Leveraging Inter-rater Agreement for Classification in the Presence of Noisy Labels

Maria Sofia Bucarelli [2] * Lucas Cassano[1] Federico Siciliano [2] * Amin Mantrach[1] Fabrizio Silvestri[2, 3]
[1] Amazon [2] Sapienza University of Rome [3] ISTI-CNR, Pisa, Italy
{mariasofia.bucarelli, federico.siciliano}@uniroma1.it
{lcecasl, mantrach}@amazon.lu
fsilvestri@diag.uniroma1.it

Abstract

In practical settings, classification datasets are obtained through a labelling process that is usually done by humans. Labels can be noisy as they are obtained by aggregating the different individual labels assigned to the same sample by multiple, and possibly disagreeing, annotators. The inter-rater agreement on these datasets can be measured while the underlying noise distribution to which the labels are subject is assumed to be unknown. In this work, we: (i) show how to leverage the inter-annotator statistics to estimate the noise distribution to which labels are subject; (ii) introduce methods that use the estimate of the noise distribution to learn from the noisy dataset; and (iii) establish generalization bounds in the empirical risk minimization framework that depend on the estimated quantities. We conclude the paper by providing experiments that illustrate our findings.

1. Introduction

Supervised learning has seen enormous progress in the last decades, both theoretical and practical. Empirical risk minimization is used as a learning framework [23], which relies on the assumption that the model is trained with iid (independent and identically distributed) sampled data from the joint distribution between features and labels. As a consequence of generalization bounds, when this assumption is satisfied any desired performance can be achieved as long as enough training data is available. However in many real-world applications, due to flaws during the data collection and labeling process, the assumption that the training data is sampled from the true feature-label joint distribution does not hold. Training data is often annotated by human raters who have some non-zero probability of making mistakes. It has been reported in [21] 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 raises the question whether it is possible to learn from datasets with noisy labels while still having performance guarantees. This question has received a lot of attention lately and has already been answered in the positive in some cases [15, 16]. Indeed multiple works have introduced learning algorithms that can cope with datasets with incorrect labels while guaranteeing desirable performance through provable generalization bounds. 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. Several works [16, 26, 27] attempt to address this issue by introducing techniques to estimate such error rate. Some of these methods have the drawback of relying on assumptions that do not always hold in practice, such as the existence of anchor samples [16]. Ideally, it would be desirable to design learning algorithms that are both robust to noisy labels, and for which performance guarantees can be provided.

An approach, often used in industry to reduce the impact of errors made by human raters, is to label the same dataset multiple times by different annotators. Then the individual labels are combined to reduce the probability of erroneous labels in the dataset, two popular approaches are majority vote or soft labeling. In these cases inter-annotator agreement (IAA) scores (like Cohen's kappa [1] and Fleiss' kappa [5]) provide measurable metrics that are directly related to the probability of error present in the labels.

Since the IAA holds a direct relationship with the error rate associated with the human raters, one could potentially estimate the error rate and leverage this estimate to modify

[*] This work was done during Maria Sofia Bucarelli's and Federico Siciliano's internship at Amazon.

the learning algorithms with the objective of making them robust to the resulting noise in the labels. This is the main direction we explore in this work.

Motivation and Contributions: This work is motivated by two main points: i- to the best of our knowledge there are no published results that indicate how to leverage the IAA statistics to estimate the label noise distribution; 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 provide a methodology to estimate the label noise distribution based on the IAA statistics; (ii) we show how to leverage this estimate to learn from the noisy dataset; and (iii) we provide generalization bounds for our methods that depend on **known** quantities.

2. Related works

Our work is related to literature on three main topics: (i) robust loss function design, (ii) label aggregating and (iii) noise rate estimation.

