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 don't 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 equation (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 don't 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 $\tilde{y}^i = y_a^i$ 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 non-negative entries (which is a convex set). We can obtain such projection by solving the following optimization problem:

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

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

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 $1 - \delta$*

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

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 equation (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 $1 - \delta$

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

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 $1 - \delta$:

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

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

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:

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

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:

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

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:

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

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 Section 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 $\ell(t) = [\ell(t, e_1), \dots, \ell(t, e_C)]^T$. The backward and forward loss functions are defined in equations 21a and 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're 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 multiclass 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 $\widehat{T}$.*

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

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

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

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

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 $\lambda_{\min}(T) = \sqrt{\kappa}$. The bounds from Theorem 4.6 holds replacing $\lambda_{\min}(T)$ 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 (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 D.1. In Fig. 4 (as well as the 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.

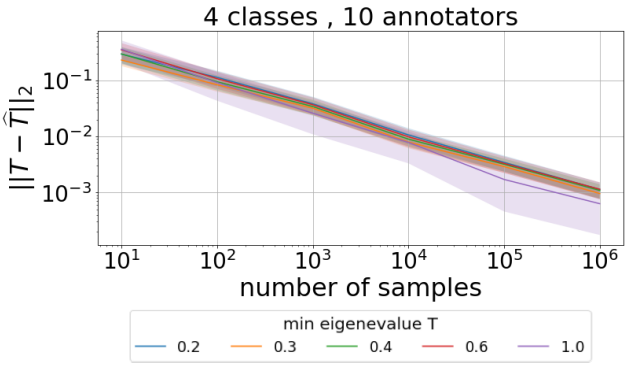


Figure 1. Error in the estimation of T for 4 classes and 10 annotators. The plots are obtained by averaging different admissible matrices T (see Appendix B) and averaged over matrices that have the same minimum eigenvalues rounded to the first decimal.

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 don't 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.

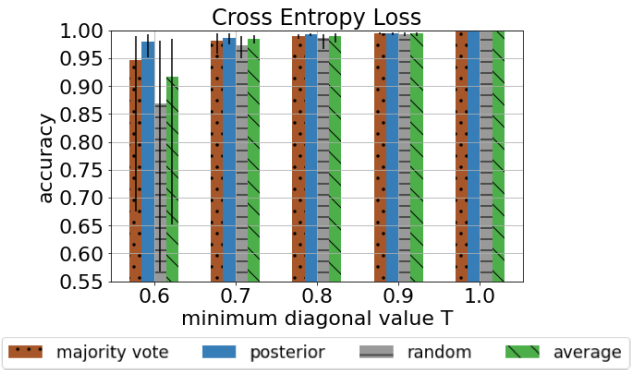


Figure 2. Comparison between performance of Cross Entropy Loss using majority vote, random aggregation method or the posteriors (posterior) and relative frequency (average) as soft labels.On the y-axis the accuracy on a clean dataset and on the x-axis the values of the minimum on the diagonal of T. Small 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.

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

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.

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

Supplementary Materials

A. Inter Annotator agreement. symmetric Noise and Symmetric ground truth distribution

Cohen's κ coefficient measures the agreement between two raters who each classify 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 [1] 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 (we will denote this quantity by p_e).

$$p_e = \sum_{c=1}^{C} \mathbb{P}(y_a = c)\mathbb{P}(y_b = c) \tag{23}$$

The Cohen's κ coefficient is defined as the difference between the true agreement and the "chance agreement" normalized by the maximum value this difference can reach

$$\kappa := \frac{p_o - p_e}{1 - p_e}, \tag{24}$$

If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters other than what would be expected by chance (i.e. $p_o = p_e$) $\kappa = 0$. It can also take negatives values. A negative κ indicates agreement worse than that expected by chance. This can be interpreted as not agreement at all between annotators. In our work we assume that the two raters are a corrupted version of a observable "clean" (ground truth) label. In this setting the label assigned by annotator a to an item and the respective uncorrupted label are not independent random variables. We found that in this setting the κ coefficient can takes only non-negative values.

B. On the hypothesis of commutativity in Lemma 4.1

In Lemma 4.1 we found how to compute T given M and D. To find this relationship we require that $D^{\frac{1}{2}}$ commutes with T. This hypothesis is satisfied when D and T have a particular structure, namely

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

That is satisfied or 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 vice-versa)) 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 instance

$$T = \begin{pmatrix} T_1 & 0 & 0 & 0 & 0 \\ 0 & T_2 & 0 & 0 & 0 \\ 0 & 0 & T_3 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & T_j \end{pmatrix} \text{ and } D = \begin{pmatrix} D_1 & 0 & 0 & 0 & 0 \\ 0 & D_2 & 0 & 0 & 0 \\ 0 & 0 & D_3 & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & D_j \end{pmatrix}$$

with

$$D_i = \begin{pmatrix} d_i & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & d_i \end{pmatrix}$$

T need not be block diagonal but must be reconducted to a block diagonal matrix by permuting the classes, for instance in the following case, we can obtain a matrix block diagonal by permuting classes 2 and 4

$$T = \begin{pmatrix} t_{11} & 0 & 0 & t_{14} \\ 0 & t_{22} & t_{23} & 0 \\ 0 & t_{23} & t_{33} & 0 \\ t_{14} & 0 & 0 & t_{44} \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & d_2 & 0 \\ 0 & 0 & 0 & d_1 \end{pmatrix}$$

Notice that T can be rewritten as follows permuting classes 2 and 4

$$T = \begin{pmatrix} t_{11} & t_{14} & 0 & 0 \\ t_{14} & t_{44} & 0 & 0 \\ 0 & 0 & t_{33} & t_{23} \\ 0 & 0 & t_{23} & t_{22} \end{pmatrix}$$

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 [18] 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 other 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 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 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 higher probability. In general we can fall back to an approximation of this case by reducing the samples.

