When Annotators Disagree: A Principled Approach to Learning with Noisy Labels

Maria Sofia Bucarelli[1,⋆], Antonio Purificato[1,2], Andrea Bacciu[2], Lucas Cassano[3], Federico Siciliano[1], Anil Nelakanti[4], Amin Mantrach[2], and Fabrizio Silvestri[1]

***Abstract*—In practical settings, classification datasets are often labeled by humans, leading to potential noise due to varying annotations from different individuals. The exact noise distribution impacting these labels is typically unknown; however, one quantity we can measure and attempt to exploit is inter-rater agreement. Building on this, our work makes key contributions: we (i) demonstrate how inter-annotator statistics can be used to estimate the label noise distribution; (ii) propose methods that leverage these estimates to train models on noisy data; and (iii) derive generalization bounds within the empirical risk minimization framework that depend on the estimated noise characteristics. Finally, we present experiments that support our findings. Our code can be found at https://github.com/amazon-science/learning_under_noisy_labels.**

***Index Terms*—Noisy Labels, Crowdsourcing, Supervised Learning**

Impact Statement

This research addresses a fundamental challenge in artificial intelligence: the reliability of human-labeled datasets used to train AI systems. When multiple people label the same data differently, it creates uncertainty that can affect AI model performance. Our work provides practical tools to measure and manage this uncertainty, making AI systems more reliable and trustworthy.

I. Introduction

Supervised learning has made significant strides in recent decades, both theoretically and practically. Empirical risk minimization serves as a common learning framework (Vapnik, 1998), which assumes that models are trained on data independently and identically distributed (*i.i.d.*) from the joint distribution of features and labels. Under this assumption, the generalization bounds suggest that, with sufficient training data, any desired performance can be achieved. However, in many real-world scenarios, the assumption of *i.i.d.* sampling from the true feature-label distribution is often violated due to imperfections in data collection and labeling. Training data is frequently annotated by human annotators who have a non-zero probability of making errors. It has been reported in Song et al. (2020) that the ratio of corrupted labels in some real-world datasets is between 8.0% and 38.5%. As a consequence of the presence of incorrect labels in the training dataset, the aforementioned assumption is violated and hence performance guarantees based on generalization bounds no longer hold.

This gap between theory and practice prompts the question of whether it is possible to learn from datasets with noisy labels while still ensuring performance guarantees. This issue has received significant attention recently and has been positively addressed in some cases (Natarajan et al., 2013; Patrini et al., 2017). Indeed, multiple works have introduced learning algorithms that can cope with datasets with incorrect labels while guaranteeing desirable performance through provable generalization bounds (Yao et al., 2021). However, these solutions do not solve the entirety of the problem due to the fact that they rely on precise knowledge of the error rate to which the labels are subject, which is often unknown in practice. Various approaches have been proposed to estimate this error rate (Patrini et al., 2017; Xia et al., 2019; Yao et al., 2020), but many depend on assumptions that may not hold in practice, such as the presence of anchor samples (Patrini et al., 2017). Ideally, it would be desirable to develop learning algorithms that are both robust to noisy labels and come with performance guarantees.

An approach, often used in industry, to mitigate errors from human raters is to have the same dataset labeled multiple times by different annotators, then combine these labels using methods like majority vote or soft labeling (Purificato et al., 2025). Inter-Annotator Agreement (IAA) scores (like Cohen's kappa (Cohen, 1960) and Fleiss' kappa (Fleiss et al., 1971)) serve as metrics directly related to the probability of labeling errors.

Given the direct link between IAA and the error rate among raters, this estimate can be used to adjust learning algorithms, making them more robust to label noise. This is the primary focus of our work.

Motivation and Contributions: This work is motivated by two main points: (i) the lack of established methodologies to directly leverage readily available Inter-Annotator Agreement (IAA) statistics for precise label noise distribution estimation; and (ii) the generalization bounds of existing noise tolerant training methods often rely on **unknown** quantities (like the true noise distribution) instead of on quantities that can be measured (like the IAA statistics).

Our contributions are the following: (i) we introduce a method to estimate label noise distribution using IAA statistics; (ii) we show how to use this estimate to train on noisy datasets; (iii) we establish generalization bounds for

This paragraph of the first footnote will contain the date on which you submitted your paper for review.

⋆ Corresponding Author (bucarelli@diag.uniroma1.it).

[1]Sapienza University of Rome, [2]Amazon, [3]Independent Researcher, [4]IIIT Hyderabad

Anil Nelakanti completed this work while at Amazon. He is now with IIIT Hyderabad.

Maria Sofia Bucarelli completed this work while at Sapienza University of Rome. She is now with CNRS.

This paragraph will include the Associate Editor who handled your paper.

our methods that rely on **known** quantities; (iv) we quantify the "distributional shift" in the expected loss of a classifier between noisy and true-label distributions can be bounded by the spectral gap of the noise transition matrix and the class-prior matrix; and (v) we provide experiments across different tasks to validate the proposed theory. A preliminary version of this work appeared in Bucarelli et al. (2023).

II. Related works

Our work is related to literature on three main topics: (a) robust loss function design, (b) label aggregation and (c) noise rate estimation.

a) Robust loss functions: Loss functions known as *symmetric*—where the sum of risks across all categories is constant for any example—are robust to label noise (Ghosh et al., 2017). Examples include the $0-1$ loss, Ramp Loss, and (softmax) Mean Absolute Error (MAE). While MAE is noise-tolerant, unlike categorical cross-entropy (CCE), it may underperform when training deep neural networks (DNNs) in challenging domains (Zhang and Sabuncu, 2018). To leverage the robust properties of the MAE and the efficiency in training of the CCE , Zhang and Sabuncu (2018) propose a loss function that can be seen as a generalization of MAE and CCE. Zhu et al. (2023) propose a temperature-controlled loss function that can transition between cross-entropy and a symmetric loss based on the temperature parameter value. Natarajan et al. (2013) present a robust loss correction method, which is extended to multiclass by Patrini et al. (2017). More recently Wani et al. (2024) proposed a robust loss function that leverages the distance to class centroids in the latent space and incorporates a discounting mechanism, aiming to diminish the influence of samples that lie distant from all class centroids.

b) Label aggregation: Supervised learning often involves multiple imperfect annotators labeling the data. Their separate labels are typically aggregated into a single label before training models. Various approaches have been introduced to handle noise from multiple annotators during training (Sheng and Zhang, 2019).

For aggregating labels from multiple annotators, majority voting is the simplest approach but may perform poorly with high noise levels or in multiclass settings. Probabilistic generative methods, with a stronger theoretical foundation, often perform better. These methods estimate the true label posterior to determine the actual labels or adjust the loss function weighting (Dawid and Skene, 1979). Inspired by Dawid and Skene (1979), early methods use the Expectation-Maximization (EM) algorithm for statistical inference. Raykar et al. (2010) introduce Bayesian estimation to model workers' sensitivity and specificity, improving binary biased labeling performance. GLAD (Whitehill et al., 2009) considers as parameters the expertise of each labeler, and the difficulty of each sample. EM algorithm is an effective tool to solve the MLE or MAP estimates with latent variables, but setting initial model parameter values correctly is crucial for inference accuracy (Sheng and Zhang, 2019).

To address the limitations of probabilistic generative methods, researchers have introduced different techniques. Karger et al. (2011) introduce a consensus algorithm that determines the correct answer for each task using worker and task messages. The method uses belief propagation to derive both annotators' reliabilities and an estimate for samples' classes. In Li and Yu (2014), the authors propose an iterative weighted majority voting (IWMV) method that optimizes the error rate bound and approximates the oracle MAP rule. Peterson et al. (2019) exploit the availability of multiple human annotations to construct soft labels and conclude that this increases performance in terms of generalization to out-of-training-distribution test datasets and robustness to adversarial attacks.

c) Noise rate estimation: A number of approaches have been proposed for estimating the noise transition matrix (i.e. the probabilities that correct labels are changed for incorrect ones) (Patrini et al., 2017; Zhu et al., 2022). Usually, these methods use a small number of anchor points (that are samples that belong to a specific class with probability one) (Hendrycks et al., 2018). In particular, Patrini et al. (2017) propose a noise estimation method based on anchor points, with the intent to provide an "end-to-end" noise-estimation-and-learning method. Due to the lack of anchor points in real data, some works focused on a way to detect anchor points in noisy data (Yao et al., 2020; Xia et al., 2019). Yao et al. (2020) introduce an intermediate class to avoid directly estimating the noisy class posterior. They factorize the original transition matrix into two easier-to-estimate transition matrices, effectively changing the problem from estimating the noisy class posterior to fitting the noisy labels. Zhang et al. (2021) also propose an iterative noise estimation heuristic that aims to partly correct the error and point out that the methods introduced by Patrini et al. (2017) and Yao et al. (2020) have an error in computing anchor points, and provide conditions on the noise under which the methods work or fail. Xia et al. (2019) provide a solution that can infer the transition matrix without anchor points. Indeed, they use the instances with the highest class posterior probabilities for noisy data as anchor points. In contrast to previous methods that utilize anchor points, our work does not require anchor points or a validation set to learn the noise rate and we only rely on noisy data for training our model. Moreover, we address the generalization properties of our proposed model, deriving bounds that depend on the estimated noise transition matrix, which most prior works do not explore.

III. Problem formulation

A. Notation

In this paper, we adopt the following notation. Matrices and sets are denoted by upper-case and calligraphic letters, respectively. The space of d-dimensional feature vectors is denoted by $\mathcal{X} \subset \mathbb{R}^d$.

We denote by C the number of classes and by e_j the j-th standard canonical vector in $\mathbb{R}^C$, namely the vector that has 1 in the j-th position and zero in all the other positions. $\mathcal{Y} = \{e_1, \dots, e_C\} \subset \{0,1\}^C$ is the label set. Feature vectors and labels are denoted by x and y, respectively. $\mathcal{D}$ is the joint distribution of the feature vectors and labels, i.e. $(x, y) \sim \mathcal{D}$. The sampled dataset of size n is denoted by $\widehat{\mathcal{D}} = \{(x_i, y_i)\}_{i=1}^n$. $f(x)$ denotes the output of the classifier f for feature vector x and is a C-dimensional vector. All vectors are column vectors.

We denote by $\ell(t, y)$ a generic loss function for the classification task that takes as input C dimensional vectors t and y. In practice t will contain the prediction of the model, and y will be the ground-truth label as a one-hot encoded vector. Namely $\ell : [0, 1]^C \times \mathcal{Y} \to \mathbb{R}$.

B. Background

We consider the classification problem within the supervised learning framework, where the ultimate goal is to minimize the ℓ -risk $R_{\ell,\mathcal{D}}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)]$, for some loss function ℓ. We denote by $\mathcal{D}$ the joint distribution of feature vectors x and labels y. In practice, since the distribution is unknown instead of minimizing $R_{\ell,\mathcal{D}}(f)$ we minimize an empirical risk over some sampled dataset $\widehat{\mathcal{D}}$:

$$\widehat{R}_{\ell,\widehat{\mathcal{D}}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) = \mathbb{E}_{(x,y)\sim\widehat{\mathcal{D}}}[\ell(f(x), y)].$$

In this work, we assume that the true labels y_i are unknown and consider two scenarios, both relying on H annotators.

1) Scenario I: In this scenario we have access to the H labels provided by the annotators for each sample, where $y_{i,a}$ refers to the label provided by the a-th annotator for the i-th sample. For a given feature vector x_i, the distribution of labels provided by annotator a is given by its noise transition matrix T_a, which is defined as follows:

$$(T_a)_{i,j} := \mathbb{P}(y_a = j | y = i) \tag{1}$$

Assumption 1. *We assume that all annotators have the same noise transition matrix (i.e. $T_a = T \,\forall\, a$), that T is symmetric and that its diagonal elements are larger than 0.5 (i.e. $\mathbb{P}(y_a = i|y = i) > 0.5, \forall i \in \{1, \dots C\}$).*

Note that by definition T is right stochastic and hence also doubly stochastic. It is also strictly diagonally dominant and therefore non-singular.

