The Majority Vote Paradigm Shift: When Popular Meets Optimal

Antonio Purificato[1,2,◇] **Maria Sofia Bucarelli**[2,3,4,5,◇,♠] **Anil Nelakanti**[6,♣]

Andrea Bacciu[1] **Fabrizio Silvestri**[2] **Amin Mantrach**[1]

[1]Amazon, [2]Sapienza University of Rome, [3]CNRS, [4]Inria, [5]i3s, [6]IIIT Hyderabad

Abstract

Reliably labelling data typically requires annotations from multiple human workers. However, humans are far from being perfect. Hence, it is a common practice to aggregate labels gathered from multiple annotators to make a more confident estimate of the true label. Among many aggregation methods, the simple and well-known Majority Vote (MV) selects the class label receiving the highest number of votes. However, despite its importance, the optimality of MV's label aggregation has not been extensively studied. We address this gap in our work by characterising the conditions under which MV achieves the theoretically optimal lower bound on label estimation error. Our results capture the tolerable limits on annotation noise under which MV can optimally recover labels for a given class distribution. This certificate of optimality provides a more principled approach to model selection for label aggregation as an alternative to otherwise inefficient practices that sometimes include *higher experts*, *gold* labels, etc., that are all marred by the same human uncertainty despite huge time and monetary costs. Experiments on both synthetic and real-world data corroborate our theoretical findings.

1 INTRODUCTION

Data labeling or annotation is crucial for numerous real-world machine learning tasks like, for example, search and retrieval (Gligorov et al., 2013; Shah and Wainwright, 2018), medical imaging (Wang et al., 2019), and translation and summarization (Sarhan and Spruit, 2020). If humans were perfect at labelling data, a single round of annotation would suffice, however, worker reliability varies widely across its population. Hence it is common to gather a noisy label set (Li et al., 2014) by requesting annotations from distinct workers for each task before aggregating them (Huang et al., 2023). The most commonly used aggregation method is *Majority Vote* (MV). MV estimation selects the class label that receives the highest number of votes from distinct annotators for a given task. Despite being easy to implement, it fails to recover true labels when incorrect annotations begin to dominate. The theoretical lower bound on the minimum achievable error given the true worker noise was described by Nitzan and Paroush (1981). Assuming there is an oracle revealing the true worker noise and class distribution, it reduces to the *maximum a posteriori* estimate with each vote weighted by the odds ratio of the worker's reliability and the class distribution. Correspondingly, we refer to this estimate as the oracle MAP (oMAP) that is practically unknown without access to such an oracle. This makes the problem ill-posed (Spurling et al., 2021; Shi et al., 2021) for which numerous methods have been proposed in literature. The earliest from Dawid and Skene (1979) (DS) uses expectation-maximization (EM), followed by several iterative (Whitehill et al., 2009; Hovy et al., 2013) and non-iterative methods (Li et al., 2019b; Yang et al., 2024a). Some of these methods offer guarantees on the expected error rate $\mathbb{E}[\mathcal{R}]$ of their estimate $\hat{y}$ of the true label y where $\mathcal{R} = \frac{1}{|\{y,\hat{y}\}|} \sum_{\{y,\hat{y}\}} \mathbb{I}[\hat{y} \neq y]$. They help understand *"How many more annotations would I need for* $\mathbb{E}[\mathcal{R}]$ *to be arbitrarily small?"* or say *"What is minimum* $\mathbb{E}[\mathcal{R}]$ *reachable given my annotation budget?"* However,

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

◇Equal contribution.
♠Work done while at Sapienza.
♣Work done while at Amazon.
Correspondence to: purifian@amazon.com

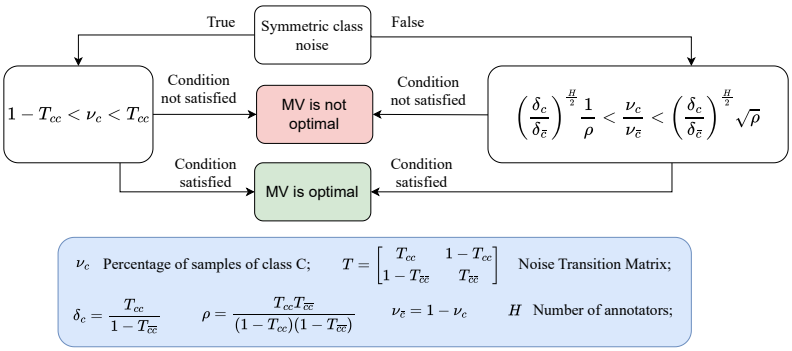


Figure 1: Flow-chart summarizing conditions for optimality of MV with its label estimates matching that of oMAP.

they do not help decide *"Which method is to be preferred for label aggregation on my data?"* or *"Does MV achieve the lowest achievable error for my annotation job?"*.

Precisely, this is the question we answer by deriving the parameter configurations where MV is sample-wise optimal, *i.e.,* when its label aggregate matches oMAP for all instances. We present a mechanism to verify if these conditions hold true for any given real-data even while the true parameters, *i.e.,* annotators' confusion matrix and class distribution are unknown. Specifically, our contributions address the following research questions:

- **RQ1**: *Can MV achieve the theoretically optimal label estimation error and under what conditions?* We identify a range of reliability and class imbalance parameters where label recovery using MV is optimal for each instance with its estimate matching that of oMAP. We derive these conditions for informative, non-adversarial workers who share the same class confusion matrix when annotating binary tasks. We also extend our results to scenarios with uniformly perturbed noise transition matrices and to specific cases involving two distinct categories of annotators. See Fig. 1 for summary.
- **RQ2**: *Are MV's optimality conditions verifiable on real-data and how tight are they?* The listed optimality conditions require an oracle to reveal the true parameters to verify if it holds. However, approximating true parameters with estimates works quite well in most cases though it comes with no certificate of correctness due to estimation errors. To address this, we derive stricter conditions on the estimated parameters that can be verified offering a high-probability guarantee of the true values lying within this regime. Our empirical results on synthetic and real data justify utility of checking the condition with estimated quantities to verify MV optimality.

  We also explore false-discovery and sensitivity trade-off in the two verification approaches we propose. Practitioners must choose from them depending on the requirements of their specific use-case.

2 RELATED WORK

In practice, MV is the most popular method due to ease of implementation. A popular alternative is a group of iterative solvers that update estimates of task labels and model priors (of class distribution and worker reliability) in successive loops. Dawid and Skene (1979) and more recent variants of iterative methods (Karger et al., 2014; Li and Yu, 2014; Li et al., 2019a; Shah et al., 2021; Chen et al., 2023) are examples of these methods. The parameters are estimated via an EM algorithm. Spectral algorithms that use the eigenspace of the worker-label matrix of annotations have also been proposed (Ghosh et al., 2011; Zhang et al., 2014; Tenzer et al., 2022) that have some similarities to iterative solvers (see Karger et al. (2014)).