Robust loss functions In classification tasks, the goal is to obtain the lowest probability of classification error. The $0-1$ loss counts how many errors a classifier makes on a given dataset and is often used in the evaluation of the classifier. However, it is rarely used in optimization procedures because it is non-differentiable and non-continuous. To overcome this, many learning strategies use some convex *surrogates* of the $0-1$ loss function (*e.g.* hinge loss, squared error loss, cross-entropy).

It was proved ([6], [7]) that *symmetric* loss functions, that are functions for which the sum of the risks over all categories is equivalent to a constant for each arbitrary example, are robust to label noise. Examples of symmetric loss functions include the $0-1$ loss, the Ramp Loss and (softmax) Mean Absolute Error (MAE). In [29] authors show that even if MAE is noise tolerant and cathegorical cross entropy (CCE) is not, MAE can perform poorly when used to train DNN in challenging domains. They also propose a loss function that can be seen as a generalization of MAE and CCE. Several other loss functions that do not strictly satisfy the symmetry condition have also been proposed to be robust against label noise when training deep neural networks [4, 13, 24].

[15] presents two methods to modify the surrogate loss in the presence of class-conditional random label noise. The first method introduces a new loss that is an unbiased estimator for a given surrogate loss, and the second method introduces a label-dependent loss. The paper provides generalization bounds for both methods, which depend on the noise rate of the dataset and the complexity of the hypothesis space.

Labels aggregation When constructing datasets for supervised learning, data is often not labeled by a single annotator, rather multiple imperfect annotators are asked to assign labels to documents. Typically, separate labels are aggregated into one before learning models are applied [3,20]. In our work, we propose to exploit a measure of the agreement between annotators to explicitly calculate the noise of the dataset. Recently some works revisited the choice of aggregating labels. In [19] authors explore how to train LETOR models with relevance judgments distributions instead of single-valued relevance labels. They interpret the output of a LETOR model as a probability value or distribution and define different KL divergence-based loss functions to train a model. The loss they proposed can be used to train any ranking model that relies on gradient-based learning (in particular they focused on transformer-based neural LETOR models and on the decision tree-based GBM model). However, the authors do not directly estimate the noise rates in the annotations or study how learning from these noisy labels affects the generalization error of the models trained with the methods they introduce. In [25] the authors analyze the performance of both label aggregation and non-aggregation approaches in the context of empirical risk minimization for a number of popular loss functions, including those designed specifically for the noisy label learning problem. They conclude that label separation is preferable to label aggregation when noise rates are high or the number of labelers/annotations is insufficient. [17] and [22] exploit the availability of multiple human annotations to construct soft labels and concludes that this increases performance in terms of generalization to out-of-training-distribution test datasets, and robustness to adversarial attacks. [2] focus on efficiently eliciting soft labels from individual annotators.

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) [12, 16, 31]. Usually these methods use a small number of anchor points (that are samples that belong to a specific class with probability one) [8]. In particular, [16] proposed 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, [26, 27]. In [27] the authors propose to introduce an intermediate class to avoid directly estimating the noisy class posterior. [28] also propose an iterative noise estimation heuristic that aims to partly correct the error and pointed out that the methods introduced by [16] and [27] have an error in computing anchor points, and provide conditions on the noise under which the methods work or fail. [26] provides 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. Our work differs from the mentioned work that use anchor points because we do not need to assume the existence of anchor points or to have a validation set to learn the noise rate and we only use noisy data to train our model, moreover we neither aim to detect anchor points in the noisy data. Also most of these works do not study the generalization properties of the proposed models, while we also address this problem and find bound that depend on the estimated noise transition matrix.

Another approach is based on the clusterability condition, that is an example belongs to the same true class of its nearest-neighbors representations. [30] presented a method that relies on statistics of high-order consensuses among the 2 nearest-neighbors noisy labels.

3. Problem formulation

3.1. Notation

In this paper we follow 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 $\ell : [0,1]^C \times \mathcal{Y} \to \mathbb{R}$.

