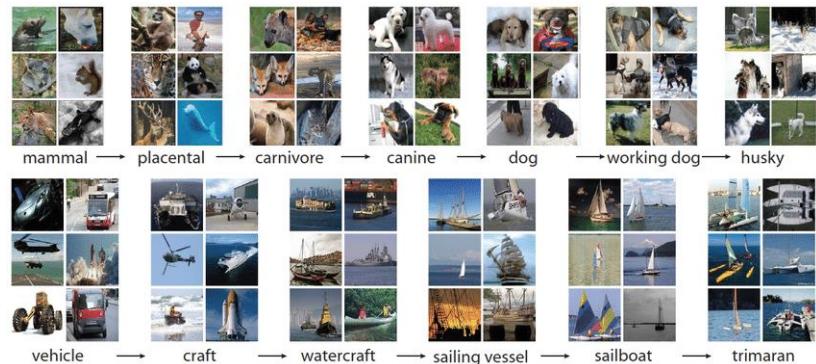
- ❑ Ukrainians used cell phone app to spot deadly Russian drones and missiles in Crimea



Opportunities and challenges

□ Opportunities

- From unlabeled to labeled datasets
- No expert supervision needed
- Low-cost and efficient learning
- LLM alignment



[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

□ Challenges

- Lack of ground-truth labels
- Annotators can be unreliable and even adversarial
- Sparsity of responses per annotator

□ This tutorial:

- Crowdsourcing from signal processing, and machine learning points of view
- Statistical learning framework

Crowdsourcing for classification

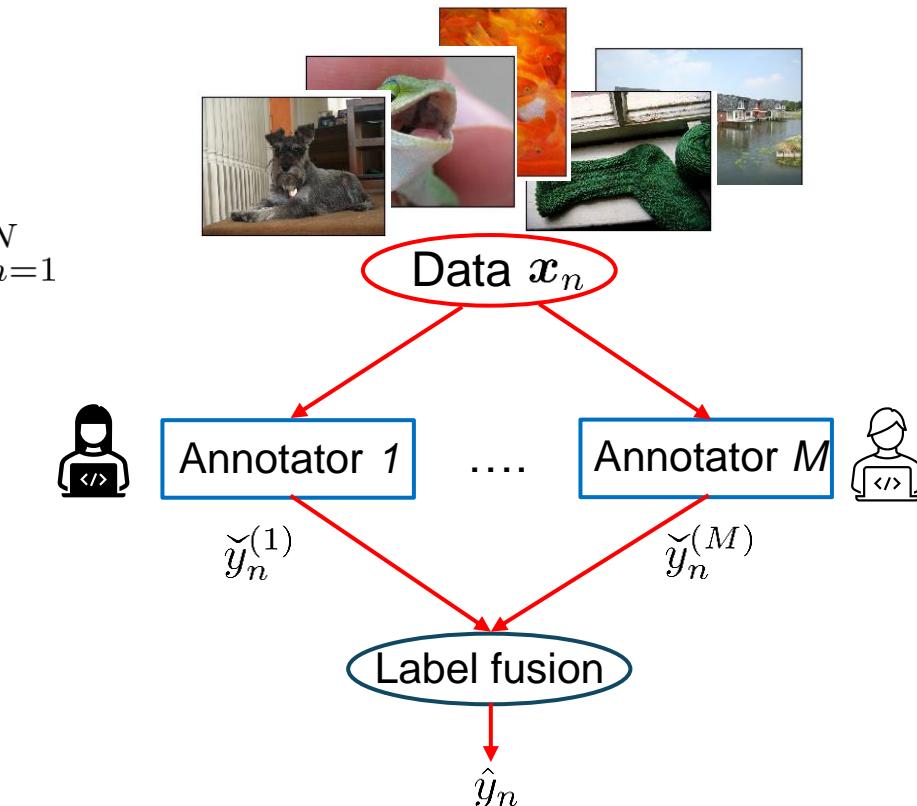
❑ N data $\{(x_n, y_n)\}_{n=1}^N$, K classes

❑ M annotators observe **subsets** of $\{x_n\}_{n=1}^N$

❑ Annotators provide **noisy labels**

$$\{\tilde{y}_n^{(m)}\}_{n,m=1}^{N,M}$$

Goal: Given $\{\tilde{y}_n^{(m)}\}_{n,m=1}^{N,M}$, find $\{\hat{y}_n\}_{n=1}^N$



Q1. Which annotators are reliable? **Q2.** How can we combine (un)reliable answers?

❑ Inference of labels is sought in a “**noisy**” unsupervised setting!

Outline

□ Motivation and problem statement

□ **Part I: Combining crowdsourced labels**

- Majority voting and annotator models
- Estimating model parameters via Expectation-Maximization (EM)
- Moment matching methods
- Identifiability

□ **Part II: End-to-end (E2E) learning with crowdsourced labels**

□ **Part III: Other aspects of crowdsourcing**

□ **Conclusions and open issues**

Majority voting or averaging

- Simplest solution to crowdsourcing
 - Per datum n , pick most common answer

$$\hat{y}_n = \arg \max_{k \in \{1, \dots, K\}} \sum_{m=1}^M \mathbb{I}[\check{y}_n^{(m)} = k]$$



Caveat: Majority voting (MV) assumes all annotators are “the same” and “independent”

- Q.** Can we do better than MV? **A.** Yes, by modeling annotator behavior!
- Simple extension: **weighted** MV

$$\hat{y}_n = \arg \max_{k \in \{1, \dots, K\}} \sum_{m=1}^M w^{(m)} \mathbb{I}[\check{y}_n^{(m)} = k]$$

Q. How to find weights?

Annotator model with independent responses

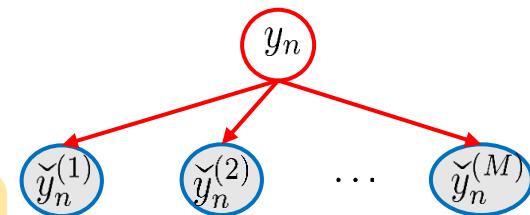
- Dawid-Skene (DS) data model $\{(x_n, y_n)\}_{n=1}^N \stackrel{\text{i.i.d.}}{\sim} \mathcal{P}$
- Maximum a posteriori (MAP) fusion of annotator labels $\check{\mathcal{Y}}$

$$\hat{y}_n = \arg \max_{k \in \{1, \dots, K\}} \Pr(y_n = k | \check{\mathcal{Y}}) = \arg \max_{k \in \{1, \dots, K\}} \Pr\left(\check{y}_n^{(1)} = k_1, \dots, \check{y}_n^{(M)} = k_M | y_n = k\right) \Pr(y_n = k)$$

(as1) Given ground-truth label y_n , annotator responses $\{\check{y}_n^{(m)}\}_{m=1}^M$ are independent

$$\Pr\left(\check{y}_n^{(1)} = k_1, \dots, \check{y}_n^{(M)} = k_M | y_n = k\right) = \prod_{m=1}^M \Pr\left(\check{y}_n^{(m)} = k_m | y_n = k\right)$$

$$\hat{y}_n = \arg \max_{k \in \{1, \dots, K\}} \log \Pr(y_n = k) + \sum_{m=1}^M \sum_{k'=1}^K \log \Pr\left(\check{y}_n^{(m)} = k' | y_n = k\right) \mathbb{I}[\check{y}_n^{(m)} = k']$$



Theorem. Under **(as1)** there are constants $\alpha, \beta > 0$ so that the error probability of the MAP classifier satisfies

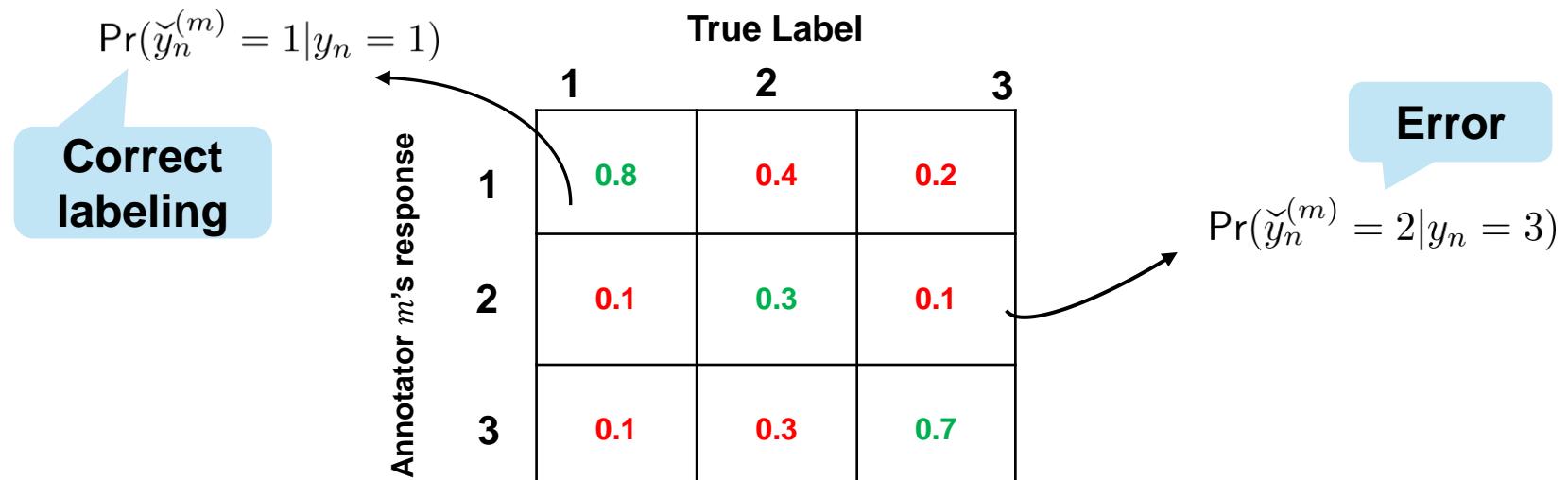
$$\mathcal{P}_e \leq \alpha e^{-M\beta}$$

- Performance

Parametric model

- MAP classifier needs $d(k) := \Pr(y_n = k)$ and $A_m(k', k) := \Pr(\tilde{y}_n^{(m)} = k' | y_n = k)$

- Confusion matrices $A_m \in \mathbb{R}^{K \times K}$, $m = 1, \dots, M$ $A_m := [\mathbf{a}_{m,1}, \dots, \mathbf{a}_{m,K}]$



- Ideal annotator $A_m = I$
- Model parameters $\{A_m\}_{m=1}^M, d$ must be estimated from annotator responses $\tilde{\mathcal{Y}}$
- d lies on a $K-1$ simplex; and likewise $\{\mathbf{a}_{m,k}\}_{k=1,m=1}^{K,M}$ lie on $K-1$ simplexes
 - Number of parameters: $d : K - 1$ and $A_m : K(K - 1)$ Total: $\mathcal{O}(MK^2)$

Alternative models

- For $K=2$, we need to estimate 2 parameters per annotator m
 - Sensitivity $A_m(1, 1) = \Pr(\check{y}_n^{(m)} = 1 | y_n = 1)$
 - Specificity $A_m(2, 2) = \Pr(\check{y}_n^{(m)} = 2 | y_n = 2)$
- Parsimonious models of $\{A_m\}_{m=1}^M$
 - One parameter (one-coin) or homogeneous D-S (one unknown per annotator m)
$$A_m = \left(w^{(m)} - \frac{1 - w^{(m)}}{K - 1} \right) \mathbf{I} + \frac{1 - w^{(m)}}{K - 1} \mathbf{1}\mathbf{1}^\top, \quad 0 \leq w^{(m)} \leq 1$$
 - Confusion vector $w^{(m)} := [w_1^{(m)} \dots w_K^{(m)}]^\top \quad w_k^{(m)} := \Pr(\check{y}_n^{(m)} = k | y_n = k)$
$$A_m(k, k') = \begin{cases} w_k^{(m)} & k = k' \\ \frac{1 - w_k^{(m)}}{K - 1} & k \neq k' \end{cases}$$
 - Significant reduction in the number of parameters $\mathcal{O}(M)$ and $\mathcal{O}(MK)$, respectively
 - More parsimonious however, implies less expressive modeling
- Item difficulty can be incorporated [Whitehill et al'09]

Outline

□ Motivation and problem statement

□ **Part I: Combining crowdsourced labels**

- Majority voting and annotator models
- Estimating model parameters via Expectation-Maximization (EM)
- Moment matching methods
- Identifiability

□ **Part II: End-to-end (E2E) learning with crowdsourced labels**

□ **Part III: Other aspects of crowdsourcing**

□ **Conclusions and open issues**

Estimating model parameters via EM

- Popular for maximum likelihood estimation of parametric mixture models [Dempster-Laird-Rubin'77]
 - Observations $\check{\mathcal{Y}}$
 - N latent true labels \mathbf{y}
 - Wanted parameters $\psi := [\{\mathbf{A}_m\}_{m=1}^M, \mathbf{d}]$

- Wish to maximize $L(\psi) = \log \Pr(\check{\mathcal{Y}}; \psi) = \log \left(\sum_{\mathbf{y}} \Pr(\mathbf{y}, \check{\mathcal{Y}}; \psi) \right)$ Typically intractable

- Use a tractable surrogate pdf q to maximize an expected lower bound with a two-step **iteration**

$$\mathbb{E}_{q(\mathbf{y})} \left[\log \Pr \left(\mathbf{y}, \check{\mathcal{Y}}; \psi \right) \right] = \log \Pr(\check{\mathcal{Y}}; \psi) - D_{KL} \left(q(\mathbf{y}) || \Pr(\mathbf{y} | \check{\mathcal{Y}}; \psi) \right)$$

S1. Expectation (E-)step $Q(\psi; \psi^i) = \mathbb{E}_{\mathbf{y} | \check{\mathcal{Y}}; \psi^i} [\log \Pr(\mathbf{y}, \check{\mathcal{Y}}; \psi)]$

Latent variables updated given observed ones and parameters

S2. Maximization (M-)step $\psi^{i+1} = \arg \max_{\psi} Q(\psi; \psi^i)$

Parameters updated given estimated latent variables

- Nondecreasing sequence of $L(\psi)$'s converges at least to a stationary point – Initialization is critical

EM for crowdsourcing

S0. Initialize $\psi^0 := [d^0, A_1^0, \dots, A_M^0]$

S1. E-step

$$Q(\psi; \psi^i) = \mathbb{E}_{\mathcal{Y} \sim \Pr(\mathcal{Y}|\check{\mathcal{Y}}; \psi^t)} [\log \Pr(\check{\mathcal{Y}}, \mathcal{Y}; \psi)] = \sum_{n=1}^N \sum_{k=1}^K q(y_n = k; \psi^i) \log \Pr(y_n = k, \check{y}_n^{(1)}, \dots, \check{y}_n^{(M)}; \psi)$$

$$q(y_n = k; \psi^i) := \Pr(y_n = k | \check{y}_n^{(1)}, \dots, \check{y}_n^{(M)}; \psi^t) = \frac{\exp \left(\log d^i(k) + \sum_{m=1}^M \sum_{k'=1}^K \log A_m^i(k', k) \mathbb{I}[\check{y}_n^m = k'] \right)}{Z}$$

Bayes rule

S2. M-step

$$\psi^{i+1} = \arg \max_{\psi} Q(\psi; \psi^i) \Rightarrow$$

$A_m^{i+1}(k', k) = \frac{\sum_{n=1}^N q(y_n = k; \psi^i) \mathbb{I}[\check{y}_n^m = k']}{\sum_{k''=1}^K \sum_{n=1}^N q(y_n = k; \psi^i) \mathbb{I}[\check{y}_n^m = k'']} \quad \forall m, k', k$
 $d^{i+1}(k) = \frac{\sum_{n=1}^N q(y_n = k; \psi^i)}{\sum_{k'=1}^K \sum_{n=1}^N q(y_n = k'; \psi^i)} \quad \forall k$

Theorem. If M and N sufficiently large, and EM is **properly initialized** then correct labels y_n are selected w.h.p.

How about prior information?

Bayesian approaches view \mathbf{d}, \mathbf{A} as random, and rely on Gibbs sampling, variational inference (VI)

- Iterative solvers guarantee local optimality – global only with reliable initialization
- Both EM and Bayesian approaches can account for a few ground-truth labels, if available

Majority voting (MV)
simple, but reliable?

$$\hat{y}_n = \arg \max_{k \in \{1, \dots, K\}} \sum_{m=1}^M \mathbb{I}(\check{y}_n^{(m)} = k)$$

Initializing reliability
parameters, but
performance?

$$\mathbf{d}^0(k) = \frac{\sum_{m=1}^M \sum_{n=1}^N \mathbb{I}(\check{y}_n^{(m)} = k)}{N_1 + \dots + N_M}, \quad k = 1, \dots, K$$
$$\mathbf{A}_m^0(k, k') = \frac{\sum_{n=1}^N \mathbb{I}(\check{y}_n^{(m)} = k', \hat{y}_n = k)}{N_m}, \quad \forall m, k', k$$

Resemble
histogram
estimates

- MV offers simple initialization, but how about nonuniformly unreliable annotators?

Outline

□ Motivation and problem statement

□ **Part I: Combining crowdsourced labels**

- Majority voting and annotator models
- Estimating model parameters via Expectation-Maximization (EM)
- Moment matching methods
- Identifiability

□ **Part II: End-to-end (E2E) learning with crowdsourced labels**

□ **Part III: Other aspects of crowdsourcing**

□ **Conclusions and open issues**

Estimating parameters from annotator statistics

Conditionally independent
annotators

□ Pairwise statistics of annotator responses

$$\begin{aligned} \mathbf{R}_{m,m'}(k_m, k_{m'}) &:= \Pr(\check{y}_n^{(m)} = k_m, \check{y}_n^{(m')} = k_{m'}) \\ &= \sum_{k=1}^K \Pr(y_n = k) \Pr(\check{y}_n^{(m)} = k_m | y_n = k) \Pr(\check{y}_n^{(m')} = k_{m'} | y_n = k) \\ &\quad \quad \quad \mathbf{A}_m(k_m, k) \qquad \qquad \qquad \mathbf{A}_{m'}(k_{m'}, k) \end{aligned}$$

$$\mathbf{R}_{m,m'} = \mathbf{A}_m \mathbf{D} \mathbf{A}_{m'}^\top, \mathbf{D} = \text{diag}(\mathbf{d})$$

□ Triplet statistics of annotator responses

$$\begin{aligned} \underline{\mathbf{T}}_{m,m',m''}(k_m, k_{m'}, k_{m''}) &:= \Pr(\check{y}_n^{(m)} = k_m, \check{y}_n^{(m')} = k_{m'}, \check{y}_n^{(m'')} = k_{m''}) \\ &= \sum_{k=1}^K \Pr(y_n = k) \Pr(\check{y}_n^{(m)} = k_m | y_n = k) \Pr(\check{y}_n^{(m')} = k_{m'} | y_n = k) \Pr(\check{y}_n^{(m'')} = k_{m''} | y_n = k) \\ \underline{\mathbf{T}}_{m,m'm''} &= \sum_{k=1}^K \mathbf{d}(k) \mathbf{A}_m(:, k) \circ \mathbf{A}_{m'}(:, k) \circ \mathbf{A}_{m''}(:, k) = [[\mathbf{A}_m \mathbf{D}, \mathbf{A}_{m'}, \mathbf{A}_{m''}]] \end{aligned}$$

PARAFAC/CPD tensor

$\{\mathbf{A}_m\}_{m=1}^M, \mathbf{d}$ recoverable from statistics of annotator responses via **moment matching**

Estimating annotator statistics

□ Estimates can be readily computed from \check{y}

- M first-order moments

$$\hat{\mu}_m(k) = \frac{1}{N_m} \sum_{n=1}^N \mathbb{I}[\check{y}_n^{(m)} = k]$$

no. of responses
from annotator m

- $\binom{M}{2}$ second-order moments

$$\hat{R}_{m,m'}(k_m, k_{m'}) = \frac{1}{N_{m,m'}} \sum_{n=1}^N \mathbb{I}[\check{y}_n^{(m)} = k_m, \check{y}_n^{(m')} = k_{m'}]$$

- $\binom{M}{3}$ third-order moments

$$\hat{T}_{m,m',m''}(k_m, k_{m'}, k_{m''}) = \frac{1}{N_{m,m',m''}} \sum_{n=1}^N \mathbb{I}[\check{y}_n^{(m)} = k_m, \check{y}_n^{(m')} = k_{m'}, \check{y}_n^{(m'')} = k_{m''}]$$

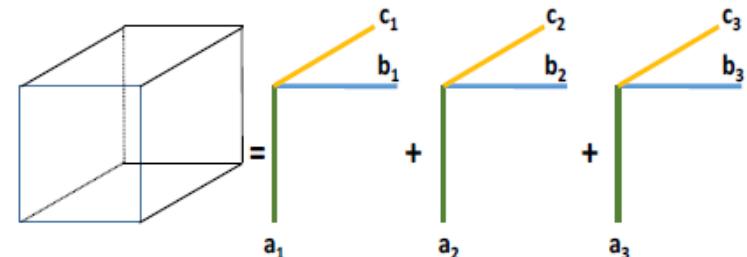
- Sparse annotator responses
- Some moments not available

Heteroskedastic noise

Canonical Polyadic Decomposition in one slide

- Data tensor $\underline{T} : I \times J \times N$

- PARAFAC/CPD Tensor model: $\underline{T} = \sum_{f=1}^F \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f$



$$\mathbf{A} := [\mathbf{a}_1, \dots, \mathbf{a}_F] : I \times F \quad \mathbf{B} := [\mathbf{b}_1, \dots, \mathbf{b}_F] : J \times F \quad \mathbf{C} := [\mathbf{c}_1, \dots, \mathbf{c}_F] : N \times F$$

- $\mathbf{A}, \mathbf{B}, \mathbf{C}$ essentially unique

- Recoverable up to scaling and common permutation ambiguities

$$\hat{\mathbf{A}} = \mathbf{A}\boldsymbol{\Pi}\boldsymbol{\Lambda}_1, \hat{\mathbf{B}} = \mathbf{B}\boldsymbol{\Pi}\boldsymbol{\Lambda}_2, \hat{\mathbf{C}} = \mathbf{C}\boldsymbol{\Pi}\boldsymbol{\Lambda}_3, \quad \boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_3 = \mathbf{I}$$

permutation matrix

$$\min_{\mathbf{A}, \mathbf{B}, \mathbf{C}} \left\| \underline{T} - \sum_{f=1}^F \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f \right\|_F^2$$

NP-hard!