C. Proofs

C.1. Proof of lemma 4.1

Proof. From equation 5 we get:

$$M = TDT = D^{\frac{1}{2}}TTD^{\frac{1}{2}} \rightarrow D^{-\frac{1}{2}}MD^{-\frac{1}{2}} = T^2 \tag{25}$$

Note that T and $D^{\frac{1}{2}}MD^{\frac{1}{2}}$ are positive definite (because D and M are positive definite) and hence they have eigenvalue decompositions of the following form:

$$T = U_T \Lambda_T U_T^T \tag{26}$$

$$D^{-\frac{1}{2}}MD^{-\frac{1}{2}} = U_M \Lambda_M U_M^T \tag{27}$$

where U_x are orthogonal matrices and Λ_x are diagonal positive definite matrices. It then follows that:

$$T^2 \overset{(a)}{=} U_T \Lambda_T^2 U_T^T = U_M \Lambda_M U_M^T \tag{28}$$

where in (a) we used the fact that U_T is orthogonal. Since $U_M \Lambda_M U_M^T$ is an eigenvalue decomposition of T^2 we conclude that:

$$T = U_M \Lambda_M^{\frac{1}{2}} U_M^T, \qquad T^{-1} = U_M \Lambda_M^{-\frac{1}{2}} U_M^T \tag{29}$$

□

C.2. Proof Lemma 4.2: Bounds error on the estimation of M,

Proposition C.0.1. *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). For every $\epsilon > 0$ it holds that*

$$\mathbb{P}^n(|(M_{a,b})_{ij}) - (\widehat{M_{a,b}})_{ij}| < \epsilon) \geq 1 - 2e^{-2\epsilon^2 n}.$$

And

$$\mathbb{P}^n\left(\forall i,j \in \{1,C\}^2 \; |(M_{a,b})_{ij}) - (\widehat{M_{a,b}})_{ij}| < \epsilon \right) \geq 1 - 2C^2 e^{-2\epsilon^2 n}.$$

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 $\mathbb{P}$.

To simplify the notation we will omit the dependency from the annotators in the matrices: $M = M_{a,b}$ and $\widehat{M} = \widehat{M}_{a,b}$ $M_{ij} = \mathbb{P}(y_a = i, y_b = j)$ and $\widehat{M}_{ij} = \frac{1}{n}\sum_{h=1}^{n} \mathbb{1}((y_a)_h = i, (y_b)_h = j)$.

Proof. To prove the claim we only need to apply the Hoeffding's inequality to the random variables $X_h^{ij} = \mathbb{1}_{y_{a\,h}=i,=y_{b\,h}=j}$. Indeed it holds that $0 \leq X^{ij} \leq 1$ 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.rf

$$\mathbb{P}(|\mathbb{E}[X_h^{ij}] - \frac{1}{n}\sum_{h=1}^{n} X_h^{ij}| > \epsilon) \leq 2e^{-2\epsilon^2 n}. \tag{30}$$

From the previous equation, using union bounds we can obtain that

$$\mathbb{P}\Big(\forall (i,j) \in \{1,C\}^2 \; |\mathbb{E}[X_h^{ij}] - \frac{1}{n}\sum_{h=1}^{n} X_h^{ij}| < \epsilon \Big) \geq 1 - 2C^2 e^{-2\epsilon^2 n}. \tag{31}$$

Namely

$$\mathbb{P}\Big(\forall (i,j) \in \{1,C\}^2 \; |M_{ij} - \widehat{M}_{ij}| < \epsilon \Big) \geq 1 - 2C^2 e^{-2\epsilon^2 n}. \tag{32}$$

□

Lemma C.1. *Let A be a matrix in $\mathbb{R}^{CxC}$ so that it exists $\epsilon > 0$ for all i, j $|A_{ij}| \leq \epsilon$. For every $p \in [1, \infty]$, if $||.||_p$ denotes the matrix norm induced by the p-vector norm,*

$$||A||_p \leq C\epsilon.$$

Proof.

$$||A||_p := \sup_{x:||x||_p=1} ||Ax||_p$$

Let x be a vector of p-norm 1. $(Ax)_i = \sum_{j=1}^{C} A_{ij}x_j$

$$||Ax||_p = \Big(\sum_{i=1}^{C}\Big|\sum_{j=1}^{C} A_{ij}x_j\Big|^p\Big)^{\frac{1}{p}} \leq \Big(\sum_{i=1}^{C}\Big(\sum_{j=1}^{C}|A_{ij}x_j|\Big)^p\Big)^{\frac{1}{p}} \leq \epsilon\Big(\sum_{i=1}^{C}\Big(\sum_{j=1}^{C}|x_j|\Big)^p\Big)^{\frac{1}{p}}$$

Now, denoting by $\mathbf{1}$ the vector with all ones, using Hölder inequality we can obtain :

$$\sum_{j=1}^{C}|x_j| = ||\mathbf{1}x||_1 \leq ||x||_p||\mathbf{1}||_{\frac{p}{p-1}} = ||x||_p C^{\frac{p-1}{p}}$$

So

$$||Ax||_p \leq \epsilon\Big(\sum_{i=1}^{C}||x||^p C^{p-1}\Big)^{\frac{1}{p}} = \epsilon C||x||_p = \epsilon C$$

□

Proof Lemma 4.2. For the previous Lemma it holds that if all the elements of the matrix are less or equal than ϵ, the p norm is bounded by ϵC

So we can derive that

$$\mathbb{P}(||M_{a,b} - \widehat{M}_{a,b}||_p > \epsilon) \geq \mathbb{P}\Big(\forall (i,j) \in \{1,C\}^2 \; |M_{ij} - \widehat{M}_{ij}| < \frac{\epsilon}{C} \Big) \geq 1 - 2C^2 e^{-2\frac{\epsilon^2}{C^2}n}. \tag{33}$$

□

C.3. Proof of Theorem 4.3:Bound error on the estimation of T,

We start by introducing the following helpful remark and lemmas.