Proposition III.0.1. *T is positive definite.*

Proof: Since T is symmetric it follows that all eigenvalues are real. Combining the fact that it is strictly diagonally dominant with Gershgorin's theorem we conclude that all eigenvalues lie in the range $(0, 1]$ and hence T is positive definite. ■

Assumption 2. *We assume that the annotators are conditionally independent on the true label y:*

$$\mathbb{P}(y_a, y_b|y) = \mathbb{P}(y_a|y)\mathbb{P}(y_b|y).$$

We now define the IAA matrix M_{ab} between annotators a and b as follows:

$$(M_{ab})_{i,j} := \mathbb{P}(y_a = i, y_b = j) \tag{2}$$

Proposition III.0.2. *Leveraging Assumption 2 the agreement matrix $M_{a,b}$ can be written as follows:*

$$\begin{aligned} M_{a,b} &= {T_a}^T D T_b \\ D &:= diag\{\nu\} \\ \nu &:= [\mathbb{P}(y = 1), \cdots, \mathbb{P}(y = C)]^T. \end{aligned} \tag{3}$$

Due to Proposition III.0.1 and the fact that D is positive definite, it follows that all matrices $M_{a,b}$ are invertible.

Assumption 3. *We assume that the class probabilities (and hence D) are known.*

Due to Assumption 1, all annotators share the same noise transition matrix T. Therefore M_{ab} is independent of a and b, and from now on, we remove this dependency in the notation (i.e. we get $M = T^T DT$). Furthermore, since T is invertible and D is diagonal and positive definite, it follows that M is also positive definite.

Note that since we have access to all the labels provided by the H annotators for all the samples, we can obtain an estimate of M, which we denote $\widehat{M}$.

Assumption 4. *We assume that $\widehat{M}$ is a consistent estimator.*

For the case of two annotators, one possible consistent estimator $\widehat{M_{a,b}}$ that exploits its symmetry condition is given by:

$$(\widehat{M_{a,b}})_{i,j} = \sum_{k=1}^{n} \frac{\mathbb{1}(y_{a,k}{=}i, y_{b,k}{=}j) + \mathbb{1}(y_{a,k}{=}j, y_{b,k}{=}i)}{2n} \tag{4}$$

If the annotators have the same transition matrix, M will be the same for all pairs of annotators. So we can estimate M, in the case of $H \geq 2$ by averaging the estimators $\widehat{M}_{ab}$ obtain by Eq. (4) for all possible pairs of annotators. The estimator in this case can be written as:

$$(\widehat{M})_{i,j} = \frac{1}{H(H{-}1)} \sum_{a=1}^{H} \sum_{\substack{b=1 \\ b\neq a}}^{H} \sum_{h=1}^{n} \frac{\mathbb{1}(y_{a,h}{=}i, y_{b,h}{=}j)}{n}. \tag{5}$$

2) Scenario II: In the second scenario, for each i-th sample we are given a unique label $\tilde{y}_i$ produced by aggregating the H individual labels according to some known aggregating policy (like majority vote). In this case, not having access to the individual annotations we assume that $\widehat{M}$ is provided.

The probability that label y_i is corrupted to some other label $\tilde{y}_i$ is given by the *aggregated noise transition matrix* $\Gamma \in [0, 1]^{C\times C}$, where $\Gamma_{ij} := \mathbb{P}(\tilde{y} = j|y = i)$ is the probability of the true label i being flipped into a corrupted label j and C is the number of classes. Note that by definition Γ is a right stochastic matrix that is determined by T, the amount of annotators H and the aggregating policy. We will study both the case where $\Gamma = T$, and the case in which there exists a generic Lipschitz function ϕ so that $\Gamma^{-1} = \phi(T)$.

There are different choices to construct the dataset that lead to $\Gamma = T$. If we decide to use only one annotator, for instance a, to build the final dataset, namely for each sample $\tilde{y}^i = y^i_a$ we have $\Gamma = T_a$. Or if annotators are homogeneous, i.e. they have the same noise transition matrix T, and to build the final dataset we decide to randomly select the label of one of the annotators we have that $\Gamma = T$.

Even restricting ourselves to the case of homogeneous annotators, depending on the rule with which we build the dataset we can have a more complex relationship between the matrix T and Γ.

We also obtain generalization bounds in the case where an estimate of the agreement matrix M is not available and

we only have access to a scalar representation of the inter-annotator agreement, in particular we consider the case where the Cohen's κ is given.

3) Objective: The objective in both scenarios is to: (i) use $\widehat{M}$ to estimate the noise transition matrices (T and Γ); (ii) leverage these estimates to be able to learn from the noisy dataset in a more robust manner; and (iii) obtain generalization bounds for the resulting learning methods.

Our main contributions are divided as follows: In the first section, we show how to estimate the noise transition matrices. Next, we indicate how to leverage these estimates to learn from datasets with noisy labels. Finally, we obtain generalization bounds that depend on the Rademacher complexity of the function class, for a bounded and Lipschitz loss function, given the estimated noise transition matrices.

C. Estimation of the noise transition matrices

We start by stating the following Lemma that allows us to write the unknown matrix T (and its inverse), as a function of D and M.

Lemma III.1. *If $D^{\frac{1}{2}}$ commutes with T we have that:*

$$T = U\Lambda^{\frac{1}{2}}U^T \tag{6}$$

$$T^{-1} = U\Lambda^{-\frac{1}{2}}U^T \tag{7}$$

$$D^{-\frac{1}{2}}MD^{-\frac{1}{2}} = U\Lambda U^T \tag{8}$$

where $U\Lambda U^T$ is the eigenvalue decomposition of $D^{-\frac{1}{2}}MD^{-\frac{1}{2}}$ (i.e. U is some orthogonal matrix and Λ is a diagonal positive definite matrix).

Proof: From Eq. (3) we have that:

$$M = TDT = D^{\frac{1}{2}}TTD^{\frac{1}{2}} \Rightarrow D^{-\frac{1}{2}}MD^{-\frac{1}{2}} = T^2.$$

Note that T and $D^{\frac{1}{2}}MD^{\frac{1}{2}}$ are positive definite since D and M are and thus have eigenvalue decompositions of the form:

$$T = U_T\Lambda_T U_T^T \quad D^{-\frac{1}{2}}MD^{-\frac{1}{2}} = U_M\Lambda_M U_M^T.$$

with U_x orthogonal and Λ_x diagonal positive definite matrices. We obtain:

$$T^2 = U_T\Lambda_T^2 U_T^T \text{ and } T^2 = U_M\Lambda_M U_M^T.$$

Since $U_M\Lambda_M U_M^T$ is an eigenvalue decomposition of T^2 we conclude that:

$$T = U_M\Lambda_M^{\frac{1}{2}}U_M^T, \qquad T^{-1} = U_M\Lambda_M^{-\frac{1}{2}}U_M^T$$

■

a) Commutativity Hypothesis: The commutativity hypothesis is satisfied when D and T are so that:

$$\frac{\sqrt{d_i}}{\sqrt{d_j}}t_{ij} = t_{ij} \;\; \forall i \text{ and } j.$$

This condition is satisfied either if $d_i = d_j$ or if $t_{ij} = 0$, namely every class so that the probability of going from class i to class j (and viceversa) is not zero is equiprobable.

So T has to be block diagonal, or better reducible by a permutation of the classes to a block diagonal matrix and D has to have all equal elements on indices relatives to the same block in T. For example:

$$T = \begin{pmatrix} T_1 & 0 & 0 \\ 0 & T_2 & 0 \\ 0 & 0 & T_3 \end{pmatrix}, D = \begin{pmatrix} D_1 & 0 & 0 \\ 0 & D_2 & 0 \\ 0 & 0 & D_3 \end{pmatrix}, D_i = d_i I.$$

Even if T is not block-diagonal, it must be reducible to that form via class permutation. From the technical point of view, we have noticed that solving this equation is extremely complicated without making such assumptions. Another assumption we could have used, also required by Potter (1966) to solve the same problem, is requiring that the matrix $D^{\frac{1}{2}}T$ has diagonal Jordan decomposition. However, this assumption is more complicated to translate at the level of the structure of the matrices T and D.

From a practical point of view, making such an assumption means that there are classes that annotators can confuse with one another while they never swap between them, other classes. For example, if the problem is to classify images and the classes are "cat", "lynx", "bats", "bird", "cougar"; we can think that the annotators have a non-zero probability of confusing with each other the feline classes "lynx", "cat", "cougar", while they have zero probability of assigning a picture of a lynx the label "bird". Commutativity is guaranteed in the case of a uniform distribution over the classes. There are many applications where we expect the distribution over the classes to be uniform and not to have any class with a higher probability. In general, we can fall back to an approximation of this case by reducing the samples.

We could use Lemma III.1 to estimate T as follows:

$$\widehat{T} = \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T \tag{9}$$

where $\widehat{U}\widehat{\Lambda}_M\widehat{U}^T$ is the eigenvalue decomposition of $D^{-\frac{1}{2}}\widehat{M}D^{-\frac{1}{2}}$. However such estimate can result in matrices that are not doubly stochastic, or diagonally dominant due to estimation errors. A more accurate estimate of T could be obtained as $\widehat{T} = \pi(\widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T)$ where π is a projection operator to the set of doubly stochastic, positive definite matrices with diagonal elements greater than 0.5 and non-negative entries (which is a convex set). We can obtain such projection by solving the following optimization problem:

$$\widehat{T} = \pi(\widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T) = \underset{B\in\mathcal{C}}{\operatorname{argmin}} \left\| B - \widehat{U}\widehat{\Lambda}_M^{1/2}\widehat{U}^T \right\|_2^2 \tag{10}$$

where $\mathcal{C}$ is the set of symmetric, row-stochastic, non-negative matrices with diagonal entries ≥ 0.5:

$$\mathcal{C} = \left\{ B \in \mathbb{R}^{n\times n} \middle| B = B^T, \textstyle\sum_j B_{i,j} = 1, B_{i,j} \geq 0, B_{i,i} \geq 0.5 \right\} \tag{11}$$

Note that this optimization problem is convex because the constraints are linear and for symmetric matrices it holds that:

$$||\widehat{T} - \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T||_2^2 = \lambda_{\max}(\widehat{T} - \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T)$$

W hich is a convex function of $\widehat{T}$.

To summarize, T can be estimated as follows. First, obtain an estimate of M. Then obtain the eigenvalue decomposition of $D^{-\frac{1}{2}}\widehat{M}D^{-\frac{1}{2}} = \widehat{U}\widehat{\Lambda}\widehat{U}^T$ (note that this decomposition always exists because $D^{-\frac{1}{2}}\widehat{M}D^{-\frac{1}{2}}$ is symmetric). Finally obtain the estimate as: $\widehat{T} := \pi(\widehat{U}\widehat{\Lambda}^{\frac{1}{2}}\widehat{U}^T)$.

Note that once the estimate of $\widehat{T}$ is obtained, $\widehat{\Gamma}$ can be obtained since we assumed the label aggregation policy to be known.

Lemma III.2. *Let $M_{a,b}$ be the agreement matrix for annotators a and b defined in Eq. (2) and $\widehat{M_{a,b}}$ be the estimated agreement matrix defined in Eq. (4) and let $||.||_p$ be the matrix norm induced by the p vector norm. For every $p \in [1, \infty]$ and for every $\delta > 0$, with probability at least $1 - \delta$:*

$$||M_{a,b} - \widehat{M_{a,b}}||_p \leq \sqrt{\frac{C^2}{2n} \ln \frac{2C^2}{\delta}}.$$

where $\mathbb{P}^n$ denotes the probability according to which the n training samples are distributed, i.e. we are assuming that the samples are independently drawn according the probability $\mathbb{P}$.

Proof: To prove the claim we apply Hoeffding's inequality to the random variables $X_h^{ij} = \mathbb{1}_{y_{a_h}=i, =y_{b_h}=j}$. Indeed $0 \leq X^{ij} \leq 1$ and $\widehat{M}_{ij} = \frac{1}{n}\sum_{h=1}^{n} X_h^{ij}$, while $\mathbb{E}[X_h^{ij}] = M_{ij}$. Notice that the random variables $X_1^{ij} \dots X_n^{ij}$ are independent since we assume samples to be independent with respect to each other and so it follows that $(x_h, y_{a_h}, y_{b,h}), (x_k, y_{a_k}, y_{b_k})$ are independent. The proof is completed using union bounds over the (i, j) indices varying in $\{1, C\}^2$. ∎

From Lemma III.2 it follows that if $\widehat{M}$ is estimated as in Eq. (5), since $\widehat{M}$ is an average of $\widehat{M_{ab}}$ it holds that for every $p \in [1, \infty]$ and for every $\delta > 0$, with probability at least $1-\delta$:

$$||M - \widehat{M}||_p \leq \sqrt{\frac{C^2}{2n} \ln \frac{2C^2}{\delta}}. \tag{12}$$

Theorem III.3. *Let T be the noise transition matrix defined as in Eq. (1) and $\widehat{T}$ its estimate (defined as in Eq. (10)).*

With probability at least $1 - \delta$:

$$||T - \widehat{T}||_2 \leq \frac{C(\sqrt{C}+1)\lambda_{\max}(D)}{\lambda_{\min}(\widehat{T})} \sqrt{\frac{1}{2n} \ln \frac{2C^2}{\delta}}$$

$$||T^{-1} - \widehat{T}^{-1}||_2 \leq \frac{9C(\sqrt{C}+1)\lambda_{\max}(D)}{\lambda_{\min}(\widehat{T})^2} \sqrt{\frac{1}{2n} \ln \frac{2C^2}{\delta}}$$

for $n > \frac{C^2(\sqrt{C}+1)^2(\ln(2C^2)^2}{2\lambda_{\min}(\widehat{T})^2}$.