- Solved via alternating optimization / Gradient descent (GD/SGD) etc.
- Off-the-shelf solvers: *N-way toolbox*, *Tensorlab*, *AO-ADMM*, *SPLATT*

Moment matching with third-order statistics

Available triplet moments

Coupled CPD

$$\text{find } \{\mathbf{A}_m\}_{m=1}^M, \mathbf{d}$$

$$\text{s.t. } \hat{\mathbf{T}}_{m,m',m''} = [[\mathbf{A}_m \mathbf{D}, \mathbf{A}_m, \mathbf{A}_{m''}]], \quad (m, m', m'') \in \Omega,$$

$$\mathbf{A}_m \geq \mathbf{0}, \mathbf{1}^\top \mathbf{A}_m = \mathbf{1}^\top, \mathbf{1}^\top \mathbf{d} = 1, \mathbf{d} \geq \mathbf{0}.$$

Sample triplet moments

□ Relaxation

$$\min \sum_{(m,m',m'') \in \Omega} \|\hat{\mathbf{T}}_{m,m',m''} - [[\mathbf{A}_m \mathbf{D}, \mathbf{A}_m, \mathbf{A}_{m''}]]\|_F^2$$

$$\text{s.t. } \mathbf{A}_m \geq \mathbf{0}, \mathbf{1}^\top \mathbf{A}_m = \mathbf{1}^\top, \mathbf{1}^\top \mathbf{d} = 1, \mathbf{d} \geq \mathbf{0}.$$

□ First- and second-order moments can be included

Theorem. If solution \mathcal{S}^* relies on ensemble moments, and solution \mathcal{S}^N on sample moments, then as $N \rightarrow \infty$

$$\mathcal{D}(\mathcal{S}^*, \mathcal{S}^N) \rightarrow 0 \quad \text{almost surely.}$$

➤ Reduced complexity via orthogonal tensor decompositions [Zhang et al'16]

Confusion matrix identifiability

Q. Can we identify $\{A_m\}_{m=1}^M, d$ from $T_{m,m',m''}$, $(m, m', m'') \in \Omega$?

□ Coupled tensor factorization inherits CPD identifiability guarantees

- Recovery of $\{A_m\}_{m=1}^M, d$ up to common permutation ambiguity
- ≥ 3 full rank A_m 's required
- Annotators have to **co-label** “sufficient” samples

□ Additional assumption required to resolve permutation ambiguity

(as) Most annotators are “better than random”

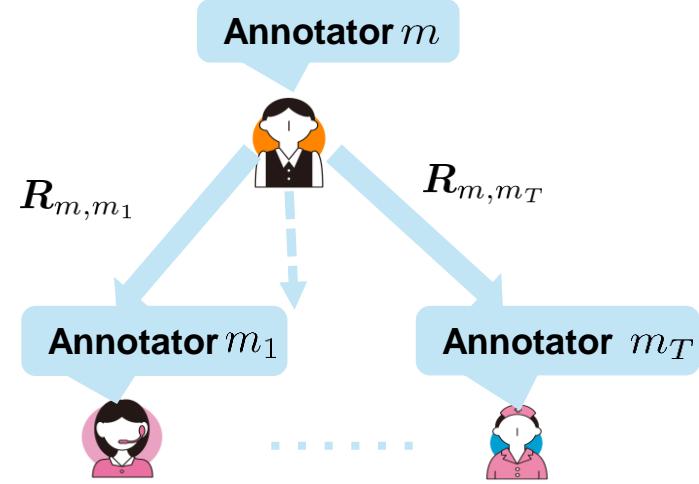
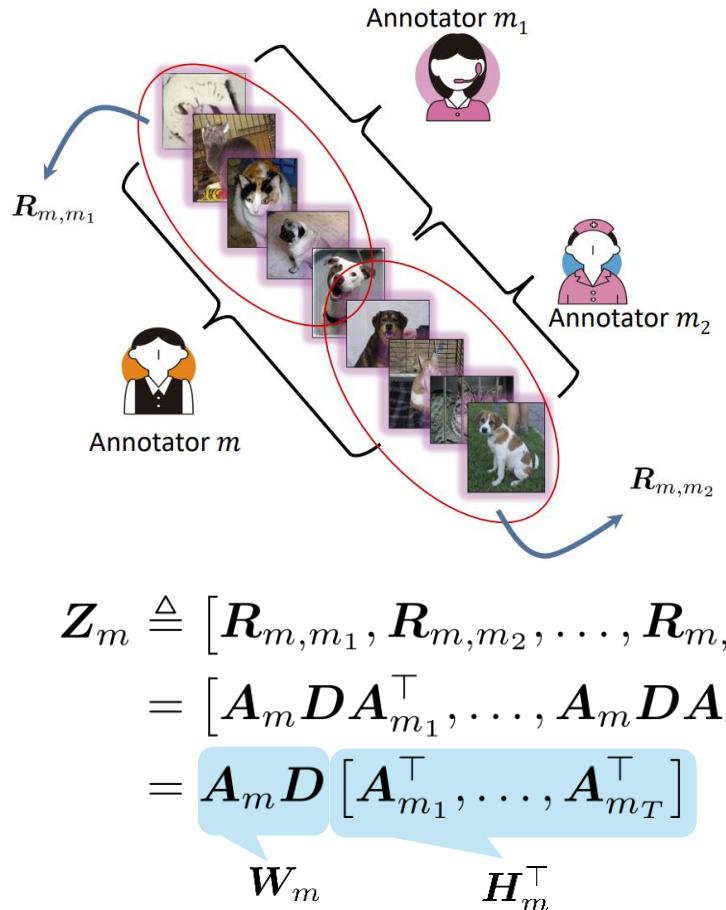
Diagonally dominant A_m

□ **Caveats**

- Complexity increases with M
- Accurate estimation of third order moments requires a lot of data

Learning from Pairwise Statistics

Assume that annotator m co-labels with annotators m_1, \dots, m_T



Stack all available pairwise statistics

$$Z_m = W_m H_m^\top, \forall m$$

Z_m observed, but W_m and H_m are unknown

Nonnegative Matrix Factorization (NMF) model

$$\mathbf{Z}_m = \mathbf{W}_m \mathbf{H}_m^\top, \forall m$$

- ❑ Generally not unique
- ❑ Identifiability can be established under some conditions
 - Separability
 - Sufficiently scattered condition (SSC)

$$\mathbf{Z}_m = \mathbf{W}_m Q Q^{-1} \mathbf{H}_m^\top$$

\mathbf{W}_m^* $\mathbf{H}_m^{*\top}$

identifiability challenge

Separability

$$\mathbf{H}_m = \begin{bmatrix} \mathbf{A}_{m_1} \\ \vdots \\ \mathbf{A}_{m_t} \\ \vdots \\ \mathbf{A}_{m_T} \end{bmatrix}$$

The diagram shows the matrix \mathbf{H}_m as a vertical stack of horizontal vectors $\mathbf{A}_{m_1}, \dots, \mathbf{A}_{m_t}, \dots, \mathbf{A}_{m_T}$. Two dashed arrows point from the bottom row \mathbf{A}_{m_T} to the right, both labeled I_K , indicating that the bottom row is a unit vector.

Suppose there exists indices Λ such that

$$\mathbf{H}_m(\Lambda, :) = I_K$$

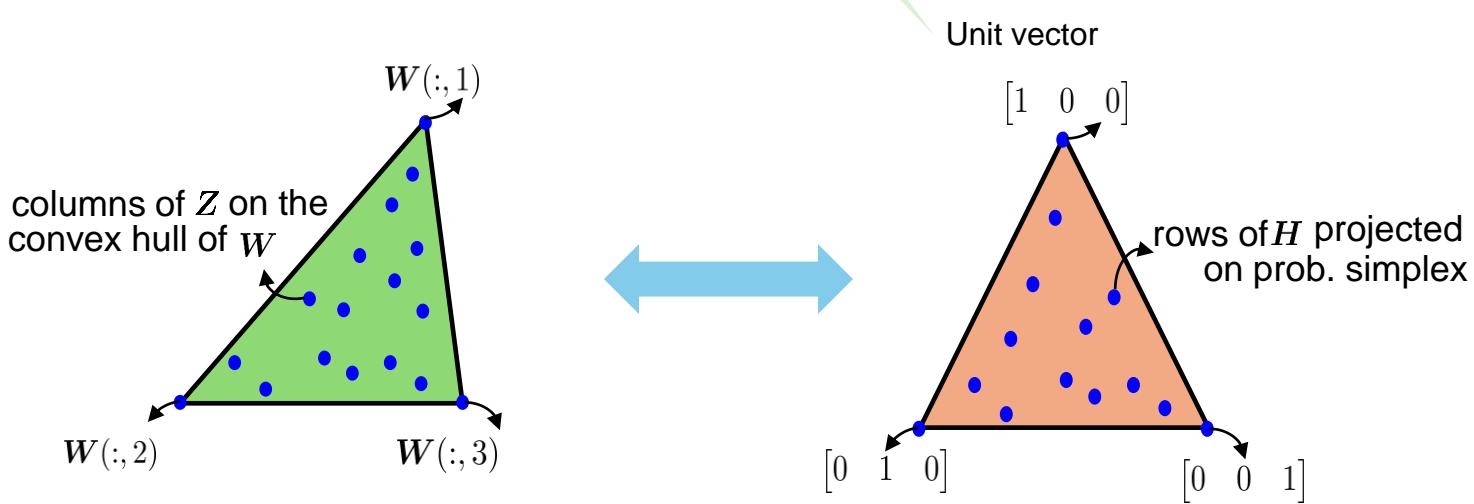
Identity Matrix

Separability in NMF

- Separability of H : H has a diagonal submatrix [Donoho and Stodden'03]

- $\exists \Lambda = \{\ell_1, \dots, \ell_K\}$ such that

$$Z = WH^\top; \quad H(\ell_i, :) = \alpha_i e_i, \quad \alpha_i > 0, \quad i = 1, \dots, K$$

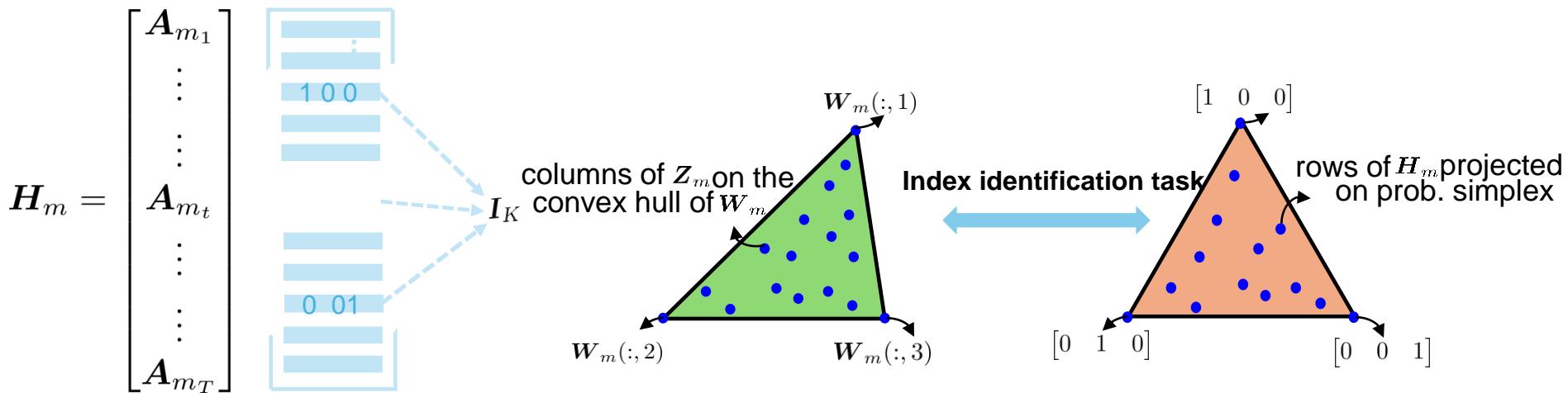


- Solvable via Gram-Schmidt-like algorithms
 - successive projection algorithm (SPA)
[Araújo et al'01, Gillis and Vavasis'14, Fu et al'15]
- Intuition: **taller** H yields higher chance of separability holding

Identifiability Result

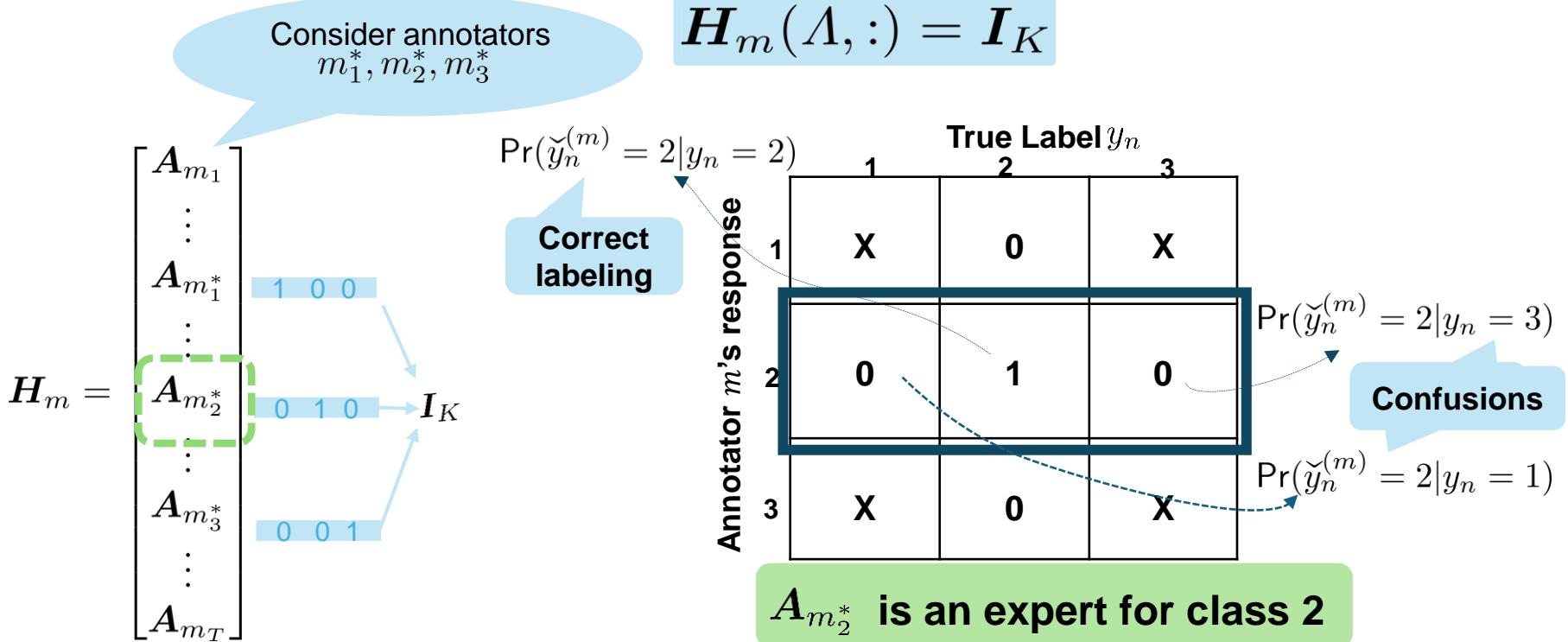
Theorem. If H_m satisfies separability, there exists a greedy algorithm to recover confusion matrices from $Z_m = W_m H_m^\top$.

Separability



- Equivalent to estimating vertices of simplex formed by columns of Z_m

A Closer Look at Separability



Specialized annotators exist \rightarrow separability holds

- Approximate specialized annotators exist \rightarrow Near-separability holds $H_m(\Lambda, :) \approx I_K$
 - Confusion matrix identified using K Gram-Schmidt iterations
 - MultiSPA algorithm [Ibrahim et al'19]

Identifiability Result

Q. Do more annotators help?

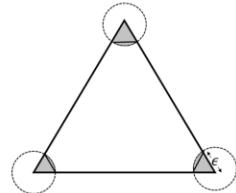
Proposition 1: Let $\rho > 0, \varepsilon > 0$, and assume that the rows of $\bar{\mathbf{H}}_m$ (row normalized version of \mathbf{H}_m) are generated within the $(K - 1)$ -probability simplex uniformly at random. If the number of annotators satisfies

$$M = \Omega\left(\frac{\varepsilon^{-2(K-1)}}{K} \log\left(\frac{K}{\rho}\right)\right),$$

then, with probability greater than or equal to $1 - \rho$, there exist rows of $\bar{\mathbf{H}}_m$ indexed by q_1, \dots, q_K such that

$$\|\bar{\mathbf{H}}_m(q_k, :) - \mathbf{e}_k^\top\|_2 \leq \varepsilon, \quad k = 1, \dots, K.$$

- Larger M : more rows in $\bar{\mathbf{H}}_m$
- **Takeaway: Large number of annotators assists identifiability**



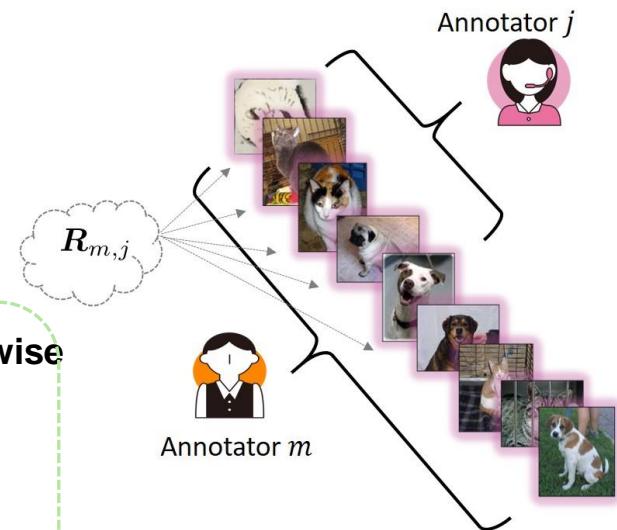
Enhanced Identifiability Guaranteed Approach

- For some annotators $Z_m = W_m H_m^\top$ may not be identifiable

Coupled NMF

$$\begin{aligned} & \text{find } \{A_m\}_{m=1}^M, d \\ & \text{s.t. } \hat{R}_{m,m'} = A_m D A_{m'}^\top, (m, m') \in \Omega, \\ & \quad A_m \geq 0, 1^\top A_m = 1^\top, 1^\top d = 1, d \geq 0. \end{aligned}$$

Available pairwise statistics



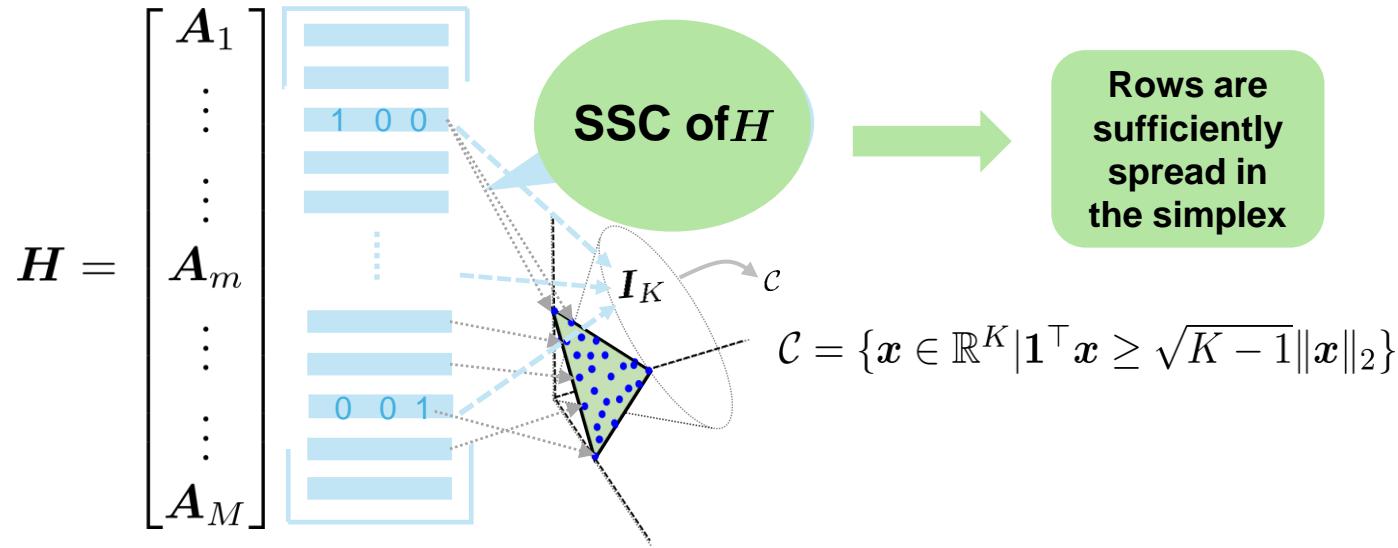
Idea: find a feasible solution that fits whatever data available

