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 [5] 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 Eq. (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 of 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. 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 [2]). The bounded operator $F_{\sqrt{\cdot}} : \mathcal{S} \to \mathcal{S}$ defined as follow $F_{\sqrt{\cdot}} : A = \sqrt{A}$, 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 $F_{\sqrt{\cdot}}$ 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 are interested in convergence properties of $\hat{T}$, so we are 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 :

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

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

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

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

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

□

C.4. Proof of 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

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

□

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 [4] and by the Lipschitz composition property of Rademacher averages, Theorem 7 in [3] 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 of 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.

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

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 of T

From Fig. 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 Fig. 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.

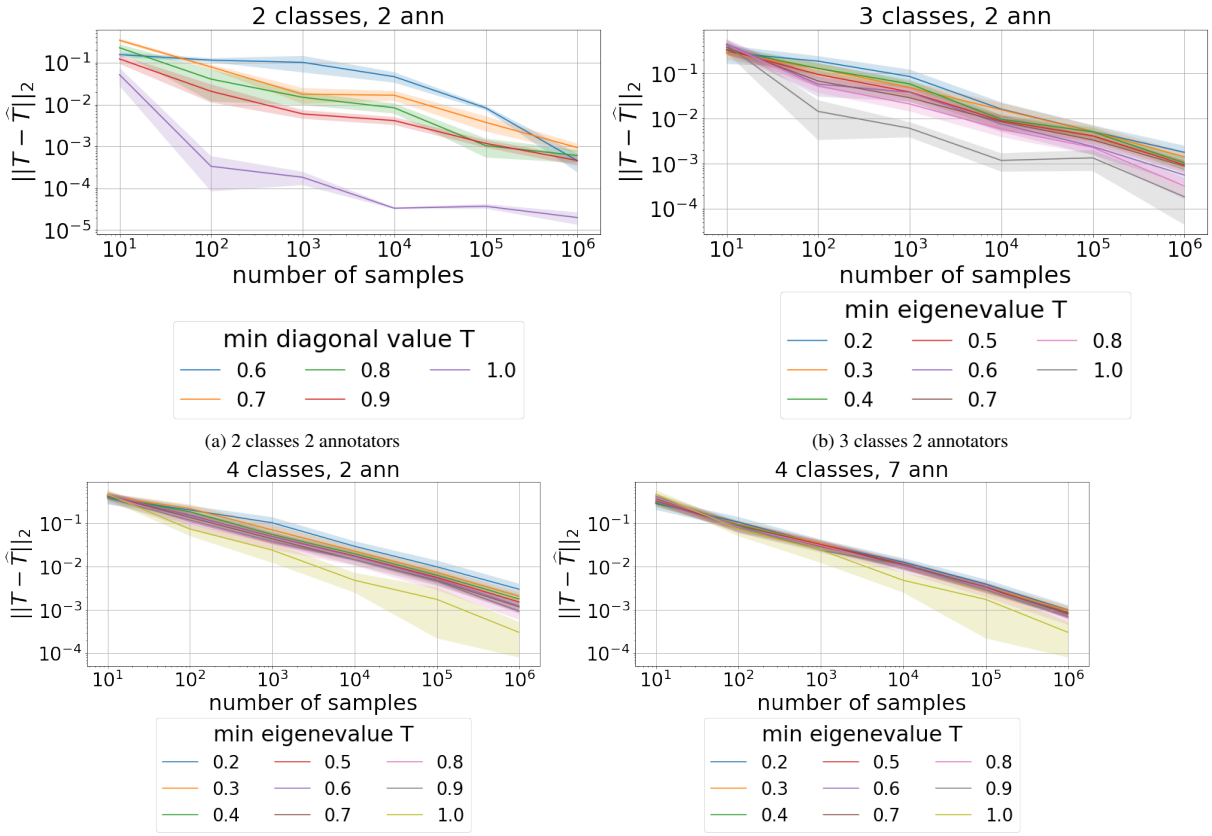
With Fig. 4 we wanted to see if datasets with a higher level of noise have higher approximation errors than less noisy datasets. The plots show a minor trend: as the noise decreases, the estimation error also decreases. The trend is not particularly noticeable perhaps due to the large number of annotators.

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 Fig. 5 for an example. Our dataset comprises 10000 samples. In Fig. 6 we see, for different



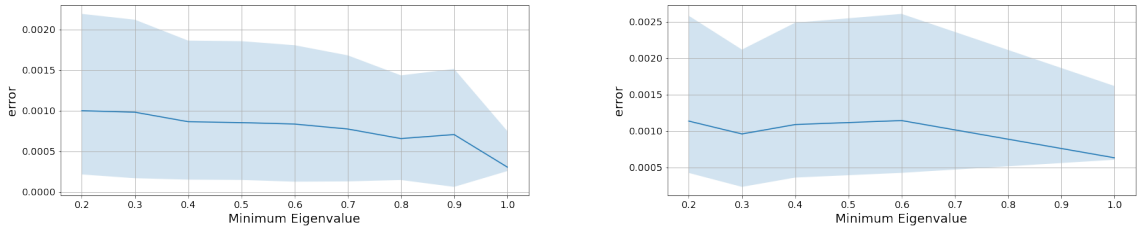
(a) 2 classes 2 annotators

(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 Fig. 5, a linear classifier is not able to reach perfect accuracy.

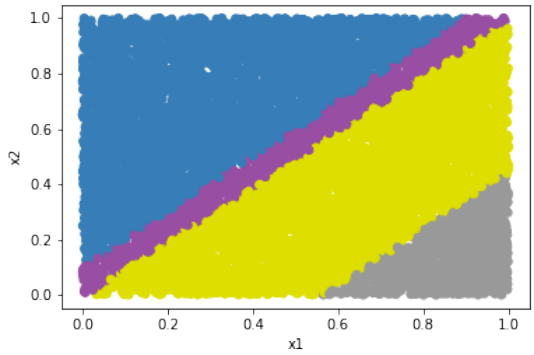


Figure 5. Synthetic data for 4 classes with distribution (0.4,0.1,0.4,0.1)

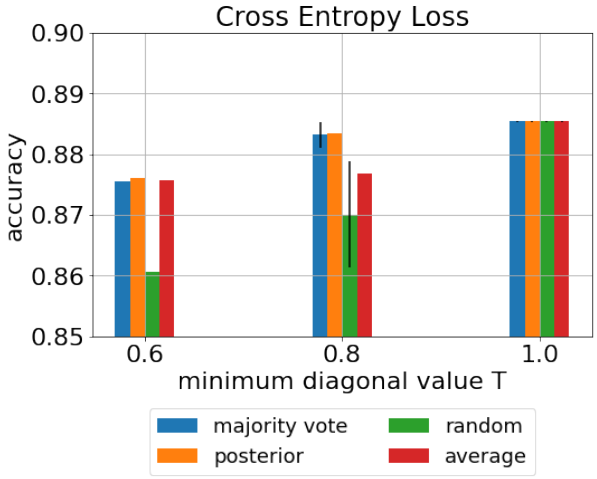


Figure 6. 5 annotators, 4 classes, no hidden layer.

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.

References

[1] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. 8

[2] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, USA, 2nd edition, 2012. 11

[3] R. Meir and T. Zhang. Generalization error bounds for bayesian mixture algorithms. *J. Mach. Learn. Res.*, 4:839–860, dec 2003. 14

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

[5] J. E. Potter. Matrix quadratic solutions. *SIAM Journal on Applied Mathematics*, 14(3):496–501, 1966. 9