Methods that leverage other principles like Bayesian formulation have also been studied (Soto et al., 2016). Li et al. (2019b) use the mean-field variational method to mine latent source relationships through tensor decomposition and object clustering. Yang et al. (2024a) view the label aggregation task as a dynamic system, with task identifiers serving as time-slices. Using a Dynamic Bayesian Network to model this system, they develop two label aggregation algorithms. Li et al. (2019a) proposed a Bayesian model based on conjugate prior and iterative EM reasoning for highly redundant labeled data. Similarities can be drawn among various methods and where feasible, they theoretically bound $\mathbb{E}[\mathcal{R}]$ as $\exp(-\mathcal{O}(vr))$ for some measure of crowd reliability r and annotation volume v. They fundamen-

Code for reproducing the results can be found at `https://github.com/amazon-science/majority_vote_paradigm_shift`.

tally differ in the implicit assumptions on the priors in model specification and the solver used that are necessary for the bounds to hold true. For example, Karger et al. (2014) assumes sparse assignment of samples to workers, while Dalvi et al. (2013) requires large eigengap for worker-label annotation matrix. Designing annotation jobs where the assumptions hold is non-trivial since they are not easily verifiable making it a trial-and-error to evaluate them suitability in practice.

Certain classes of methods explicitly estimate only the noise transition matrix of workers, sometimes with guarantees on optimality, from which label estimates could be inferred subsequently. For example, Bonald and Combes (2017) employ correlations of worker triplets to estimate workers' noise transition with ℓ_∞-norm of the error bounded as $\mathcal{O}(\frac{1}{v\sqrt{N}})$, the minimax lower bound for their setting with N being the number of samples or tasks labelled. Similarly, Bucarelli et al. (2023) solve an optimisation problem that uses pairwise annotator similarities with a guarantee on p-norm of estimation error as $\mathcal{O}(\frac{1}{\sqrt{N}})$. The methods of DS and others like Li et al. (2019b) estimate this as a by-product of their algorithm. We describe how these noise estimates can be leveraged to derive optimality certificates on real-data.

In contrast to above related literature, our work does not propose a new aggregation method but instead studies the conditions under which the simplest aggregation method optimally recovers labels. A relevant work to our discussion is that of Nitzan and Paroush (1981) that establishes the optimality of a suitably weighted majority vote. A special case of their method is that of uniform class distribution with equally reliable workers[1] where it reduces to MV aligning with our finding for this scenario. However, the equivalence of MV's label recovery to that of oMAP applies to a larger more general range of parameters that we establish in our work.

Notwithstanding some differences like the distinct but symmetric worker noise of Nitzan and Paroush (1981) as against non-symmetric shared noise that we study, the assumption of access to an oracle renders it impractical for real-world use. We address this shortcoming in our work, describing a method to verify with high probability but without access to an oracle if a given parameter configuration is in the regime where MV is optimal. This result is central to practical application of our theory that is validated in our experiments. For all such configurations, label estimate for each sample from MV is guaranteed to match that of oMAP with no incentive for exploring more complex aggregation alternatives.

[1]See Corollary 2b in (Nitzan and Paroush, 1981).

3 GAP IN MV AND MAP METHODS

Notation. The number of tasks, annotators, and classes are represented by N, H, and C, respectively. ν denotes class distribution with subscript i when referring to class with label i, *i.e.*, $\nu_i \coloneqq \mathbb{P}(y=i)$ (Barber, 2012). The noise transition matrix for annotator a, $T^a_{ij} \coloneqq \mathbb{P}(x^a_k = j | y_k = i)$ is the probability of incorrectly assigning label j for the task x^a_k that has true label $y_k = i$. Let x denote the worker annotations for a task with true label y of which y_φ is an estimate from method φ. Correspondingly, $(y_{\text{MV}})_k, (y_{\text{MAP}})_k$ refer to the label obtained through majority vote and maximum *a posteriori* aggregation respectively for task k, *i.e.*,

$$(y_{\text{MV}})_k = \underset{c\in\{1,\dots,C\}}{\operatorname{argmax}} \sum_{h=1}^{H} \mathbb{1}\left[x^h_k = c\right],$$

$$(y_{\text{MAP}})_k = \underset{c\in\{1,\dots,C\}}{\operatorname{argmax}} \mathbb{P}(y_k = c \mid x^{h_1}_k, x^{h_2}_k, \dots, x^{h_H}_k).$$

Oracle MAP refers to the case where both ν and the true T are known while estimated MAP uses the estimate $\widetilde{T}$ and $\tilde{\nu}$.

Problem statement. We assume binary classification tasks with labels $y \in \{0, 1\}$ (Patrini et al., 2017; Karger et al., 2014; Chen et al., 2021). Given a set of task annotations $\{x_1, x_2 \dots x_N\}$, with $x_k \in \{0,1\}^H$, we characterize the parameter space (ν, T) where the gap in probabilities of MV and oracle MAP of recovering the true labels $\{y_1, y_2 \dots y_N\}$ *i.e.*, the quantity $(\mathbb{P}(y_{\text{MV}}=y) - \mathbb{P}(y_{\text{oMAP}}=y))$ vanishes. We choose oMAP because it establishes the lower bound for best achievable error rate given true (ν, T) and is the minimizer of the expected error rate under the 0-1 loss (Li and Yu, 2014), see Proposition 3.1 below.

Proposition 3.1. *The oracle MAP estimator minimizes the expected* 0-1 *loss.*

Proof. Expected 0-1 loss is $\mathbb{E}[\mathcal{L}_{0\text{-}1}(y = c, y_{\text{oMAP}} = \hat{c}|x, \theta)] = \mathbb{E}[\mathbb{I}(c \neq \hat{c}|x, \theta)] = \sum_{c\in C} \mathbb{I}(c \neq \hat{c})\mathbb{P}(c|x, \theta)$
$= 1 - \sum_{c\in C}(c = \hat{c})\mathbb{P}(c|x, \theta) = 1 - \mathbb{P}(\hat{c} \mid x, \theta)$.

Hence, the minimizer of the 0-1 loss is the same as the maximizer of the oracle posterior $\mathbb{P}(\hat{c}|x)$ derived from true parameters $\theta = (T, \nu)$ and observed labels x. □

Correspondingly, given that oracle MAP has the lowest achievable expected 0-1 loss, we use it as the benchmark for MV in quantifying its effectiveness in recovering the true labels. We note that our study, however, will focus not on expected loss but on the probability of MV recovering the true label for each sample or instance relative to that of oMAP *i.e.*,

$$\mathbb{P}(y_{\text{MV}} = y) = \mathbb{P}(y_{\text{oMAP}} = y). \tag{1}$$

The noise introduced in the annotations by workers is described through a transition matrix, $T = \left(\begin{smallmatrix} T_{00} & T_{01} \\ T_{10} & T_{11} \end{smallmatrix}\right)$. We begin with the assumption that T is identical for all annotators and relax it later in Section 3.4. A one-parameter noise model has symmetric error rate for the two classes with $T_{00} = T_{11}$ while a two-parameter model admits distinct error rates, *i.e.*, $T_{00} \neq T_{11}$. Eq. (1) can be rewritten by observing that worker's votes can be modelled using a suitably defined Binomial distribution with parameters (H, T) (see Lemmas 3.1 and 3.2). Similarly, T^{φ} captures the class-confusion matrix over the aggregated labels recovered from applying the method φ that are defined as follows for T^{MV} and T^{oMAP}.