From the previous theorem we can notice that the error in estimation of T decays as $\frac{1}{\sqrt{n}}$ as a function of n.

Before stating the proof of Theorem III.3 we start by introducing the following helpful remark and Lemmas.

Remark 1. *We defined $\widehat{T} = \underset{B}{\operatorname{argmin}} ||B - \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T||_2^2$, with B that satisfies all the constraints in Eq. (11). We know that the matrix T we want to approximate satisfies all the constraints in Eq. (11), so by definition $||\widehat{T} - \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T||_2^2 \leq ||T - \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T||_2^2$ from which it follows that $||T - \widehat{T}||_2^2 \leq 2||T - \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T||_2^2$ so any bound we will find for $||T - \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T||_2^2$ holds also for $\widehat{T}$ estimated as in Eq. (10) with a coefficient 2.*

Lemma III.4. *Let A be a square, symmetric, positive definite matrix, in $\mathbb{R}^{C \times C}$ and let $\sqrt{A}$ the unique positive definite symmetric, matrix so that $\sqrt{A}\sqrt{A} = A$ (On the existence of this matrix, see Theorem 7.2.6 at p. 439 in Horn and Johnson (2012)). The bounded operator $F_{\surd} : \mathcal{S} \to \mathcal{S}$ defined as follow $F_{\surd} : A = \sqrt{A}$, where we denote by $\mathcal{S}$ the space of symmetric positive definite matrix, is differentiable and it hold the following upper bound for the induced 2 norm of the derivative:*

$$||D[\sqrt{A}]||_2 \leq \frac{1}{2\sqrt{\lambda_{\min}(A)}} ||vec(A)||_2.$$

Proof: Let us consider the vector space of square matrices $M_C(\mathbb{R})$ with the 2 norm and let $D[\sqrt{A}]$ denote the operator that is the derivative of $F_{\surd}$ in this space and $D[A]$ the derivative of A. From the fact that $\sqrt{A}\sqrt{A} = A$ it follows that:

$$D[\sqrt{A}]\sqrt{A} + \sqrt{A}D[\sqrt{A}] = D[A]. \tag{14}$$

Eq. (14) is a special case of Sylvester equation, and using that $\sqrt{A}$ is symmetric can be rewritten as:

$$(I_C \otimes \sqrt{A} + \sqrt{A} \otimes I_C)\text{vec}(D[\sqrt{A}]) = \text{vec}(D[A]).$$

It follows that:

$$\begin{aligned}\text{vec}(D[\sqrt{A}]) &= (I_C \otimes \sqrt{A} + \sqrt{A} \otimes I_C)^{-1}\text{vec}(D[A]) \\ &= (I_C \otimes \sqrt{A} + \sqrt{A} \otimes I_C)^{-1}\text{vec}(A).\end{aligned}$$

Notice that the eigenvalues of the square root of a symmetric, positive def matrix are the square root of the eigenvalues of the original matrices. Indeed if A can be decomposed as $A = U\Lambda U^T$, with U orthogonal matrix, it holds that $\sqrt{A} = U\sqrt{\Lambda}U^T$. Now the eigenvalues of $\sqrt{A} \otimes I_C + I_C \otimes \sqrt{A}$ are $\sqrt{\lambda_i} + \sqrt{\lambda_j}$ with $1 \leq i, j \leq C$. The minimum eigenvalue of $\sqrt{A} \otimes I_C + I_C \otimes \sqrt{A}$, that is the maximum eigenvalue of $(\sqrt{A} \otimes I_C + I_C \otimes \sqrt{A})^{-1}$ is $2\lambda_{\min}(\sqrt{A})$. Thus:

$$\begin{aligned}||D[\sqrt{A}]||_2 &\leq ||(I \otimes \sqrt{A} + \sqrt{A} \otimes I)^{-1}||_2 \cdot ||\text{vec}(A)||_2 \\ &\leq \frac{1}{2\sqrt{\lambda_{\min}(A)}} ||\text{vec}(A)||_2.\end{aligned}$$

∎

Let T and $\widehat{T}$ be defined as in Eq. (1) and Eq. (9).

The following Lemma holds for two general double stochastic matrices.

Lemma III.5. *Let T and $\widehat{T}$ be two symmetric, stochastic matrices, it holds that:*

$$||T - \widehat{T}||_2 \leq \frac{\sqrt{C}||T^2 - \widehat{T}^2||}{\lambda_{\min}(T^2) - ||T^2 - \widehat{T}^2||_2} \quad \textit{and}$$

$$||T - \widehat{T}||_2 \leq \frac{\sqrt{C}||T^2 - \widehat{T}^2||}{\lambda_{\min}(\widehat{T}^2) - ||T^2 - \widehat{T}^2||_2}$$

Proof: Using the differentiability of the matrix square root (for instance, Proposition A.2 from Andrews and Hopper (2010)), defining $A_\theta := \theta T^2 + (1-\theta)\widehat{T}^2$, it holds that:

$$||T - \widehat{T}||_2 = ||\sqrt{T^2} - \sqrt{\widehat{T}^2}||_2 \leq ||T^2 - \widehat{T}^2||_2 \cdot \sup_{\theta \in [0,1]} ||D[\sqrt{A_\theta}]||_2,$$

For Weyl's inequality $\lambda_{\min}(\theta T^2 + (1-\theta)\widehat{T}^2) \leq \lambda_{\min}(\theta T^2) + \lambda_{\min}((1-\theta)\widehat{T}^2) = \theta\lambda_{\min}(T^2) + (1-\theta)\lambda_{\min}(\widehat{T}^2)$.

$$\sup_{0\leq\theta\leq 1} ||D\sqrt{\theta T^2+(1-\theta)\widehat{T}^2}||_2$$
$$\leq \frac{1}{2}\sup_{0\leq\theta\leq 1}\frac{||\text{vec}(\theta T^2)+(1-\theta)\widehat{T}^2)||_2}{\theta\lambda_{\min}(T^2)+(1-\theta)\lambda_{\min}(\widehat{T}^2)}$$
$$\leq \sup_{0\leq\theta\leq 1}\frac{\sqrt{C}}{\theta\lambda_{\min}(T^2)+(1-\theta)\lambda_{\min}(\widehat{T}^2)} \quad (15)$$

In the last inequality we used that T and $\widehat{T}$ are doubly stochastic, so $||\text{vec}(T^2)||_2 = \left(\sum_{i=1}^C\sum_{j=1}^C T_{ij}^2\right)^{\frac{1}{2}} \leq \sqrt{C}$. Moreover deriving $\frac{1}{\theta\lambda_{\min}(T^2)+(1-\theta)\lambda_{\min}(\widehat{T}^2)}$ with respect to θ we find that:

$$\sup_{0\leq\theta\leq 1}\frac{1}{\theta\lambda_{\min}(T^2)+(1-\theta)\lambda_{\min}(\widehat{T}^2)} =$$
$$= \frac{1}{\min(\lambda_{\min}(\widehat{T}^2),\lambda_{\min}(T^2))}.$$

Since $T^2-\widehat{T}^2$ is symmetric:

$$||T^2-\widehat{T}^2||_2 = |\lambda_{\max}(T^2-\widehat{T}^2)|$$

It holds that:

$$\min(\lambda_{\min}(\widehat{T}^2),\lambda_{\min}(T^2)) \geq \lambda_{\min}(T^2)-||T^2-\widehat{T}^2||_2.$$

In the previous equations we use that $|\lambda_{\min}(T^2)-\lambda_{\min}(\widehat{T}^2)| \leq |\lambda_{\max}(T^2-\widehat{T}^2)|$. We now prove that it is true. Suppose without loss of generality that $\lambda_{\min}(T^2) > \lambda_{\min}(\widehat{T}^2)$. If it is the case $\lambda_{\min}(T^2)-\lambda_{\min}(\widehat{T}^2) = \lambda_{\min}(T^2)+\lambda_{\max}(-\widehat{T}^2) \leq \lambda_{\max}(T^2-\widehat{T}^2) \leq |\lambda_{\max}(T^2-\widehat{T}^2)|$, where we used Weyl's inequality.

It follows that $\lambda_{\min}(T^2)-\lambda_{\min}(\widehat{T}^2) < ||T^2-\widehat{T}^2||_2$. ■

Proof Theorem III.3:

From Lemma III.5 we know that

$$||T-\widehat{T}||_2 \leq \frac{\sqrt{C}||T^2-\widehat{T}^2||}{\lambda_{\min}(T^2)-||T^2-\widehat{T}^2||_2}$$

It follows that:

$$\mathbb{P}\big(||T-\widehat{T}||_2<\epsilon\big) = \mathbb{P}\left(||T^2-\widehat{T}^2||_2 < \lambda_{\min}(\widehat{T}^2)\frac{\epsilon}{\sqrt{C}+\epsilon}\right)$$
$$\geq \mathbb{P}\left(||T^2-\widehat{T}^2||_2 < \frac{\lambda_{\min}(\widehat{T}^2)}{\sqrt{C}+1}\epsilon\right)$$

Since we are interested in convergence properties of $\widehat{T}$, we want to find these bounds for small ϵ we can assume $\epsilon \leq 1$.

Now $T^2-\widehat{T}^2 = D^{1/2}(M-\widehat{M})D^{1/2}$.

So $||T^2-\widehat{T}^2||_2 \leq ||M-\widehat{M}||_2\lambda_{\max}(D)$. As a consequence:

$$\mathbb{P}(||T-\widehat{T}||_2<\epsilon) \geq \mathbb{P}\left(||M-\widehat{M}||_2 < \frac{\lambda_{\min}(\widehat{T}^2)}{(\sqrt{C}+1)\lambda_{\max}(D)}\epsilon\right)$$
$$\geq 1-2C^2e^{-\frac{\epsilon^2}{C^2(\sqrt{C}+1)^2}\frac{\lambda_{\min}(\widehat{T}^2)^2}{\lambda_{\max}(D)^2}n}$$

For the inverse: $T^{-1}-\widehat{T}^{-1} = T^{-1}(\widehat{T}-T)\widehat{T}^{-1}$.

So, $||T^{-1}-\widehat{T}^{-1}||_2 \leq (\lambda_{\min}(T)\lambda_{\min}(\widehat{T}))^{-1}||\widehat{T}-T||_2$

We obtain:

$$\frac{1}{\lambda_{\min}(T)\lambda_{\min}(\widehat{T})} \leq \frac{1}{\min(\lambda_{\min}(T^2),\lambda_{\min}(\widehat{T}^2))}$$
$$\leq \frac{1}{\lambda_{\min}(\widehat{T}^2)-|\lambda_{\min}(T^2)-\lambda_{\min}(\widehat{T}^2)|}$$

Using that T and $\widehat{T}$ are doubly stochastic:

$$||T^2-\widehat{T}^2||_2 \leq ||T(T-\widehat{T})+(T-\widehat{T})\widehat{T}||_2 \leq 2||T-\widehat{T}||_2$$

From which it follows:

$$||T^{-1}-\widehat{T}^{-1}||_2 \leq \frac{||T-\widehat{T}||_2}{\lambda_{\min}(\widehat{T}^2)-2||T-\widehat{T}||_2}$$

Finally we obtain:

$$\mathbb{P}\Big(||T^{-1}-\widehat{T}^{-1}||_2 \leq \epsilon\Big) \geq \mathbb{P}\left(||T-\widehat{T}||_2 \leq \epsilon\frac{\lambda_{\min}(\widehat{T})}{1+2\epsilon}\right)$$
$$\geq \mathbb{P}\left(||T-\widehat{T}||_2 \leq \frac{\epsilon}{3}\lambda_{\min}(\widehat{T})\right)$$
$$\geq 1-2C^2e^{-\frac{\epsilon^2}{9C^2(\sqrt{C}+1)^2}\frac{\lambda_{\min}(\widehat{T}^2)^4}{\lambda_{\max}(D)^2}n}$$

■

D. Learning from noisy labels

In this section, we show how to leverage the estimates of the error rates to train the models.