- Realized via minimize $\sum_{m,m' \in \Omega} \text{KL}(\hat{R}_{m,m'} || A_m D A_{m'}^\top)$

Q. Identifiability?

Towards more realistic conditions

- ❑ Sufficiently scattered condition (SSC) [Huang et al'14]



- ❑ SSC on H implies that the rows are “sufficiently spread”
 - A second-order cone \mathcal{C} can be contained inside conic hull spanned by H
$$\mathcal{C} \subseteq \text{cone}(H^\top)$$
- ❑ Implications for NMF $R = WH^\top$:
 - If both factors satisfy SSC, model identifiable up to permutation
 - Relaxed condition compared to separability; more rows help

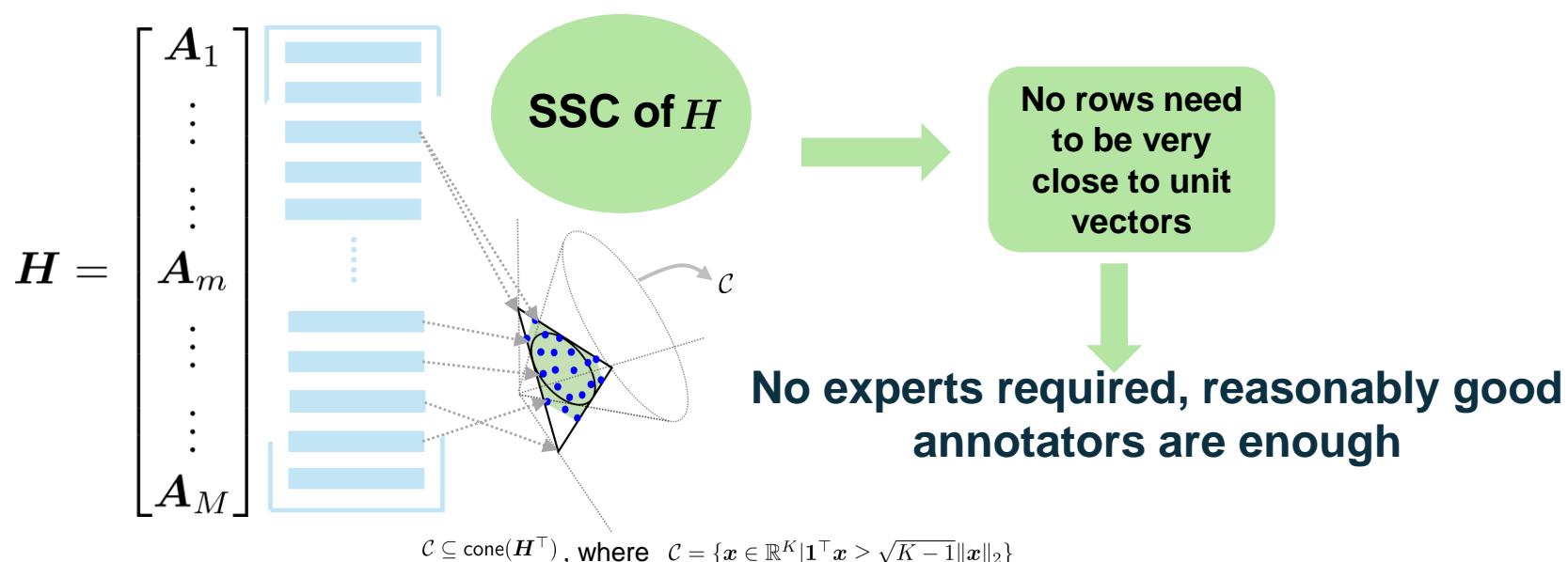
Identifiability via the SSC

Q. How does SSC improve the identifiability conditions?

Idea:

$$\begin{bmatrix} \mathbf{R}_{1,3} & \mathbf{R}_{1,4} \\ \mathbf{R}_{2,3} & \mathbf{R}_{2,4} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix}}_{\mathbf{H}^{(1)}} \mathbf{D} \underbrace{[\mathbf{A}_3^\top, \mathbf{A}_4^\top]}_{\mathbf{H}^{(2)\top}}$$

- If both matrices satisfy SSC, model is identifiable
- If $\mathbf{R}_{2,5} = \mathbf{A}_2 \mathbf{D} \mathbf{A}_5^\top$ is available, \mathbf{A}_5 identifiable via least squares ($\text{rank}(\mathbf{A}_2) = K$)



Towards More Realistic Conditions

Proposition

$$\begin{bmatrix} \mathbf{R}_{1,1} & \dots & \mathbf{R}_{1,M} \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{M,1} & \dots & \mathbf{R}_{M,M} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_M \end{bmatrix}}_H \mathbf{D} \underbrace{[\mathbf{A}_1^\top, \dots, \mathbf{A}_M^\top]}_{\mathbf{H}^\top}$$

Even if there are no class experts,
we can learn
the model parameters via

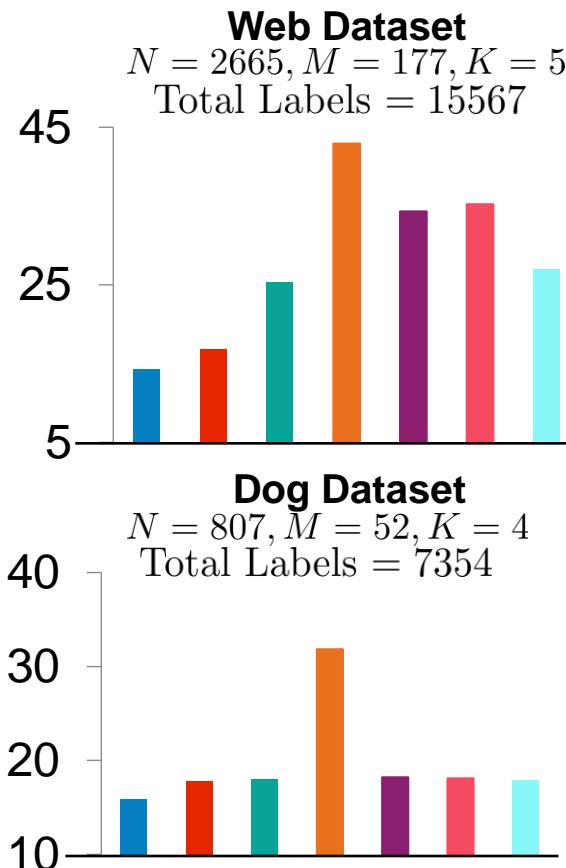
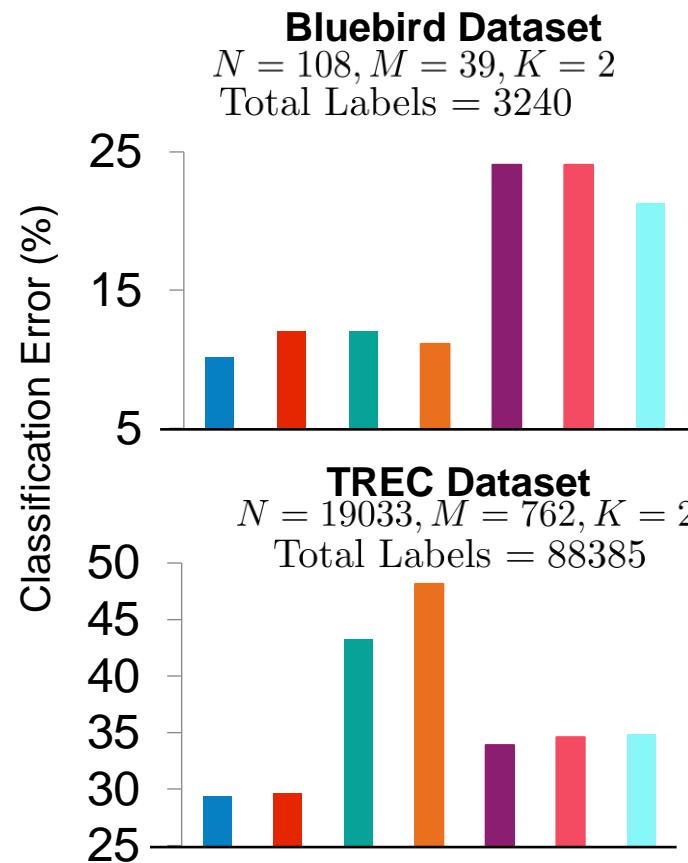
$$\begin{aligned} &\text{find } \{\mathbf{A}_m\}_{m=1}^M, \mathbf{d} \\ &\text{s.t. } \hat{\mathbf{R}}_{m,m'} = \mathbf{A}_m \mathbf{D} \mathbf{A}_{m'}^\top, (m, m') \in \Omega, \\ &\quad \mathbf{A}_m \geq \mathbf{0}, \mathbf{1}^\top \mathbf{A}_m = \mathbf{1}^\top, \mathbf{1}^\top \mathbf{d} = 1, \mathbf{d} \geq \mathbf{0}. \end{aligned}$$

Available
pairwise
statistics

- Missing blocks handled via imputation (matrix completion)
- Symmetric NMF problem
- Faster, convergence guaranteed algorithm

Empirical Results

Noisy labels from workers



- Model Identifiability
 - CNMF
 - Spectral-DS
 - TensorADMM
- No Model Identifiability
 - KOS
 - PG-TAC
 - CRIA
 - Majority Voting

Methods guaranteeing model identifiability consistently perform well

Other Algorithms

- For $K=2$, second-order moments (inter-annotator covariance) are enough!
 - Sensitivity and specificity estimated via eigendecomposition [Parisi et al'14, Jaffe et al'14]
- Second-order moments sufficient for one-coin model
 - [Karger et al'11, Ghosh et al'11, Dalvi et al'13, Ma et al'18]

Generalization. Constrained optimization with $\mathbf{w} = [w_1 \dots w_M]$ capturing annotator reliability

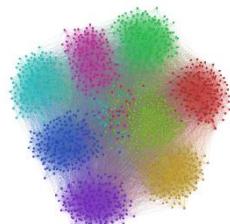
$$\min_{\mathbf{y}, \mathbf{w}} g(\mathbf{y}, \mathbf{w}) = \sum_{m=1}^M w_m \sum_{n=1}^N d(y_n, \check{y}_n^{(m)})$$

subject to $r(\mathbf{w}) = 1$

- d : loss; e.g., $d(y_n, \check{y}_n^{(m)}) = \mathbb{I}(y_n = \check{y}_n^{(m)})$
- $r(\mathbf{w})$: regularization; e.g., for sparsity
- alternating minimization

- Minimax conditional entropy [Zhou et al'12]
- Maximum margin MV [Tian-Zhu'15] - resembles SVM
- Data dependencies can be captured
 - Sequential data [Nguyen et al'17, Simpon-Gurevych'19, Sabetpour et al'21, Traganitis-Giannakis'22, Marinnan et al'24]
 - Graph data [Traganitis-Giannakis'22]

But Google ORG is starting from behind. The company made a late push into hardware, and Apple ORG's Siri PRODUCT, available on iPhones PRODUCT, and Amazon ORG's Alexa PRODUCT software, which runs on its Echo PRODUCT and Dot PRODUCT devices, have clear leads in consumer adoption.

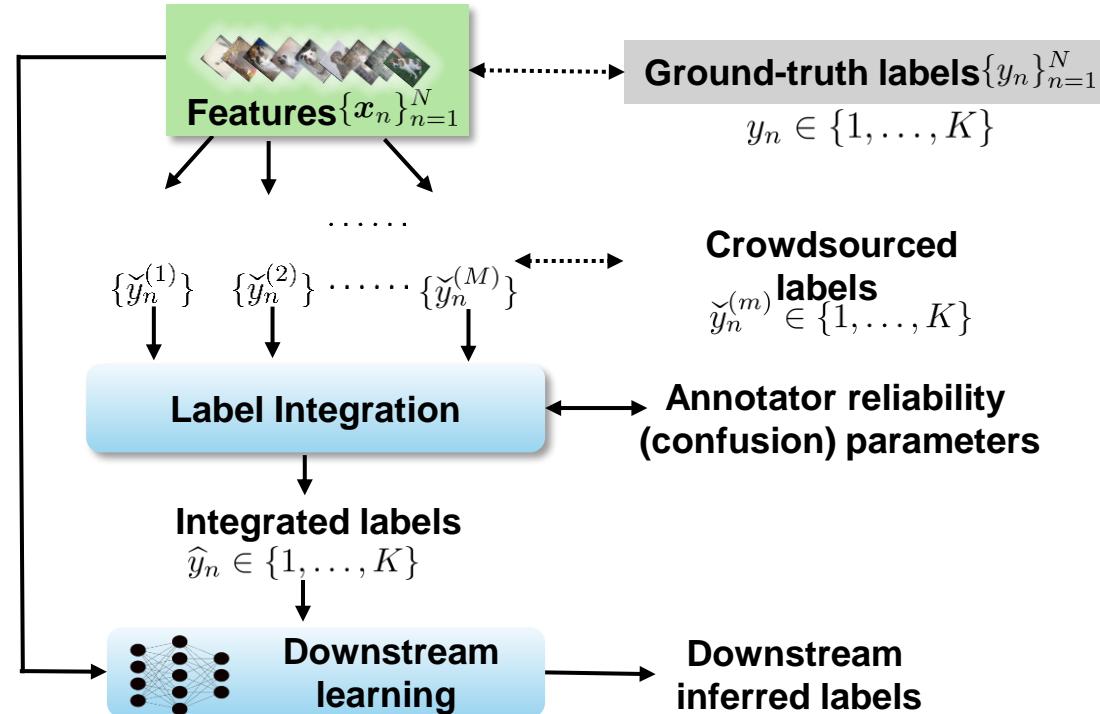


Two-stage Approach: Key Takeaways

- ❑ **DS model:** core model for label aggregation
 - Several variants available
 - Identifiability extensively studied
- ❑ Model identifiability often key performance indicator
- ❑ Sample complexity is crucial when selecting algorithm
- ❑ Moment-based methods
 - ✓ **Label efficiency**
 - ✓ **Computational efficiency**
 - ✓ **Model identifiability**

Challenges

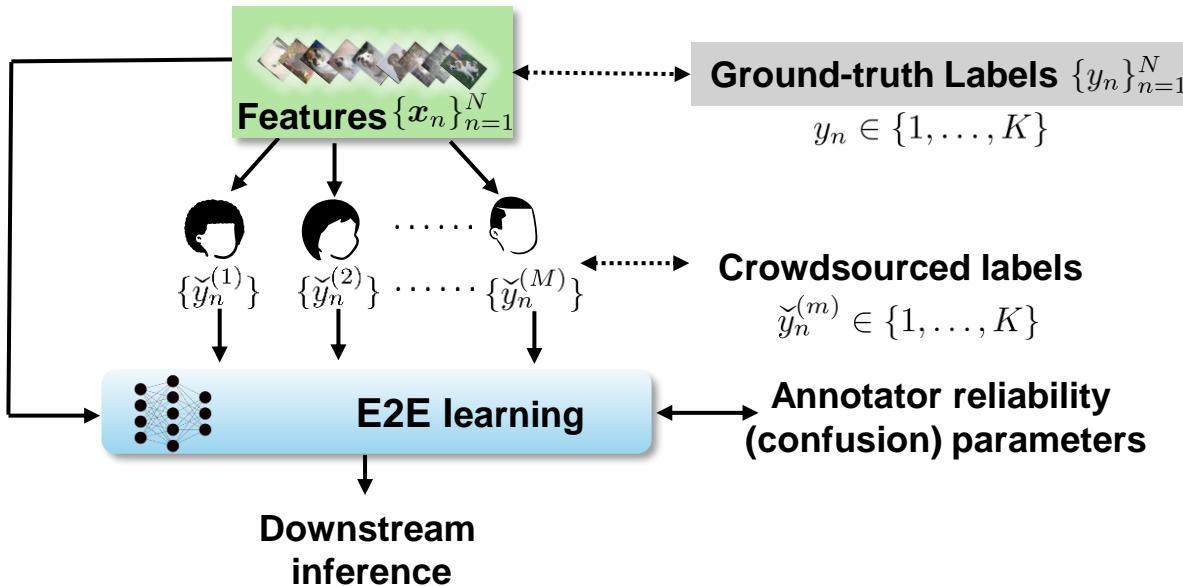
- ❖ **Ignore data features**
- ❖ **Error propagation**
- ❖ **Conditional independence**



Outline

- Motivation and problem statement
- Part I: Combining crowdsourced labels
- Part II: End-to-end (E2E) learning with crowdsourced labels
 - E2E learning with EM
 - Deep learning-based E2E crowdsourcing
 - Instance-dependent E2E crowdsourcing
- Part III: Other aspects of crowdsourcing
- Conclusions and open issues

End-To-End (E2E) Learning



[Raykar et al'10]
[Rodrigues & Pereira'18]
[Khetan et al'18]
[Tanno et al'19]
[Li et al'20]
[Chen et al'20]
[Chu et al'21]
[Wei et al'22]

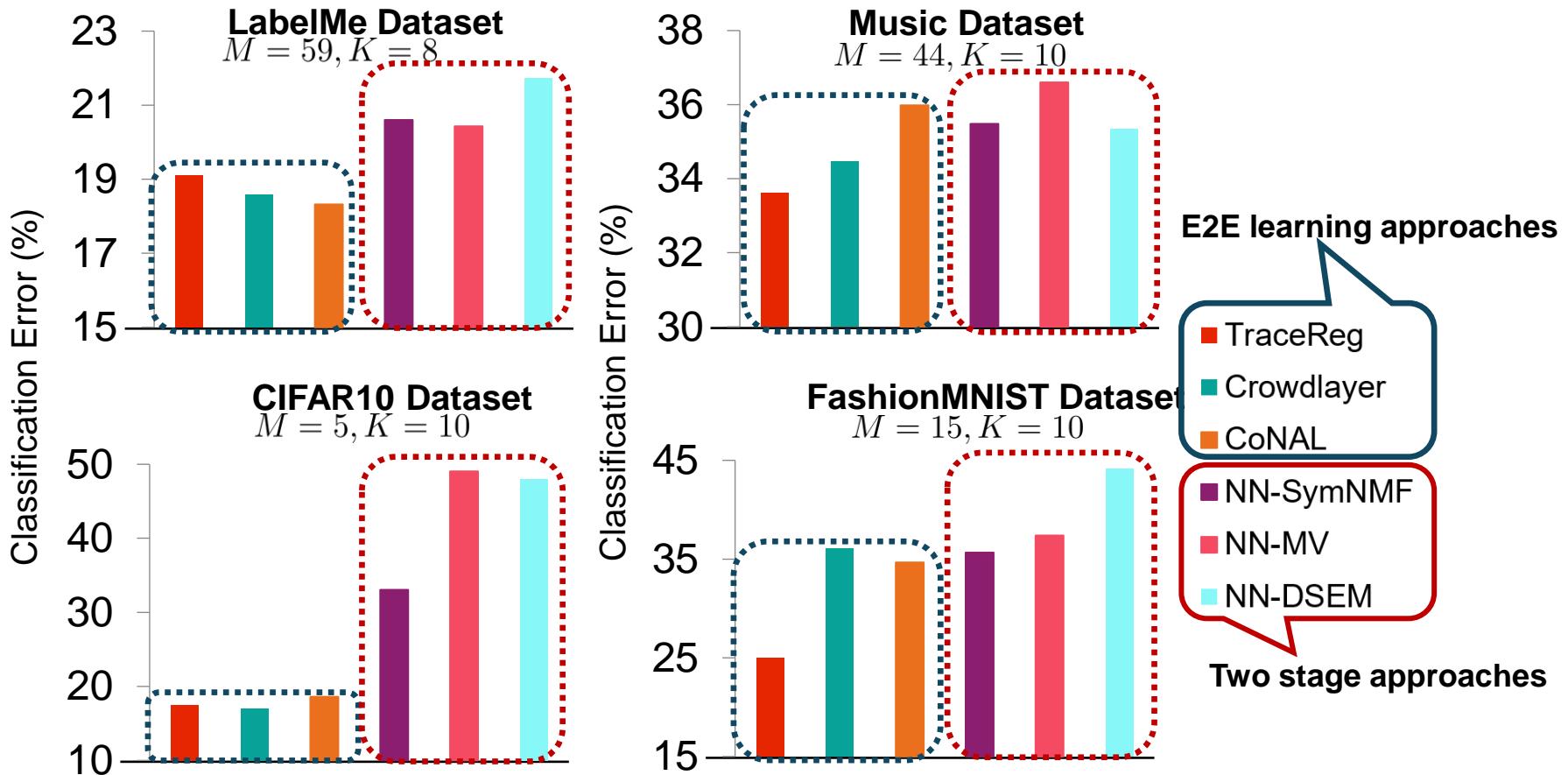
- ❑ Joint label integration and downstream learning
 - For classification, E2E systems aim at

$$\hat{f}_{\theta} \leftarrow \arg \min_{\theta, \eta} \ell(\{x_n\}, \{\tilde{y}_n^{(m)}\}, \theta, \eta)$$

additional model parameters, e.g., annotator confusion matrices

- The goal of the E2E learning is to learn \hat{f}_{θ} corresponds to the optimal classifier