Remark 2. *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.* (15). *We know that the matrix* T *we want to approximate satisfies all the constraints in Eq.* (15), *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 found for $||T - \widehat{U}\widehat{\Lambda}_M^{\frac{1}{2}}\widehat{U}^T||_2^2$ *holds also for* $\hat{T}$ *estimated as in Eq.* (14) *with a coefficent* 2.

Lemma C.2. *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 [9]). The bounded operator* $F_{\sqrt{}} : \mathcal{S} \to \mathcal{S}$ *defined as follow* $F_{\sqrt{}} : A = \sqrt{A}$, *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*

$$||D[\sqrt{A}]||_2 \leq \frac{1}{2\sqrt{\lambda_{\min}(A)}}||vec(A)||_2. \tag{34}$$

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 $F_{\sqrt{}}$ in this space and $D[A]$ the derivative of A. From the fact that $\sqrt{A}\sqrt{A} = A$ it follows that

$$D[\sqrt{A}]\sqrt{A} + \sqrt{A}D[\sqrt{A}] = D[A]. \tag{35}$$

Eq. 35 is a special case of Sylvester equation, and using that $\sqrt{A}$ is symmetric can be rewritten as

$$(I_C \otimes \sqrt{A} + \sqrt{A} \otimes I_C)\text{vec}(D[\sqrt{A}]) = \text{vec}(D[A]). \tag{36}$$

It follow that

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

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 $1 \leq i, j \leq C$, with λ_i eigenvalue of A. The minimum eigenvalue of a symmetric positive def matrix B is the minimum eigenvalue of the inverse, indeed id $B = VDV^T$, with V orthogonal, $B^{-1} = VD^{-1}V^T$. So 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 $2\lambda_{\min}(\sqrt{A})$. It follows that

$$\begin{aligned}
||(I_C \otimes \sqrt{A} + \sqrt{A} \otimes I_C)^{-1}||_2 &= \sqrt{\lambda_{\max}((I_C \otimes \sqrt{A} + \sqrt{A} \otimes I_C)^{-2})} \\
&= \sqrt{\lambda_{\min}((I_C \otimes \sqrt{A} + \sqrt{A} \otimes I_C)^{2})} \\
&= \lambda_{\min}((I_C \otimes \sqrt{A} + \sqrt{A} \otimes I_C)) \\
&= \frac{1}{2\sqrt{\lambda_{\min}(A)}}.
\end{aligned}$$

So $||\text{vec}(D[\sqrt{A}])||_2 \leq \frac{1}{2\sqrt{\lambda_{\min}(A)}}||\text{vec}(A)||_2$. $||\text{vec}(A)||_2^2 = \sum_{k=1}^{C^2} a_k^2$ for every vecto x of norm 1 (this implies $x_i < 1$)

$$||Ax||_2^2 = \sum_{k=1}^{C}\sum_{i=1}^{C} a_{ki}^2 x_i^2 \leq \sum_{k=1}^{C}\sum_{i=1}^{C} a_{ki}^2 = ||\text{vec}(A)||_2^2.$$

It follows that the induce 2 norm of the derivative $||D[\sqrt{A}]||_2 \leq \frac{1}{2\sqrt{\lambda_{\min}(A)}}||\text{vec}(A)||_2$ □

Let T and $\hat{T}$ be defined as in Eq. (29) and Eq. (13).

The following Lemma holds for two general double stochastic matrices.

Lemma C.3. *Let T and $\widehat{T}$ be two symmetric, stochastic matrices, it holds that :*

$$||T-\hat{T}||_2 \leq \frac{\sqrt{C}||T^2-\hat{T}^2||}{\lambda_{\min}(T^2)-||T^2-\hat{T}^2||_2} \quad \textit{and} \quad ||T-\hat{T}||_2 \leq \frac{\sqrt{C}||T^2-\hat{T}^2||}{\lambda_{\min}(\widehat{T}^2)-||T^2-\hat{T}^2||_2} \tag{37}$$

Proof. From the previous lemma and the mean absolute value

$$||\sqrt{A}-\sqrt{B}||_2 \leq ||A-B||_2 \sup_{0\leq\theta\leq 1} ||D[\sqrt{\theta A+(1-\theta)B}]||_2$$

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

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

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

$$\sup_{0\leq\theta\leq 1}\frac{1}{\theta\lambda_{\min}(T^2)+(1-\theta)\lambda_{\min}(\widehat{T}^2)} = \begin{cases} \frac{1}{\lambda_{\min}(T^2)} & \text{if} \quad \lambda_{\min}(T^2) < \lambda_{\min}(\widehat{T}^2) \\ \frac{1}{\lambda_{\min}(\widehat{T}^2)} & \text{if} \quad \lambda_{\min}(T^2) > \lambda_{\min}(\widehat{T}^2) \end{cases}$$

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

Now,

$$\min(a,b) = \begin{cases} a = b-|b-a| & \text{if } a<b \\ b & \text{if } b\leq a \end{cases} \tag{38}$$

We notice that for symmetric matrices $||A||_2 = \sqrt{\lambda_{\max}(A)^2} = \sqrt{(\lambda_{\max}(A))^2} = |\lambda_{\max}(A)|$. So we can Since $T^2-\hat{T}^2$ is symmetric: $||T^2-\hat{T}^2||_2 = |\lambda\max(T^2-\hat{T}^2)|$.