Lemma 3.1 (Noise transition matrix T^{MV}, also Lemma 2.1 (Wei et al., 2023))**.** *Assume H is odd and binary label $c \in \{0, 1\}$, the noise transition matrix of MV aggregated labels is:*

$$T_{cc}^{MV} = \sum_{i=\lceil \frac{H}{2} \rceil}^{H} \binom{H}{i} T_{cc}^{i} \left(1 - T_{cc}\right)^{H-i}. \tag{2}$$

Lemma 3.2 (Noise transition matrix T^{oMAP})**.** *Assume H is odd, binary label $c \in \{0, 1\}$, then noise transition matrix of MAP aggregated labels is:*

$$T_{cc}^{oMAP} = \sum_{k=\lfloor A_c + 1 \rfloor}^{H} \binom{H}{k} T_{cc}^{k} (1 - T_{cc})^{H-k}. \tag{3}$$

with $A_c = \frac{\log \frac{\nu_{\bar{c}}}{\nu_c} + H \log \frac{T_{\bar{c}\bar{c}}}{1 - T_{cc}}}{\log \frac{T_{cc} T_{\bar{c}\bar{c}}}{(1 - T_{cc})(1 - T_{\bar{c}\bar{c}})}}$ *where $\bar{c} = 1 - c$ and $\nu_c, \nu_{\bar{c}}$ are priors for classes $c, \bar{c}$.*

Details of their derivation are deferred to Appendix A.1. Note that the lower bound of the summation in T^{MV} is from $\lceil \frac{H}{2} \rceil$. This is because MV only requires that majority of workers vote for the correct class to recover the true label. We may now rewrite Eq. (1) as follows:

$$T_{00}^{\text{MV}} + T_{11}^{\text{MV}} \frac{1 - \nu_0}{\nu_0} = T_{00}^{\text{oMAP}} + T_{11}^{\text{oMAP}} \frac{1 - \nu_0}{\nu_0}. \tag{4}$$

Plugging in expressions of T^{MV} and T^{oMAP} from Lemmas 3.1 and 3.2 into Eq. (4) and analysing the various cases that arise, we can deduce how MV and oMAP vary across the parameter space.

3.1 Characterisation for symmetric class noise

We restrict our analysis in this section to one-parameter models with $\varrho = T_{01} = T_{10}$ for ease of analysis and defer the more general two-parameter case to Section 3.2. Substituting T^{MV} and T^{oMAP} in Eq. (1) with their expressions, we get the equality conditions as a function of parameters ν and H. For all other cases that fail this condition oMAP is better than MV with a non-zero gap.

Theorem 3.3 (MV optimality criterion for one-parameter T for binary tasks)**.** *If the probability of a label flip, $\varrho = T_{01} = T_{10}$, is less than 0.5 and denoting by $\nu = \mathbb{P}(y = 0)$ the class zero distribution we have that:*

$$\mathbb{P}(y_{oMAP} = y) = \mathbb{P}(y_{MV} = y) \quad \textit{iff} \quad \varrho < \nu < 1 - \varrho. \tag{5}$$

$\varrho < 0.5$ translates to better than random labellers that is the most general case of non-adversarial workers with no other constraints (Karger et al., 2013; Li et al., 2019b; Bucarelli et al., 2023). Notice that if $\varrho < \nu < 1 - \varrho$ it also holds that $\varrho < 1 - \nu < 1 - \varrho$. The condition in this theorem can be rewritten as $\frac{\varrho}{1-\varrho} < \frac{\nu}{1-\nu} < \frac{1-\varrho}{\varrho}$. This indicates that the class imbalance ratio must fall within the range defined by the odds ratio associated with the noise rate of the dataset and its inverse. For a perfectly balanced dataset, this condition is always satisfied. However, as a dataset becomes more imbalanced, the maximum noise rate at which MV remains optimal decreases. The proof of Theorem 3.3 can be found in Appendix A.2. The condition for equality of MV and MAP is a rather simple and elegant boundary in Theorem 3.3 beyond which their probability gap begins to widen. It is interesting to note that if this bound is known to hold true for a given parameter configuration (ν, T), then MV's label estimate exactly matches that of oMAP for each individual instance, unlike Proposition 3.1 that applies on average, as is common in literature. Correspondingly, this result can be used for optimal label estimates without access to an oracle if only we knew that (ν, T) satisfies Theorem 3.3. However, this verification of the bounds itself assumes access to the true parameters and we show in Section 3.3 how this can be achieved without the oracle but only estimators.

3.2 Extension to asymmetric class noise

We extend Theorem 3.3 to the more general case of $(T_{01} \neq T_{10})$ that corresponds to admitting noise models with unequal sensitivity and specificity. It translates to workers confusing class 1 to class 0 and vice-versa with different error rates. In this case, the necessary and sufficient condition for probability of MV recovering the same label as that of oMAP aggregation is provided by the following theorem:

Theorem 3.4 (MV optimality criterion for two-parameter T for binary tasks)**.** *Assuming better than random workers' reliability (Karger et al., 2013; Li et al., 2019b; Bucarelli et al., 2023), namely $T_{00}, T_{11} > 0.5$, labeling binary tasks, we have that:*

$$\mathbb{P}(y_{oMAP} = y) = \mathbb{P}(y_{MV} = y) \textit{ if and only if}$$
$$\left(\frac{\delta_c}{\delta_{1-c}}\right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}} < \frac{\nu_{1-c}}{\nu_c} < \left(\frac{\delta_c}{\delta_{1-c}}\right)^{\frac{H}{2}} \sqrt{\rho}. \tag{6}$$

where $\rho = \frac{T_{00} T_{11}}{(1 - T_{00})(1 - T_{11})}$ *and* $\delta_c = \frac{T_{cc}}{1 - T_{\bar{c}\bar{c}}}$.

The proof of this theorem can be found in Appendix A.2. The condition in Theorem 3.4 holds for ν_0 if and only if it holds for ν_1. Therefore, it is sufficient to verify the condition for just one of them. We emphasize that for $T_{01} = T_{10}$ we recover the condition of Theorem 3.3 and the inferences from Section 3.1 generalize to this case.

3.3 Verification without an oracle

Theorems 3.3 and 3.4 characterise the optimality criterion for MV under the one and two-coin case respectively. However, they require true T and ν for verifiability, both of which are usually unknown. Nevertheless, with only their estimates, we will still be able to verify the bounds with high confidence. We define $f(T) = \left(\frac{\delta_c}{\delta_{1-c}}\right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}}$, $h(T) = \left(\frac{\delta_c}{\delta_{1-c}}\right)^{\frac{H}{2}} \sqrt{\rho}$ and $g(\nu) = \frac{\nu_{1-c}}{\nu_c}$.