3.2. 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 $\widehat{\mathcal{D}}$:

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

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

3.2.1 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:

$$(T_a)_{i,j} := \mathbb{P}(y_a = j | y = i) \quad (2)$$

Assumption 1. *We assume that all annotators have the same noise transition matrix (i.e. $T_a = T$ for all 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 3.0.1. *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:*

$$\mathbb{P}(y_a, y_b|y) = \mathbb{P}(y_a|y)\mathbb{P}(y_b|y). \quad (3)$$

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

$$(M_{ab})_{i,j} := \mathbb{P}(y_a = i, y_b = j) \quad (4)$$

Proposition 3.0.2. *Leveraging Assumption 2 the agreement matrix $M_{a,b}$ can be written as follows:*

$$M_{a,b} = {T_a}^T D T_b \quad (5)$$

$$D : = diag\{\nu\} \quad (6)$$

$$\nu : = [\mathbb{P}(y=1), \cdots, \mathbb{P}(y=C)]^T. \quad (7)$$

Due to Proposition 3.0.1 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 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:

$$(\widehat{M_{a,b}})_{i,j} = \sum_{k=1}^{n} \frac{\mathbb{1}(y_{a,k}{=}i, y_{b,k}{=}j) + \mathbb{1}(y_{a,k}{=}j, y_{b,k}{=}i)}{2n} \tag{8}$$

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. (8) for all possible pairs of annotators. The estimator in this case can be written as

$$(\widehat{M})_{i,j} = \frac{1}{H(H{-}1)} \sum_{a=1}^{H} \sum_{\substack{b=1 \\ b\neq a}}^{H} \sum_{h=1}^{n} \frac{\mathbb{1}(y_{a,h}{=}i, y_{b,h}{=}j)}{n}. \tag{9}$$

3.2.2 Scenario II

In the second scenario, for each i-th sample we are given a unique label $\tilde{y}_i$ that is produced by aggregating the H individual labels according to some known aggregating policy (like majority vote). In this case, since we do not have 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 policy 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 were 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.

3.2.3 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.

4. Main results

We divide the main contributions in three sections. In the first section we show how to estimate the noise matrices T Next we indicate how to leverage these estimates to learn for the datasets with noisy labels. Finally we obtain bounds,depending on the Rademacher complexity of the class of functions, on the generalization gap for a bounded and Lipschitz loss function

4.1. Estimation of the noise transition matrices

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

Lemma 4.1. *If $D^{\frac{1}{2}}$ commutes with T we have that:*

$$T = U\Lambda^{\frac{1}{2}}U^T \tag{10}$$

$$T^{-1} = U\Lambda^{-\frac{1}{2}}U^T \tag{11}$$

$$D^{-\frac{1}{2}}MD^{-\frac{1}{2}} = U\Lambda U^T \tag{12}$$

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).

A detailed discussion of when the commutativity assumption is satisfied is included in Appendix B. The proof of the previous Lemma can be find in Appendix C.1.

Note that we could use Lemma 4.1 to estimate T as follows:

$$\widehat{T} = \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T \tag{13}$$

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 nonnegative entries (which is a convex set). We can obtain such projection by solving the following optimization problem:

$$\widehat{T} = \pi(\widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T) = \underset{B}{\operatorname{argmin}} ||B - \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T||_2^2 \tag{14}$$

$$\text{s.t.} \quad \begin{aligned} & B = B^T \\ & \sum_j B_{i,j} = 1 \qquad \forall i \\ & B_{i,j} \geq 0 \qquad \forall i, j \\ & B_{i,i} \geq 0.5 \qquad \forall i \end{aligned}$$

Note that this optimization problem is convex because the constraints are linear and for symmetric matrices it holds that $||\widehat{T} - \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T||_2^2 = \lambda_{\max}(\widehat{T} - \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T)$, which 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 aggregating policy to be known.