It follows that

$$\begin{aligned}
\min(\lambda_{\min}(\hat{T}^2),\lambda_{\min}(T^2)) &\geq \lambda_{\min}(T^2)-|\lambda_{\min}(T^2)-\lambda_{\min}(\hat{T}^2)| && (39)\\
&\geq \lambda_{\min}(T^2)-|\lambda_{\min}(T^2)-\lambda_{\min}(\hat{T}^2)| && (40)\\
&\geq \lambda_{\min}(T^2)-|\lambda_{\min}(T^2-\hat{T}^2)| && (41)\\
&\geq \lambda_{\min}(T^2)-|\lambda_{\max}(T^2-\hat{T}^2)| && (42)\\
&= \lambda_{\min}(T^2)-||T^2-\hat{T}^2||_2. && (43)
\end{aligned}$$

In the previous equations we use that $|\lambda_{\min}(T^2)-\lambda_{\min}(\hat{T}^2)| \leq |\lambda_{\max}(T^2-\hat{T}^2)|$. 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 $\lambda_{\min}(T^2)-\lambda_{\min}(\hat{T}^2) = \lambda_{\min}(T^2)+\lambda_{\max}(-\hat{T}^2) \leq \lambda_{\max}(T^2-\hat{T}^2) \leq |\lambda_{\max}(T^2-\hat{T}^2)|$, where we used Weyl's inequality.

If the $\lambda_{\min}(T^2) > \lambda_{\min}(\widehat{T}^2)$ following the same path we obtain $\lambda_{\min}(\widehat{T}^2)-\lambda_{\min}(T^2)| \leq |\lambda_{\max}(\hat{T}^2-T)|$.

it follow that $\lambda_{\min}(T^2)-\lambda_{\min}(\hat{T}^2) < ||T^2-\hat{T}^2||_2$

□

Proof Theorem Theorem 4.3. From Lemma C.3 we know that

$$||T-\hat{T}||_2 \leq \frac{\sqrt{C}||T^2-\hat{T}^2||}{\lambda_{\min}(T^2)-||T^2-\hat{T}^2||_2} \tag{44}$$

Now, in general

$$\frac{\sqrt{C}x}{b-x} < \epsilon \text{ iif } x < b\frac{\epsilon}{\sqrt{C}+\epsilon}.$$

It follows that

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

or

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

Since we can assume $\epsilon \leq 1$ (if $n > \frac{C^2(\sqrt{C}+1)^2(\ln(2C^2)^2}{2\lambda_{\min}(\hat{T})^2}$. Notice that we're interested in convergence properties of $\hat{T}$, so we're interested in founding these bounds for small ϵ.

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

So $||T^2-\widehat{T}^2||_2 \leq ||M-\hat{M}||_2||D^{1/2}||_2^2 = ||M-\hat{M}||_2||D||_2 = ||M-\hat{M}||_2\lambda_{\max}(D)$. As a consequence :

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

For the inverse:

$$T^{-1}-\widehat{T}^{-1} = T^{-1}(\widehat{T}-T)\widehat{T}^{-1} \tag{45}$$

So,

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

Following what we did for the κ in

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

Than for Eq. (39)

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

So

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

Where we used that

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

because T and $\widehat{T}$ doubly stochastic.

So

$$\mathbb{P}\left(||T^{-1}-\widehat{T}^{-1}||_2 \leq \epsilon\right) \geq \mathbb{P}\left(||T-\widehat{T}||_2 \leq \epsilon\frac{\lambda_{\min}(\widehat{T})}{1+2\epsilon}\right) \tag{46}$$

$$\geq \mathbb{P}\left(||T-\widehat{T}||_2 \leq \frac{\epsilon}{3}\lambda_{\min}(\widehat{T})\right) \tag{47}$$

$$\geq 1-2C^2 e^{-\frac{\epsilon^2}{9C^2(\sqrt{C}+1)^2}\frac{\lambda_{\min}(\widehat{T}^2)^4}{\lambda_{\max}(D)^2}n} \tag{48}$$

□

C.4. Proof Theorem 4.6 :Generalization gap bounds

Proposition C.3.1. *Let $\ell(t,y)$ be any bounded loss function and let $l(t,y)$ be the backward loss function defined in Eq. (21a).*

We define $\hat{l}(t,y)$ as the loss obtained using $\widehat{\Gamma}^{-1} := \widehat{T}^{-1}$. If μ is the constant that bounded the loss ℓ , i.e. $\sup_{(t,y)\in[0,1]^C\times\mathcal{Y}}\ell(t,y)\leq\mu$. For every ϵ

$$\mathbb{P}(|l(t,y)-\hat{l}(t,y)|\geq\epsilon) \leq 2C^2 e^{-2\frac{\epsilon^2}{C^2\mu^2 L_{\phi,p}}n} \tag{49}$$

Proof of Proposition C.3.1. Using Cauchy–Schwarz inequality and the fact that ℓ is bounded by μ and that we obtain:

$$\begin{aligned}
|l(t,y)-\hat{l}(t,y)| &= |(T^{-1}\cdot\ell(t)-\widehat{T}^{-1}\cdot\ell(t))_y| \\
&= |[(T^{-1}-\widehat{T}^{-1})\ell(t)]\cdot e_y| \\
&\leq ||(T^{-1}-\widehat{T}^{-1})\ell(t)||_2||e_y||_2 \\
&\leq ||T^{-1}-\widehat{T}^{-1}||_2||\ell(t)||_2 \\
&\leq \mu||T^{-1}-\widehat{T}^{-1}||_2
\end{aligned}$$

So

$$\mathbb{P}\big(|l(t,y)-\hat{l}(t,y)|\leq\epsilon\big) \geq 1-2C^2 e^{-\frac{\epsilon^2}{\mu^2 9C^2(\sqrt{C}+1)^2}\frac{\lambda_{\min}(\widehat{T}^2)^4}{\lambda_{\max}(D)^2}n}$$