End-To-End (E2E) Learning



- ❑ E2E methods in general outperform two-stage approaches
- ❑ Key research directions:
 - Understanding the performance of E2E crowdsourcing
 - Enhancing performance with principled design

E2E crowdsourcing via maximum likelihood

- Given dataset $\mathcal{D} := \{\mathbf{x}_n, \{\check{y}_n^{(m)}\}_{m=1}^M\}_{n=1}^N$, consider

$$\Pr(\mathcal{D}) = \prod_{n=1}^N \Pr(\mathbf{x}_n, \check{y}_n^{(1)}, \dots, \check{y}_n^{(M)}) = \prod_{n=1}^N \sum_{y_n=1}^K \Pr(y_n | \mathbf{x}_n) \prod_{i=1}^M \Pr(\check{y}_n^{(m)} | y_n, \mathbf{x}_n)$$

Classifier

$$[f(\mathbf{x}_n)]_k = \Pr(y_n = k | \mathbf{x}_n)$$

Confusion Matrix

$$[A_m]_{k,k'} = \Pr(\check{y}_n^{(m)} = k | y_n = k')$$

Common assumption:
Confusion matrices are not
data dependent

[Li et al'21, Rodrigues & Pereira'18, Tanno et al'19]

$$= \prod_{n=1}^N \sum_{y_n=1}^K \Pr(y_n | \mathbf{x}_n) \prod_{m=1}^M \Pr(\check{y}_n^{(m)} | y_n)$$

$$f(\mathbf{x}_n)$$

$$A_m$$

Goal : learn f and $A_m, \forall m$

EM for E2E crowdsourcing

$$\Pr(\mathcal{D}; \psi) = \prod_{n=1}^N \sum_{y_n=1}^K [f_\theta(x_n)]_{y_n} \prod_{m=1}^M A_m(\tilde{y}_n^{(m)}, y_n)$$

Learnable parameters

$$\psi = [A_1, \dots, A_M, \theta]$$

Classifier parameterized by θ

□ E-Step

Expected value of complete log-likelihood

$$Q(\psi; \psi^t) = \mathbb{E}_{\mathcal{Y} \sim \Pr(\mathcal{Y}; \mathcal{D}, \psi^t)} [\log \Pr(\mathcal{D}, \mathcal{Y}; \psi)] = \sum_{n=1}^N \sum_{k=1}^K q(y_n = k; \psi^t) \log \Pr(x_n, y_n = k, \tilde{y}_n^{(1)}, \dots, \tilde{y}_n^{(M)}; \psi)$$

□ M-Step

Estimates ψ by maximizing $Q(\psi; \psi^t)$

$$A_m^{t+1}(k', k) = \frac{\sum_{n=1}^N q(y_n = k; \psi^t) \mathbb{I}[\tilde{y}_n^m = k']}{\sum_{k''=1}^K \sum_{n=1}^N q(y_n = k; \psi^t) \mathbb{I}[\tilde{y}_n^m = k'']}$$

$$\theta^{t+1} = \arg \max_{\theta} Q(\theta; (\theta^t, \{A_m^{t+1}\})).$$

- Similar to EM for DS model
- Several classifier variants
 - logistic regression [Raykar et al'10]
 - neural network [Rodrigues & Pereira'18]
 - conditional random field [Rodrigues et al'14]

Deep learning with Crowd Layer

$$\Pr(\check{y}_n^{(m)} = k | \mathbf{x}_n) = \sum_{k'=1}^K \Pr(\check{y}_n^{(m)} = k | y_n = k') \Pr(y_n = k' | \mathbf{x}_n)$$

$$\mathbf{p}_n^{(m)}$$

$$\mathbf{p}_n^{(m)} = \mathbf{A}_m \mathbf{f}(\mathbf{x}_n), \forall m, n$$

$$\check{y}_n^{(m)} \sim \text{categorical}(\mathbf{p}_n^{(m)})$$

Annotator's confusion

$$\mathbf{A}_m$$

Label predictor

$$\mathbf{f}(\mathbf{x}_n)$$

Goal : learn \mathbf{f} and $\mathbf{A}_m, \forall m$

- Model parameters estimated via

$$\underset{\mathbf{f}_{\theta}, \{\mathbf{A}_m\}}{\text{minimize}} \frac{1}{|\mathcal{S}|} \sum_{(m,n) \in \mathcal{S}} \text{CE}(\mathbf{A}_m \mathbf{f}_{\theta}(\mathbf{x}_n), \check{y}_n^{(m)})$$

subject to $\mathbf{f}_{\theta} \in \mathcal{F}, \mathbf{A}_m \in \mathcal{A}, \forall m.$

Indices of observed labels

Neural network function class

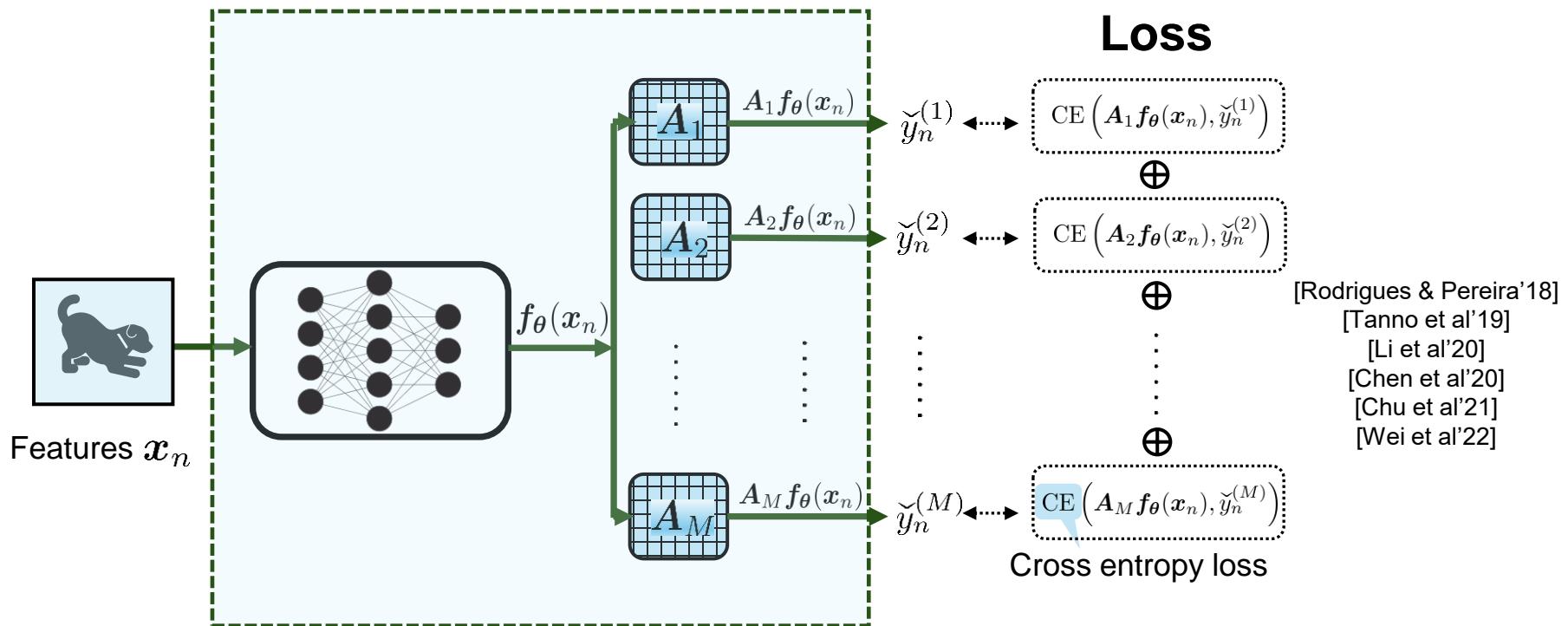
Constrained set
 $\{\mathbf{A} \in \mathbb{R}^{K \times K} | \mathbf{A} \geq 0, \mathbf{1}^\top \mathbf{A} = \mathbf{1}^\top\}$

Coupled Cross Entropy Minimization (CCEM)

□ E2E learning criterion

$$\underset{\mathbf{f}_{\theta}, \{\mathbf{A}_m\}}{\text{minimize}} \quad \frac{1}{|\mathcal{S}|} \sum_{(m,n) \in \mathcal{S}} \text{CE}(\mathbf{A}_m \mathbf{f}_{\theta}(x_n), \tilde{y}_n^{(m)})$$

subject to $\mathbf{f}_{\theta} \in \mathcal{F}, \mathbf{A}_m \in \mathcal{A}, \forall m.$



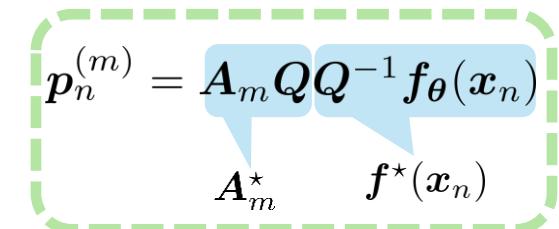
Identifiability in E2E crowdsourcing

Q. Identifiability in E2E crowdsourcing?

- When $N \rightarrow \infty$, CCEM returns $p_n^{(m)} = A_m f_\theta(x_n)$

$$\underset{\{f_\theta, \{A_m\}}}{\text{minimize}} \quad \frac{1}{|\mathcal{S}|} \sum_{(m,n) \in \mathcal{S}} \text{CE}(A_m f_\theta(x_n), \check{y}_n^{(m)})$$

subject to $f_\theta \in \mathcal{F}, A_m \in \mathcal{A}, \forall m.$



- There may exist many possible nonsingular matrices s.t. $p_n^{(m)} = A_m Q Q^{-1} f_\theta(x_n)$

Q. How can we understand performance of CCEM?

Q. Identifiability under CCEM?

Identifiability for CCEM

$$\underset{\mathbf{f}_{\theta}, \{\mathbf{A}_m\}}{\text{minimize}} \quad \frac{1}{|\mathcal{S}|} \sum_{(m,n) \in \mathcal{S}} \text{CE}(\mathbf{A}_m \mathbf{f}_{\theta}(\mathbf{x}_n), \check{y}_n^{(m)})$$

subject to $\mathbf{f}_{\theta} \in \mathcal{F}, \mathbf{A}_m \in \mathcal{A}, \forall m.$

CCEM

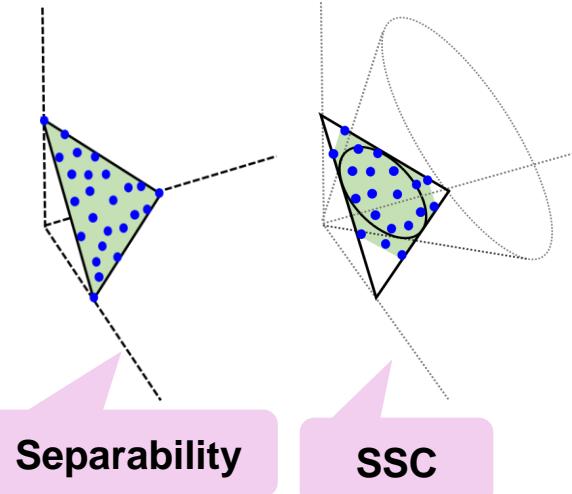
- If both latent factors satisfy **separability condition or (SSC)**, NMF model is essentially unique up to a permutation matrix

$$\|\widehat{\mathbf{A}}_m \boldsymbol{\Pi} - \mathbf{A}_m\| \rightarrow 0, \quad \|\boldsymbol{\Pi}^\top \widehat{\mathbf{f}}_{\theta}(\mathbf{x}) - \mathbf{f}(\mathbf{x})\| \rightarrow 0,$$

- Separability on \mathbf{W} : class experts for each class
- Separability on \mathbf{F} : existence of anchor points for each class $\{\mathbf{x}_{n_1}, \dots, \mathbf{x}_{n_K}\}$ s.t.

$$\mathbf{f}(\mathbf{x}_{n_k}) = \mathbf{e}_k \Leftrightarrow \Pr(y_{n_k} = k | \mathbf{x}_{n_k}) = 1$$

- SSC relaxes these conditions



Model identifiability under CCEM

Theorem 1 [Ibrahim et al., 2023]. Suppose that the Assumptions 1-5 hold true, Assume that each $[\mathbf{A}_m^\natural \mathbf{f}^\natural(\mathbf{x}_n)]_k$ and $[\mathbf{A}_m \mathbf{f}(\mathbf{x}_n)]_k$, $\forall \mathbf{A}_m \in \mathcal{A}, \forall \mathbf{f} \in \mathcal{F}$ are at least $(1/\beta)$. Also assume that $\sigma_{\max}(\mathbf{A}_m^\natural) \leq \sigma$, $\forall m$, for a certain $\sigma > 0$. Then, for any $\alpha > 0$, with probability greater than $1 - K/N^\alpha$, any optimal solution $\widehat{\mathbf{A}}_m$ and $\widehat{\mathbf{f}}$ of the CCEM satisfies the following relations:

$$\min_{\boldsymbol{\Pi}} \|\widehat{\mathbf{A}}_m - \mathbf{A}_m^\natural \boldsymbol{\Pi}\|_{\text{F}}^2 = K\sigma^2(\eta + \xi_1 + \xi_2), \quad \forall m \in [M],$$

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[\min_{\boldsymbol{\Pi}} \|\widehat{\mathbf{f}}(\mathbf{x}) - \boldsymbol{\Pi}^\top \mathbf{f}^\natural(\mathbf{x})\|_2^2 \right] = K(\eta + \xi_1 + \xi_2),$$

where $\eta^2 = \mathcal{O} \left(\beta MN^\alpha / \sqrt{S} \left(\sqrt{M \log S} + (\|\mathbf{X}\|_{\text{F}} \mathcal{R}_{\mathcal{F}})^{\frac{1}{4}} \right) + \beta \sqrt{K} MN^\alpha \nu \right)$, $\boldsymbol{\Pi} \in \{0, 1\}^K$ is a permutation matrix, and $\mathbf{X} = [\mathbf{x}_{n_1}, \dots, \mathbf{x}_{n_S}]$, $(m_s, n_s) \in \mathcal{S}$, if conditions $\xi_1, \xi_2 \leq 1/K$, $\nu \leq 1/\beta K^2 M^2 N^\alpha$, and $S = |\mathcal{S}| = \Omega \left(\beta^2 M^2 N^{2\alpha} K^2 \max(M \log S, \sqrt{\|\mathbf{X}\|_{\text{F}} \mathcal{R}_{\mathcal{F}}}) \right)$ hold.

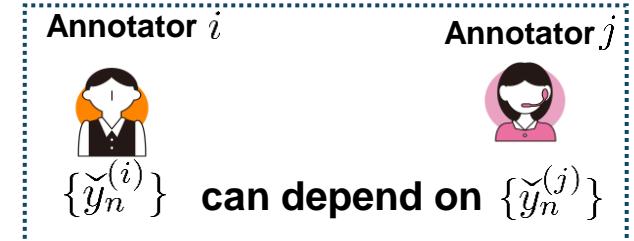
CCEM correctly learns the true confusions and the true classifier, under the assumptions

1 **Anchor point condition**  Per class, there is a data sample belonging to that class with prob. close to 1

2 **Class expert condition**  Per class, there is an expert

 **no conditional independence required**

 **finite number of data and missing labels**



Towards more realistic settings

1

Anchor point condition



Per class, there is a data sample belonging to that class with prob. close to 1

Reflects diversity of data

Easier to satisfy for very large datasets



Natural Images



Social Media Sentiment Data



Social Network Data

Labels often obtained from non-experts, e.g., via AMT



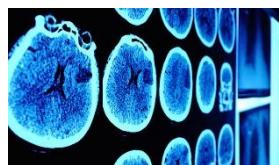
2

Class expert condition



At least one expert, per class

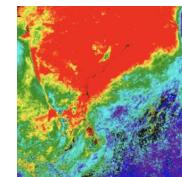
Reflects wisdom of the crowd



Medical Data



Ecological Data



Geospatial Data

Requires domain experts



Identifiability by volume minimization

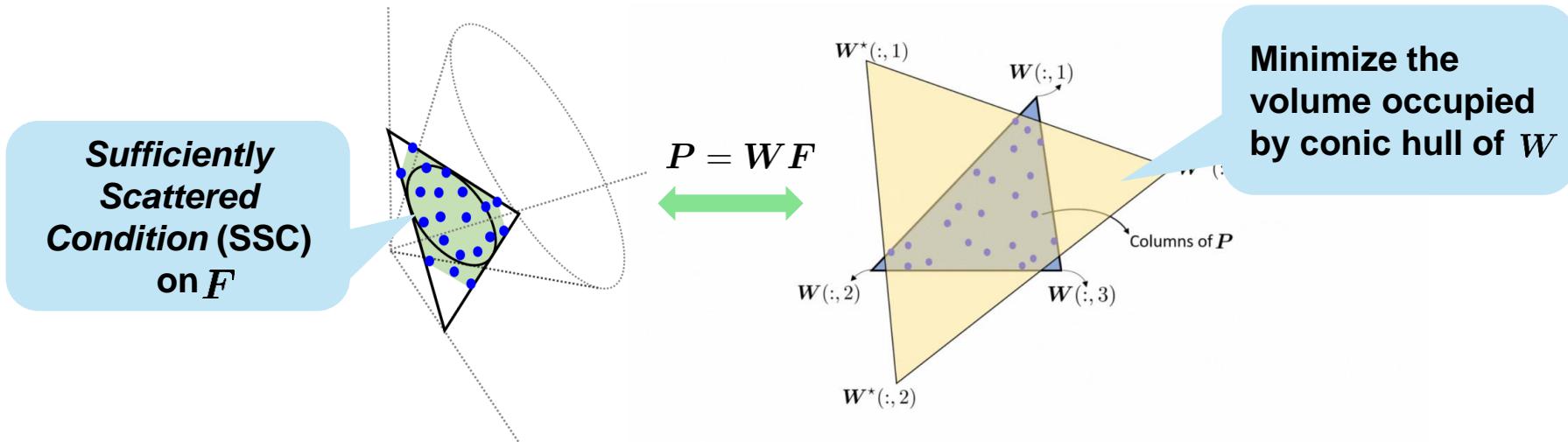
- Identifiability of vanilla CCEM requires conditions on
 - latent factors • **class experts** • **anchor points**
- Conditions on one of the latent factors can be relaxed via regularization

$$\begin{bmatrix} \mathbf{p}_1^{(1)} & \dots & \mathbf{p}_N^{(1)} \\ \vdots & \ddots & \vdots \\ \mathbf{p}_1^{(M)} & \dots & \mathbf{p}_N^{(M)} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_M \end{bmatrix} [f(\mathbf{x}_1) \quad \dots \quad f(\mathbf{x}_N)]$$

$P \qquad \qquad W \qquad \qquad F$

NMF Volume Minimization

If one of the latent factors satisfy the SSC and other is full rank, finding the minimum volume data enclosing simplex guarantees NMF identifiability [Fu et al'15]



Enhanced identifiability under CCEM

Regularized CCEM:

If F satisfies the SSC

(we have more data, but may have no experts to label)

GeoCrowdNet (F)

$$\underset{\mathbf{f}_{\theta}, \{\mathbf{A}_m\}}{\text{minimize}} \quad \frac{1}{|\mathcal{S}|} \sum_{(m,n) \in \mathcal{S}} \text{CE}(\mathbf{A}_m \mathbf{f}_{\theta}(\mathbf{x}_n), \tilde{y}_n^{(m)}) - \lambda \log \det \mathbf{F} \mathbf{F}^{\top}$$

subject to $\mathbf{f}_{\theta} \in \mathcal{F}, \mathbf{A}_m \in \mathcal{A}, \forall m.$

Volume minimization-based regularization

$$\begin{bmatrix} \mathbf{p}_1^{(1)} & \dots & \mathbf{p}_N^{(1)} \\ \vdots & \ddots & \vdots \\ \mathbf{p}_1^{(M)} & \dots & \mathbf{p}_N^{(M)} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_M \end{bmatrix} \begin{bmatrix} \mathbf{f}(\mathbf{x}_1) & \dots & \mathbf{f}(\mathbf{x}_N) \end{bmatrix}^T$$



amazon
mechanical turk

Enhanced identifiability under CCEM

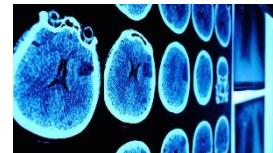
Regularized CCEM:

If W satisfies the SSC
(we may have less data, but have experts to label)

GeoCrowdNet (W)

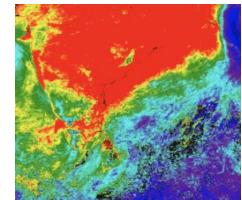
$$\underset{\mathbf{f}_{\theta}, \{\mathbf{A}_m\}}{\text{minimize}} \quad \frac{1}{|\mathcal{S}|} \sum_{(m,n) \in \mathcal{S}} \text{CE}(\mathbf{A}_m \mathbf{f}_{\theta}(x_n), \tilde{y}_n^{(m)}) - \lambda \log \det \mathbf{W}^{\top} \mathbf{W}$$

subject to $\mathbf{f}_{\theta} \in \mathcal{F}, \mathbf{A}_m \in \mathcal{A}, \forall m.$

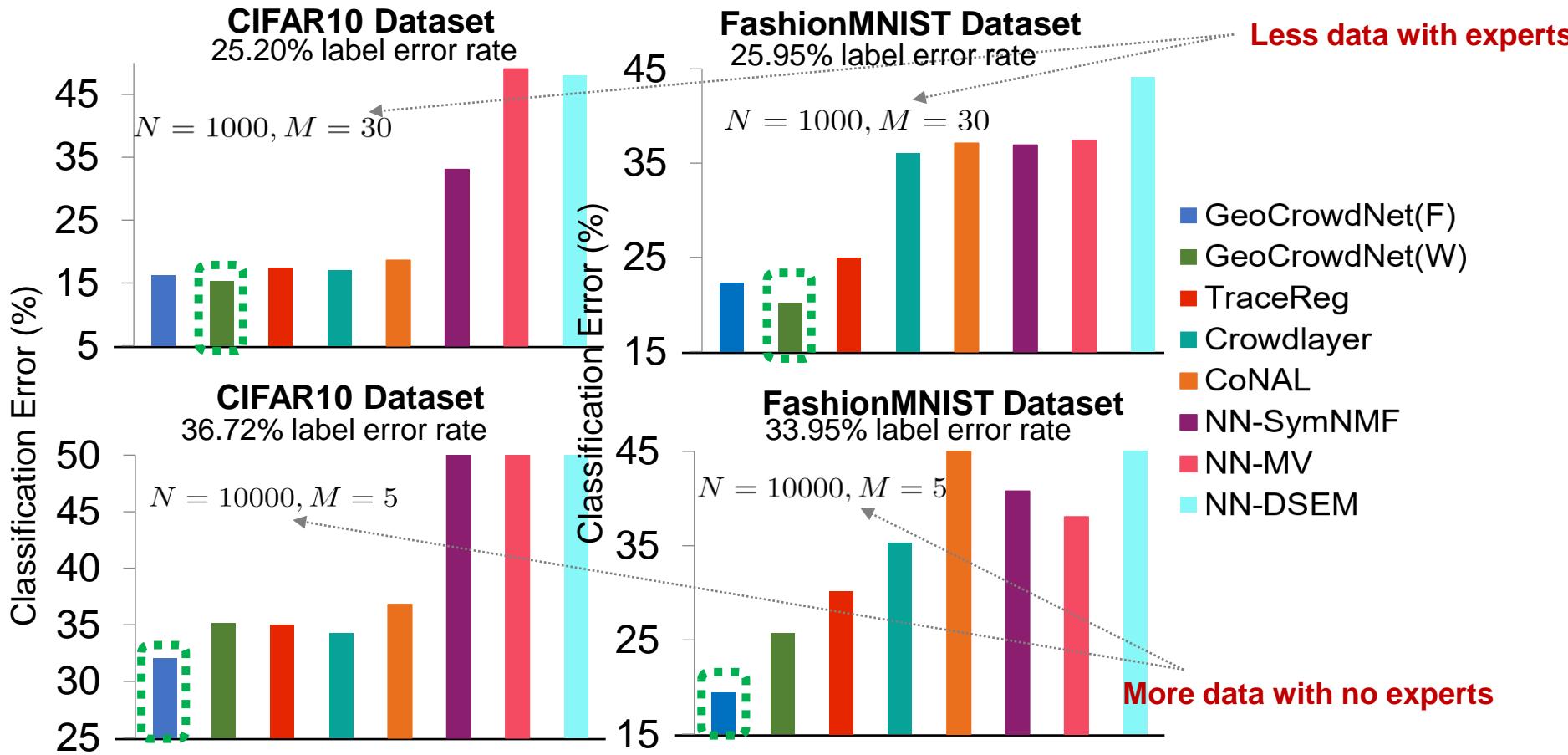


Volume minimization-based regularization

$$\begin{bmatrix} \mathbf{p}_1^{(1)} & \dots & \mathbf{p}_N^{(1)} \\ \vdots & \ddots & \vdots \\ \mathbf{p}_1^{(M)} & \dots & \mathbf{p}_N^{(M)} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_M \end{bmatrix} [\mathbf{f}(\mathbf{x}_1) \quad \dots \quad \mathbf{f}(\mathbf{x}_N)] \mathbf{W}$$



Empirical results



- Around 57,000 training samples; any annotator only labels 10%.
- The annotators are randomly sampled from a pool of machine classifiers whose accuracy ranges from 15% to 85%

Other E2E crowdsourcing approaches

□ CCEM with Trace Regularization [Tanno et al'19]

$$\underset{\mathbf{f}_{\theta}, \{\mathbf{A}_m\}}{\text{minimize}} \quad \frac{1}{|\mathcal{S}|} \sum_{(m,n) \in \mathcal{S}} \text{CE}(\mathbf{A}_m \mathbf{f}_{\theta}(\mathbf{x}_n), \check{\mathbf{y}}_n^{(m)}) + \beta \sum_{m=1}^M \text{trace}(\mathbf{A}_m)$$

subject to $\mathbf{f}_{\theta} \in \mathcal{F}, \mathbf{A}_m \in \mathcal{A}, \forall m.$

- If average of confusion matrices diagonally dominant – identifiable!

□ Agreement-based Model [Peng et al'19]

- Maximizes agreement between label predictor and predictor trained on $\check{\mathcal{Y}}$

Average agreement
between annotators

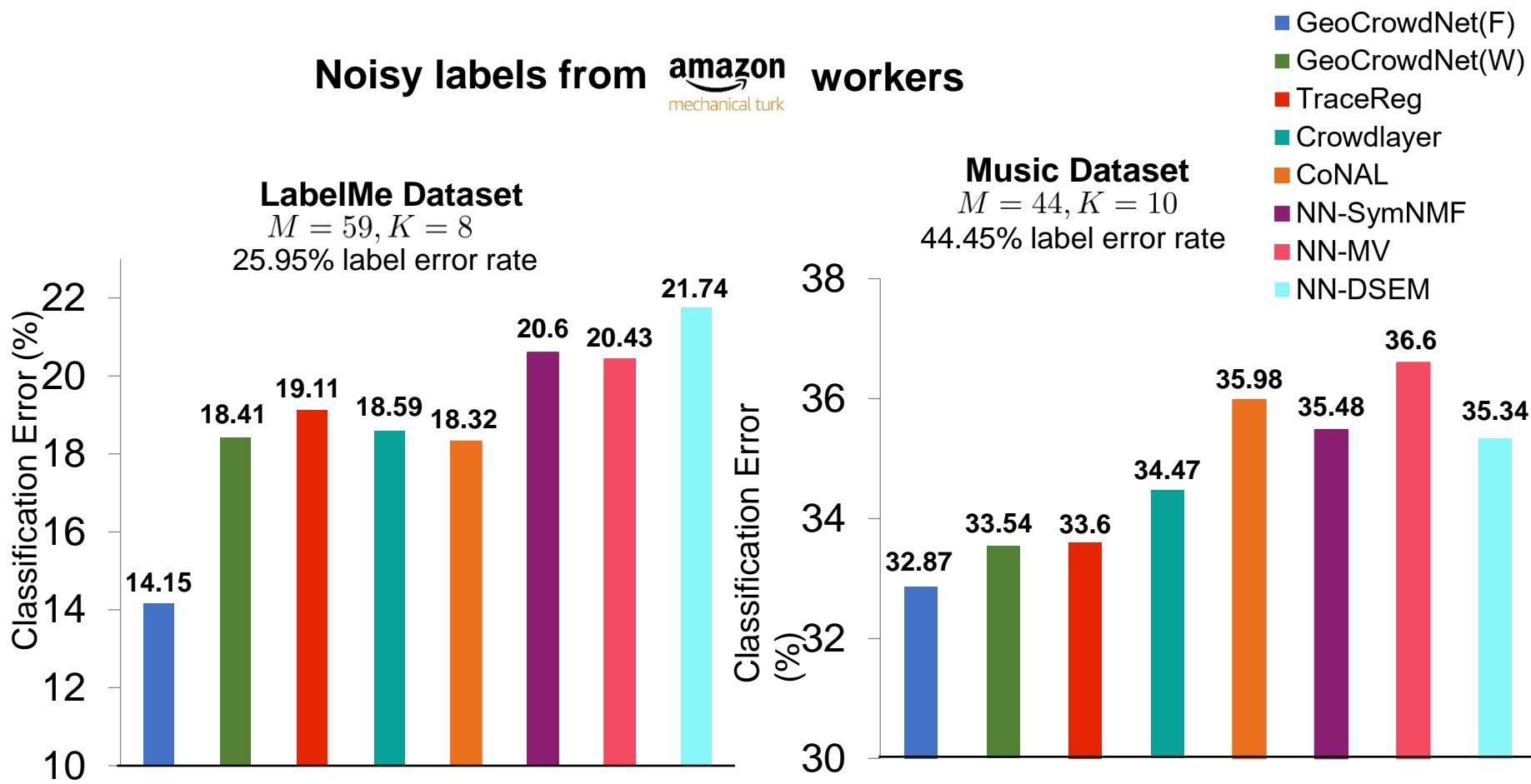
$$\underset{\theta, \phi}{\text{maximize}} \text{ MIG}^f(\mathbf{f}_{\theta}, g_{\phi}; \{\mathbf{x}_n\}, \{\check{\mathbf{y}}_n^{(m)}\})$$

$$g_{\phi}(\{\check{\mathbf{y}}_n^{(m)}\}) = \text{softmax} \left(\sum_{m=1}^M \mathbf{W}^{(m)} \check{\mathbf{y}}_n^{(m)} + \mathbf{b} \right)$$

one hot encoding
of annotator labels

- Asymptotic optimality under conditional independence

Empirical results



Criterion designed for “no experts case” shows edge in practice

Instance-dependent Label Noise

- Annotator errors can be instance-dependent in practice

$$\Pr(\check{y}_n^{(m)} = k | \mathbf{x}_n) = \sum_{k'=1}^K \Pr(\check{y}_n^{(m)} = k | y_n = k', \mathbf{x}_n) \Pr(y_n = k' | \mathbf{x}_n)$$

$p_n^{(m)}$

$$p_n^{(m)} = T_m(\mathbf{x}_n) f(\mathbf{x}_n), \forall m, n$$

Data-dependent confusion

$T_m(\mathbf{x}_n)$

Label predictor

$f(\mathbf{x}_n)$

Goal : learn $f(\mathbf{x})$ and $T_m(\mathbf{x}), \forall m$

- Several approaches for single-annotator scenario
 - Part-dependent label noise [Xia et al'20]
 - Similarity regularization-based [Cheng et al'22]
 - Bayes label-based [Yang et al'22]
 - Sample selection-based [Cheng et al'21, Berthon et al'21]
 - Graphical model-based [Yao et al'21, Wang et al'22]
- Identifiability challenging even with single annotator

Methods for instance-dependent label noise

- Common approach for multiple annotators: **learn two functions**

$$\Pr(\check{y}_n^{(m)} = k | \mathbf{x}_n) = \sum_{k'=1}^K \Pr(\check{y}_n^{(m)} = k | y_n = k', \mathbf{x}_n) \Pr(y_n = k' | \mathbf{x}_n)$$

$\mathbf{p}_n^{(m)}$ $\mathbf{T}_m^\phi : \mathbb{R}^D \rightarrow \mathbb{R}^{K \times K}$ $\mathbf{T}_m(\mathbf{x}_n)$ $\mathbf{f}_\theta : \mathbb{R}^D \rightarrow \mathbb{R}^K$ $\mathbf{f}(\mathbf{x}_n)$

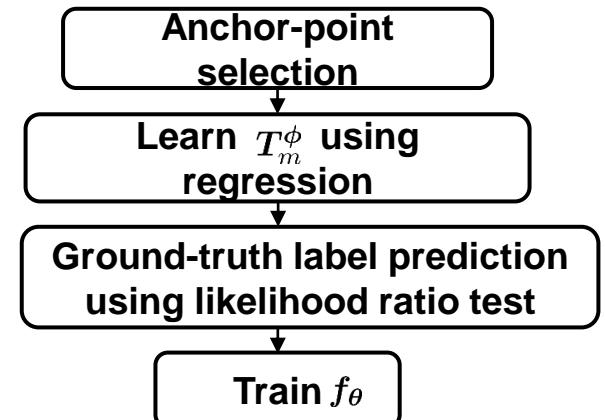
- [Zhang et al'20] extended trace regularization

$$\underset{\{\mathbf{T}_m^\phi \in \mathcal{A}\}, \mathbf{f}_\theta \in \Delta_K}{\text{minimize}} -\frac{1}{|\mathcal{S}|} \sum_{(m,n) \in \mathcal{S}} \sum_{k=1}^K \mathbb{I}[\check{y}_n^{(m)} = k] \log[\mathbf{T}_m^\phi(\mathbf{x}_n) \mathbf{f}_\theta(\mathbf{x}_n)]_k + \beta \sum_{m=1}^M \sum_{n=1}^N \text{trace}(\mathbf{T}_m^\phi(\mathbf{x}_n))$$

- [Guo et al'23] used a multi-stage strategy:

- ❖ Learn \mathbf{T}_m^ϕ using anchor points
- ❖ Train \mathbf{f}_θ

- Identifiability still remains as a critical challenge



Instance-dependent noise as outliers

- Instance-dependent label noise may happen *sparingly* [Nguyen et al., 2014]

$$\mathbf{T}_m(\mathbf{x}_n) = \mathbf{A}_m + \mathbf{E}_m(\mathbf{x}_n)$$

Instance-independent
confusion matrix

$$\mathbf{p}_n^{(m)} = \mathbf{T}_m(\mathbf{x}_n)\mathbf{f}(\mathbf{x}_n), \forall n$$

Instance-dependent
perturbation matrix

$$\mathbf{p}_n^{(m)} = \begin{cases} \mathbf{A}_m\mathbf{f}(\mathbf{x}_n) + \mathbf{e}_n^{(m)}, & \forall n \in \mathcal{I} \\ \mathbf{A}_m\mathbf{f}(\mathbf{x}_n), & \forall n \in \mathcal{I}^c, \end{cases}$$

Outlier index set with
 $\mathbf{E}_m(\mathbf{x}_n) \neq \mathbf{0}$

$$\mathbf{e}_n^{(m)} = \mathbf{E}_m(\mathbf{x}_n)\mathbf{f}(\mathbf{x}_n)$$

(as) For many data samples $\mathbf{E}_m(\mathbf{x}_n) = \mathbf{0}$



Intuition: ➤ Nominal images (left) exhibit similar labeling difficulty
➤ Special/outlier images (right) display a range of labeling challenges

Outlier-robust E2E crowdsourced learning

- One annotator insufficient to detect instance-dependent outliers or learning the ground-truth label classifier [Nguyen et al'24]

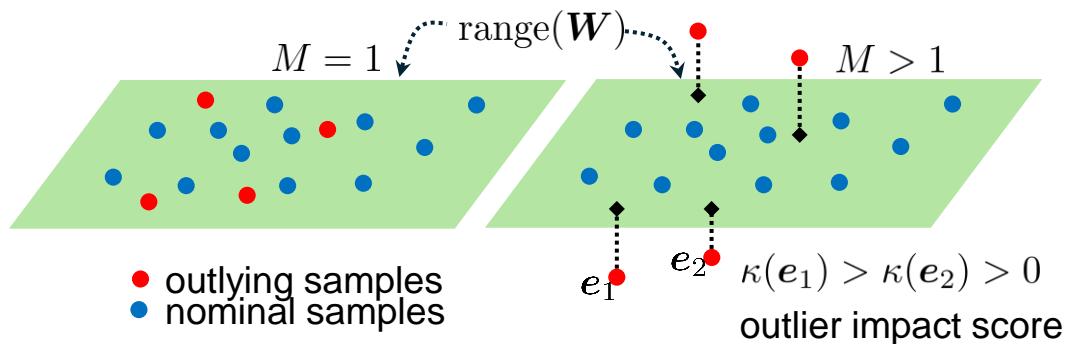
$$\underset{\{\mathbf{A}_m \in \mathcal{A}\}, \{\mathbf{e}_n^{(m)} \in \mathcal{E}\}, \mathbf{f} \in \mathcal{F}}{\text{minimize}} \quad \frac{1}{S} \sum_{(m,n) \in \mathcal{S}} \text{CE}(\mathbf{A}_m \mathbf{f}_{\theta}(\mathbf{x}_n) + \mathbf{e}_n^{(m)}, \check{y}_n^{(m)}),$$

$$\text{subject to } \sum_{n=1}^N \mathbb{I} \left\{ \sum_{m=1}^M \|\mathbf{e}_n^{(m)}\|_2 > 0 \right\} \leq C,$$

Outlier no. estimate
(instance-dependent
label noise samples)

Wisdom of the Crowd
helps in instance-
dependent label noise
learning!!

$$\mathbf{W} = [(\mathbf{A}_1)^{\top}, \dots, (\mathbf{A}_M)^{\top}]^{\top}, \quad \mathbf{e}_n = [(\mathbf{e}_n^{(1)})^{\top}, \dots, (\mathbf{e}_n^{(M)})^{\top}]^{\top}$$



Identifiability under Instance-dependent Label Noise

Theorem 3.5 [Nguyen et al., 2024] (Informal) Let $(\{\widehat{\mathbf{A}}_m\}, \{\widehat{\mathbf{e}}_n^{(m)}\}, \widehat{\mathbf{f}})$ be any optimal solution of (1). The following result holds with probability greater than $1 - 2/S - K/T^\alpha$:

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}_x} \left[\min_{\boldsymbol{\Pi}} \|\widehat{\mathbf{f}}(\mathbf{x}) - \boldsymbol{\Pi}^\top \mathbf{f}(\mathbf{x})\|_2^2 \right] \leq K(\eta + \xi_1 + \xi_2),$$

$$\min_{\boldsymbol{\Pi}} \|\widehat{\mathbf{A}}_m - \mathbf{A}_m \boldsymbol{\Pi}\|_F^2 = K\sigma^2(\eta + \xi_1 + \xi_2), \quad \forall m,$$

where $\eta^2 = \mathcal{O}(\beta M T^\alpha / \sqrt{S} (\sqrt{M \log S} + (\|\mathbf{X}\|_2 \mathcal{R}_F)^{0.25}))$, $\boldsymbol{\Pi}$ a permutation matrix, \mathbf{X} is defined as before, and $T = N - |\mathcal{I}|$. In addition, we have exact outlier detection, i.e., $\widehat{\mathcal{I}} = \mathcal{I}$

Outlier-robust E2E crowdsourced learning

$$\underbrace{\begin{bmatrix} \mathbf{p}_1^{(1)} & \dots & \mathbf{p}_N^{(1)} \\ \vdots & \ddots & \vdots \\ \mathbf{p}_1^{(M)} & \dots & \mathbf{p}_N^{(M)} \end{bmatrix}}_P = \underbrace{\begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_M \end{bmatrix}}_W \underbrace{\begin{bmatrix} \mathbf{f}(\mathbf{x}_1) & \dots & \mathbf{f}(\mathbf{x}_N) \end{bmatrix}}_F + \underbrace{\begin{bmatrix} \mathbf{e}_1^{(1)} & \dots & \mathbf{e}_N^{(1)} \\ \vdots & \ddots & \vdots \\ \mathbf{e}_1^{(M)} & \dots & \mathbf{e}_N^{(M)} \end{bmatrix}}_E$$

$$\mathbf{P} = \mathbf{WF} + \mathbf{E},$$

Robust
NMF

- Identifiability via
 - ❖ Anchor point condition
 - ❖ Class expert condition
 - ❖ Scarce outliers – lying outside the simplex spanned by \mathbf{W}

Empirical results

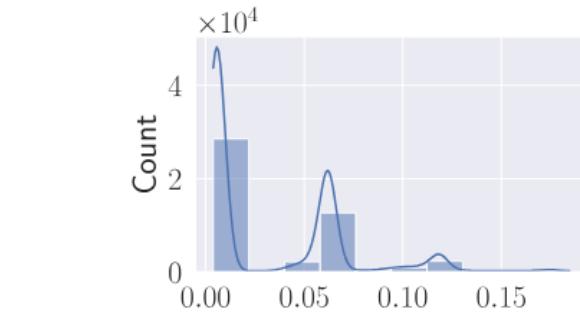
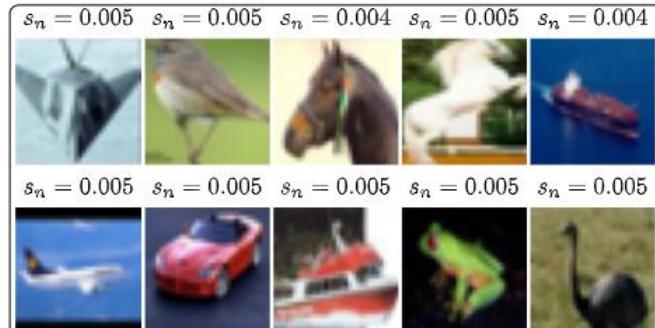
Implementation: COINNet

$$\underset{\mathbf{A}_m \in \mathcal{A}, \mathbf{e}_n^{(m)} \in \mathcal{E}, \mathbf{f} \in \mathcal{F}}{\text{minimize}} \quad \mathcal{L}_{\text{ce}} + \mu_1 \mathcal{L}_{\text{outlier}} + \mu_2 \mathcal{L}_{\text{vol}}$$

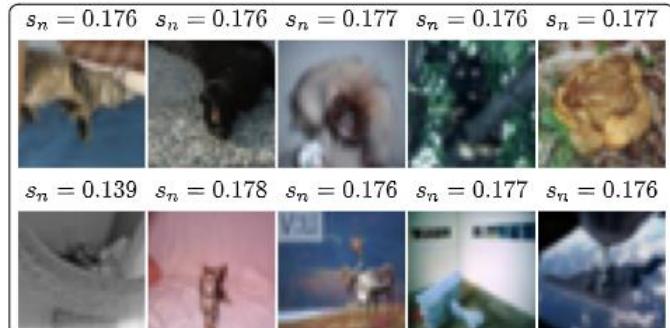
$$\begin{aligned}\mathcal{L}_{\text{CE}} &= \frac{1}{S} \sum_{(m,n) \in \mathcal{S}} \text{CE}(\mathbf{A}_m \mathbf{f}_{\theta}(\mathbf{x}_n) + \mathbf{e}_n^{(m)}, \tilde{y}_n^{(m)}) \\ \mathcal{L}_{\text{outlier}} &= \sum_{n=1}^N \left(\sum_{m=1}^M \|\mathbf{e}_n^{(m)}\|_2^2 + \zeta \right)^{\frac{p}{2}} \\ \mathcal{L}_{\text{vol}} &= -\log \det(\mathbf{F} \mathbf{F}^\top)\end{aligned}$$

Dataset: CIFAR10-N [Wei, et al., 2022];
 $N = 60000, K = 10, M = 3$;
 Annotator error rates: 17.23%, 18.12%, 17.64%.