Lemma 4.2. *Let $M_{a,b}$ be the agreement matrix for annotators a and b defined in Eq. (4) and $\widehat{M_{a,b}}$ be the estimated agreement matrix defined in Eq. (8) 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 $1 - \delta$*

$$||M_{a,b} - \widehat{M_{a,b}}||_p \leq \sqrt{\frac{C^2}{2n} \ln \frac{2C^2}{\delta}}. \tag{15}$$

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. The proof can be found in Appendix C.2. □

From Lemma 4.2 it follows that if $\widehat{M}$ is estimated as in Eq. (9), since $\widehat{M}$ is an average of $\widehat{M_{ab}}$ it also holds that for every $p \in [1, \infty]$ and for every $\delta > 0$, with probability at least $1 - \delta$

$$||M - \widehat{M}||_p \leq \sqrt{\frac{C^2}{2n} \ln \frac{2C^2}{\delta}}. \tag{16}$$

Theorem 4.3. *Let T be the noise transition matrix defined as in Eq. (2) and $\widehat{T}$ its estimate (defined as in Eq. (14)). With probability at least $1 - \delta$:*

$$||T - \widehat{T}||_2 \leq \frac{C(\sqrt{C}+1)\lambda_{\max}(D)}{\lambda_{\min}(\hat{T})} \sqrt{\frac{1}{2n} \ln \frac{2C^2}{\delta}} \tag{17a}$$

$$||T^{-1} - \widehat{T}^{-1}||_2 \leq \frac{9C(\sqrt{C}+1)\lambda_{\max}(D)}{\lambda_{\min}(\hat{T})^2} \sqrt{\frac{1}{2n} \ln \frac{2C^2}{\delta}} \tag{17b}$$

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

Proof. The proof can be found in Appendix C.3. □

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.

4.2. Learning from noisy labels

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

4.2.1 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:

$$\underbrace{\mathbb{P}(y_i = c | y_{1,i}, \ldots, y_{H,i})}_{:=p_{c,i}} \propto \nu_c \prod_{h=1}^{H} \underbrace{\mathbb{P}(y_{h,i} | y_i = c)}_{=T_{c,y_{h,i}}} \tag{18}$$

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 4.4. *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. See Appendix C.5. □

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

$$\ell(f(x_i), y_{1,i}, \ldots, y_{H,i}) = \ell(f(x_i), \bar{p}_i) \tag{19}$$

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:

$$\ell(f(x_i), y_{1,i}, \ldots, y_{H,i}) = \sum_{c=1}^{C} p_{c,i} \ell(f(x_i), e_c) \tag{20}$$

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 correspond, but in general they define two different loss functions.

Note that these soft-labels are different from the ones obtained by averaging the annotators labels as is done in [25]. The method using the posteriors exploits the T matrix and thus more information than the simple mean of the values of the losses among annotators. We therefore expect this to yield better results than the aggregation using the mean proposed in [25]. These considerations are supported by the empirical results we obtained on synthetic datasets (see Sec. 6).

4.2.2 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 [15, 16]. Let $\ell(t, y)$ be a generic loss function for the classification task, with a little abuse of notation we define $\ell(t) = [\ell(t, e_1), \ldots, \ell(t, e_C)]^T$. The backward and forward loss functions are defined in Eq. (21a) and Eq. (21b), respectively.

$$l_b(t, y) = (\widehat{\Gamma}^{-1}\ell(t))y \tag{21a}$$

$$l_f(t, y) = (\ell(\widehat{\Gamma}^T t))y \tag{21b}$$

To explain the notation in Eq. (21a) 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.

4.3. Generalizations gap bounds

In this section we derive generalization gap bounds for the backward loss that depend on the noise transition matrix estimated in Eq. (14). Since we are only addressing the problem for the backward loss, from now on we will denotethe backward loss by l.

Remark 1. *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 will have a factor $L_{\phi,p}$.