□

Proof Lemma 4.5 . For every f we have

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

So using union bounds and by the classic results on Rademacher complexity bounds [14] and by the Lipschitz composition property of Rademacher averages, Theorem 7 in [11] it follows that

$$\mathbb{P}^n\Big(\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-R_{l,\mathcal{D}}(f)| \leq L||T^{-1}||_2\mathfrak{R}_n(\mathcal{F})+\frac{\epsilon}{2}\Big) \geq \tag{50}$$

$$\mathbb{P}^n\Big(\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-\hat{R}_l(f)|+\sup_{f\in\mathcal{F}}|\hat{R}_l(f)-R_{l,\mathcal{D}}(f)| \leq L||T^{-1}||_2\mathfrak{R}_n(\mathcal{F})+\frac{\epsilon}{2}\Big) \geq \tag{51}$$

$$1-\mathbb{P}^n\Big(\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-\hat{R}_l(f)|>\frac{\epsilon}{4}\Big)-\mathbb{P}^n\Big(\sup_{f\in\mathcal{F}}|\hat{R}_l(f)-R_{l,\mathcal{D}}(f)| \leq L||T^{-1}||_2\mathfrak{R}_n(\mathcal{F})+\frac{\epsilon}{4}\Big) \tag{52}$$

$$\geq 1-\mathbb{P}^n\Big(\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-\hat{R}_l(f)|>\frac{\epsilon}{4}\Big)-2e^{-\frac{n}{2}\left(\frac{\epsilon}{4\mu}\right)^2} \tag{53}$$

Now,

$$\mathbb{P}^n\Big(\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-R_{l,\mathcal{D}}(f)|\leq L||\widehat{T^{-1}}||_2\mathfrak{R}_n(\mathcal{F})+\epsilon\Big)=$$

$$\mathbb{P}^n\Big(\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-R_{l,\mathcal{D}}(f)|\leq L||\widehat{T^{-1}}||_2\mathfrak{R}_n(\mathcal{F})+\epsilon\Big)$$

$$\mathbb{P}^n\Big(\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-R_{l,\mathcal{D}}(f)|\leq L||T^{-1}||_2\mathfrak{R}_n(\mathcal{F})+(||\widehat{T^{-1}}||_2-||T^{-1}||_2)\mathfrak{R}_n(\mathcal{F})+\epsilon\Big)\geq$$

$$1-\mathbb{P}^n\Big(\{\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-R_{l,\mathcal{D}}(f)|\leq L||T^{-1}||_2\mathfrak{R}_n(\mathcal{F})+\frac{\epsilon}{2}\}\text{ and }\{(||\widehat{T^{-1}}||_2-||T^{-1}||_2)\mathfrak{R}_n(\mathcal{F})\leq\frac{\epsilon}{2}\}\Big)\geq$$

$$1-\mathbb{P}^n\Big(\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-R_{l,\mathcal{D}}(f)|\leq L||T^{-1}||_2\mathfrak{R}_n(\mathcal{F})+\frac{\epsilon}{2}\Big)-\mathbb{P}^n\Big((||\widehat{T^{-1}}||_2-||T^{-1}||_2)\mathfrak{R}_n(\mathcal{F})\leq\frac{\epsilon}{2}\Big)\geq$$

$$1-2e^{-\frac{n}{2}\left(\frac{\epsilon}{4\mu}\right)^2}-\mathbb{P}^n\Big((||\widehat{T^{-1}}-T^{-1}||_2)\leq\frac{\epsilon}{2\mathfrak{R}_n(\mathcal{F})}\Big)-\mathbb{P}^n\Big(\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-\hat{R}_l(f)|>\frac{\epsilon}{4}\Big)\geq$$

$$1-2e^{-\frac{n}{2}\left(\frac{\epsilon}{4\mu}\right)^2}-2C^2e^{-\frac{\epsilon^2}{4\mathfrak{R}_n(\mathcal{F})^2 9C^2(\sqrt{C}+1)^2}\frac{\lambda_{\min}(\hat{T}^2)^4}{\lambda_{\max}(D)^2}n}-2C^2e^{-\frac{\epsilon^2}{4\mu^2 9C^2(\sqrt{C}+1)^2}\frac{\lambda_{\min}(\hat{T}^2)^4}{\lambda_{\max}(D)^2}n}$$

$$\geq 1-2e^{-\frac{n}{2}\left(\frac{\epsilon}{4\mu}\right)^2}-4C^2e^{-\frac{1}{\max(\mathfrak{R}_n(\mathcal{F}),\mu)^2}\frac{\epsilon^2}{36C^2(\sqrt{C}+1)^2}\frac{\lambda_{\min}(\hat{T}^2)^4}{\lambda_{\max}(D)^2}n}$$

$$\geq 1-4e^{-\left[\min\left(\frac{1}{8},2\ln(C)\frac{1}{9\mathfrak{R}_n(\mathcal{F})^2C^2}\frac{\lambda_{\min}(\hat{T}^2)^4}{(\sqrt{C}+1)^2\lambda_{\max}(D)^2}\right)\right]\frac{\epsilon^2}{4\mu^2}n}$$

$$\geq 1-4Ce^{-\left(\frac{1}{9\mathfrak{R}_n(\mathcal{F})^2C^2}\frac{\lambda_{\min}(\hat{T}^2)^4}{(\sqrt{C}+1)^2\lambda_{\max}(D)^2}\right)\frac{\epsilon^2}{2\mu^2}n}$$