1) Posterior distribution of true labels as soft-labels: It is noteworthy that if we have access to the labels provided by all annotators, the posterior probabilities of the true labels can be calculated leveraging T and Bayes' Theorem as follows:

$$\underbrace{\mathbb{P}(y_i=c|y_{1,i},\ldots,y_{H,i})}_{:=p_{c,i}} \propto \nu_c\prod_{h=1}^H\underbrace{\mathbb{P}(y_{h,i}|y_i=c)}_{=T_{c,y_{h,i}}} \quad (16)$$

we recall that $\nu_c = \mathbb{P}(y_i=c)$ and that the conditional probabilities on the r.h.s. are given by T. In our case, we can use our noisy transition estimates to estimate the posterior probabilities of the true labels, and afterwards, we can use these posteriors to train the classifier.

Lemma III.6. *For infinite annotators, the posterior distribution over every sample calculated using the true T converges to the Dirac delta distribution centered on the true label almost surely (i.e. $\lim_{H\to\infty} p_{c,i} \overset{a.s.}{=} \mathbb{1}(y_i=c)$).*

Proof of Lemma III.6:

$$p_{c,i} = \frac{\mu_c\prod_{h=1}^H T_{c,y_{h,i}}}{\sum_{j=1}^C\mu_j\prod_{h=1}^H T_{j,y_{h,i}}}$$
$$\prod_{h=1}^H T_{c,y_{h,i}} = \prod_{j=1}^C T_{c,j}^{N_{i,j}}$$

where $N_{i,j}$ is the amount of annotators that labeled sample i as class j. Note that as a consequence of the strong law of large numbers for the conditional random variables that are independent with the same conditional distribution we have that the following equation is true almost surely:

$$\lim_{H\to\infty}\frac{N_{i,j}}{H}=\lim_{H\to\infty}\frac{\sum_{a=1}^{H}\mathbb{1}_{\{y_{a,i}=j\}}}{H}$$
$$=\mathbb{E}[\mathbb{1}_{\{y_{a,i}=j\}}|y=j]=T_{y_i,j}$$

Combining we get:

$$\lim_{H\to\infty}p_{c,i}=\lim_{H\to\infty}\frac{\mu_c\prod_{j=1}^{C}T_{c,j}^{N_{i,j}}}{\sum_{k=1}^{C}\mu_k\prod_{j=1}^{C}T_{k,j}^{N_{i,j}}}$$
$$=\lim_{H\to\infty}\frac{\mu_c\left(\prod_{j=1}^{C}T_{c,j}^{T_{y_i,j}}\right)^H}{\sum_{k=1}^{C}\mu_k\left(\prod_{j=1}^{C}T_{k,j}^{T_{y_i,j}}\right)^H}$$
$$=\lim_{H\to\infty}\frac{1}{1+\sum_{\substack{k=1\\k\neq c}}^{C}\frac{\mu_k}{\mu_c}\left(\prod_{j=1}^{C}\left(\frac{T_{k,j}}{T_{c,j}}\right)^{T_{y_i,j}}\right)^H}$$
$$\overset{(a)}{=}\mathbb{1}(y_i=c)$$

where in (a) we used the fact that due to the assumption that T is strictly dominant, then the term $\prod_{j=1}^{C}T_{k,j}^{T_{y_i,j}}$ is maximized when $k=y_i$ and this term is strictly larger than all the other ones, this fact can be derived by the fact that Kullback–Leibler divergence is non negative and is zero only if the two distributions are equal.

If $y_i\neq c$ it means that y_i is one of the values k can assume and since that one is the max, it means that for sure it will be greater than $\prod_{j=1}^{C}(T_{c,j})^{T_{y_i,j}}$. Otherwise, if $y_i=c$ it means that $\prod_{j=1}^{C}(T_{c,j})^{T_{y_i,j}}>\prod_{j=1}^{C}(T_{k,j})^{T_{y_i,j}}$ so all elements are less than 1 and the limit goes to 1. ∎

We can use the posterior distributions as soft-labels defining the following loss for the i-th sample:

$$\ell(f(x_i),y_{1,i},\dots,y_{H,i})=\ell(f(x_i),\bar{p}_i)\tag{17}$$

where $\bar{p}_i=[p_{1,i},\cdots,p_{C,i}]^T$. Or we can use the posterior distributions to weight the loss function at the i-th sample evaluated at each of the possible labels:

$$\ell(f(x_i),y_{1,i},\dots,y_{H,i})=\sum_{c=1}^{C}p_{c,i}\ell(f(x_i),e_c)$$

where e_c is the vector in $\mathbb{R}^C$ with 1 in the c-th position. Notice that for categorical cross-entropy loss, the two functions defined above coincide, but in general, they define two different loss functions.

2) Robust loss functions: Another way to leverage the estimate of T is to use robust loss functions, like the forward and backward loss functions presented in Natarajan et al. (2013); Patrini et al. (2017). Let $\ell(t,y)$ be a generic loss function for the classification task, with a little abuse of notation we define $\ell(t)=[\ell(t,e_1),\dots,\ell(t,e_C)]^T$. The backward and forward loss functions are defined in Eq. (18a) and Eq. (18b), respectively.

$$l_b(t,y)=(\widehat{\Gamma}^{-1}\ell(t))y\tag{18a}$$
$$l_f(t,y)=(\ell(\widehat{\Gamma}^{T}t))y\tag{18b}$$

To explain the notation in Eq. (18a) we are first doing the dot product between the matrix Γ^{-1} and the vector $\ell(t)$ and then the dot product of the resulting vector with y. These losses leverage aggregated labels and therefore different aggregating techniques can be used, like majority vote. Another possible aggregating technique that leverages the posterior probabilities is to assume that the true label is the one that corresponds to the class that has the highest posterior probability.

E. Generalizations gap bounds

In this section, we derive generalization gap bounds for the backward loss that depends on the noise transition matrix estimated in Eq. (10). Since we are only addressing the problem for the backward loss, from now on we will denote the backward loss by l.

Remark 2. *If $\ell(t,y)$ is Lipschitz with constant L, the loss function $l(t,y)$ is Lipschitz with Lipschitz constant $||\Gamma^{-1}||_2 L$.*

We will prove the following theorem in the case of $\Gamma=T$. We emphasize that all the results apply also when $\Gamma^{-1}=\phi(T^{-1})$ and that the function that associate Γ^{-1} and T^{-1} ,ϕ is Lipschitz with respect to the norm p, i.e. there exists a Lipschitz constant $L_{\phi,p}$ s.t.:

$$||\phi(T^{-1})-\phi(\widehat{T}^{-1})||_p\leq L_{\phi,p}||T^{-1}-\widehat{T}^{-1}||_p.$$

The only difference is that in the bound we have a factor $L_{\phi,p}$.

It has been proved, first in Natarajan et al. (2013) (Lemma 1) for the binary classification task and then in general for the multi-class case in Patrini et al. (2017) (Theorem 1) that $l(t,y)$ is an unbiased estimator for ℓ, i.e.:

$$\mathbb{E}_{\tilde{y}|y}[l(t,\tilde{y})]=\ell(t,y).$$

Lemma III.7. *Let ℓ be a bounded loss function, so that the image of ℓ is in $[0,\mu]$, and s.t. ℓ is Lipschitz in the first argument with Lipschitz constant L. Let $\widehat{R}_l(f)$ be the empirical risk for the loss l and let $R_{l,\mathcal{D}}$ be the risk for a loss l under the distribution $\mathcal{D}$, with l unbiased estimator for the loss ℓ. We denote by $\hat{l}$ the backward loss obtained using $\widehat{T}$.*

$$\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-R_{l,\mathcal{D}}(f)|$$
$$\leq\left[L\lambda_{\min}(\widehat{T}^2)+\frac{\mu\lambda_{\min}(D)}{\lambda_{\min}(\widehat{T})^2}\sqrt{\frac{1}{n}\ln\left(\frac{4C}{\delta}\right)}\right]\mathfrak{R}_n(\mathcal{F})g(C).$$

with $g(C)=6C^2(\sqrt{C}+1)$

Sketch of the Proof of Lemma III.7: By the triangle inequality, for any $f\in\mathcal{F}$,

$$\left|\hat{R}_{\hat{\ell}}(f)-R_{l,\mathcal{D}}(f)\right|\leq\left|\hat{R}_{\hat{\ell}}(f)-\hat{R}_{\ell}(f)\right|+\left|\hat{R}_{\ell}(f)-R_{l,\mathcal{D}}(f)\right|.$$

Taking the supremum over $\mathcal{F}$ and applying a union bound:

$$\sup_{f\in\mathcal{F}}\left|\hat{R}_{\hat{\ell}}(f)-R_{l,\mathcal{D}}(f)\right|\leq\sup_{f\in\mathcal{F}}\left|\hat{R}_{\hat{\ell}}(f)-\hat{R}_{\ell}(f)\right|+$$
$$+\sup_{f\in\mathcal{F}}\left|\hat{R}_{\ell}(f)-R_{l,\mathcal{D}}(f)\right|.$$

a) Uniform convergence for the clean loss.: Since $f \mapsto \ell(T^{-1}f(x), y)$ is $L\|T^{-1}\|$–Lipschitz in the sample, by union bounds and by the classic results on Rademacher complexity bounds (Mohri et al., 2012) and Theorem 7 in Meir and Zhang (2003) we obtain that with probability at least $1 - \frac{\delta}{2}$:

$$\sup_{f\in\mathcal{F}} \left|\hat{R}_\ell(f) - R_{l,\mathcal{D}}(f)\right| \;\le\; L\,\|T^{-1}\|\,\mathfrak{R}_n(\mathcal{F}) \;+\; \mu\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

b) Bounding the noise-estimation term.: Noting that:

$$\hat{R}_{\hat{\ell}}(f) - \hat{R}_\ell(f) \;=\; \mathbb{E}_n\Big[\big(\widehat{T}^{-1} - T^{-1}\big)\,\ell(f(x), y)\Big],$$

one shows:

$$\sup_{f\in\mathcal{F}} \left|\hat{R}_{\hat{\ell}}(f) - \hat{R}_\ell(f)\right| \;\le\; \left\|\widehat{T}^{-1} - T^{-1}\right\|\,\mathfrak{R}_n(\mathcal{F}).$$

A matrix-concentration inequality (e.g.Tropp (2012)) yields, with probability at least $1 - \frac{\delta}{2}$:

$$\left\|\widehat{T}^{-1} - T^{-1}\right\| \;\le\; C\sqrt{\frac{\ln(2C/\delta)}{n}}.$$

c) Combine the two bounds.: By a union bound, with probability at least $1 - \delta$:

$$\sup_{f\in\mathcal{F}} \left|\hat{R}_{\hat{\ell}}(f) - R_{l,\mathcal{D}}(f)\right| \;\le\; L\,\|T^{-1}\|\,\mathfrak{R}_n(\mathcal{F}) + \\ + C\,\mathfrak{R}_n(\mathcal{F})\sqrt{\frac{\ln(2C/\delta)}{n}} \;+\; \mu\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

Absorbing constants into a single function $g(C) = 6C^2(\sqrt{C}+1)$ gives the claimed result. ∎

Theorem III.8. *Let l be an unbiased estimator for ℓ defined as in Eq. (18a), Denoting $\hat{f} = \operatorname{argmin}_f(\widehat{R}_{\hat{l}}(f))$. It holds that:*

$$R_{\ell,\mathcal{D}}(\hat{f}) - \min_{f\in\mathcal{F}} R_{\ell,\mathcal{D}}(f) \\ \le \left[2L\lambda_{\min}(\widehat{T}^2) + \frac{\mu\lambda_{\min}(D)}{\lambda_{\min}(\widehat{T})^2}\sqrt{\frac{1}{n}\ln\big(\frac{4C}{\delta}\big)}\right]\mathfrak{R}_n(\mathcal{F})g(C)$$

with $g(C) = 6C^2(\sqrt{C}+1)$

Proof of Theorem III.8 : By the unbiasedness of l we have that $R_{\ell,\mathcal{D}}(\hat{f}) = R_{l,\mathcal{D}}(\hat{f})$. Moreover since $\hat{f} = \operatorname{argmin}_f(\widehat{R}_{\hat{l}}(f))$ we have $\widehat{R}_{\hat{l}}(\hat{f}) \le \widehat{R}_{\hat{l}}(g)\ \forall g \in \mathcal{F}$.