Images with high score values show more instance-dependent confusion characteristics (such as background noise and blurring)



Histogram of $s_n = \sum_{m=1}^M \|\hat{\mathbf{e}}_n^{(m)}\|_2^2$



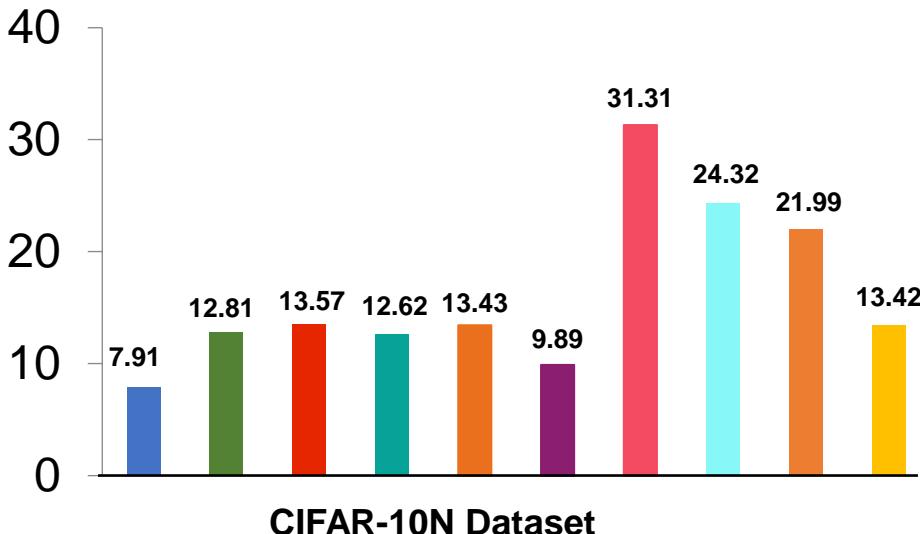
Empirical results

Dataset: ImageNet-15N [Nguyen et al'24]; annotations collected by AMT workers annotating images from ImageNet;
Average annotator error rate: 42.68%.

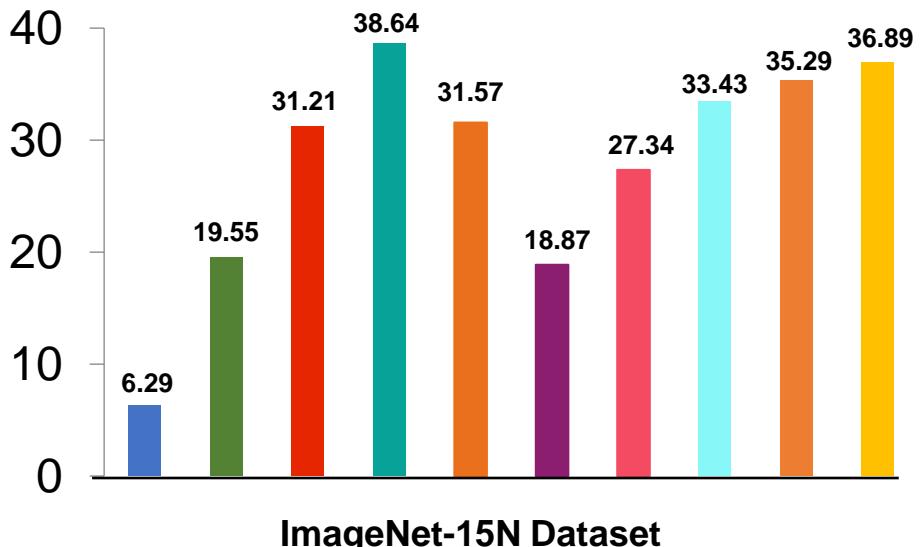
$$N = 2514, K = 15, M = 100;$$



Empirical results

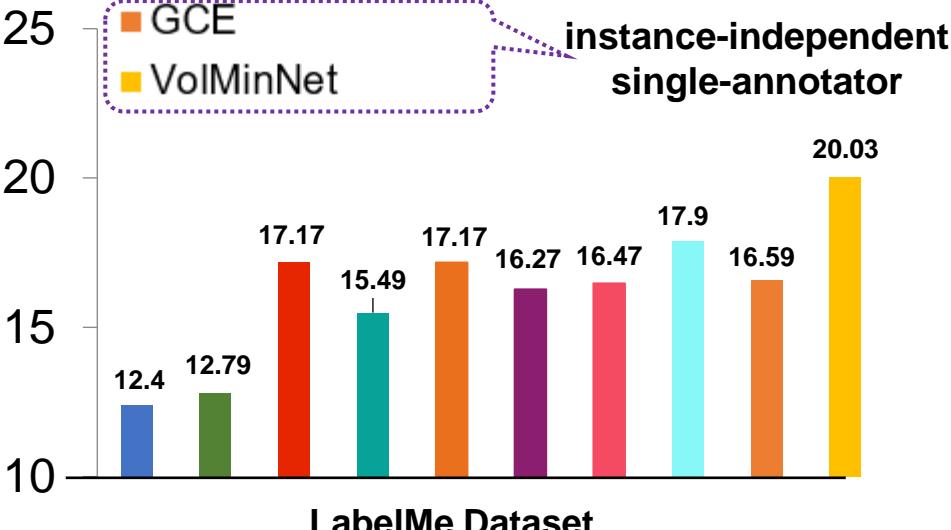
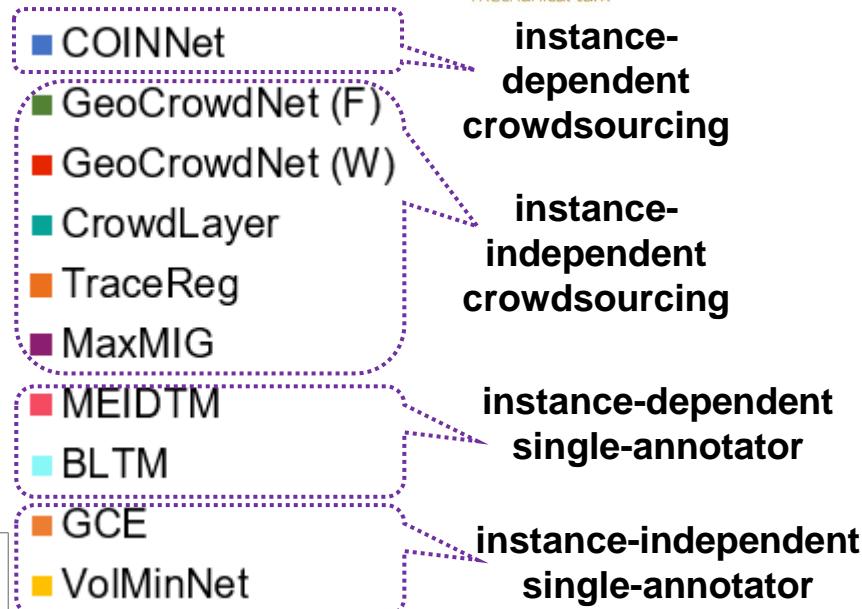


CIFAR-10N Dataset



ImageNet-15N Dataset

Noisy labels from workers



LabelMe Dataset

Outline

- Motivation and problem statement
- Part I: Combining crowdsourced labels
- Part II: End-to-end (E2E) learning with crowdsourced labels
- **Part III: Other aspects of crowdsourcing**
 - Other types of annotations
 - Adversary aware crowdsourcing
 - Active crowdsourcing
 - Bias and fairness issues
- Conclusions and open issues

Crowdsourced regression

- Ground-truth labels and annotations take continuous real values $y_n, \{\check{y}_n^{(m)}\} \in \mathbb{R}$
 - e.g., bounding box or timestamp annotations
- Simple aggregation approach: **Averaging** $\hat{y}_n = \frac{1}{|\mathcal{M}_n|} \sum_{m \in \mathcal{M}_n} \check{y}_n^{(m)}$
- Gaussian priors for ground-truth labels and annotator responses
$$y_n \sim \mathcal{N}(\mu_n, \sigma_n^2) \quad \check{y}_n^{(m)} | y_n = \alpha \sim \mathcal{N}(\alpha, \sigma_m^2)$$

Annotator
response noise
- Solved via Bayesian iterative algorithm [Ok et al'19]
- Moment-based approach [Tenzer et al'22]
 - Sparse + low rank decomposition of annotator covariance matrix
$$[C]_{i,j} = \mathbb{E}[(\check{y}_n^{(i)} - \mu_i)(\check{y}_n^{(j)} - \mu_j)], \mu_i = \mathbb{E}[\check{y}_n^{(i)}]$$
- E2E approach – CrowdLayer for regression [Rodrigues et al'18]

Similarity annotations

- Annotators indicate the similarity of data items

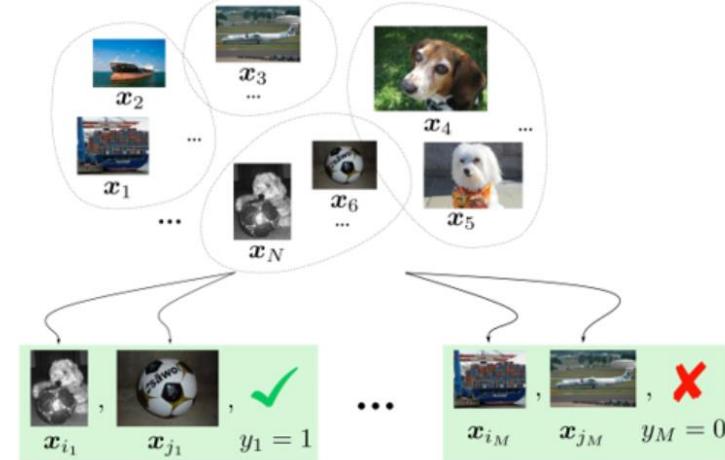
➤ E2E learning can be formulated as graph clustering

$$\mathbf{G} \in \{0, 1\}^{N \times N} \quad \text{Graph adjacency}$$

$$\mathbf{G}(i, j) \sim \text{Bernoulli}(\mathbf{f}(\mathbf{x}_i)^\top \mathbf{f}(\mathbf{x}_j)) \quad \text{modeling similarity}$$

$$[\mathbf{f}(\mathbf{x}_n)]_k = \Pr(y_n = k | \mathbf{x}_n)$$

- Requires lower expertise level from annotators



Maximum likelihood Estimation (MLE)

$$\underset{\mathbf{f}_\theta \in \Delta_K}{\text{minimize}} \sum_{(i,j) \in \Omega} [\mathbb{I}[\mathbf{G}(i,j) = 1] \log (\mathbf{f}_\theta(\mathbf{x}_i)^\top \mathbf{f}_\theta(\mathbf{x}_j)) + \mathbb{I}[\mathbf{G}(i,j) = 0] \log (1 - \mathbf{f}_\theta(\mathbf{x}_i)^\top \mathbf{f}_\theta(\mathbf{x}_j))]$$

- Identifiability of f established by treating model as quantized NMF problem
- Similarity annotations easily acquired by the crowd
- Often referred to as ***crowdclustering***

Pairwise annotations

Ranking

- Popular model for pairwise preference is the **Bradley-Terry (BT)** one:

$$\Pr(x_n \succ x_{n'}) = \frac{e^{s_n}}{e^{s_n} + e^{s_{n'}}}, \quad \text{ranking score: } s_n$$

$x_n \succ x_{n'} \rightarrow x_n$ is preferred over $x_{n'}$

- Annotator-specific preference can be modeled as [Chen et al'13]

$$\Pr(x_n \succ_m x_{n'}) = \Pr(x_n \succ_m x_{n'} | x_n \succ x_{n'}) \Pr(x_n \succ x_{n'}) + \Pr(x_n \succ_m x_{n'} | x_n \prec x_{n'}) \Pr(x_n \prec x_{n'})$$

$\omega_m \qquad \qquad \qquad \frac{e^{s_n}}{e^{s_n} + e^{s_{n'}}}$

- MLE-based optimization to jointly learn both ω_m and s_n
- In E2E learning, s_n represented using a NN $f_\theta(x_n)$ [Chhan et al'24]

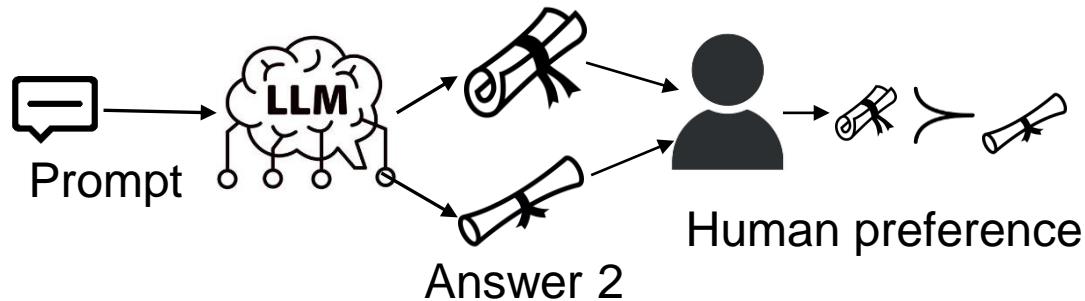
- Ordinal ranking: convert preference order into set of binary labels [Raykar et al'10]

$$[\check{y}_m^{(m)}]_k = \begin{cases} 1 & \text{if } \check{y}_n^{(m)} > k \\ 0 & \text{otherwise} \end{cases}, \quad k = 1, \dots, K-1 \quad K-1 \text{ binary labels}$$

Preference annotations

- LLM alignment relies on human preference annotations Answer 1

- Reinforcement Learning from Human Feedback (RLHF)
 - Direct Preference Optimization (DPO)



- Single annotator setting: Robust DPO model-preference noise modeled as [Chowdhury et al'24]

$$\epsilon := \Pr(x_n \succsim x_{n'} | x_n \succ x_{n'})$$

Preference
flipping rate

- RLHF with **multiple annotators**: Weighted MV [Li et al'24]

- Alternative: preferences integrated via mixture of reward models [Chakraborty et al'24]

LLM alignment using crowd feedback remains an active research area with exciting open directions

Outline

- Motivation and problem statement
- Part I: Combining crowdsourced labels
- Part II: End-to-end (E2E) learning with crowdsourced labels
- **Part III: Other aspects of crowdsourcing**
 - Other types of annotations
 - Adversary aware crowdsourcing
 - Active crowdsourcing
 - Bias and fairness issues
- Conclusions and open issues

Crowdsourcing under attack

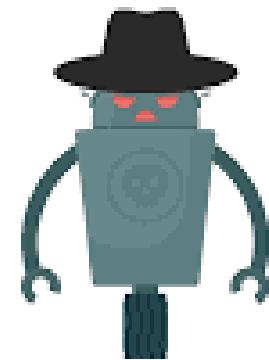
❑ Adversarial attacks

- Adversaries may hide as legitimate annotators
- Adversaries manipulate labels to hamper performance or drain resources
- Adversaries poison crowdsourcing datasets



Q1. Which are the worst adversarial attacks?

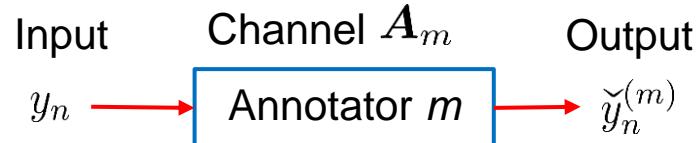
Q2. Can we identify such attacks and discard corresponding annotators?



- ❑ Two types of adversaries considered: **spammers** and **colluding adversaries**
-

Information-theoretic characterization of spammers

Idea. View annotators as independent information-bearing channels



- Per annotator m , performance can be benchmarked by channel capacity

$$C^{(m)} := \max_{\pi} I(y_n, \check{y}_n^{(m)}) \geq 0 \quad \triangleright \text{ Overall capacity } C = \sum_{m=1}^M C^{(m)}$$



- Worst spammer annotator: $C^{(m)} = 0$, output not related to input y_n

$$A_m(k', k) := \Pr(\check{y}_n^{(m)} = k' | y_n = k) = \Pr(\check{y}_n^{(m)} = k') := s_{k'}^{(m)}$$

Irrelevant to y_n ,
 $\Rightarrow C^{(m)} = 0$

➤ Zero-capacity confusion matrix $A_m = s^{(m)} \mathbf{1}^\top$

- Two groups of annotators
 - Spammers $m \in \mathcal{S}$ to be removed
 - Honest $m \in \mathcal{H}$ follow DS annotator model
- Spammer detection while fusing labels via modified EM [Raykar-Yu'12]

Testing the effect of spammers

- Synthetic dataset, $N=10,000$, $K = 4$, $M = 20$

- αM annotators as adversaries, $(1-\alpha)M$ honest

- Oracle: classifier knows $\{A_m\}_{m=1}^M, d$

- MV: majority voting

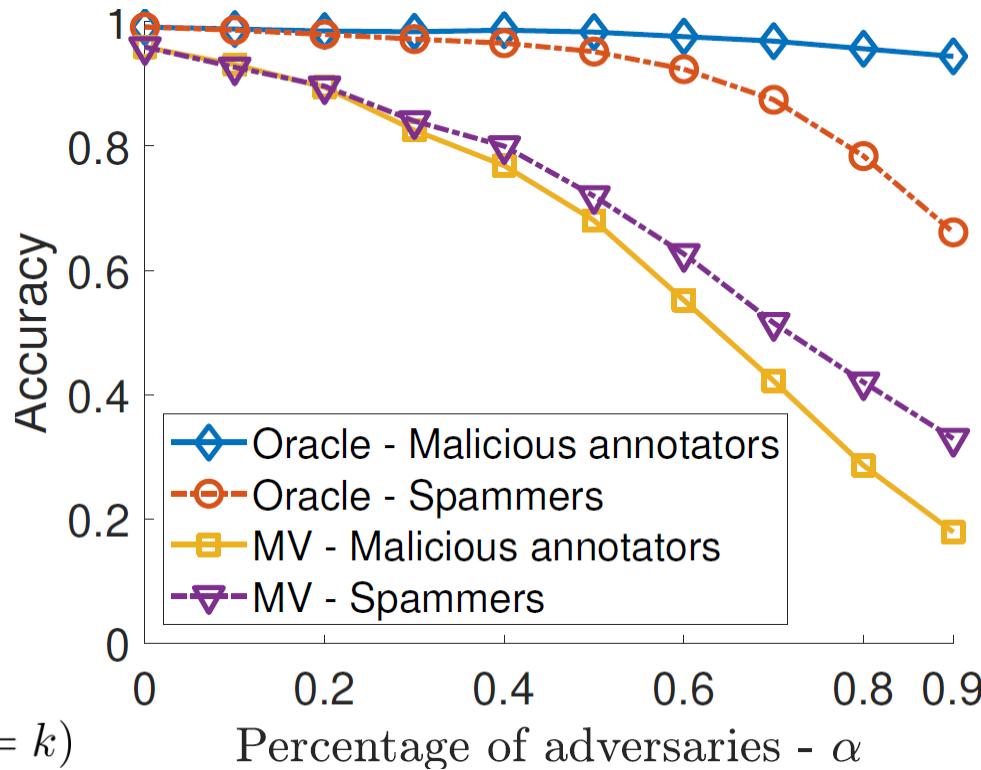
- Adversaries

- Spammers give random responses

$$\Pr(\check{y}_n^{(m)} = k' | y_n = k) \stackrel{\text{?}}{=} \Pr(\check{y}_n^{(m)} = k')$$