So with probability at least $1-\delta$

$$\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-R_{l,\mathcal{D}}(f)|\leq 2L\lambda_{\min}(\widehat{T}^2)\mathfrak{R}_n(\mathcal{F})+\frac{6\mu\mathfrak{R}_n(\mathcal{F})\lambda_{\min}(D)C^2(\sqrt{C}+1)}{\lambda_{\min}(\hat{T})^2}\sqrt{\frac{1}{n}\ln\left(\frac{4C}{\delta}\right)}.$$

Or

$$\sup_{f\in\mathcal{F}}|\hat{R}_{\hat{l}}(f)-R_{l,\mathcal{D}}(f)|\leq\left[2L\lambda_{\min}(\widehat{T}^2)+\frac{\mu\lambda_{\min}(D)}{\lambda_{\min}(\hat{T})^2}\sqrt{\frac{1}{n}\ln\left(\frac{4C}{\delta}\right)}\right]\mathfrak{R}_n(\mathcal{F})g(C).\tag{54}$$

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

□

Theorem 4.6 . By the unbiasedness of l we have that $R_{\ell,\mathcal{D}}(\hat{f})=R_{l,\mathcal{D}}(\hat{f})$. Moreover since $\hat{f}=\underset{f}{\operatorname{argmin}}(\widehat{R}_{\hat{l}}(f))$ we have $\widehat{R}_{\hat{l}}(\hat{f})\leq\widehat{R}_{\hat{l}}(g)\ \forall g\in\mathcal{F}$.

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

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

□

C.5. Proof of Lemma 4.4

Lemma C.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. $\lim_{H\to\infty} p_{c,i} \overset{a.s.}{=} \mathbb{I}(y_i = c)$).*

Proof.

$$p_{c,i} = \frac{\mu_c \prod_{h=1}^{H} T_{c,y_{h,i}}}{\sum_{j=1}^{C} \mu_j \prod_{h=1}^{H} T_{j,y_{h,i}}} \tag{55}$$

$$\prod_{h=1}^{H} T_{c,y_{h,i}} = \prod_{j=1}^{C} T_{c,j}^{N_{i,j}} \tag{56}$$

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:

$$\lim_{H\to\infty} \frac{N_{i,j}}{H} = \lim_{H\to\infty} \frac{\sum_{a=1}^{H} \mathbb{1}_{\{y_{a,i}=j\}}}{H} = \mathbb{E}[\mathbb{1}_{\{y_{a,i}=j\}} | y = j] = T_{y_i,j} \tag{57}$$

Combining we get:

$$\lim_{H\to\infty} p_{c,i} = \lim_{H\to\infty} \frac{\mu_c \prod_{j=1}^{C} T_{c,j}^{N_{i,j}}}{\sum_{k=1}^{C} \mu_k \prod_{j=1}^{C} T_{k,j}^{N_{i,j}}} \tag{58}$$

$$= \lim_{H\to\infty} \frac{\mu_c \left(\prod_{j=1}^{C} T_{c,j}^{T_{y_i,j}}\right)^H}{\sum_{k=1}^{C} \mu_k \left(\prod_{j=1}^{C} T_{k,j}^{T_{y_i,j}}\right)^H} \tag{59}$$

$$= \lim_{H\to\infty} \frac{1}{1 + \sum_{\substack{k=1 \\ k\neq c}}^{C} \frac{\mu_k}{\mu_c} \left(\prod_{j=1}^{C} \left(\frac{T_{k,j}}{T_{c,j}}\right)^{T_{y_i,j}}\right)^H} \tag{60}$$

$$\overset{(a)}{=} \mathbb{1}(y_i = c) \tag{61}$$

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. □

C.6. Proof Proposition 5.1: Relationship between ρ and κ.

Proof.

$$p_o = \mathbb{P}(y_a = y_B) = \sum_{k,h=1}^{C} \mathbb{P}(y_A = k, y_B = k | y = h)\mathbb{P}(y = h)$$

$$= \sum_{k,h=1}^{C} \mathbb{P}(y_A = k|y = h)\mathbb{P}(y_B = k|y = h)\nu_h = \sum_{k,h=1}^{C} T_{h,k}^2 \nu_h$$

$$= \sum_{h=1}^{C} (1-p)^2 c_h + \sum_{h=1}^{C} \left(\frac{p}{C-1}\right)^2 (C-1)c_h = (1-p)^2 + \frac{p^2}{C-1}$$

Now

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

In the previous equation we used that T is symmetric.

$$\begin{aligned} p_e &= \sum_{k=1}^{C} \mathbb{P}(y_A = k)\mathbb{P}(y_B = k) = \sum_{k=1}^{C} \mathbb{P}(y_A = k)\mathbb{P}(y_B = k) = c^T T^2 c \\ &= 2\frac{p}{C-1} - \frac{Cp^2}{(C-1)^2} + \Big(1 - \frac{Cp}{C-1}\Big)^2 \nu^T \nu \end{aligned} \tag{62}$$

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 so

$$\kappa = \frac{C^2 p^2 - 2C(C-1)p + (C-1)^2}{(C-1)^2} \tag{63}$$

From which it follows that

$$p = (1 - C^{-1})(1 - \sqrt{\kappa}) \tag{64}$$

□

D. Experiments

D.1. Estimation T

From Figure 3 we can notice that the error in the estimation decreases as $\frac{1}{\sqrt{n}}$ the n number of samples increases. The results with respect to the minimum eigenvectors and with respect to the maximum diagonal value are consistent with each other and very similar.

The results were obtained from a synthetic, generated dataset in which we generate the classes predicted by the annotators according to various T matrices, choosing 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. We can notice in Figure 3 that as the number of annotators increase the estimation becomes more precise.