It has been proved, first in [15] (Lemma 1) for the binary classification task and then in general for the multi-class case in [16] (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 4.5. *Let ℓ be a bounded loss function, with $\ell \in [0, \mu]$, s.t. there exists a Lipschitz function α, with Lipschitz constant L, so that $\ell(f(x), y) = \alpha|f(x) - y|$. Let $\widehat{R}_l(f)$ be the empirical risk for the loss l and let $R_{l,\mathcal{D}}$ be the risk for 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 $\widehat{T}$.*

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

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

Theorem 4.6. *Let l be an unbiased estimator for ℓ defined as in Eq. (21a), Denoting $\hat{f} = \underset{f}{\text{argmin}}(\widehat{R}_{\hat{l}}(f))$. It holds that*

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

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

The proofs of Lemma 4.5 and Theorem 4.6 can be find in Appendix C. We observe that in all the previous theorems, the bounds found are always decreasing as one over the square root of the number of samples. The above theorem gives us a performance bound for the classifier found minimizing the backward loss l, i.e. the unbiased estimator of the loss ℓ on the noisy dataset. The bounds found depend on, the Rademacher complexity of the function space and the Lipschitz constant of the loss function.The importance of these bounds lies in the fact that they allow us to obtain performance bounds for a model trained with noisy data that depends on values that we can estimate from the noisy dataset.In particular, there is no dependence on the true noise transition matrix of the annotators, as in other work [15] which is instead a quantity that cannot be known a priori having access only to the training data. More in detail the bound depends on the estimate noise transition matrix, the number of classes in the dataset, the Rademacher complexity and the Lipschitz constant, which we can take as known a priori and on the distribution of ground truth, which in many cases it makes sense to assume uniform.

5. 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.

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 5.1 (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:*

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

with κ the Cohen's kappa coefficient of the two annotators (see Appendix A).

Proof. The proof can be found in Appendix C.6. □

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. (22) we get that $\lambda_{\text{min}}(T) = \sqrt{\kappa}$. The bounds from Theorem 4.6 holds replacing $\lambda_{\text{min}}(T)$ with $\sqrt{\kappa}$. This allows us to obtain bound for the generalization gap of a classifier trained with backward loss even in the case where a single statistic on agreement between annotators is provided.

6. Experimental results

We performed experiments to validate the effectiveness of the method we propose for estimating $\widehat{T}$ by studying the error in the estimation as a function of the number of samples. We also performed experiments to show how the estimated T can be leveraged to train classifiers in the presence of noise labels. In particular we performed experiments for a classification task on a synthetic dataset and on the CIFAR10-N dataset, comparing the performance of a classifier trained using labels obtained by some baseline aggregation method with the performance of a classifier trained using the distribution of posteriors obtained from the estimation of T (Eq. (18)) as soft-labels.

Estimation of T With these experiments we aim to validate the theoretical results of Sec. 4.1. We generate various matrices T that are symmetric, stochastic and diagonally dominant, the exact details about the generation of T can be found in Appendix D.1. 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, all the other plots are in Appendix D.1. In Fig. 4 (as well as the plots in the Appendix) we can be observed 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 4.3. We also observed that, as expected, the estimation becomes more accurate as the number of annotators increases.

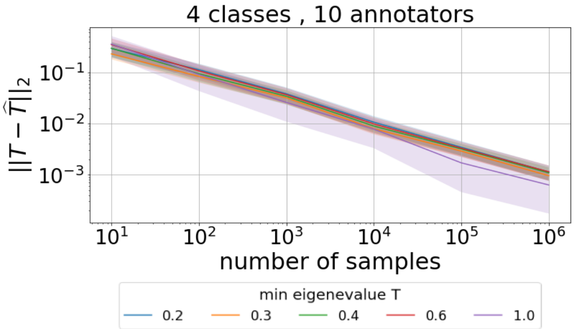

Figure 1. Error in the estimation of T for 4 classes and 10 annotators. The plots are obtained by averaging different admissible matrices T (see Appendix B) and averaged over matrices that have the same minimum eigenvalues rounded to the first decimal.