- Malicious annotators $k' \neq k$

$$\Pr(\check{y}_n^{(m)} = k' | y_n = k) > \Pr(\check{y}_n^{(m)} = k | y_n = k)$$



- Spammers challenge oracle approaches; we can “flip” malicious responses

Flagging spammers with cross-covariances

- ❑ Covariance of annotators $m \in \mathcal{H} \cup \mathcal{S}$ and $m' \in \mathcal{S}$

$$c_{m,m'} := \mathbf{E}[\tilde{y}_n^{(m)} \tilde{y}_n^{(m')}] - \tilde{\mu}_m \tilde{\mu}_{m'} = \mathbf{E}[\tilde{y}_n^{(m)}] \mathbf{E}[\tilde{y}_n^{(m')}] - \tilde{\mu}_m \tilde{\mu}_{m'} = 0$$

Can be used to flag spammers

- ❑ Sample cross-covariance matrix has low rank + diagonal structure

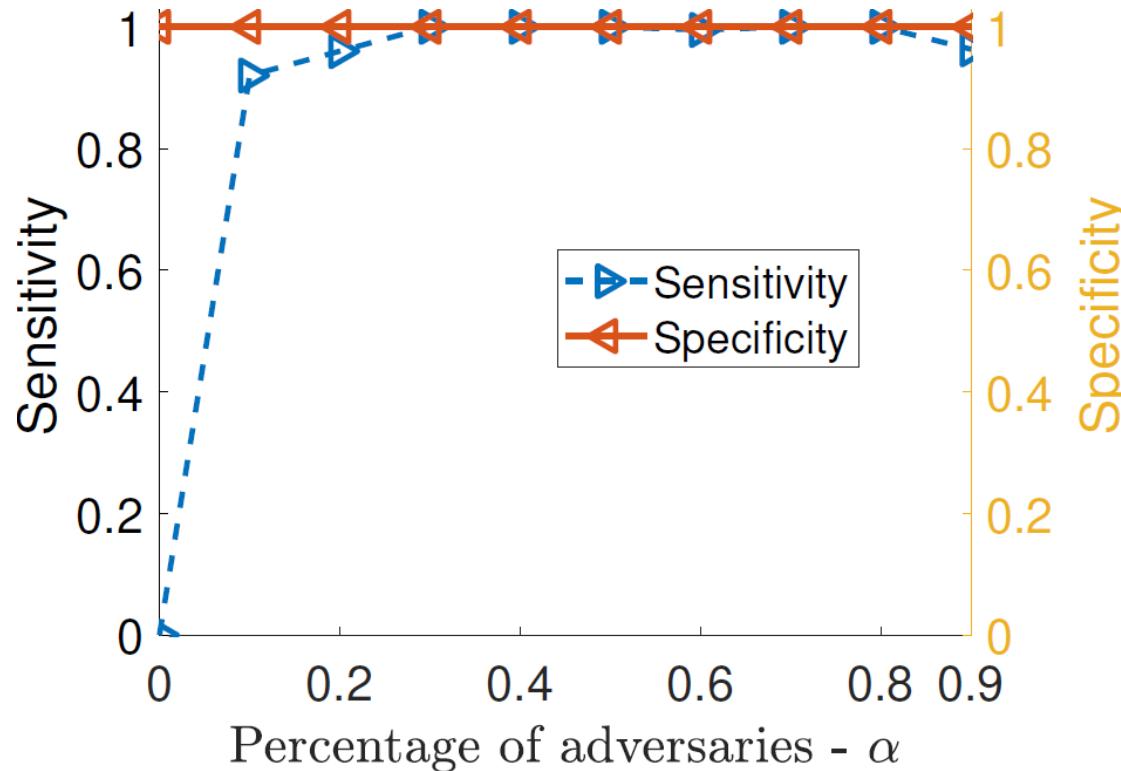
$$\hat{\mathbf{C}} = \mathbf{L} + \mathbf{D} = \mathbf{V}\mathbf{V}^\top + \mathbf{D}, \quad \mathbf{V} \in \mathbb{R}^{M \times K}$$

- Rows of \mathbf{V} close to $\mathbf{0}$ indicate spammers
- ❑ Spammer identification algorithm
 - S1.** Recover \mathbf{V} from sample cross-covariance matrix
 - S2.** Cluster rows of \mathbf{V} using e.g., k -means
 - S3.** Identify spammers as the cluster heads with norm closest to 0

Spammer detection performance

□ Synthetic dataset with $N=10,000$, $K = 4$, $M = 20$

- Figures of merit: Sensitivity (true positive rate) and Specificity (true negative rate)



Bottomline. A few spammers misclassified as honest; all honest classified as honest

Real crowdsourced data

- ❑ Bluebird dataset $N=108$, $K = 2$, $M=39$
- ❑ Dog dataset $N=807$, $K = 4$, $M=109$
- ❑ Web dataset $N=2,655$, $K = 5$, $M=177$



- Annotators deemed **spammers** were **removed** from dataset
- **Alg. 1:** Covariance-based spammer identification algorithm

Classification accuracy

Dataset	MV	EM	Alg. 1 + MV	Alg. 1 + EM
Bluebird	0.759	0.88	0.852(22)	0.899(22)
Dog	0.817	0.834	0.819(12)	0.834(12)
Web	0.776	0.871	0.841(158)	0.91(158)

Parentheses indicate **number of pruned annotators**

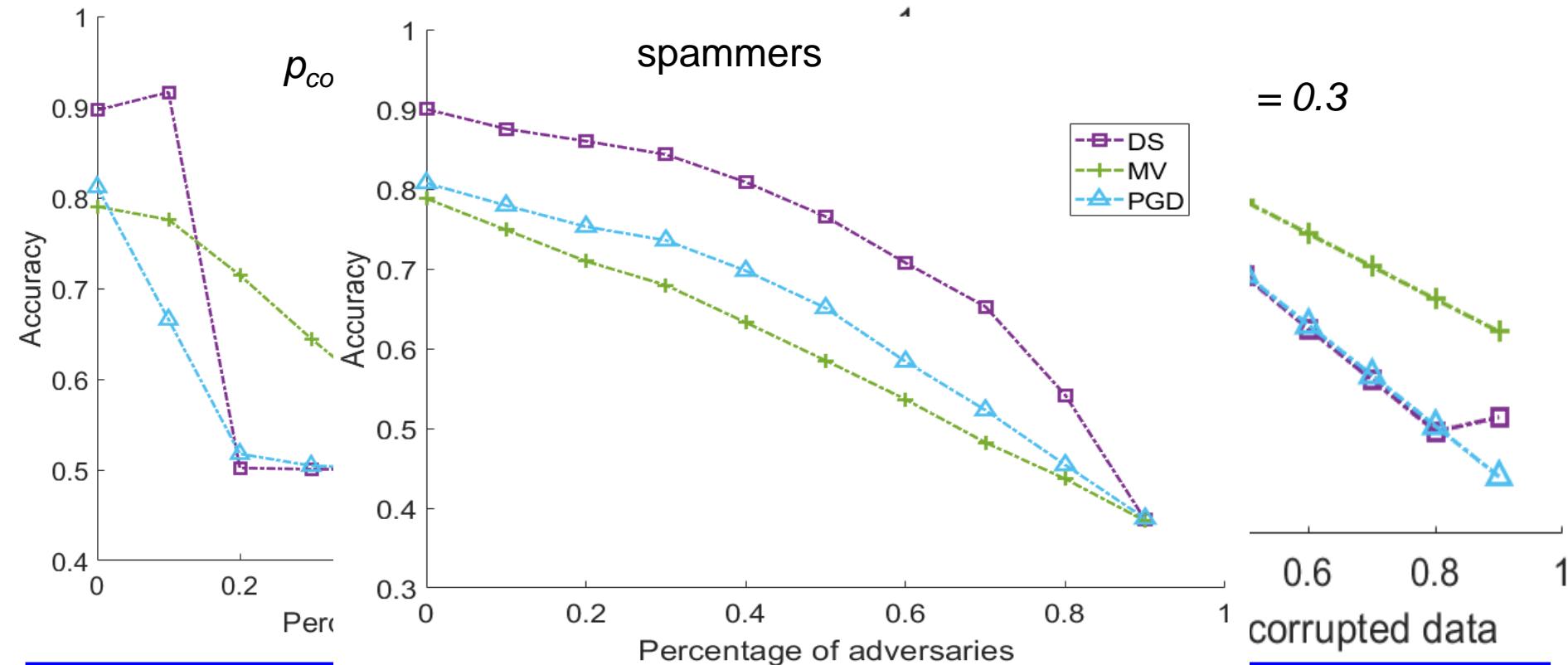
Bottomline. Identifying and removing spammers boosts crowdsourcing performance

Effect of colluding adversaries

Q. What if adversaries cooperate?

□ Synthetic dataset with $N = 5,000$, $M=60$, $K=3$. Percentage of adversaries = α

- Colluding adversaries provide wrong response w.p. p_{corr} , and ground-truth label w.p. $1 - p_{corr}$
 - PGD [Ma '18]: Moment matching for one-parameter model
- Colluding adversaries hamper performance more severely than spammers



Colluding adversaries

- Adversaries cooperate to change labels of some data
 - Model mismatch as DS model no longer applies!
- Two groups of annotators
 - Adversaries **deviate from presumed model** $m \in \mathcal{A}$, $\text{card}(\mathcal{A}) = M_{\mathcal{A}}$
 - Honest follow presumed model $m \in \mathcal{H}$, $\text{card}(\mathcal{H}) = M_{\mathcal{H}}$

Q. How can colluding adversaries be identified?

- Adversaries alter characteristics associated with the expected model
 - Arbitrary adversaries under one-coin model, not exceeding 50% of annotators
[Jagabathula et al'17, Kleindessner-Awasthi'18, Ma-Olshevsky'20]
 - Arbitrary adversaries under DS model [Traganitis-Giannakis'21]
 - Graph-based detection of adversaries and attacked data [Karaaslanli et al'25]

□ E2E learning

This ICASSP

- Honest annotators – corrupted data x_n [Chen et al'22]

Annotator agreement matrix

- Agreement rate of honest annotators m, m'

$$\sigma_{m,m'} := \Pr\left(\check{y}_n^{(m)} = \check{y}_n^{(m')}\right) = \text{tr}\left(\mathbf{A}_m \mathbf{D} {\mathbf{A}_{m'}}^\top\right) = \mathbf{u}^{(m)\top} \mathbf{u}^{(m')}$$
$$\mathbf{u}^{(m)} := \text{vec}(\mathbf{D}^{1/2} {\mathbf{A}_m}^\top) : K^2 \times 1$$

$$\sigma_{m,m} := \Pr(\check{y}_n^{(m)} = \check{y}_n^{(m)}) = 1$$

- Annotator agreement matrix $[\Sigma]_{m,m'} = \sigma_{m,m'} : M \times M$

(as) $M_{\mathcal{H}} > K^2$

- W.l.o.g. First $M_{\mathcal{H}}$ annotators are honest and the rest adversarial

$$\Sigma = \mathbf{C} + \mathbf{I} = \left[\begin{array}{c|c} \mathbf{C}_{\mathcal{H}} & \mathbf{C}_{\mathcal{H},\mathcal{A}} \\ \hline \mathbf{C}_{\mathcal{A},\mathcal{H}} & \mathbf{C}_{\mathcal{A}} \end{array} \right] + \left[\begin{array}{cc} \mathbf{I}_{\mathcal{H}} & \\ & \mathbf{I}_{\mathcal{A}} \end{array} \right]$$

Low rank + Diagonal

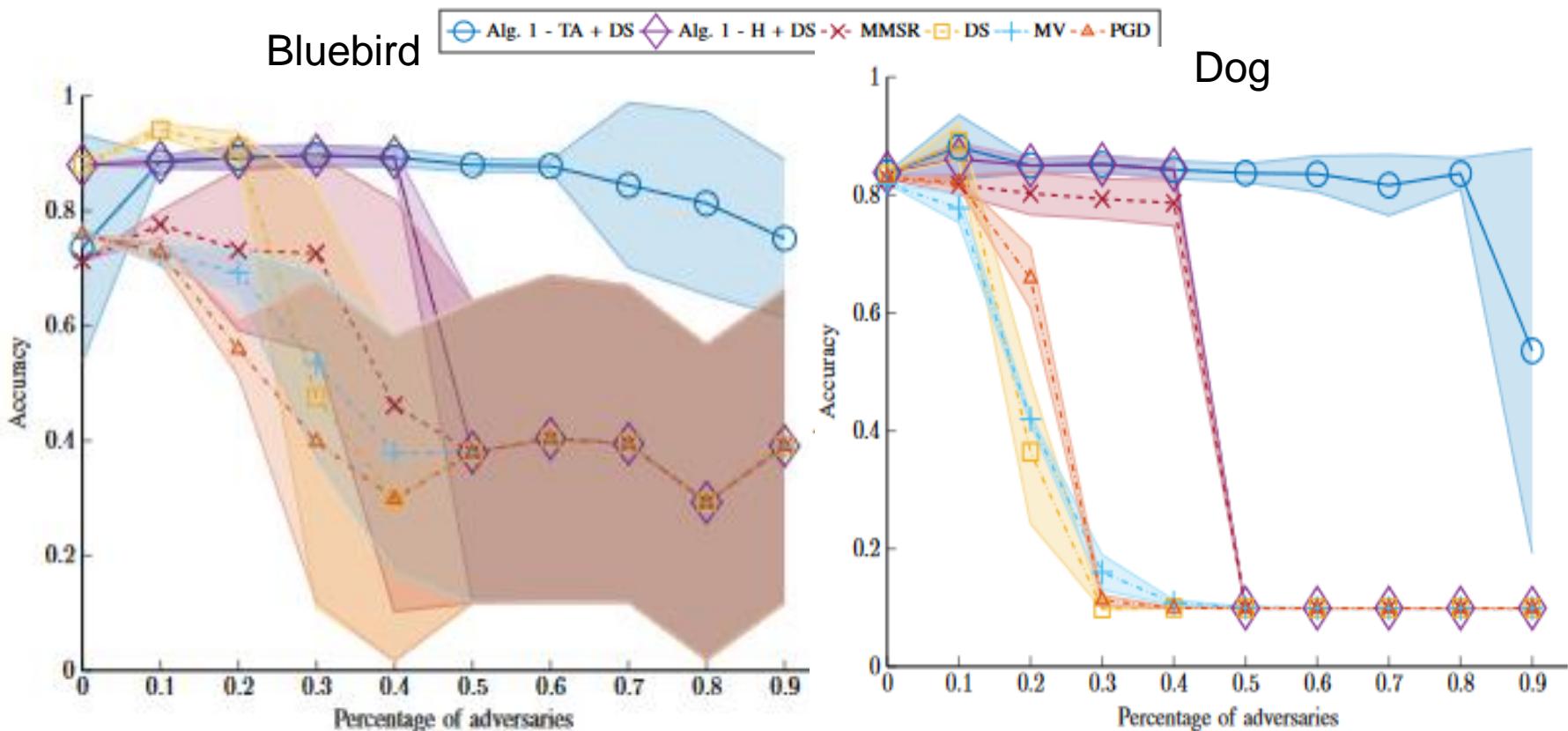
rank $\left(\begin{bmatrix} \mathbf{C}_{\mathcal{H}} \\ \mathbf{C}_{\mathcal{A},\mathcal{H}} \end{bmatrix} \right) \leq K^2$

unknown

Real data tests

Dataset	Bluebird	Sentence Polarity	Dog	Web
N	108	5,000	807	2,665
M	39	203	109	177
K	2	2	5	5

- MMSR: State-of-the-art adversary aware method under one-coin model $p_{corr} = 0.9$
- Alg. 1 – TA: [Traganitis-Giannakis'21] with knowledge of one trusted annotator
- Alg. 1 – H: [Traganitis-Giannakis'21] assuming at least 50% of annotators are honest



Crowdsourcing with dependent annotators

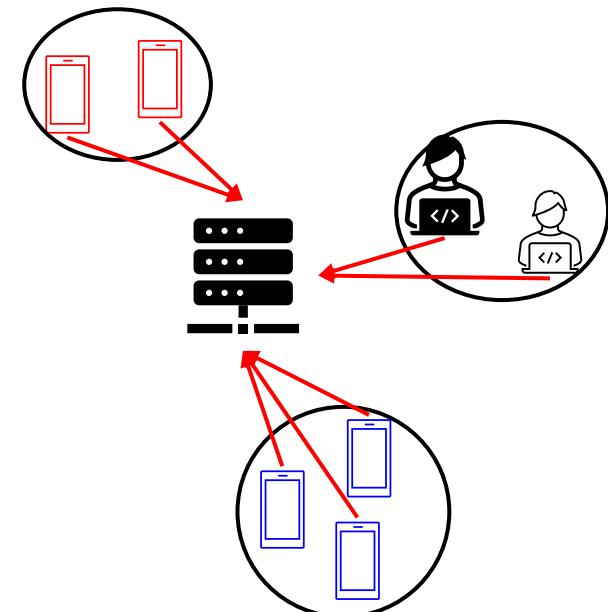
- ❑ So far, annotators were considered **independent**
 - Not always the case: nearby sensors; collaboration; similar devices ...

Q1. Can label integration methods be broadened to deal with dependent annotators?

Q2. Can we identify this dependence structure?

- ❑ Two types of approaches:

- EM and Bayesian methods
- Moment-based methods



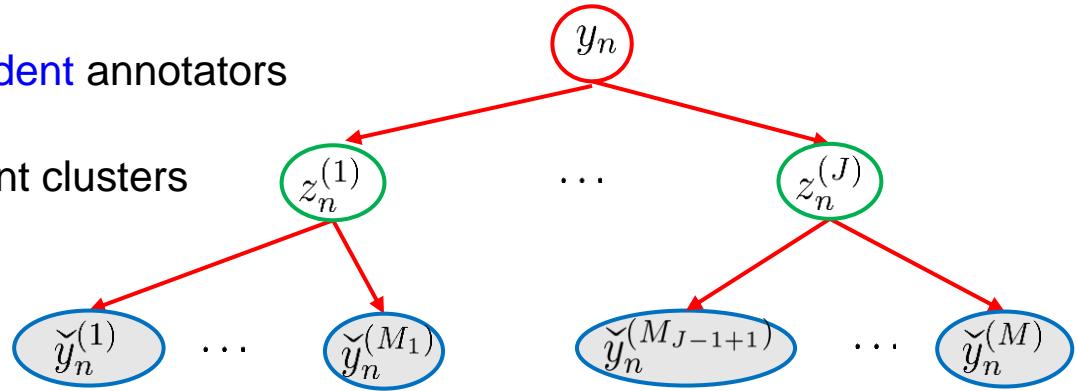
DS model for dependent classifiers

- DS model does not hold for **dependent** annotators

- Annotators grouped in $J < M$ disjoint clusters

➤ Indices collected in $\{\mathcal{C}_j\}_{j=1}^J$

$$M_j = |\mathcal{C}_j|$$



Dependencies captured by a hidden variable $z_n^{(j)} \in \{1, \dots, K\}$ per group

(as) Given $z_n^{(j)}$ annotator responses, $\{\check{y}_n^{(m)}\}_{m \in \mathcal{C}_j}$ are independent

(as) Given y_n hidden variables $\{z_n^{(j)}\}_{j=1}^J$ are independent

- Confusion matrices for annotators and hidden variables

$$\Xi^{(j)}(k, k') = \Pr(z_n^{(j)} = k | y_n = k') \quad \tilde{A}_m(k, k') = \Pr(\check{y}_n^{(m)} = k | z_n^{(j)} = k'), m \in \mathcal{C}_j$$

- Hierarchical label fusion

Caveat. Group indices are unknown, and must be estimated!