Let f^* be so that $\min_{f\in\mathcal{F}} R_{\ell,\mathcal{D}}(f) = R_{\ell,\mathcal{D}}(f^*)$. It follows:

$$\begin{aligned} &R_{\ell,\mathcal{D}}(\hat{f}) - \min_{f\in\mathcal{F}} R_{\ell,\mathcal{D}}(f) \\ &\quad= R_{l,\mathcal{D}}(\hat{f}) - \min_{f\in\mathcal{F}} R_{l,\mathcal{D}}(f) \\ &\quad= R_{l,\mathcal{D}}(\hat{f}) - \widehat{R}_{l,\mathcal{D}}(\hat{f}) + \widehat{R}_{l,\mathcal{D}}(\hat{f}) - R_{\ell,\mathcal{D}}(f^*) \\ &\quad\ge R_{l,\mathcal{D}}(\hat{f}) - \widehat{R}_{\hat{l},\mathcal{D}}(\hat{f}) - (R_{\ell,\mathcal{D}}(f^*) - \widehat{R}_{\hat{l},\mathcal{D}}(f^*)) \\ &\quad\ge 2\max_{f\in\mathcal{F}} |R_{\ell,\mathcal{D}}(f) - \widehat{R}_{\hat{l},\mathcal{D}}(f)| \end{aligned}$$

∎

The previous theorems demonstrate that the performance bounds decrease as one over the square root of the number of samples. The above theorem provides a performance bound for the classifier obtained by minimizing the backward loss l, which is an unbiased estimator of the true loss ℓ on the noisy dataset. These bounds depend on the Rademacher complexity of the function space and the Lipschitz constant of the loss function. Importantly, these bounds allow us to obtain performance guarantees for models trained on noisy data without relying on the true noise transition matrix of the annotators, which is typically unknown (Natarajan et al., 2013). Instead, the bounds depend on quantities that can be estimated from the noisy dataset, such as the estimated noise transition matrix, the number of classes, the Rademacher complexity, and the Lipschitz constant. Additionally, the bounds rely on the assumption of a uniform distribution over the ground truth labels, which is a reasonable assumption in many cases.

F. Not homogeneous annotators

The problem we are studying can be generalized to the case when the annotators have different noise transition matrices. In this case, given empirical agreement matrices $\widehat{M}_{ab}$ estimated from the data, and the matrix D, the goal is to recover the set of confusion matrices $\{T_a\}_{a=1}^H$ by solving the coupled constrained least-squares problem:

$$\min_{\{T_a\}_{a=1}^H, T_a\in\mathcal{C}} \quad \sum_{a<b} \left\|\widehat{M}_{ab} - T_a^\top D\,T_b\right\|_F^2, \tag{19}$$

where $\mathcal{C}$ is the set of symmetric, row-stochastic, non-negative matrices with diagonal entries ≥ 0.5 described by Eq. (11). This formulation generalizes the homogeneous case (obtained when $T_a = T$ for all a) to the setting of heterogeneous annotators. The problem (19) can be solved by *alternating constrained least squares*, *projected stochastic gradient*, or related block-coordinate methods. Each subproblem in T_a is convex and admits efficient updates under the row-stochasticity and nonnegativity constraints. The main theoretical challenges relative to the homogeneous case are: (i) the loss of symmetry of M_{ab}, which complicates spectral estimation; (ii) the increase in dimensionality, as the number of parameters grows linearly with H; and (iii) the joint objective becomes more non-convex and more sensitive to noise in the empirical agreement matrices, making the estimation procedure more prone to local minima and numerical instability.

Example with 3 annotators. We estimate the annotator confusion matrices $T_1, T_2, T_3 \in \mathbb{R}^{C\times C}$ given the empirical pairwise agreement matrices $\widehat{M}_{12}, \widehat{M}_{13}, \widehat{M}_{23}$. With $D = \operatorname{diag}(\nu)$ given, the optimization problem is:

$$\begin{aligned} \min_{T_1,T_2,T_3} \quad & \|\widehat{M}_{12} - T_1^\top D T_2\|_F^2 + \|\widehat{M}_{13} - T_1^\top D T_3\|_F^2 \\ & + \|\widehat{M}_{23} - T_2^\top D T_3\|_F^2, \\ \text{s.t.}\forall a \in \{1,2,3\}\ & T_a \text{ symmetric, row-stochastic, non-negative,} \\ & \text{with diagonal entries} \ge 0.5 \end{aligned} \tag{20}$$

When updating, for instance, T_1 with T_2 and T_3 fixed, the corresponding subproblem can be conveniently vectorized. So the problem can be written in a standard vectorized least-squares form subject to linear equality and inequality constraints.

This constrained least-squares problem can then be efficiently solved using quadratic programming (QP) or projected gradient methods. The same procedure is repeated for T_2 and T_3, and the process iterates until the total fit stabilizes, that is, until the objective in Eq. (20) no longer decreases significantly.

G. Example

Suppose three radiologists independently label chest X-rays as *Pneumonia* (1) or *Healthy* (2), and they are assumed to be equally reliable with the same error rate p. Their shared noise transition matrix is:

$$T = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}.$$

The inter-annotator agreement matrix between any two annotators a, b is given by:

$$M_{ab} = TDT,$$

where D is the class prior matrix (e.g., $D = \mathrm{diag}(0.5, 0.5)$ for balanced classes). Given the empirical estimate $\widehat{M}$ of the agreement matrix from the data, we can perform the eigenvalue decomposition of the normalized agreement matrix:

$$D^{\frac{1}{2}}\widehat{M}D^{\frac{1}{2}} = U\Lambda U^{\top}.$$

The estimate of the noise transition matrix is then given by Theorem III.1.

$$\widehat{T} = U\Lambda^{\frac{1}{2}}U^{\top}.$$

This estimate $\widehat{T}$ may not always be a valid noise matrix due to estimation errors. A projection onto the set of valid noise transition matrices with constraints such as symmetry, row-stochasticity, and diagonal dominance can be used to get a valid estimate. With this $\widehat{T}$, the posterior distribution over true labels for a sample can be computed via Bayes rule by combining annotator labels $\hat{y}_1, \ldots, \hat{y}_H$:

$$P(y = c \mid \hat{y}_1, \ldots, \hat{y}_H) \propto \overbrace{P(y=c)}^{D_{cc}} \prod_{a=1}^{H} \widehat{T}_{c,\hat{y}_a}.$$

This posterior provides a noise-aware soft label used in training, and the training loss can be defined as Eq. (17).

IV. Cohen's κ

We can also consider the case where an estimate of the IAA matrix M is not available and we only have access to a scalar representation of the inter-annotator agreement like Cohen's κ. In this case, we can only estimate one parameter and hence the matrix T has to be parameterized by a single parameter that can be estimated. Cohen's κ measures agreement between two raters classifying n items into C mutually exclusive categories. We define the agreement among raters a and b as p_o: $p_o = \sum_{c=1}^{C} \mathbb{P}(y_a = c \cap y_b = c)$. Cohen and others (Cohen, 1960) suggest comparing the actual agreement (p_o) with the "chance agreement" that could be obtained if the labels assigned by the two annotators were independent $p_e = \sum_{c=1}^{C} \mathbb{P}(y_a = c)\mathbb{P}(y_b = c)$. Cohen's κ is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e}.$$

It ranges from 1 (perfect agreement) to 0 (chance agreement), and can be negative when agreement is worse than the one expected by chance. In our work, we assume that raters are noisy versions of an underlying ground-truth label, making them dependent. Under this assumption, κ is always non-negative.

One particular example is the case where the noise is uniform among classes. Under these hypotheses, T is a matrix with all values $1-p$ on the diagonal and $\frac{p}{C-1}$ off the diagonal.

Lemma IV.1 (Relationship between p and κ)**.** *In the case of classification with uniform noise for two homogeneous annotators with noise rate p, i.e if a is one annotator,* $\mathbb{P}(y_a = i|y = j) = p$ *if* $i \neq j$*. If the distribution of the ground-truth labels is uniform, it holds that:*

$$p = (1 - C^{-1})(1 - \sqrt{\kappa}) \tag{21}$$

with κ *the Cohen's kappa coefficient of the two annotators.*

Proof of Proposition IV.1: relationship between ρ *and* κ*.:*

$$\begin{aligned} p_o &= \mathbb{P}(y_a = y_B) \\ &= \sum_{k,h=1}^{C} \mathbb{P}(y_A = k, y_B = k|y = h)\mathbb{P}(y = h) \\ &= \sum_{k,h=1}^{C} \overbrace{\mathbb{P}(y_A = k|y = h)}^{T_{h,k}} \overbrace{\mathbb{P}(y_B = k|y = h)}^{T_{h,k}} \nu_h \\ &= \sum_{h=1}^{C}(1-p)^2 c_h + \sum_{h=1}^{C}\Big(\frac{p}{C-1}\Big)^2 (C-1)c_h \\ &= (1-p)^2 + \frac{p^2}{C-1} \end{aligned}$$

Now:

$$\mathbb{P}(y_B = k) = \sum_{h=1}^{C} \overbrace{\mathbb{P}(y_B = k|y = h)}^{T_{hk}} \overbrace{\mathbb{P}(y = h)}^{\nu_h} = (T\nu)_k$$

In the previous equation, we used that T is symmetric.

$$\begin{aligned} p_e &= \sum_{k=1}^{C} \mathbb{P}(y_A = k)\mathbb{P}(y_B = k) = \sum_{k=1}^{C} \mathbb{P}(y_A = k)\mathbb{P}(y_B = k) \\ &= c^T T^2 c = 2\frac{p}{C-1} - \frac{Cp^2}{(C-1)^2} + \Big(1 - \frac{Cp}{C-1}\Big)^2 \nu^T\nu \end{aligned}$$

If the distribution of the true label y is symmetric, the probability vector $\nu = (\frac{1}{C}, \ldots, \frac{1}{C})$. So $\nu^T\nu = \frac{1}{C}$ and:

$$\kappa = \frac{C^2p^2 - 2C(C-1)p + (C-1)^2}{(C-1)^2}$$

From which it follows that:

$$p = (1 - C^{-1})(1 - \sqrt{\kappa}) \tag{22}$$

■

If T is assumed to be of the form described above (with all diagonal elements equal to $1-p$ and all off-diagonal entries equal), it has one eigenvalue equal to 1 and all the rest are equal to $1 - pC(C-1)^{-1}$ (this follows from the fact that in this case T can be written as a weighted summation of the identity and a rank-one matrix). Hence using Eq. (21) we get that $\lambda_{\min}(T) = \sqrt{\kappa}$. The bounds from Theorem III.8 holds replacing $\lambda_{\min}(T)$ with $\sqrt{\kappa}$. This allows us to obtain a bound for the generalization gap of a classifier trained with backward loss even in the case where a single statistic on the agreement between annotators is provided.

a) Non Homogeneous Case and Pairwise Cohen's κ.: In this paragraph we will consider the case of non-homogeneous annotators each characterized by a uniform noise rate $p_a \in [0,1]\ \forall a \in \{1,\dots,H\}$. In this case we can define a pairwise Cohen's κ_{ab}. For annotators a, b, the probability that annotators a and b give the same label is:

$$\begin{aligned} p_o(a,b) &= \Pr(y_a = y_b) \\ &= \sum_{h=1}^{C} \nu_h \sum_{k=1}^{C} \Pr(y_a = k \mid y = h)\ \Pr(y_b = k \mid y = h) \\ &= \sum_{h=1}^{C} \nu_h \Big[(1-p_a)(1-p_b) + (C-1)\Big(\tfrac{p_a}{C-1}\Big)\Big(\tfrac{p_b}{C-1}\Big)\Big] \\ &= (1-p_a)(1-p_b) + \frac{p_a p_b}{C-1}. \end{aligned}$$

The marginal label distributions for annotators a and b are:

$$\pi_a = T_a^\top \nu, \qquad \pi_b = T_b^\top \nu,$$

and the expected chance agreement is:

$$p_e(a,b) = \sum_{k=1}^{C} \pi_{a,k}\pi_{b,k} = \pi_a^\top \pi_b = \nu^\top T_a T_b \nu,$$

since each T_a is symmetric. Using the structure of T_a we obtain:

$$p_e(a,b) = \frac{p_a + p_b}{C-1} - \frac{C\, p_a p_b}{(C-1)^2} + \Big(1 - \frac{C(p_a+p_b)}{2(C-1)}\Big)^2 \|\nu\|_2^2,$$

where $\|\nu\|_2^2 = \sum_{c=1}^{C} \nu_c^2$. If the true label distribution is uniform, $\nu = \left(\frac{1}{C}, \dots, \frac{1}{C}\right)$ so that $\|\nu\|_2^2 = \frac{1}{C}$, Cohen's coefficient for the annotator pair (a,b) becomes:

$$\kappa_{ab} = \frac{C^2 p_a p_b - 2C(C-1)(p_a+p_b)/2 + (C-1)^2}{(C-1)^2}. \quad (23)$$

When $p_a = p_b = p$, Eq. (23) reduces to the homogeneous result obtained earlier.