Classification task with synthetic data We consider a classification task with a synthetic dataset. The features are generated uniformly in $[0, 1]^2$. 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 More information on how the class distributions are generated can be found in Appendix D.2.

For each dataset annotations are generated according to the noise transition matrix T. Various combinations of T are tested that respect the assumptions of symmetry, stochasticity and diagonally dominance, as well as being commutative with D (more details can be found in Appendix B).The number of annotators is variable in the set $\{3, 5\}$. See Appendix D.2 for implementations details.

Losses We use categorical cross entropy as loss function. We use 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 and majority vote. In random aggregation the final label is randomly picked from the labels of the annotators. In 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. As soft-labels we consider the relative frequency among annotators and the posterior distribution according to Eq. (18). 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. 2 we report the results for 4 classes with distribution $(0.4, 0.1, 0.4, 0.1)$ and 3 annotators.

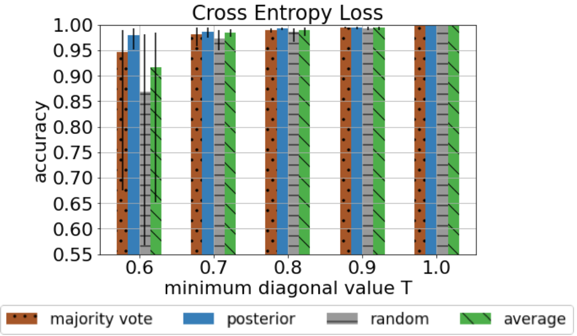

Figure 2. Comparison between performance of Cross Entropy Loss using majority vote, random aggregation method or the posteriors (posterior) and relative frequency (average) as soft labels.On the y-axis the accuracy on a clean dataset and on the x-axis the values of the minimum on the diagonal of T. Small 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 ammissible matrices T (see Appendix B) that have the same minimum diagonal values rounded to the first decimal. The error bands show the maximum and minimum performance for each method.

We use accuracy with respect to a clean dataset as performance metric. Our results show that using the posteriors distribution ,as soft labels, allows for better performance than using the average of the labels assigned by annotators and than using majority vote or random aggregation.

Our method is shown to be more robust to the noise and is also the one with less variance in the results. This confirms our hypotheses that by leveraging the matrix $\widehat{T}$ better classification accuracy can be achieved.

Experiments on CIFAR10-N The CIFAR10-N dataset[1] contains CIFAR-10 train images with noisy labels annotated by humans using Amazon Mechanical Turk. Each image is labelled by three independent annotators. Table 1 shows the accuracy achieved using the different aggregation methods. For this experiment we used Resnet34 [10] with and without pre-training. In both cases, our approach of aggregation 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.

Aggregation Method	Pretrained	Not-Pretrained
random	0.718 ± 0.035	0.579 ± 0.023
majority vote	0.740 ± 0.017	0.590 ± 0.006
average	0.762 ± 0.012	0.637 ± 0.016
posteriors (ours)	$\mathbf{0.794 \pm 0.005}$	$\mathbf{0.652 \pm 0.014}$

Table 1. Test Accuracy on CIFAR10-N with Resnet34

7. Concluding remarks

We have addressed the problem of learning from noisy labels in the case where the dataset is labeled by annotators that occasionally make mistakes. We have introduced a methodology to estimate the noise transition matrix T of the annotators given the IAA. We further showed different techniques to leverage this estimate to learn from the noisy dataset in a robust manner. We have shown theoretically that the methods we introduce are sound. We supported our methodology with some experiments that confirms our estimation of the noise transition matrix is valid and that this can be leveraged in the learning process to obtain better performance.