Theorem 3.5. *Given an approximation of the noise transition matrix $\widetilde{T}$, such that $||T - \widetilde{T}||_2 \leq \epsilon$ holds with probability at least $1-\gamma$, we define $\tilde{\nu} = \widetilde{T}^{-1}\hat{\nu}$, where $\hat{\nu}$ is an approximation of the noisy label distribution. If ν_0 and ν_1 are so that $\eta \leq \nu_c \leq 1-\eta$, for $0 < \eta < 1$, $\frac{1}{2} < T_{cc} \leq 1-\xi$ for $0.5 < \xi < 1$ and the following inequalities hold:*

$$\begin{cases} g(\tilde{\nu}) - f(\widetilde{T}) > \psi + \epsilon\chi \\ h(\widetilde{T}) - g(\tilde{\nu}) > \psi + 4\epsilon\chi \end{cases}$$

then it follows that:

$$f(T) < g(\nu) < h(T)$$

with probability $1 - 2\gamma - 2e^{-\epsilon^2 N}$, where: $\psi = \frac{\epsilon}{\lambda_{min}(\widetilde{T})}\left[\frac{1}{\lambda_{min}(\widetilde{T})-\epsilon} + \sqrt{C}\right]\frac{1}{\min(\eta,1-\eta)^2}$ *and* $\chi = \sqrt{\max(\frac{1}{2}, \frac{H-1}{2} - H\xi)^2 + \max(\frac{1}{2}, \frac{H+1}{2} - \xi H)^2}$.

The proof of this theorem is in Appendix A.4. This bound depends on the quality of $\widetilde{T}$ and $\hat{\nu}$ via the parameter ϵ. The conditions on η and ξ describe the amount of imbalance in class distribution and the worker noise that can be tolerated for the theorem to hold. Specifically, η is lower bound on the fraction of samples from the minority class and annotators reliability should be better than random by a margin of ξ.

This theorem states that given estimates $(\hat{\nu}, \widetilde{T})$, it is feasible to determine with high probability if (ν, T) satisfies Theorem 3.4 enabling a practical method for the verification of MV's optimality. Suitable estimators $(\hat{\nu}, \widetilde{T})$ from literature with the required estimation guarantee can be chosen. For example, Bonald and Combes (2017) use agreement between worker triplets from a pool of workers with better than random average reliability with at least three informative workers. Similarly, Bucarelli et al. (2023) leverage pairwise worker agreement assuming all workers are better than random.

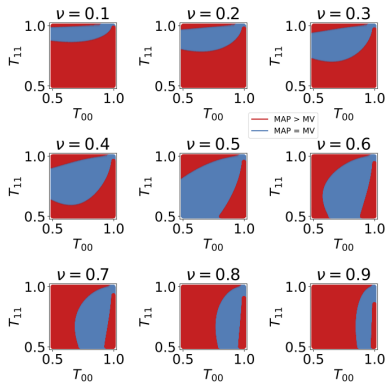


Figure 2: Illustration of Theorem 3.4 on simulations. We analyze the optimality of MV compared to oMAP verifying if the condition in Eq. (4) is satisfied for ten different ν_0 values as T_{11} and T_{00} vary between 0.5 and 1. Blue points denote where MV is equal to oMAP.

3.4 Beyond equal reliability condition

In this section, we relax the equal reliability assumption and present two different settings.

Uniformly perturbed T. In this setting annotators' noise transition matrices are not fixed but are sampled from a distribution. Specifically, for a fixed σ, the noise transition matrix of annotator h is given by:

$$T_h = \begin{bmatrix} T_{00} & T_{01} \\ T_{10} & T_{11} \end{bmatrix} + \sigma_h \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}, \ \sigma_h \sim \text{Unif}[-\sigma, \sigma]$$

We seek the conditions under which the expected performance of MV equals that of oMAP, or, more formally:

$$\mathbb{E}[\mathbb{P}(y^{\text{MV}} = y)] = \mathbb{E}[\mathbb{P}(y^{\text{oMAP}} = y)]. \tag{7}$$

We take expectation over annotator distributions, so the derived conditions will depend only on the number of annotators and not on their specific matrices. We can prove that, if σ is small enough, the conditions of Theorem 3.4 ensure Eq. (7) holds. Specifically, we require: $\sigma \leq \frac{\log \rho}{H} R_c \min(A_c - \lfloor A_c \rfloor, 1 - A_c - \lfloor A_c \rfloor)$, where $R_c = \left[\frac{2}{T_{\bar{c}\bar{c}}} + \frac{2}{T_{c\bar{c}}} + \frac{1}{T_{cc}} + \frac{1}{T_{\bar{c}c}}\right]^{-1}$ with A_c is as defined in Eq. (3).

Annotators of two categories. Here we consider two groups of annotators A and B with different reliabilities with noise transition matrices respectively T_A and T_B. *W.l.o.g.* we can assume $|A| < |B|$ and $|A| < \lceil \frac{H}{2} \rceil$. Under the assumption $\rho_A = \rho_B$, (which holds for instance, when $(T_A)_{cc} = (T_B)_{\bar{c}\bar{c}}$), we derive conditions under which MV is equivalent to oMAP. These conditions are given by:

$$\left(\frac{(\delta_B)_c}{(\delta_B)_{\bar{c}}}\right)^{\frac{H}{2}} \sqrt{\frac{1}{\rho_B}} \zeta_{B,A} < \frac{\nu_{\bar{c}}}{\nu_c} \leq \left(\frac{(\delta_B)_c}{(\delta_B)_{\bar{c}}}\right)^{\frac{H}{2}} \sqrt{\rho_B} \zeta_{B,A}$$

with $\zeta_{B,A} = \left(\frac{(\delta_B)_{\bar{c}}}{(\delta_A)_{\bar{c}}}\right)^{|A|}$ and δ_A, δ_B, ρ_A and ρ_B following the notation of Theorem 3.4. Notice that in the case of one set only we recover precisely the condition of Theorem 3.4. For additional details refer to Appendices C.1 and C.2.

3.5 Multiclass case with uniform noise

Regarding the extension to the multiclass setting, our current framework already provides the necessary ingredients to generalize the theory beyond the binary case. In particular, the results established in Theorem 3.3 naturally extend to multiclass classification.

In this setting, we have $C > 2$. Given the C-sized vector of the votes received for a sample by the various classes n_c , MV selects class $c \in C$ iff.:

$$n_c > n_j \forall j \neq c \Rightarrow n_c \geq n_j + 1,$$

while oMAP selects class $c \in C$ iff.:

$$\log(\nu_c) + \sum_{l=1}^{C} n_l log(T_{cl}) > \log(\nu_j) + \sum_{l=1}^{C} n_l log(T_{jl}) \, \forall j \neq c.$$

Uniform Noise and Uniform Priors Here, the T matrix can be parametrized by a single parameter.

$$T_{ij} = \begin{cases} 1-\rho & \text{if } i = j \\ \frac{\rho}{C-1} & \text{if } i \neq j \end{cases} \quad \nu_i = \frac{1}{C} \quad \forall i \in \{1, \ldots, C\}$$

Notice that in this case $\nu_j = \nu_c \; \forall j \neq c$. Apart from element cc and jj for all the other ℓ, $T_{c\ell} = T_{j\ell}$, so the oMAP constraint becomes:

$$\log \frac{(1-\rho)(C-1)}{\rho} \cdot (n_c - n_j) \geq 0$$

Since $\log \frac{(1-\rho)(C-1)}{\rho} \geq 1$ this becomes $n_c \geq n_j$. In this case oMAP and MV are equivalent (we are excluding ties).