V. Distribution Shift

In order to understand how annotation noise skews our empirical risk, we now quantify the "distributional shift" between the clean joint (x,y) and the noisy joint. Concretely, under the *common instance-independent* noise model, the observed labels arise via a fixed transition matrix T from the true labels. We will show that the difference between the expected loss of the classifier f under the noisy-label distribution and its expected loss under the true (clean)-label distribution i.e.:

$$\Big| R_{\ell,\tilde{\mathcal{D}}}(f) - R_{\ell,\mathcal{D}}(f) \Big|$$

can be bounded in terms of the spectral gap of T and the class-prior matrix D. This bound tells us that when T is well-conditioned (i.e. $\lambda_{min}(T)$ is bounded away from zero), the bias introduced by noisy labels remains controlled.

Lemma V.1. *Let $x \in \mathcal{X}$ and $y, \tilde{y} \in \{1,\dots,C\}$. Let $(x,\tilde{y}) \sim \tilde{\mathcal{D}}$ and $(x,y) \sim \mathcal{D}$, so that conditioned distribution of $\tilde{y}$ w.r.t y is the described by a noise transition matrix T, i.e. $\mathbb{P}(\tilde{y} = j | y = i) = T_{ij}$. Let ℓ be a loss function so that for the distribution $\mathcal{D}$ it holds $\mathbb{E}_{\{x\}\sim\mathcal{D}}[\max_{i\in\{1,\dots,C\}} \ell(f(x),i)] < \infty$. Using the usual notation of the paper:*

$$\Big| R_{\ell,\tilde{\mathcal{D}}}(f) - R_{\ell,\mathcal{D}}(f) \Big| \le \lambda_{\max}(D)(1-\lambda_{\min}(T))\mathbb{E}_x[||\boldsymbol{\ell}(f(x))||_2]$$

with $\boldsymbol{\ell}(f(x))$ be the vector $[\ell(f(x),1),\dots,\ell(f(x),C)]^T$. We emphasize that we don't need the loss ℓ to be bounded we only require that the expected value of the 2 norm of the vector $\boldsymbol{\ell}$ is bounded. This condition is satisfied when $\mathbb{E}_x[|\ell(f(x),i)|] < \infty \quad \forall i \in \{1,\dots C\}$ that is of course satisfied if the loss is bounded.

Proof: By definition of expected value, we can rewrite:

$$\begin{aligned} &\Big| R_{\ell,\tilde{\mathcal{D}}}(f) - R_{\ell,\mathcal{D}}(f) \Big| = \\ &= \int_x \sum_{j=1}^{C} \ell(f(x),j)[\tilde{p}(x,j) - p(x,j)]dx \end{aligned}$$

Under the instance-independent noise assumption we have by definition $p(\tilde{y}|x,y) = p(\tilde{y}|y)$.

As a consequence $p(x,\tilde{y}|y) = p(\tilde{y}|x,y)p(x|y) = p(\tilde{y}|y)p(x|y)$. Re-writing everything in matrix form: $\boldsymbol{\ell}((f(x)) \in \mathbb{R}^C$, $\boldsymbol{p}(x|C) = [p(x|1), p(x|2)\dots p(x|C)]^T \in \mathbb{R}^C$. Then $\tilde{p}(x,y) = TD\boldsymbol{p}(x|C)$ and $p(x,y) = D\boldsymbol{p}(x|C)$.

Namely, the integrand is $\boldsymbol{\ell}((f(x))^T(T-I)D\boldsymbol{p}(x|C)$. Apply Holder and spectral-norm bounds:

$$\begin{aligned} &|\boldsymbol{\ell}((f(x))^T(T-I)D\boldsymbol{p}(x|C)| \\ &\qquad \le \|\boldsymbol{\ell}((f(x))\|_2 \|T-I\|_2 \|D\|_2 \|\boldsymbol{p}(x|C)\|_2 \end{aligned}$$

Now using that $||(T-I)||_2 = 1 - \lambda_{\min}(T)$, $||D||_2 = \lambda_{\max}(D)$ and that $||\boldsymbol{p}(x|C)||_2 \le ||\boldsymbol{p}(x|C)||_1 = 1$,

Putting everything together:

$$\begin{aligned} &\Big| \int_{\mathcal{X}} \boldsymbol{\ell}(f(x))(T-I)D\mathbf{p}(x)dx \Big| \\ &\le \int_{\mathcal{X}} \big| \boldsymbol{\ell}(f(x))(T-I)D\mathbf{p}(x)dx \big| \\ &= \lambda_{\max}(D)(1-\lambda_{\min}(T)) \int_{\mathcal{X}} \sum_{i=1}^{C} p(x,i)||\boldsymbol{\ell}(f(x))||_2 dx \Big| = \\ &\le \lambda_{\max}(D)(1-\lambda_{\min}(T))\ \mathbb{E}_x[\max_{i\in\{1,\dots C\}} |\ell(f(x),i)|] \end{aligned}$$

If we want a bound that depends on the minimum entry of T and not on its minimum eigenvalue we could obtain a bound a similar bound of this form

$$\left|R_{\ell,\tilde{\mathcal{D}}}(f) - R_{\ell,\mathcal{D}}(f)\right| \leq \max_{i,j\in\{1,\dots C\}} (I_C - T_{ij}) D_{ii}\, \mathbb{E}_x[||\boldsymbol{\ell}(f(x))||_1]$$

A bound of similar, but slightly different form was obtained in Wei et al. (2023) only for two classes and bounded losses:

$$\left|R_{\ell,\tilde{\mathcal{D}}}(f) - R_{\ell,\mathcal{D}}(f)\right| \leq (D_{22}T_{21} + D_{11}T_{12})(\sup(\ell) - \inf(\ell)).$$

■

Building on the previous Lemma we now ask: what happens when we only have an empirical estimate $\widehat{T}$ of the transition matrix, and when we replace the true noisy risk by its sample analogue?

Applying Weyl's inequality to bound the latter term by $\|T - \widehat{T}\|_2$, and then invoking a matrix-Hoeffding concentration on $\|T - \widehat{T}\|_2$, one shows with high probability:

$$\left|R_{\ell,\tilde{\mathcal{D}}}(f) - R_{\ell,\mathcal{D}}(f)\right| \;\leq\; \left(1 - \lambda_{\min}(\widehat{T})\right) \lambda_{\max}(D)\, R_{\ell,\mathcal{D}}(f) + \epsilon.$$

Finally, combining the above bias control with a standard Rademacher-complexity bound on $\left|\widehat{R}_{\ell,\tilde{\mathcal{D}}}(f) - R_{\ell,\tilde{\mathcal{D}}}(f)\right|$ (see Proposition V.1.1) and applying a union bound yields, with high probability:

$$\left|\widehat{R}_{\ell,\tilde{\mathcal{D}}}(f) - R_{\ell,\mathcal{D}}(f)\right| \;\leq\; \epsilon \;+\; L\,\mathfrak{R}_n(\mathcal{F}).$$

More precisely, we obtain:

Proposition V.1.1 (High-probability bias bound)**.** *Under the assumptions of Lemma V.1, let $\widehat{T}$ be the estimator defined in (9). Then for any $\epsilon > 0$:*

$$\mathbb{P}\left(\left|R_{\ell,\tilde{\mathcal{D}}}(f) - R_{\ell,\mathcal{D}}(f)\right| \leq \left(1 - \lambda_{\min}(\widehat{T})\right)\lambda_{\max}(D)\, R_{\ell,\mathcal{D}}(f) + \epsilon\right) \geq 1 - 2C^2 \exp\left(-2\,\frac{\epsilon^2}{C^2\, R_{\ell,\mathcal{D}}(f)^2}\,\frac{\lambda_{\min}(\widehat{T})^4}{\lambda_{\max}(D)^4}\, n\right).$$

Proof idea for Proposition V.1.1: Starting from the deterministic bias bound of Lemma V.1, we replace the true transition matrix T by its estimator $\widehat{T}$. The difference:

$$\lambda_{\min}(T) \;-\; \lambda_{\min}(\widehat{T})$$

is controlled by Weyl's inequality in terms of the operator-norm error $\|T - \widehat{T}\|_2$. A standard matrix-concentration (e.g. Hoeffding or McDiarmid) argument then shows that $\|T - \widehat{T}\|_2$ is exponentially small in the sample size n. Combining these ingredients yields the stated high-probability bound. ■

Proposition V.1.2 (Noisy-risk generalization bound)**.** *Under the same conditions, let $\widehat{R}_{\ell,\tilde{\mathcal{D}}}(f)$ be the empirical noisy risk on n samples. Then for any $\epsilon > 0$:*

$$\mathbb{P}\Big(\left|\widehat{R}_{\ell,\tilde{\mathcal{D}}}(f) - R_{\ell,\mathcal{D}}(f)\right| \;\leq\; \epsilon \;+\; L\,\mathfrak{R}_n(\mathcal{F})\Big) \;\geq\; 1 \;-\; 2C^2 \exp\big(\zeta(\mathcal{D}, f, D, T, C)\, n\big) \;-\; 2\exp\Big(-\tfrac{n}{2}\big(\tfrac{\epsilon}{4\mu}\big)^2\Big).$$

with $\zeta(\mathcal{D}, f, D, T, C) = -2\frac{\epsilon^2}{C^2 R_{\ell,\mathcal{D}}(f)^2}\frac{\lambda_{\min}(\widehat{T}^2)^2}{\lambda_{\max}(D)^4}$

Proof idea for Proposition V.1.2: We decompose:

$$\left|\widehat{R}_{\ell,\tilde{\mathcal{D}}}(f) - R_{\ell,\mathcal{D}}(f)\right| \;\leq\; \underbrace{\left|\widehat{R}_{\ell,\tilde{\mathcal{D}}} - R_{\ell,\tilde{\mathcal{D}}}\right|}_{\text{empirical estimation error}} \;+\; \underbrace{\left|R_{\ell,\tilde{\mathcal{D}}} - R_{\ell,\mathcal{D}}\right|}_{\text{noise bias}}.$$

The first term is bounded by a Rademacher-complexity argument (plus a small slack $\epsilon/2$), and the second term is handled by Proposition V.1.1. A union bound over these two high-probability events yields the final generalization guarantee. ■

VI. Experimental results

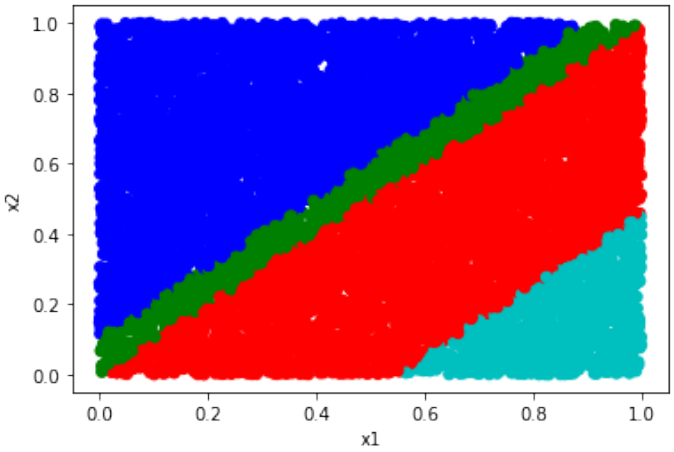

Fig. 1: Synthetic data for 4 classes with distribution (0.4,0.1,0.4,0.1)

A. Implementation details

a) Synthetic experiments: A neural network with a hyperbolic tangent activation function with one hidden layer is used for the dataset with more classes. The data are separated into train, validation and test set using a split 64%, 16%, 20%. Models are trained with the following configuration: batch size 256, learning rate 5×10^{-3}, maximum number of epochs 100, early stopping of training based on validation loss with patience of 10 epochs. Once the training is finished, the model with the lowest validation loss is retrieved.

b) Experiments on image datasets: For the experiments with CIFAR-10N, a ResNet34 is trained with the following configuration: batch size 128, learning rate 10^{-1}, with momentum (0.9) and learning rate decay (0.0005), 300 epochs and early stopping of training based on validation loss with patience of 100 epochs. For the pre-trained model, we use the model[1] provided by torch-vision. The train-val-test split is 70-20-10. The only difference for TrashNet experiments is in the learning rate, which is 2×10^{-3}.