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

Acknowledgements

We acknowledge financial support from NRRP MUR project PE0000013-FAIR. This research was partially supported by MIUR under the grant "Dipartimenti di eccellenza 2018–2022" of the Department of Computer Science and the Department of Computer Engineering at Sapienza University of Rome. It was also partially supported by the ERC Advanced Grant 788893 "AMDROMA" , the EC H2020RIA project "SoBigData++" (871042), the MIUR PRIN project "ALGADIMAR", and the project SERICS (PE00000014) under the NRRP MUR program funded by the EU-NGEU.

[1] http://www.noisylabels.com

References

[1] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. 1, 11

[2] K. M. Collins, U. Bhatt, and A. Weller. Eliciting and learning with soft labels from every annotator, 2022. 2

[3] A. P. Dawid and A. M. Skene. 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, 1979. 2

[4] L. Feng, S. Shu, Z. Lin, F. Lv, L. Li, and B. An. Can cross entropy loss be robust to label noise? In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2206–2212. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track. 2

[5] J. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. 1

[6] A. Ghosh, H. Kumar, and P. S. Sastry. 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, 2017. 2

[7] A. Ghosh, N. Manwani, and P. Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015. 2

[8] D. Hendrycks, M. Mazeika, D. Wilson, and K. Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 2

[9] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, USA, 2nd edition, 2012. 14

[10] A. Khetan, Z. C. Lipton, and A. Anandkumar. Learning from noisy singly-labeled data, 2017. 8

[11] R. Meir and T. Zhang. Generalization error bounds for bayesian mixture algorithms. *J. Mach. Learn. Res.*, 4:839–860, dec 2003. 17

[12] A. Menon, B. V. Rooyen, C. S. Ong, and B. Williamson. Learning from corrupted binary labels via class-probability estimation. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 125–134, Lille, France, 07–09 Jul 2015. PMLR. 2

[13] A. K. Menon, A. S. Rawat, S. J. Reddi, and S. Kumar. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*, 2020. 2

[14] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. 17

[15] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 1, 2, 6

[16] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu. 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, 2017. 1, 2, 6

[17] J. Peterson, R. Battleday, T. Griffiths, and O. Russakovsky. 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, Oct. 2019. Institute of Electrical and Electronics Engineers Inc. 2

[18] J. E. Potter. Matrix quadratic solutions. *SIAM Journal on Applied Mathematics*, 14(3):496–501, 1966. 12

[19] A. Purpura, G. Silvello, and G. A. Susto. Learning to rank from relevance judgments distributions, 2022. 2

[20] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322, 2010. 2

[21] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee. Learning from noisy labels with deep neural networks: A survey, 2020. 1

[22] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):173–177, Oct. 2020. 2

[23] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998. 1

[24] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey. Symmetric cross entropy for robust learning with noisy labels. In *IEEE International Conference on Computer Vision*, 2019. 2

[25] J. Wei, Z. Zhu, T. Luo, E. Amid, A. Kumar, and Y. Liu. To aggregate or not? learning with separate noisy labels, 2022. 2, 6

[26] X. Xia, T. Liu, N. Wang, B. Han, C. Gong, G. Niu, and M. Sugiyama. Are anchor points really indispensable in label-noise learning? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1, 2, 3

[27] Y. Yao, T. Liu, B. Han, M. Gong, J. Deng, G. Niu, and M. Sugiyama. 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, 2020. Curran Associates Inc. 1, 2, 3

[28] M. Zhang, J. Lee, and S. Agarwal. Learning from noisy labels with no change to the training process. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12468–12478. PMLR, 18–24 Jul 2021. 2

[29] Z. Zhang and M. R. Sabuncu. 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, 2018. Curran Associates Inc. 2

[30] Z. Zhu, Y. Song, and Y. Liu. Clusterability as an alternative to anchor points when learning with noisy labels, 2021. 3

[31] Z. Zhu, J. Wang, and Y. Liu. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27633–27653. PMLR, 17–23 Jul 2022. 2