Uniform Noise and Non-Uniform Priors Now $\log \nu_j \neq \log \nu_c$. The constraint becomes:

$$n_c \geq n_j + \frac{\log(\nu_j/\nu_c)}{\log((1-\rho)(C-1)/\rho)}$$

This means that MV is equivalent to oMAP when:

$$1 < \frac{\nu_j}{\nu_c} \leq \left(\frac{(1-\rho)(C-1)}{\rho}\right) \quad \forall j \neq c, \; \forall c \in \{1, \ldots, C\}$$

This condition with $C = 2$ gives the condition from Theorem 3.4.

Deriving and characterizing the equivalence conditions in the general multi-class setting represents a substantial theoretical contribution in its own right, as the constraints become considerably more complex with increasing C. While we provide the above derivation to demonstrate the extensibility of our framework, a complete characterization of the multi-class case, including the geometry of the feasible regions and optimality conditions, merits dedicated treatment and is left to future work. We emphasize that establishing these results rigorously even in the binary case represents a significant theoretical contribution.

4 EXPERIMENTS

We empirically verify our results comparing MV against oMAP for various parameter configurations. Simulations are used for validating Theorems 3.3 and 3.4 that require true data generating (ν, T). Theorem 3.5 that uses estimated $(\tilde{\nu}, \widetilde{T})$ is verified on simulated data. We compare estimates $(\tilde{\nu}, \widetilde{T})$ from different methods in the literature, namely, Dawid-Skene Dawid and Skene (1979), GLAD (Whitehill et al., 2009), MACE (Hovy et al., 2013), IWMV (Li and Yu, 2014), BWA (Li et al., 2019a), IAA (Bucarelli et al., 2023) and LA (Yang et al., 2024a). Annotator count H is used as available for all real data while it is set to $H = 3$ in all simulations unless stated otherwise. Here we report experiments with the same noise transition matrix shared by all annotators. Additional experiments that relax this assumption, as discussed in Section 3.4, are presented in Appendix C.

Simulated symmetric noise with oracle (RQ1). We illustrate Theorem 3.3 in Fig. 3a marking points in the parameter space where MV performs optimally in blue distinguishing them from those where it underperforms oMAP (in red). The plot obtained through simulations accurately reflects the theorem's conditions. Towards the left on x-axis are the low-noise regimes where MV is optimal even for skewed class balance. As the noise increases to the right, MV's optimality is restricted to fewer ν closer to 0.5. This, however, is when we use the true generating noise that is unavailable for real-world cases. So, for comparison, we make similar plots with its estimate from two methods, IWMV and IAA, in Figs. 3b and 3c respectively. Evidently, there is a distortion in the blue region where MV is optimal that grows with the estimation error. While IWMV largely retains the shape, IAA distorts it further with a higher number of sub-optimal parameters falsely marked as optimal for MV. We further inspect the actual difference in the probability scores of MV and oMAP from Fig. 3a on a heatmap in Fig. 3d. Large magnitude differences are restricted to extreme cases of skewed class distribution with noisy labels on the top and bottom corners to the right. Note that this is the worst-case scenario with $H = 3$, and the blue region

Dataset	ν	N	v_{tr}	H	v_{wr}
SP	[49.9, 50.1]	4999	203	5	5
SP_amt	[49, 51]	500	143	20	500
ZC_in	[78.3, 21.7]	2040	25	5	2125
MS	[11, 10, 10, 9, 10, 10, 10, 10, 11, 9]	700	44	4	67
CF_amt	[19, 23, 24, 31, 3]	300	110	20	55
D-Product	[87.8, 12.2]	8715	176	3	140

(a) Real data statistics.

Dataset	ν	T_{00}	T_{11}
α-Data	[20, 80]	0.7	0.7
β-Data	[50, 50]	0.9	0.9
γ-Data	[30, 70]	0.6	0.75
δ-Data	[90, 10]	0.6	0.9

(b) Synthetic data statistics.

Table 1: Data statistics with number of tasks, annotators, label class distribution ν (%), task rank (v_{tr}) and worker rank (v_{wr}) that are respectively, the number of annotators who provided labels for each sample and the number of samples each annotator was assigned to label. In Table 1b α, β-Data from one-parameter models satisfy and fail Eq. (5) respectively. γ, δ-Data from two-parameter models satisfy and fail Theorem 3.4 respectively.

where MV is optimal grows larger with the size of the worker pool, however, with the caveat that we need exponentially more workers. This is illustrated in Fig. 3e where we draw four distinct parameter configurations to plot the gap $\mathbb{P}(y_{\text{oMAP}}) - \mathbb{P}(y_{\text{MV}})$ for increasing H. (ν, T) for the orange curve satisfies Theorem 3.3 and is flat all through independent of H as it should be. MV is sub-optimal for the other three parameter configurations at lower H with a gap relative to oMAP that vanishes asymptotically as predicted by Li and Yu (2014).

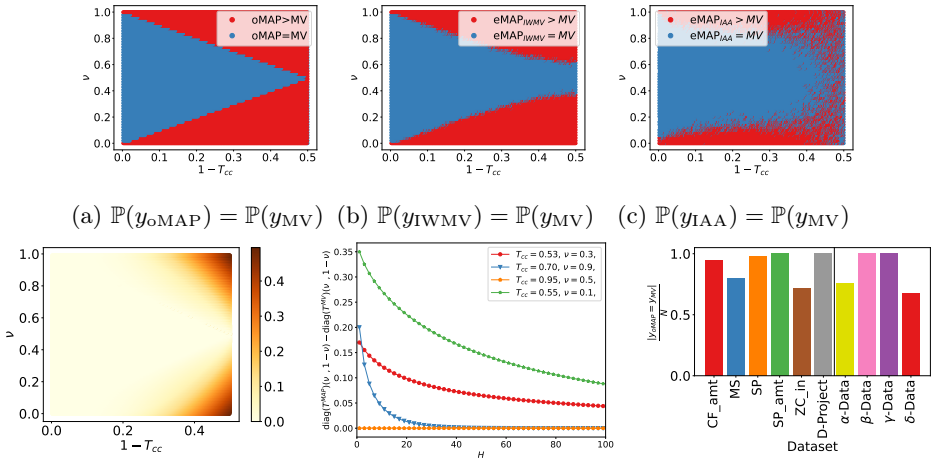
Simulated asymmetric noise (RQ1). Similar visualisations for the two-parameter model with $H = 3$ are plotted in Fig. 2 for various values of ν. Blue colored $\{(T_{00}, T_{11})\}$ pairs are MV optimal satisfying Theorem 3.4 while red regions have MV underperforming oMAP. Notice how the region where MV is optimal lying around the line $T_{00} = T_{11}$ are maximum for $\nu = 0.5$ shrinking as we move away to 0.1 or 0.9 similar to the one-parameter model in Fig. 3a.