[1] https://pytorch.org/vision/main/models/generated/torchvision.models.resnet34.html#resnet34

c) Experiments on text datasets: One model for each language was used for textual experiments. All pre-trained models are selected from HuggingFace[2]. The configuration is the following: batch size 128, learning rate 2×10^{-5}, 300 epochs for all the languages except Spanish, where 100 epochs are enough to train the model. Same data split of the images is used.

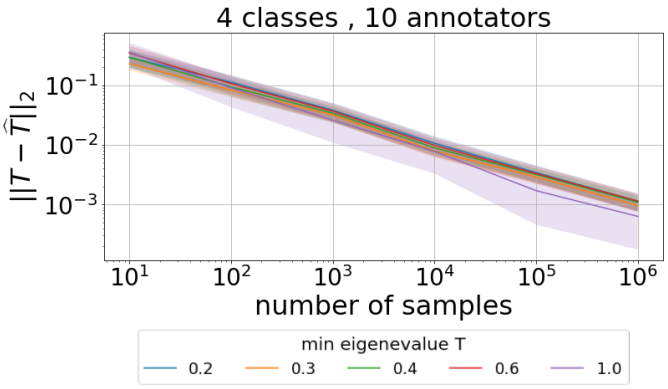

Fig. 2: Error in the estimation of T for 4 classes and 10 annotators. Plots are obtained by averaging different admissible matrices T) that have the same minimum eigenvalues rounded to the first decimal.

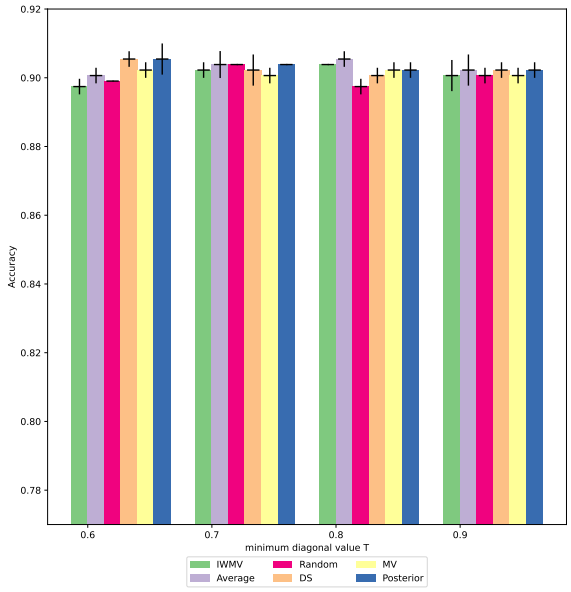

Fig. 3: Comparison between the performance of Cross Entropy Loss using majority vote, random, Dawid-Skene, IWMV aggregation method or the posteriors (posterior) and relative frequency (average) as soft labels.On the y-axis the accuracy on a clean dataset and on the x-axis the values of the minimum on the diagonal of T. Small values of the minimum diagonal value mean a noisy dataset, while the minimum is 1 in the noise-free case. The results are obtained for 3 annotators and 4 classes, by averaging on different admissible matrices T that have the same minimum diagonal values rounded to the first decimal. The error bands show the maximum and minimum performance for each method.

B. Results

We conduct experiments to evaluate the effectiveness of our method for estimating the noise transition matrix $\widehat{T}$, analyzing the estimation error as a function of the number of samples. Additionally, we train different classifiers using the estimated $\widehat{T}$ on a synthetic dataset and two real datasets with noisy labels. We compare the performance of classifiers trained using labels from a baseline aggregation method with those trained using the posterior distributions obtained from the estimated $\widehat{T}$ as soft labels on two different tasks: image classification and text classification. Lastly, we evaluate our approach on a real dataset with synthetic annotations.

a) Estimation of T: With these experiments, we aim to validate the theoretical results of Section III-C. We generate various matrices T that are symmetric, stochastic and diagonally dominant. For each annotator, we produce their prediction according to the matrix T. We run experiments for the number of annotators $H = 10, 7, 3, 2$. We report here the results for $H = 10$, and 4 classes. In Fig. 2 we observe that the error in the estimation decreases as $\frac{1}{\sqrt{n}}$ with n number of samples, which is in agreement with the bound provided in Theorem III.3. We also observe that, as expected, the estimation becomes more accurate as the number of annotators increases. The results were obtained from a synthetically generated dataset in which we generate the classes predicted by the annotators according to various T matrices, choosing from all possible (admissible) combinations that have $[0, 0.2, 0.4]$ out of the diagonal and $[0.6, 0.8, 1.0]$ on the diagonal.

For experiments with 2 and 7 annotators, we generate T as all possible symmetric, stochastic and diagonally dominant matrices with $[0.1, 0.2, 0.3, 0.4, 0.5]$ out of the diagonal and $[0.6, 0.8, 1.0]$ on the diagonal. Classes are uniformly distributed. For experiments with 10 annotators, we generate the matrices T as all possible (admissible) combinations that have $[0, 0.2, 0.4]$ out of the diagonal and $[0.6, 0.8, 1.0]$ on the diagonal. In this case, we both include uniform distribution of the true labels among the 4 classes and all the distributions so that the four classes can be partitioned into two groups of indices so that classes in the same group have the same probability. Namely, if the distributions on the classes are given by $= [d_1, d_2, d_3, d_4]$, admissible distributions are the ones for which there are two subsets if indices I and J so that $I \cup J = \{0, 1, 2, 3, 4\}$ and for all $i, k \in I : d_i = d_k$.

Results for $2, 3$ and 7 annotators were obtained by averaging over 3 runs, while for 10 over 10 runs. The error that appears on axis y in the plots is the difference in norm 2 of the true matrix T and the estimated matrix $\widehat{T}$, obtained as explained in Section III-C.

In Fig. 4 we fix the number of annotators to $H = 3$. Then we perform an analysis similar to the previous one. The main difference is the following: in this plot it can be noticed how the number of samples impacts both the estimation error and

[2]https://huggingface.co/

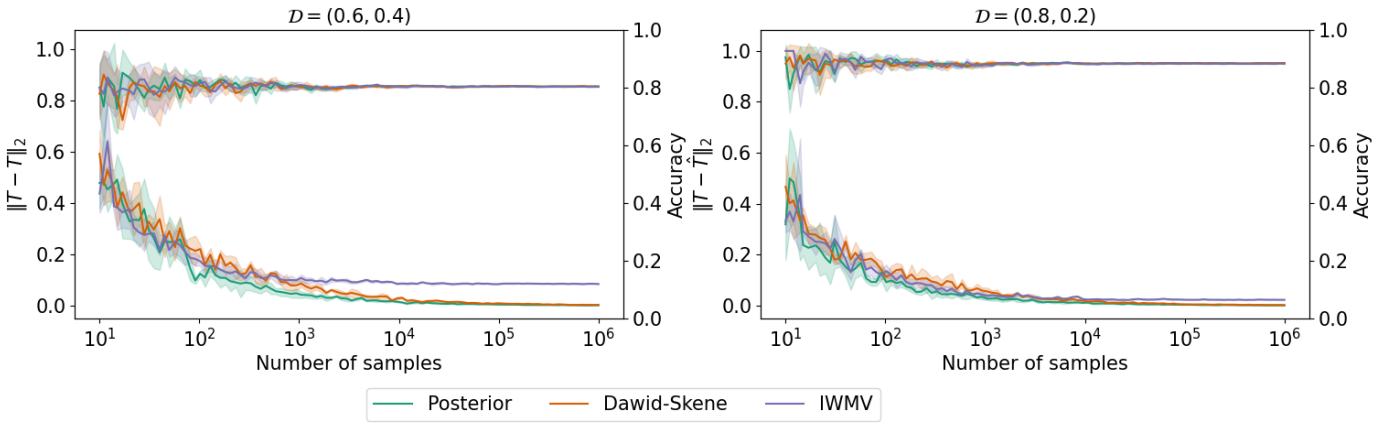

Fig. 4: Reconstruction error and accuracy by varying the number of samples with $H = 3$ annotators with 5 independent runs.

the accuracy in estimating the gold label. The results are shown for two different noise percentage values $(0.6, 0.8)$ and with 3 different aggregation methods based on the the noise transition matrix estimation (DS, MV and Posterior).

b) Classification task with synthetic data: We consider a classification task using a synthetic dataset. The features are generated uniformly in the range $[0, 1]^2$. The assignment of labels (y) is done by following the label distribution established for each experiment, separating the space with lines parallel to the bisector of the first and third quadrants. Our dataset comprises 10000 samples. See Fig. 1 for an example.

For each dataset, annotations are generated according to the noise transition matrix T. Various combinations of T are tested, ensuring they respect the assumptions of symmetry, stochasticity, diagonal dominance, and commutativity with D. As before, they are stochastic and diagonally dominant matrices with $[0.1, 0.2, 0.3, 0.4, 0.5]$ out of the diagonal and $[0.6, 0.8, 1.0]$ on the diagonal. The number of annotators is variable in the set $\{3, 5\}$. We use categorical cross entropy as loss function and both hard labels and soft labels to train the models. To train the models with hard labels an aggregation method is needed to obtain one final label from the annotators. We consider random, majority vote, Dawid-Skene and IWMV. In random aggregation, the final label is randomly picked from the labels of the annotators. In the majority vote the final label is the one with the most amount of votes (the mode), if the mode is not unique, we randomly choose one of the most voted classes. Details about Dawid-Skene and IWMV can be found in Section II. As soft labels, we consider the relative frequency among annotators and the posterior distribution according to Eq. (16). In the case of frequency for each sample we average the one-hot encoded annotations. Notice that random, majority vote and frequency soft labels do not leverage the estimate of T while the posterior does. In Fig. 3 we report the results for 4 classes with distribution $(0.4, 0.1, 0.4, 0.1)$ and 3 annotators.

We use accuracy with respect to a clean dataset as a performance metric. Our results show that using the posterior distribution, as soft labels, allows for better performance than using the average of the labels assigned by annotators and then using majority vote or all the other hard labels-based aggregation methods.

Our method is shown to be more robust to noise and exhibits the least variance in the results. This confirms our hypothesis that leveraging the matrix $\widehat{T}$ leads to better classification accuracy. In particular, this is evident when the probability of correctly classifying labels is 0.6. This is a challenging situation since the noise rate is high, but the proposed approach outperforms all the competitors in this case as well.

c) Experiments on CIFAR10-N: The CIFAR10-N dataset[3] contains CIFAR-10 train images with noisy labels annotated by humans using Amazon Mechanical Turk. Each image is labelled by three independent annotators. Table I shows the accuracy achieved using the different aggregation methods. Since each sample has 3 annotations, we will have 3 different T matrices, one per annotator. Due to space constraints, we insert the L_1 between each couple of noise transition matrices:

$$L_1^{12} = 0.270 \quad L_1^{13} = 0.219 \quad L_1^{23} = 0.248$$

Our aggregation approach achieves the best performance. Note that in this dataset there are no guarantees that the assumptions we made on T are satisfied, however, the method is still applicable with positive results.

d) Experiments on TrashNet: The TrashNet[4] dataset contains images of recyclable objects across six classes, with approximately 400-500 images in each class. However, each image is manually annotated by a single annotator. This is a common issue in many widely used datasets (Deng, 2012). To address this, we create synthetic annotations. We assign to each sample $H = 6$ annotations following a noise transition matrix T:

$$T = \begin{bmatrix} 0.7 & 0.3 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.4 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.3 & 0.1 & 0.6 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.2 & 0.7 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.1 & 0.1 & 0.6 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.1 & 0.9 \end{bmatrix}$$

[3] http://www.noisylabels.com
[4] https://github.com/garythung/trashnet

Aggregation Method	CIFAR-10N	TrashNet
IWMV	0.7354 ± 0.5277	0.6389 ± 0.0688
Average	0.7381 ± 0.4772	0.6706 ± 0.0638
Random	0.7053 ± 0.6346	0.6796 ± 0.0587
DS	0.7386 ± 0.3117	0.7222 ± 0.0155
MV	0.7363 ± 0.3375	0.6815 ± 0.0303
Posterior	**0.7405** ± 0.3707	**0.7361** ± 0.0315

TABLE I: Test Accuracy on CIFAR10-N and TrashNet using a pre-trained Resnet34.