For experiments with $2, 3$ and 7 annotators we generate T as all possible symmetric ,stochastic and diagonally dominant matrices that have $[0.1, 0.2, 0.3, 0.4, 0.5]$ 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 that are so that the four classes can be partitioned in two groups of indices so that classes in the same group have the same probability. Namely if the distributions on the classes is 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 $i, k \in I : d_i = d_k$. The probability of the classes take value in $[0.1, 0.2, 0.3, 0.4]$. This means that for instance we will find the distribution $[0.3, 0.3, 0.3, 0.1]$ or the distribution $[0.4, 0.1, 0.1.0.4]$ but not $[0.3, 0.2, 0.1, 0.4]$.

Results for $2, 3$ and 7 annotators were obtained by averaging over 3 runs. Results for 10 annotators were obtained by averaging 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 Sec. 4.1.

We recall that if the minimum eigenvalue is 1 the matrix T is the identity and thus the annotators always predict the exact class. The smaller the minimum eigenvalue the noisier the dataset will be.

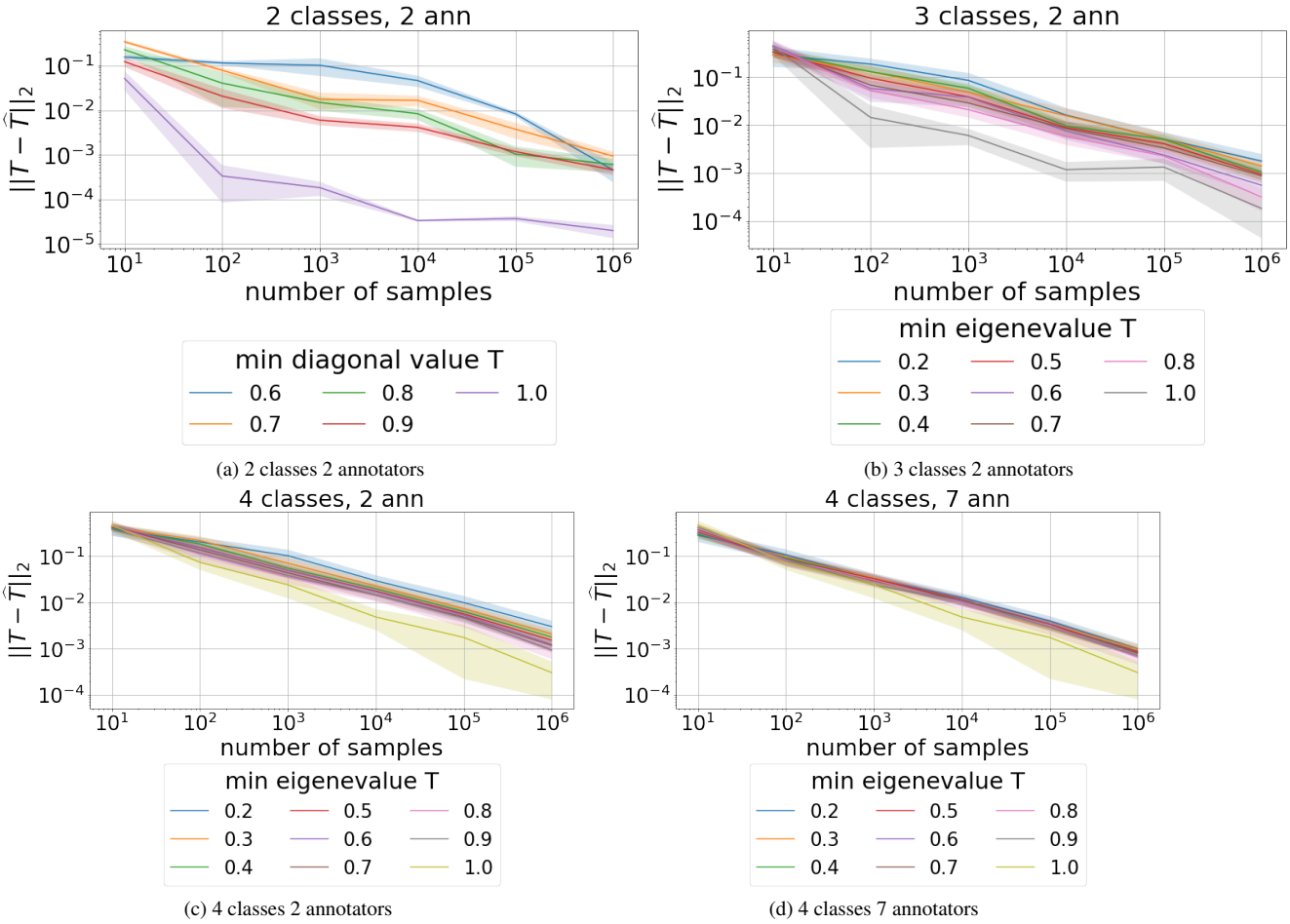
With Figure 4 we wanted to see if datasets with a higher level of noise have higher approximation errors than less noisy datasets. The plots show a minor trend: as the noise decreases, the estimation error also decreases. The trend is not particularly noticeable perhaps due to the large number of annotators.

We recall that if the minimum eigenvalue is 1 or if the maximum value of the diagonals is 1 the matrix T is the identity and thus the annotators always predict the exact class.

The smaller the minimum eigenvalue or the maximum value on the diagonal, the noisier the dataset will be.