Quantitative evaluation on simulation (RQ1). We simulate data for four different configurations, see Table 1b. Two for one-parameter models of which α-Data fails the optimality condition for MV in Eq. (5) while β-Data satisfies it. Similarly, for the two-parameter case γ-Data satisfies Theorem 3.4 while δ-Data does not. Table 2 reports the accuracy of various methods on these data. As predicted by the theory, MV recovers labels optimally for β, γ-Data while it underperforms for α, δ-Data. Fig. 3f plots the histogram with the fraction of instances for which the labels recovered using MV is exactly the same as that of oracle MAP. There is an exact match in all samples of β, γ-Data emphasizing instance optimality of the results. Note that the shortfall in oMAP from perfect accuracy score is the irreducible error and it is larger if the noise is higher. Moreover, we can notice that in the settings for which oMAP equals MV, it implies that nearly every other aggregation method performs equally well in terms of effectiveness. This again suggests that in such scenarios, employing a straightforward approach like MV, which only relies on individual sample-wise information, performs comparably to employing more complex methods that leverage data relationships across the entire dataset for aggregation.

Results on real-world data (RQ2). We use benchmark crowd-sourced datasets from Active Crowd Toolkit (Venanzi et al., 2015) and Crowdsourcing datasets repository[2] for our evaluation. Data statistics including the number of samples, annotators, classes, and their distributions are reported in Table 1a. We do not have access to oracle's true T of annotators, hence, we use a widely adopted approach (Xia et al., 2019) leveraging carefully curated gold labels or anchor points. We compute element i, j of the matrix T as the fraction of anchor points with gold label i that were marked as class j. This approximation assumes noise characteristics are consistent across workers and assigns to each one the average noise estimated from all annotations. We use this as the oracle T in experiments and report its numbers as oMAP in Table 2. Predicted labels are compared against gold labels and accuracy is reported in Table 2.

Classes are balanced in SP_amt and $T_{00} = T_{11} = 0.64$ satisfying Eq. (5) for optimality of MV. As a consequence, MV and Anchor oMAP achieve the same accuracy of 0.942 with exact match of labels for all instances as seen in histogram of Fig. 4. Data in SP also has balanced class distribution but $T_{00}(= 0.57) \neq T_{11}(= 0.65)$. Correspondingly, MV achieves an accuracy of 0.887 that is slightly lower than that of Anchor oMAP at 0.891 with a small drop from perfect match in all instances. ZenCrowd_in (ZC_in) and D-Product have skewed class balance with 80-20 and 87-13 data splits respectively. However, ZC_in with $T_{00}(= 0.58) \neq$

[2] https://dbgroup.cs.tsinghua.edu.cn/ligl/crowddata/



(a) $\mathbb{P}(y_{\text{oMAP}}) = \mathbb{P}(y_{\text{MV}})$ (b) $\mathbb{P}(y_{\text{IWMV}}) = \mathbb{P}(y_{\text{MV}})$ (c) $\mathbb{P}(y_{\text{IAA}}) = \mathbb{P}(y_{\text{MV}})$

(d) $\mathbb{P}(y_{\text{oMAP}})$–$\mathbb{P}(y_{\text{MV}})$ heatmap (e) $\mathbb{P}(y_{\text{oMAP}})-\mathbb{P}(y_{\text{MV}})$ vs H. (f) $\sum \mathbb{I}[y_{\text{oMAP}} = y_{\varphi}]$ histogram

Figure 3: Illustrations on simulated data of Theorem 3.3 comparing MV to oMAP are in (a) with similar plots for IWMV in (b) and for IAA in (c). Heatmap in (d) is of MV vs oMAP plot in (a). Curves in (e) show how that probability gap falls as H grows for four distinct parameter settings of which only orange one satisfies optimality condition for MV. Histogram (f) shows the percentage of labels aggregated using MV equal to the the ones aggregated using oMAP.

	SP ✗	SP_amt ✓	ZC_in ✗	MS *?*	CF_amt *?*	D-Product ✗	α-Data ✗	β-Data ✓	γ-Data ✓	δ-Data ✗
DS (a)	0.502	0.632	0.575	0.103	0.257	**0.940**	0.802	0.971	0.785	0.831
GLAD (b)	0.502	0.630	0.611	0.096	0.263	0.927	0.786	0.971	0.785	0.678
MACE (c)	0.503	0.632	0.631	0.097	0.270	0.695	0.786	0.971	0.785	0.678
IWMV (d)	0.904	0.942	0.750	**0.799**	0.857	0.928	0.807	0.971	0.785	0.678
BWA(e)	**0.917**	0.942	0.765	0.786	0.857	0.919	0.786	0.971	0.785	0.901
LA(f)	0.903	0.942	0.749	0.797	0.857	0.926	0.784	0.971	0.785	0.678
IAA (g)	0.882	0.942	0.744	0.703	0.857	0.905	0.786	0.971	0.785	0.678
MV (h)	0.887	0.942	0.740	0.707	0.857	0.905	0.786	0.971	0.785	0.678
oMAP	0.891^{abcfh}	$\mathbf{0.942}^{\text{abc}}$	$\mathbf{0.784}^{\text{abcdefgh}}$	0.749^{abcdefgh}	$\mathbf{0.857}^{\text{abc}}$	0.905^{abcdef}	$\mathbf{0.841}^{\text{abcdefgh}}$	0.971	0.785	$\mathbf{0.915}^{\text{abcdefgh}}$

Table 2: Accuracy (↑ is better) of different aggregation methods as measured against gold labels. Real data followed by synthetic from left to right. ✓ or ✗ indicate if Theorem 3.4 is satisfied or not, while *?* refers to multi-class datasets. Super-scripted letters indicate the methods statistically significant with respect to oMAP, determined by Wilxocon tests ($p < 0.05$) with Bonferroni correction. Best results in **bold** and second-best underlined.

$T_{11}(= 0.51)$ satisfies neither condition for MV optimality while D-Product with $T_{00}(= 0.72) \neq T_{11}(= 0.55)$ satisfies Theorem 3.4. Correspondingly, D-Product has MV accuracy on par with that Anchor oMAP, such as ZC_in. GLAD and MACE could be underperforming due to model misspecifications given that they make very specific assumptions about workers. It's important to note that Anchor oMAP isn't consistently the top-performing method, primarily because our estimation of matrix T is merely an approximation. Appendix A.1 shows the noise transition matrices estimated from the data using this method. CF_amt and MS (denoted with *?* in Table 2) are both multi-class data and, hence, the results from this work do not apply. However, for CF_amt data we empirically observe MV match the label recovery accuracy of Anchor oMAP, while it is sub-optimal for MS data.