To compute the accuracy on the test set, we use the labels provided by the dataset creators as the ground truth.

As can be noticed from Table I, the posterior approach is able to surpass all the competitors on the TrashNet dataset. This result shows how it is able to estimate the real labels also in case of synthetically-generated annotations.

e) Experiments on SentiMP: The SentiMP dataset[5] (Rodríguez-Barroso et al., 2024) contains tweets written by members of parliament in Greece, Spain, and United Kingdom in 2021 with collected 500 tweets per language. It is designed for sentiment analysis with 3 different labels (positive, negative and neutral) and each sample is annotated by at least 3 different annotators. For each sample, we have 3 annotations meaning we have 3 different T matrices. Due to space constraints, we show the T matrices for the english language:

$$T_{\text{en}}^1 = \begin{bmatrix} 0.931 & 0.010 & 0.059 \\ 0.049 & 0.939 & 0.012 \\ 0.052 & 0.060 & 0.888 \end{bmatrix} \quad T_{\text{en}}^2 = \begin{bmatrix} 0.752 & 0.149 & 0.099 \\ 0.024 & 0.967 & 0.008 \\ 0.030 & 0.045 & 0.925 \end{bmatrix} \quad T_{\text{en}}^3 = \begin{bmatrix} 0.772 & 0.149 & 0.079 \\ 0.110 & 0.873 & 0.016 \\ 0.142 & 0.082 & 0.776 \end{bmatrix}$$

For Spanish and Greek language, we insert the L_1 loss between the T matrices. For Spanish we have:

$$L_1^{12} = 1.077 \quad L_1^{13} = 0.367 \quad L_1^{23} = 1.039$$

For Greek we have:

$$L_1^{12} = 0.411 \quad L_1^{13} = 0.602 \quad L_1^{23} = 0.378$$

The proposed approach is similar to the previous one, where soft labels are used in the training and validation steps, while hard labels are used for test. Also in this different domain, our method is able to beat all the competitors on the three different languages, as can be seen from Table II.

VII. Concluding Remarks

We tackled the problem of learning from noisy labels, where the dataset is labeled by annotators who occasionally make mistakes. We proposed a methodology to estimate the noise transition matrix T of the annotators given the inter-annotator agreement. We then introduced techniques to leverage this estimated T for robust learning from the noisy dataset. Theoretical analysis showed the soundness of our methods. Experimental results confirmed the validity of our noise transition matrix estimation and demonstrated its effectiveness in improving performance when learning from noisy labels.

[5] https://huggingface.co/rbnuria

Dataset	Method	Accuracy	F1
en	IWMV	0.8125 ± 0.034	0.7732 ± 0.0446
	Average	0.8333 ± 0.0295	0.7990 ± 0.0336
	Random	0.8333 ± 0.045	0.7996 ± 0.0473
	DS	0.8056 ± 0.0428	0.7618 ± 0.0489
	MV	0.8333 ± 0.017	0.8008 ± 0.0193
	Posterior	**0.8403** ± 0.026	**0.8043** ± 0.0305
gr	IWMV	0.7619 ± 0.0509	0.736 ± 0.0553
	Average	0.7551 ± 0.0289	0.7226 ± 0.04
	Random	0.7483 ± 0.0585	0.7246 ± 0.0709
	DS	0.7619 ± 0.0694	0.7349 ± 0.0738
	MV	0.7823 ± 0.0585	0.7522 ± 0.0656
	Posterior	**0.7891** ± 0.0509	**0.7602** ± 0.046
sp	IWMV	0.7154 ± 0.0230	0.6600 ± 0.0653
	Average	0.7073 ± 0.0718	0.6601 ± 0.0891
	Random	0.7073 ± 0.0527	0.6490 ± 0.0842
	DS	0.7236 ± 0.0304	0.6652 ± 0.0691
	MV	0.7073 ± 0.0345	0.6655 ± 0.0300
	Posterior	**0.7301** ± 0.0501	**0.6751** ± 0.0415

TABLE II: Results in terms of F1-Score and Accuracy on the SentiMP dataset with three different languages. Our method is always able to surpass all the competitors on both the metrics.

Limitations: The main limitation of our current approach to estimating T is that it only considers the case where T is symmetric and D assumed to be known and commutes with T. Extending the results to the case where T might not be symmetric and different among annotators is one possible future research direction.

References

Andrews, B. and Hopper, C. (2010). *The Ricci flow in Riemannian geometry: a complete proof of the differentiable 1/4-pinching sphere theorem*. springer.

Bucarelli, M. S., Cassano, L., Siciliano, F., Mantrach, A., and Silvestri, F. (2023). Leveraging inter-rater agreement for classification in the presence of noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3439–3448.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

Fleiss, J. et al. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Ghosh, A., Kumar, H., and Sastry, P. S. (2017). Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 1919–1925. AAAI Press.

Hendrycks, D., Mazeika, M., Wilson, D., and Gimpel, K. (2018). Using trusted data to train deep networks on labels corrupted by severe noise. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Gar-

nett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, USA, 2nd edition.

Karger, D., Oh, S., and Shah, D. (2011). Iterative learning for reliable crowdsourcing systems. *Advances in neural information processing systems*, 24.

Li, H. and Yu, B. (2014). Error rate bounds and iterative weighted majority voting for crowdsourcing. *arXiv preprint arXiv:1411.4086*.

Meir, R. and Zhang, T. (2003). Generalization error bounds for bayesian mixture algorithms. *J. Mach. Learn. Res.*, 4:839–860.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. The MIT Press.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. (2013). Learning with noisy labels. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. (2017). Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952.

Peterson, J., Battleday, R., Griffiths, T., and Russakovsky, O. (2019). Human uncertainty makes classification more robust. In *Proceedings - 2019 International Conference on Computer Vision, ICCV 2019*, Proceedings of the IEEE International Conference on Computer Vision, pages 9616–9625, United States. Institute of Electrical and Electronics Engineers Inc.

Potter, J. E. (1966). Matrix quadratic solutions. *SIAM Journal on Applied Mathematics*, 14(3):496–501.

Purificato, A., Bucarelli, M. S., Nelakanti, A. K., Bacciu, A., Silvestri, F., and Mantrach, A. (2025). The majority vote paradigm shift: When popular meets optimal. *arXiv preprint arXiv:2502.12581*.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322.

Rodríguez-Barroso, N., Martínez-Cámara, E., Camacho-Collados, J., Luzón, M. V., and Herrera, F. (2024). Federated learning for exploiting annotators' disagreements in natural language processing. *Transactions of the Association for Computational Linguistics*, 11:630–648.

Sheng, V. S. and Zhang, J. (2019). Machine learning with crowdsourcing: A brief summary of the past research and future directions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9837–9843.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. (2020). Learning from noisy labels with deep neural networks: A survey.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Wani, F. A., Bucarelli, M. S., and Silvestri, F. (2024). Learning with noisy labels through learnable weighting and centroid similarity. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.

Wei, J., Zhu, Z., Luo, T., Amid, E., Kumar, A., and Liu, Y. (2023). To aggregate or not? learning with separate noisy labels. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2523–2535.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J., and Ruvolo, P. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22.

Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., and Sugiyama, M. (2019). Are anchor points really indispensable in label-noise learning? In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yao, Y., Liu, T., Gong, M., Han, B., Niu, G., and Zhang, K. (2021). Instance-dependent label-noise learning under a structural causal model. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4409–4420. Curran Associates, Inc.

Yao, Y., Liu, T., Han, B., Gong, M., Deng, J., Niu, G., and Sugiyama, M. (2020). Dual t: Reducing estimation error for transition matrix in label-noise learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Zhang, M., Lee, J., and Agarwal, S. (2021). Learning from noisy labels with no change to the training process. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12468–12478. PMLR.

Zhang, Z. and Sabuncu, M. R. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8792–8802, Red Hook, NY, USA. Curran Associates Inc.

Zhu, D., Ying, Y., and Yang, T. (2023). Label distributionally robust losses for multi-class classification: Consistency, robustness and adaptivity. In *International Conference on Machine Learning*, pages 43289–43325. PMLR.

Zhu, Z., Wang, J., and Liu, Y. (2022). Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27633–27653. PMLR.

Maria Sofia Bucarelli Maria Sofia Bucarelli is a postdoctoral researcher at Inria Sophia Antipolis, working for CNRS supervised by Emanuele Natale in the COATI team. She previously held a postdoctoral position at Sapienza University of Rome, where she also obtained her PhD under Prof. Fabrizio Silvestri, collaborating with the RSTLess group. Her research interests include the theory and applications of machine learning, as well as the foundations of machine learning.

Antonio Purificato Antonio Purificato is a PhD student in Data Science at Sapienza University of Rome in the Department of Computer, Control and Management Engineering "Antonio Ruberti". He is working under the guidance of Professor Fabrizio Silvestri and collaborating with the RSTless group. His research interests include Graph Neural Networks and Category Theory, with a focus on the environmental impact of deep learning algorithms. He also an Applied Scientist Visitor at Amazon working on the impact of noise on supervised learning.

Andrea Bacciu Andrea Bacciu is an Appliend Scientist at Amazon, working on multilingual summarization. He graduated with honors in the master's degree in Computer Science at Sapienza University of Rome, presenting a research work on Cross-lingual Semantic Role Labeling, which was published and awarded with the Outstanding Long Paper at the NAACL 2021 conference. He has a PhD from Sapienza University of Rome. His research topics include Large Language Models, Natural Language Understanding and Information Retrieval.

Lucas Cassano Lucas Cassano received his Electronics Engineer degree from Buenos Aires Institute of Technology in 2013. Afterwards he went to UCLA where he received his M.S. and Ph.D. degrees in 2015 and 2020, respectively. After obtaining his PhD he joined EPFL a postdoctoral researcher. He worked as Applied Scientist at Amazon. Now he works as Independent Researcher.

Federico Siciliano Federico Siciliano is a post-doc in the RSTLess Lab at Sapienza University of Rome under the guidance of Prof. Fabrizio Silvestri. He holds a PhD in Data Science (2024) with a thesis on Architectural Components of Trustworthy Artificial Intelligence with Prof. Fabrizio Silvestri. His current research projects include Information Retrieval, Recommender Systems and Explainable Artificial Intelligence. He had the privilege of partnering with numerous world-renowned institutions, including Cambridge University, Meta, Amazon, and UniPi.

Anil Nelakanti Anil Nelakanti is an assistant professor at IIIT Hyderabad associated with the Language Technologies Research Center. His interests lie in machine learning and its applications to perceptory and behavioral data including vision, language, speech and search. In the past he was Amazon and before that taught at a university (IIT-BHU, Varanasi). Before that he worked with SIERRA/INRIA in Paris and with MLS/XRCE in Grenoble towards a doctoral degree defending it in February of 2014.

Amin Mantrach Amin Mantrach is an Applied Science Manager at Amazon, leading a team that develops Generative AI solutions across multiple languages, with a focus on evaluation techniques and responsible AI. His team notably expanded AI-generated customer review summaries to non-English Amazon stores. He previously worked as a research scientist at Xerox Research Center, Yahoo Labs, and Criteo AI Lab. With over 15 years of experience, he holds a PhD in Machine Learning and has published in top venues like CVPR, or KDD.

Fabrizio Silvestri Fabrizio Silvestri is a Full Professor and the coordinator of the Ph.D. in Data Science, at Dipartimento di Ingegneria informatica, automatica e gestionale (DIAG) of the University of Rome, La Sapienza. His research interests lie in Artificial Intelligence, and in particular, machine learning applied to web search problems and natural language processing. He is the author of more than 150 papers in international journals and conference proceedings. He holds nine industrial patents. He is the holder of the "test-of-time" award at the ECIR 2018 conference for an article published in 2007. He is the holder of three best paper awards and other international awards. Fabrizio Silvestri spent eight years abroad in industrial research laboratories (Yahoo! and Facebook). Fabrizio Silvestri has a Ph.D. in computer science awarded by the University of Pisa.