D.2. Synthetic Datasets

The synthetic dataset consists of two-dimensional features ($\mathbf{x} = (x_1, x_2)$). To create the dataset, we generate points uniformly at random in $[0, 1]^2$. Each of these points is then assigned a label (y) based on the predetermined label distribution for each experiment. We divide the space into sections using lines parallel to the bisector of the first and third quadrants (specifically, $x_2 = x_1$). See Figure 5 for an example. Our dataset comprises 10000 samples. In Figure 6 we see, for different



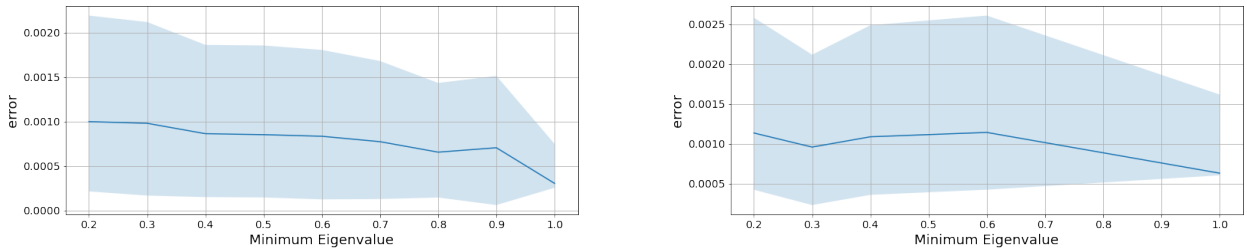
(a) 2 classes 2 annotators

(b) 3 classes 2 annotators

(c) 4 classes 2 annotators

(d) 4 classes 7 annotators

Figure 3. Error in the Estimation of T. The error is $||T - \widehat{T}||_2$. We aggregated the matrices that have the same minimum eigenvalue rounded at the first decimal.



(a) 4 classes 7 annotators

(b) 4 classes 10 annotators

Figure 4. The plots show the trend of the error estimation as the minimum eigenvalue increases

amounts of noise, the results of the different aggregation methods when using a neural network without hidden layer (i.e. a Logistic Regression) trained with Cross Entropy Loss. When noise is absent, we check that, as expected, the results are all identical. In the presence of noise (0.6 and 0.8), we notice in general that the random aggregation is the worst. The others are equivalent, except for the posterior (ours) which obtains slightly higher results. Average, on the other hand, obtains a slightly lower value with minimum diagnal value of T equal to 0.8. However, attention must be drawn to the fact that the y-scale of the graph is very narrow and that in the case of 4 classes with a dataset constructed as in figure 5, a linear classifier is not able to reach perfect accuracy.

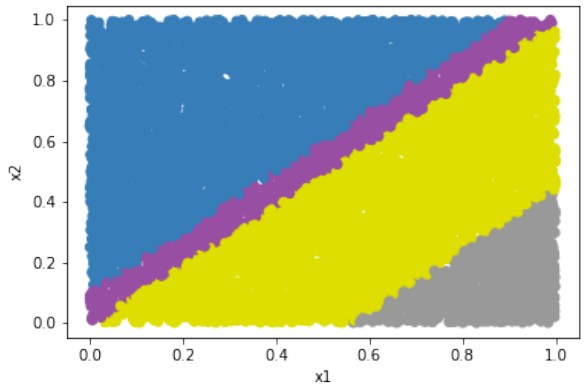


Figure 5. Synthetic data for 4 classes with distribution (0.4,0.1,0.4,0.1)

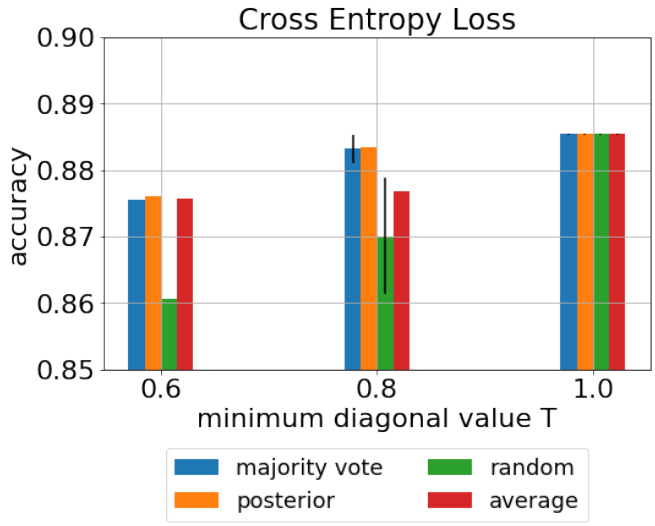


Figure 6. 5 annotators, 4 classes, no hidden layer.

Figure 7

Referring to Fig. 2 and the other figures of this section. The minimum value on the diagonal of the matrix T denotes the annotators' probability of assigning the correct label for the class in which the noise is maximum. As expected, random aggregation is the lowest performing method, and for all noise rates soft label methods perform better than methods using hard labels.

Fig. 6 shows the accuracy for the case of 4 classes and a NN with no hidden layer and 5 annotators. We can notice that even in the case where the number of hidden neurons is not enough to obtain a perfect accuracy, so the classifier is not the best possible, our approach for dataset with high noise performs better.

The posteriors distribution are computed using the estimated T.

D.3. Implementation details

Logistic Regression is used for synthetic data with 2 classes and a neural network with 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%, The models are trained with the following configuration: batch size 256, learning rate 10^{-3}, maximum number of epochs 1000, early stopping of training based on validation loss with a patience of 100 epochs. Once the training is finished, the model with the lowest validation loss is retrieved.

For the experiments with CIFARN-10, the model, Resnet 34, is trained with the following configuration: batch size 128, learning rate 10^{-3}, with momentum (0.9) and learning rate decay (0.0005) the maximum number of epochs 1000, we also used early stopping of training based on validation loss with a patience of 100 epochs. We didn't use data augmentation. For the pretrained model we used the model provided by torchvision, https://pytorch.org/vision/main/models/generated/torchvision.models.resnet34.html#resnet34.

All code is written in Python 3 Programming Language. The cvxpy package is used for the optimization of $\widehat{T}$, and the pytorch library is used for the models. All the experiments have been run on a machine with this configuration: AMD EPYC 7373 Processor, 64GB RAM and NVIDIA GeForce RTX A4000 GPU.