Optimality verification (RQ2). To apply our theorems in practical settings, one can verify optimality conditions by directly plugging in estimates $(\widetilde{\nu}, \widetilde{T})$ from any method directly into Theorem 3.4. Theorem 3.5 provides guarantees on this approach, showing that when the assumptions of Theorem 3.5 are met, the correctness of the MV optimality check is assured with high probability. Empirically, we observed using estimates from three different methods (EBCC, IWMV and IAA) that even if the conditions of Theorem 3.5 are not strictly met, checking the optimality conditions by substituting estimated quantities into Theorem 3.4 often suffices. In these cases, the estimated conditions frequently match those of the unknown true quantities (see Figs. 3b and 3c). Fig. 4 illustrates this trade-off. The plot displays bars representing the average number of instances where the estimated quantities (each from different estimation method) are plugged into Theorem 3.4 and match the condition computed with Theorem 3.4

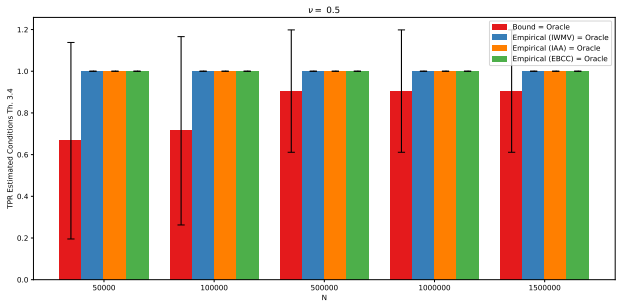


Figure 4: Non-red bars show the fraction of experiments where verification of Theorem 3.4 with estimated parameters from the candidate methods aligns with that of Theorem 3.4 using the true (ν, T), considering cases where the theorem is verified with true parameters. Red bars indicate cases where Theorem 3.5 aligns with Theorem 3.4 using true parameters. Synthetic data have various sample sizes N, and the average True Positive Rate is plotted over multiple T values, with $\nu_0 = 0.5$.

using the true, but unknown, noise rate parameters and data distribution. We consider only cases where the theorem is verified with true parameters, namely we computed the True Positive Rate (TPR). These are compared with red bars illustrating the cases where Theorem 3.5 aligns with Theorem 3.4 when calculated based on these true parameters and data distribution. Although Theorem 3.5 achieves perfect sensitivity, it often rejects MV optimality in cases where it actually holds, as the conditions it requires for confirmation are quite strict. We further show similar observation on a simulation with all four cases of true/false-positive/negative from three different estimation methods (IWMV, IAA and EBCC) on various sample sizes in Table 3 from Appendix B.2. For reference, we also report accuracy when aggregated using the three methods along with run time of each which is lowest for MV as expected.

5 DISCUSSION AND CONCLUSION

MV is often considered naive for crowdsourced labeling and, before this paper, its optimality remained unexplored. It was unknown whether it could reach the theoretically optimal oMAP bound, which uses complete knowledge of annotators' noise, and the conditions under which it would match that optimal bound. We address this problem for identical workers annotating binary tasks. We identify sufficient and necessary conditions for MV to be equivalent to oMAP. Our major finding is that when classes are well balanced, MV is robust to a wide range of worker noise levels exactly matching the oMAP bound and as the class distribution skews, worker reliability becomes increasingly important for MV to be optimal. Furthermore, we also extended our findings to some of the scenarios involving annotators with varying reliabilities.

Experiments show that verifying the conditions using estimated quantities is often sufficient, thereby making our findings applicable to real-world scenarios. Finding a suitable aggregation method currently relies on evaluating multiple hypotheses through trial-and-error on a *dev set* with *expert* generated *gold* labels that suffer from the same annotator uncertainty. It is particularly challenging for critical tasks like medical diagnostics where the accuracy of individual labels is paramount that the instance-level optimality derived in this work guarantees. This also serves as a guide for future research to focus effort on those regions of the parameter space where MV is not optimal.

Limitations Areas for improvement include extending to arbitrary noise models including adversarial workers and multiclass tasks that are theoretically challenging cases to analyse but have wider applicability.

Acknowledgments

Maria Sofia Bucarelli acknowledges partial support from the French government, through the 3IA Cote d'Azur Investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001.

Antonio Purificato and Fabrizio Silvestri acknowledge project FAIR (PE0000013), under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Bonald, T. and Combes, R. (2017). A minimax optimal algorithm for crowdsourcing. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4355–4363, Red Hook, NY, USA. Curran Associates Inc.

Boole, G. (1847). *The mathematical analysis of logic*. CreateSpace Independent Publishing Platform.

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

Chen, W., Zhu, C., and Li, M. (2023). Sample prior guided robust model learning to suppress noisy labels. In *Joint European Conference on Machine Learning*

and Knowledge Discovery in Databases, pages 3–19. Springer.

Chen, Z., Wang, H., Sun, H., Chen, P., Han, T., Liu, X., and Yang, J. (2021). Structured probabilistic end-to-end learning from crowds. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1512–1518.

Dalvi, N., Dasgupta, A., Kumar, R., and Rastogi, V. (2013). Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 285–294, New York, NY, USA. Association for Computing Machinery.

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

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27:642–669.

Ghosh, A., Kale, S., and McAfee, P. (2011). Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, page 167–176, New York, NY, USA. Association for Computing Machinery.

Gligorov, R., Hildebrand, M., Van Ossenbruggen, J., Aroyo, L., and Schreiber, G. (2013). An evaluation of labelling-game data for video retrieval. In *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings 35*, pages 50–61. Springer.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning whom to trust with MACE. In Vanderwende, L., Daumé III, H., and Kirchhoff, K., editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Huang, O., Fleisig, E., and Klein, D. (2023). Incorporating worker perspectives into MTurk annotation practices for NLP. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1028, Singapore. Association for Computational Linguistics.

Karger, D. R., Oh, S., and Shah, D. (2013). Efficient crowdsourcing for multi-class labeling. *SIGMETRICS*, 41.

Karger, D. R., Oh, S., and Shah, D. (2014). Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24.

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

Li, H., Zhao, B., and Fuxman, A. (2014). The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 165–176.

Li, Y., IP Rubinstein, B., and Cohn, T. (2019a). Truth inference at scale: A bayesian model for adjudicating highly redundant crowd annotations. In *The World Wide Web Conference*, pages 1028–1038.

Li, Y., Rubinstein, B., and Cohn, T. (2019b). Exploiting worker correlation for label aggregation in crowdsourcing. In *International conference on machine learning*, pages 3886–3895. PMLR.

Nitzan, S. and Paroush, J. (1981). The characterization of decisive weighted majority rules. *Economics Letters*, 7(2):119–124.

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

Sarhan, I. and Spruit, M. (2020). Can we survive without labelled data in nlp? transfer learning for open information extraction. *Applied Sciences*, 10(17):5758.

Shah, N. B., Balakrishnan, S., and Wainwright, M. J. (2021). A permutation-based model for crowd labeling: Optimal estimation and robustness. *IEEE Transactions on Information Theory*, 67(6):4162–4184.

Shah, N. B. and Wainwright, M. J. (2018). Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(199):1–38.

Shi, Y., Li, S.-Y., and Huang, S.-J. (2021). Learning from crowds with sparse and imbalanced annotations.

Soto, V., Suárez, A., and Martínez-Muñoz, G. (2016). An urn model for majority voting in classification ensembles. *Advances in Neural Information Processing Systems*, 29.

Spurling, M., Hu, C., Zhan, H., and Sheng, V. S. (2021). Estimating crowd-worker's reliability with interval-valued labels to improve the quality of crowdsourced work. In *2021 ieee symposium series on computational intelligence (ssci)*, pages 01–08. IEEE.

Tenzer, Y., Dror, O., Nadler, B., Bilal, E., and Kluger, Y. (2022). Crowdsourcing regression: A spectral approach. In Camps-Valls, G., Ruiz, F. J. R., and

Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5225–5242. PMLR.

Venanzi, M., Parson, O., Rogers, A., and Jennings, N. (2015). The activecrowdtoolkit: An open-source tool for benchmarking active learning algorithms for crowdsourcing research. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 3, pages 44–45.