Identifying annotator dependencies

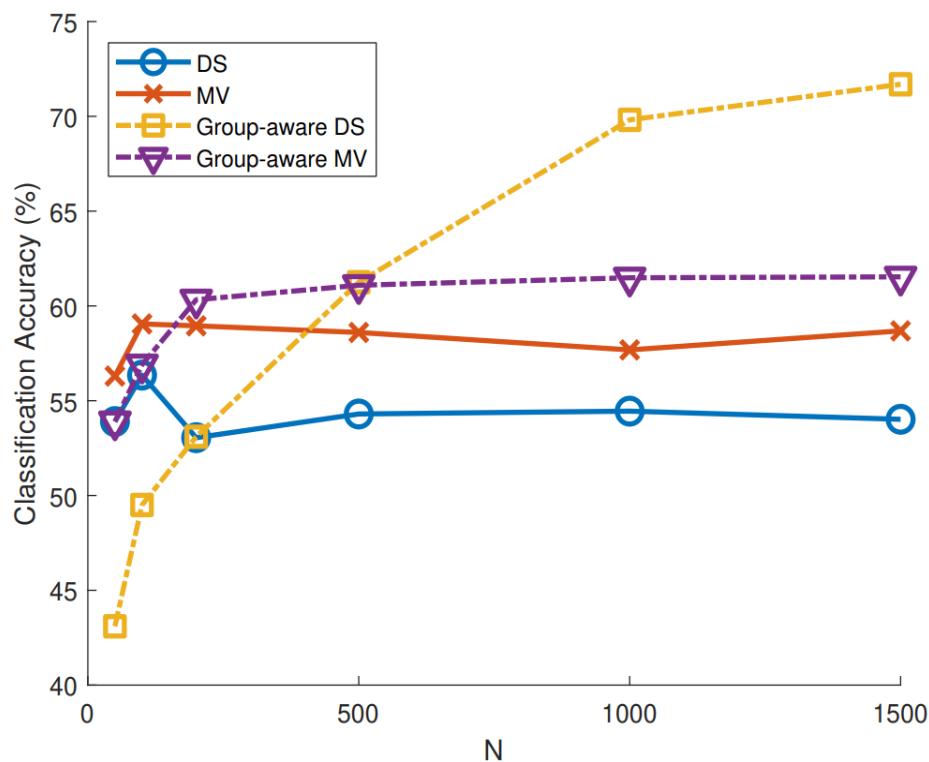
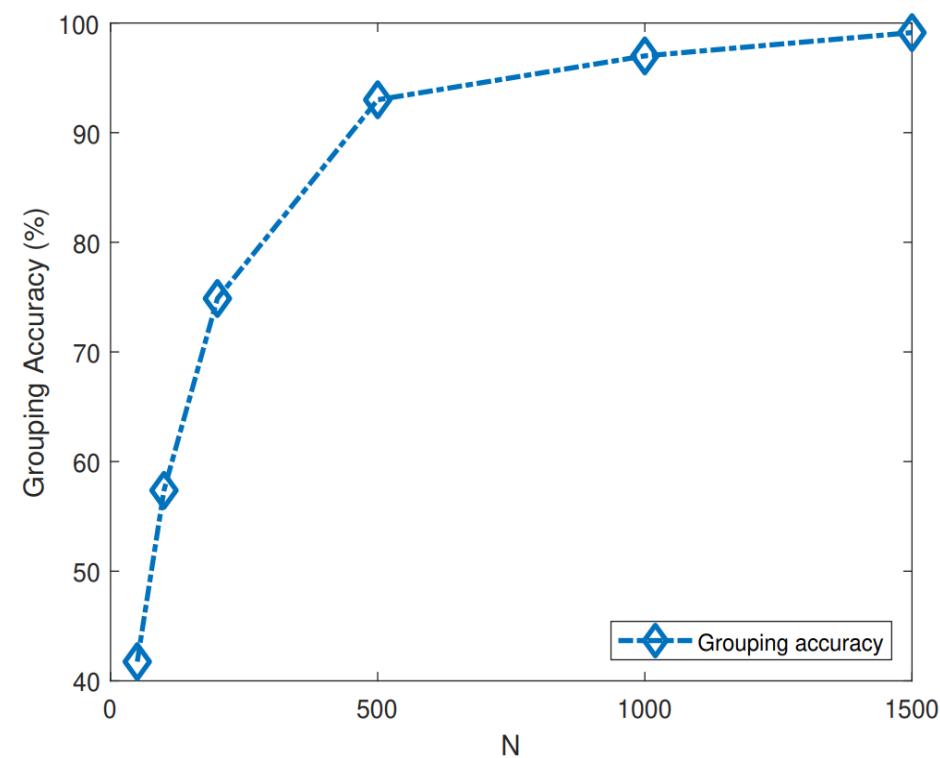
- Spectral methods for binary and general DS model
[Jaffe et al'16, Traganitis-Giannakis'22]
 - Annotator clusters identified from second-order statistics
- Worker-cluster model [Imamura et al'18]
 - J distinct confusion matrices
- Bayesian methods – joint group and label inference
 - Community Bayesian Classifier Combination [Venanzi et al'14]
 - Bayesian nonparametrics [Moreno et al'15]

Synthetic tests

□ Synthetic dataset: $N = 5,000, M=40, K=3, J = 5$

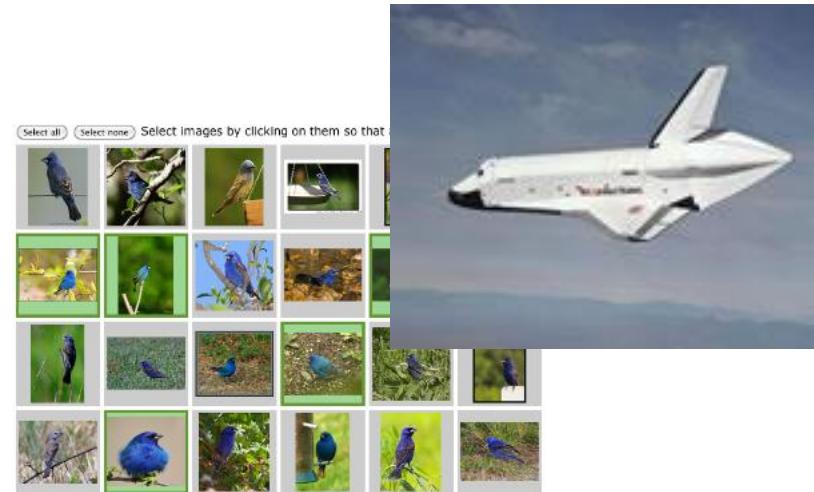
□ DS: Expectation maximization
MV: Majority voting

□ Group-aware DS (MV): Proposed algorithm + DS (MV)



Real data tests

- Bluebird, $N = 108, M = 39, K = 2$
- Shuttle, $N = 58,000, M = 15, K = 7$
- CoverType, $N = 581,012, M = 15, K = 7$



Classification accuracy

Dataset	MV	DS	Group-aware MV	Group-aware DS
Bluebird	75.92%	87.96%	72.22% ($J = 7$)	91.6% ($J = 7$)
Shuttle	98.82%	97.54%	98.93% ($J = 3$)	99.46% ($J = 3$)
CoverType	75%	48.46%	74.49% ($J = 4$)	74.09% ($J = 4$)

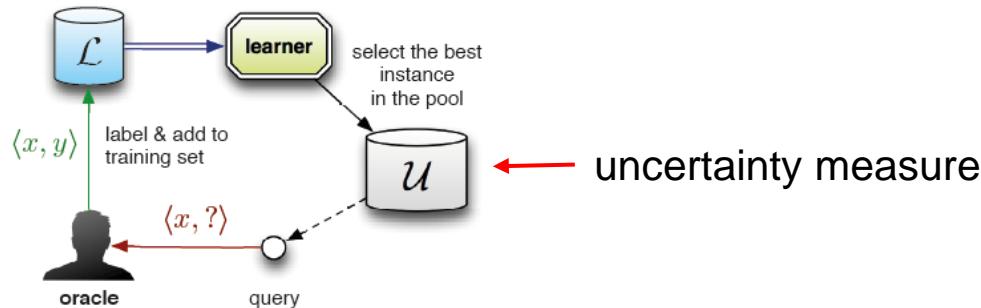
Outline

- Motivation and problem statement
- Part I: Combining crowdsourced labels
- Part II: End-to-end (E2E) learning with crowdsourced labels
- **Part III: Other aspects of crowdsourcing**
 - Other types of annotations
 - Adversary aware crowdsourcing
 - Active crowdsourcing
 - Bias and fairness issues
- Conclusions and open issues

Learning with label budget constraints

General Goal: For a given budget of queries, maximize accuracy by actively selecting which instances (data) to label (“query”).

- Standard machine learning approach: Active Learning (AL)
 - allows for sample (label) complexity reduction
- Data selection – uncertainty sampling
 - Select the most **uncertain** instance according to the learner’s model



- Crowdsourcing requires choosing which annotators to query

Active crowdsourcing

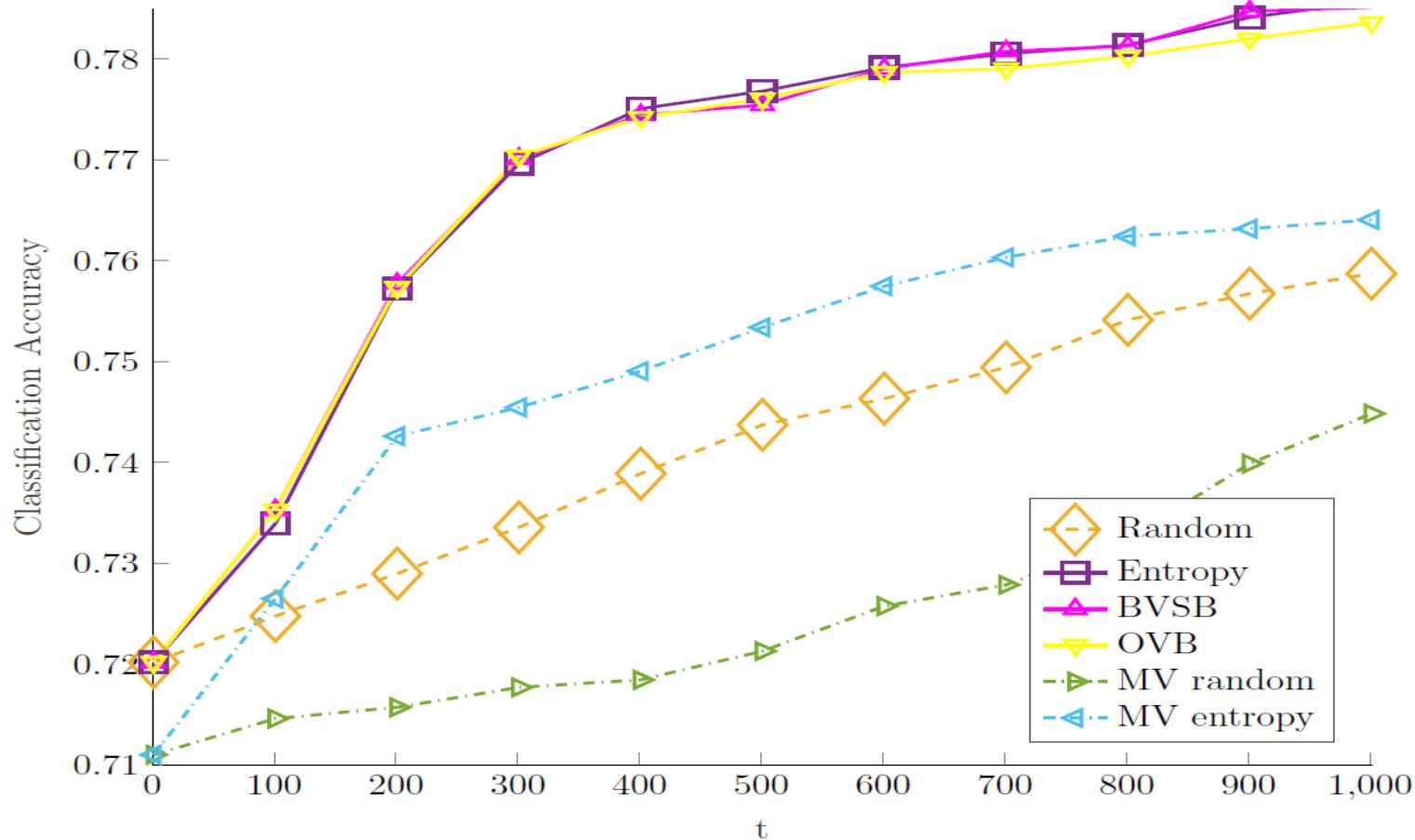
- Data selection
 - Select the most **uncertain** instance according to “crowd”
 - Uncertainty quantified by current label posterior $\Pr(y_n|\check{\mathcal{Y}})$
- Annotator selection
 - Select “best” given current annotator reliabilities
 - Confusion matrix estimates may be unreliable – **exploration** required
- Online or streaming algorithms preferable!
 - Parametric & Gaussian Process models [Welinder et al '10, Yan et al '11, Rodrigues et al '14]
 - AL w/ batch algorithms [Venanzi et al '15]
 - Online EM [Traganitis et al '20]
 - Multi-armed bandit approach [Rangi-Francescetti'18]

Numerical tests: Dog dataset

□ $N = 807, M = 109, K = 4$

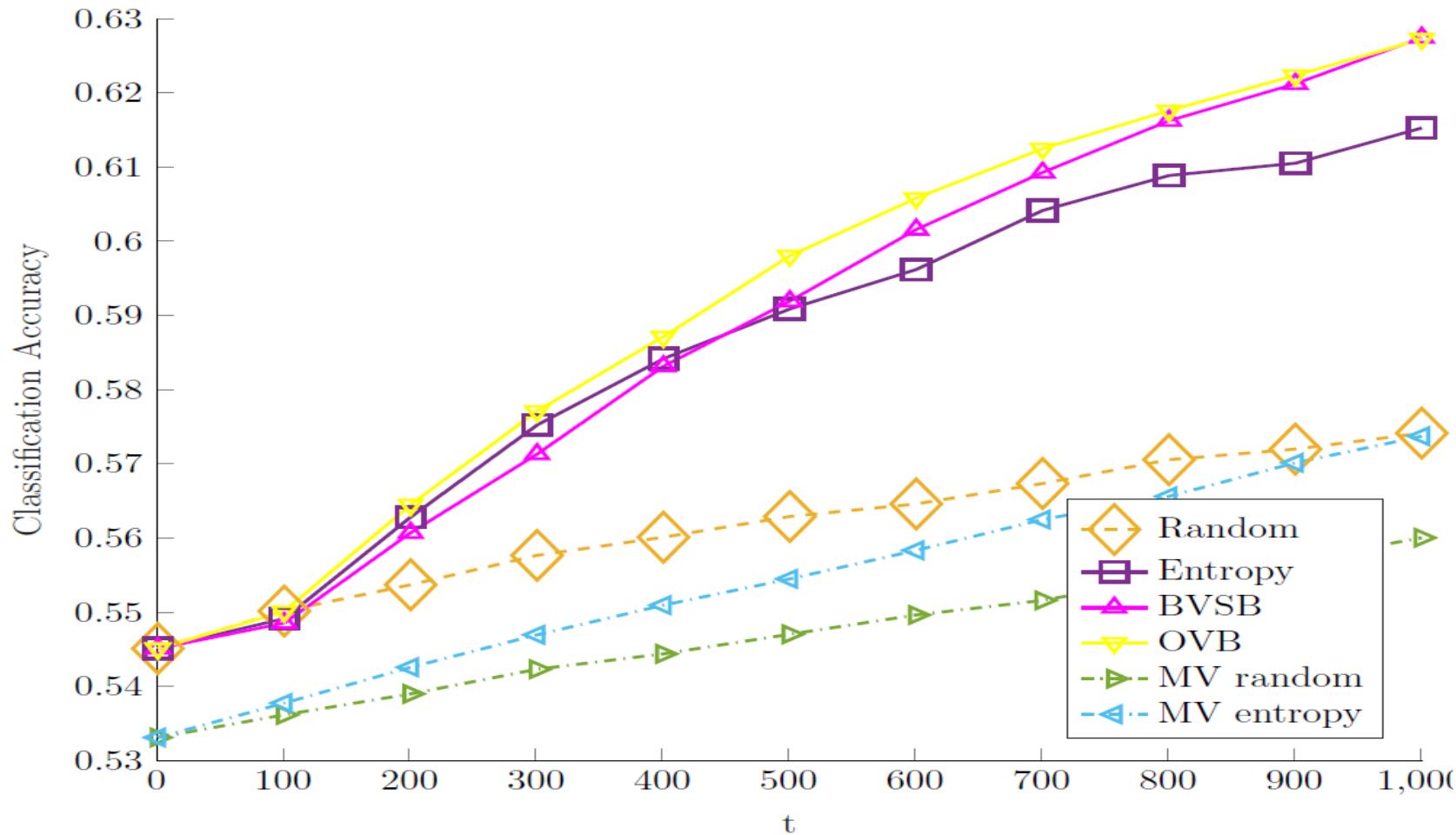
$$\lambda_t = t^{-0.6}$$

$$\varepsilon = \begin{cases} 0.5 & t = 1, \dots, 100 \\ 0.05 & t = 101, \dots, 1000 \end{cases}$$



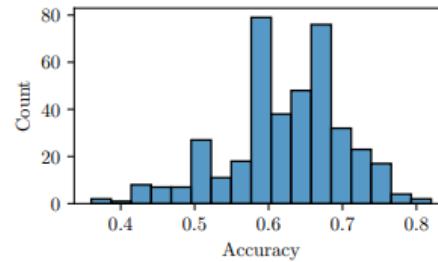
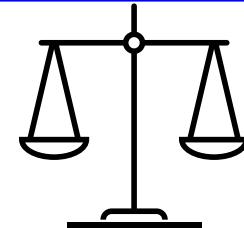
Web dataset

□ $N = 2,665, M = 155, K = 5$

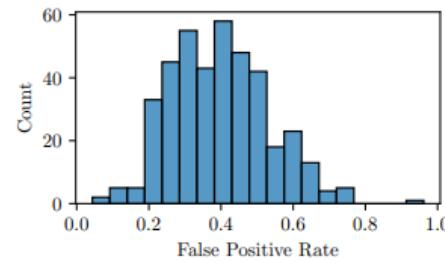


Bias and fairness in crowdsourcing

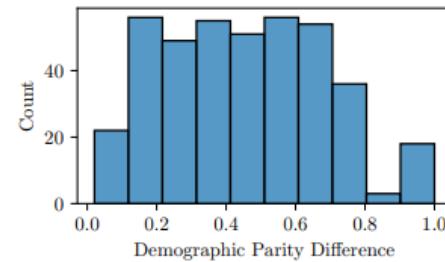
- ML/AI applied in areas w/ societal impact
 - Loans, credit score, penal system, job applications etc.
- Typical performance metrics do not capture bias against sensitive attributes
 - Race, age, affiliation etc.
 - Fairness metrics required – Demographic parity, equal opportunity etc.
- Annotators also exhibit bias



(a) Accuracy



(b) False Positive Rate



(c) Demographic Parity

Taking bias into account

- ☐ Crowdsourcing perpetuates annotator biases

- Unfair annotators yield unfair ML models

Classifier	TD Algorithm	Delta Accuracy	Delta Demo Par. Diff
Random Forest	Majority Voting	6.04	0.06
Random Forest	Dawid-Skene	7.63	0.04
Random Forest	Learning from Crowd	7.54	0.04
Logistic Regression	Majority Voting	2.55	0.04
Logistic Regression	Dawid-Skene	2.77	0.05
Logistic Regression	Learning from Crowd	2.86	0.05

- Label integration does not optimize for fairness

- ☐ Annotator behavior w.r.t. sensitive data attribute $z_n \in \{0, 1\}$ captured via

$$[\mathbf{A}_m(z_n = z)]_{k,k'} = \Pr(\tilde{y}_n^{(m)} = k | y_n = k', z_n = z)$$

- Similar to instance specific confusion matrices
 - Fair annotator assignment [Goel et al'19]

requires ground-truth labels to estimate confusion matrices

Outline

- Motivation and problem statement
- Part I: Combining crowdsourced labels
- Part II: End-to-end (E2E) learning with crowdsourced labels
- Part III: Other aspects of crowdsourcing
- Conclusions and open issues

Conclusions and open issues

Take home 1. Crowdsourcing learns from unreliable information sources

- Also assesses source reliability in learning tasks
- Two major paradigms: Two-stage and E2E
- ML/AI catalyst



Take home 2. Two-stage approaches require only annotator labels

- Available approaches tradeoff accuracy for simplicity
- Workhorse algorithms: EM or VI
- Moment-based algorithms for reliable initialization

Take home 3. E2E approaches simultaneously learn a model and annotator reliabilities

- Inputs: data features and annotator labels
- Improved performance compared to two-stage
- Flexible DL models

Open issues and future directions

❑ Application-specific crowdsourcing

- Unique challenges and design
- Collaboration of human and machine annotators

❑ Crowdsourcing in classical data/information fusion

- Robust learning from multi-view data
- Game-theoretic crowdsourcing

❑ Open research directions

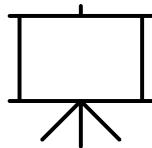
- Crowdsourcing and AI/LLMs
- Bias and fairness
- Transfer learning/meta learning

Relevant links



Learning From Crowdsourced Noisy Labels: A Signal Processing Perspective

<https://arxiv.org/abs/2407.06902>



LINK TO SLIDES

https://github.com/shahana-ibrahim/Learning-from-Crowdsourced-Noisy-Labels/blob/main/Crowdsourcing_ICASSP25.pptx



[https://github.com/shahana-ibrahim/
Learning-from-Crowdsourced-Noisy-Labels](https://github.com/shahana-ibrahim/Learning-from-Crowdsourced-Noisy-Labels)

Thank you!