Wang, X., Guo, Z., Zhang, Y., and Li, J. (2019). Medical image labelling and semantic understanding for clinical applications. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pages 260–270. Springer.

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

Weyl, H. (1912). Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J., and Ruvolo, P. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., and Sugiyama, M. (2019). Are anchor points really indispensable in label-noise learning? *Advances in neural information processing systems*, 32.

Yang, Y., Zhao, Z.-Q., Wu, G., Zhuo, X., Liu, Q., Bai, Q., and Li, W. (2024a). A lightweight, effective, and efficient model for label aggregation in crowdsourcing. *ACM Transactions on Knowledge Discovery from Data*, 18(4):1–27.

Yang, Y., Zhao, Z.-Q., Wu, G., Zhuo, X., Liu, Q., Bai, Q., and Li, W. (2024b). A lightweight, effective, and efficient model for label aggregation in crowdsourcing. *ACM Trans. Knowl. Discov. Data*, 18(4).

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. (2014). Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *Advances in neural information processing systems*, 27.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   [Yes]. The manuscript contains a clear description of the mathematical setting and assumptions in Section 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   [Yes]. The properties of each algorithm are analyzed in Section 4.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   [Yes]. The code can be downloaded from the Additional Material. All the external libraries are clearly specified.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   [Yes]. Each theoretical claim contains as statement with the full set of assumptions.
   (b) Complete proofs of all theoretical results.
   [Yes]. Each theoretical claim is related to a proof, which can be found in the main manuscript or in the Appendix.
   (c) Clear explanations of any assumptions.
   [Yes]. Each assumption is clearly explained before the corresponding claim.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   [Yes]. The code to reproduce all the experiment is fully available and all the istructions are clearly explained in the README.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   [Yes]. All the experimental details can be found or in Section 4 from the main manuscript or in Appendix B.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   [Yes]. All the statistics are clearly explained in the Appendix.

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
[Yes]. The description of the configuration is given in Appendix B from the Additional Material.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets.
   [Yes]. All the citations to code or datasets creators are in the text.

   (b) The license information of the assets, if applicable.
   [Yes]. The license of the code or datasets are clearly stated.

   (c) New assets either in the supplemental material or as a URL, if applicable.
   [Yes]. The code the reproduce all the experiments is released as Supplementary Material.

   (d) Information about consent from data providers/curators.
   [Not Applicable]. All the used data don't require consent from data curators.

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
   [Not Applicable]. There is no sensible content from the experimental data.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots.
   [No]. Also if this paper is about crowdsourcing, no participants were involved. Only synthetic or publicly available datasets were used.

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
   [No]. Also if this paper is about crowdsourcing, no participants were involved.

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
   [No]. Also if this paper is about crowdsourcing, no participants were involved.

A PROOFS

A.1 Proof of Lemma 3.2. Computation of T_{cc}^{MAP}

Given that the analysis remains consistent for both T_{00} and T_{11}, let us proceed by considering a general T_{cc}, where $c \in \{0, 1\}$.

$$T_{cc}^{oMAP} = \mathbb{P}(y_{oMAP} = c|y = c) = \mathbb{P}(\text{argmax}\, p_i = c) \tag{8}$$

Let us define $m \sim \text{Bin}(H, T_{cc})$ and $H - m \sim \text{Bin}(H, 1 - T_{cc})$. We recall that the element i-th of the posterior probabilities vector, $p_{i|\tilde{c}} = \nu_i \prod_{c'=1}^{C} T_{ic'}^{n_{c'|\tilde{c}}}$ where $n_{i|\tilde{c}}$ is the number of annotators that vote class i given that the true class is $\tilde{c}$.
The $\text{argmax}_{i\in\{1,0\}}$ of p_i is c if $p_{c|c} > p_{1-c|c}$ i.e.:

$$\nu_c(T_{cc})^m(1 - T_{cc})^{H-m} > \nu_{1-c}(1 - T_{1-c1-c})^m(T_{1-c1-c})^{H-m} \tag{9}$$

Where m is the number of annotators that vote class c given that the true class is c.

$$\left(\frac{T_{cc}}{1 - T_{1-c1-c}}\right)^m \left(\frac{1 - T_{cc}}{T_{1-c1-c}}\right)^{H-m} > \frac{\nu_{1-c}}{\nu_c} \tag{10}$$

$$\left(\frac{T_{00}T_{11}}{(1 - T_{00})(1 - T_{11})}\right)^m > \frac{1 - \nu}{\nu}\left(\frac{T_{11}}{1 - T_{00}}\right)^H \iff m > \frac{\log \frac{1-\nu}{\nu} + H \log \frac{T_{11}}{1-T_{00}}}{\log \frac{T_{00}T_{11}}{(1-T_{00})(1-T_{11})}}$$

By defining:

$$A_c = \frac{\log \frac{\nu_{1-c}}{\nu_c} + H \log \frac{T_{1-c,1-c}}{1-T_{cc}}}{\log \frac{T_{cc}T_{1-c,1-c}}{(1-T_{cc})(1-T_{1-c,1-c})}}$$

We obtain that, if $A_c \notin \mathbb{N}$:

$$T_{cc}^{oMAP} = \mathbb{P}(y_{oMAP} = c|y = c) = \mathbb{P}(m > A_c) = \sum_{k=\lceil A_c \rceil}^{H} \binom{H}{k} T_{cc}^k (1 - T_{cc})^{H-k}.$$

While if $A_c \in \mathbb{N}$:

$$T_{cc}^{oMAP} = \mathbb{P}(y_{oMAP} = c|y = c) = \mathbb{P}(m > A_c) = \sum_{k=A_c+1}^{H} \binom{H}{k} T_{cc}^k (1 - T_{cc})^{H-k}.$$

For the sake of simplicity, we opt to exclude the scenario where $A_c \in \mathbb{N}$.

A.2 Possible scenarios for exact matches. Proof of Theorems 3.3 and 3.4

We want to solve Eq. (1) that we rewrite here to make the reading easier:

$$\begin{aligned} &\mathbb{P}(y_{MV} = y|y = 0)\mathbb{P}(y = 0) + P(y_{MV} = y|y = 1)\mathbb{P}(y = 1) = \\ &\quad \mathbb{P}(y_{oMAP} = y|y = 0)\mathbb{P}(y = 0) + \mathbb{P}(y_{oMAP} = y|y = 1)\mathbb{P}(y = 1) \end{aligned} \tag{11}$$

We switch to Eq. (4), namely:

$$T_{00}^{\text{MV}} + T_{11}^{\text{MV}} \frac{1 - \nu}{\nu} = T_{00}^{\text{oMAP}} + T_{11}^{\text{oMAP}} \frac{1 - \nu}{\nu}.$$

The following list elaborates on the possible scenarios in for sign of the inequality in Eq. (4):

1. If $T_{00}^{oMAP} > T_{00}^{MV}$ and $T_{11}^{MAP} > T_{11}^{MV} \Rightarrow \mathbb{P}(y_{oMAP} = y) > \mathbb{P}(y_{MV} = y)$;
2. If $T_{00}^{oMAP} > T_{00}^{MV}$ and $T_{11}^{oMAP} < T_{11}^{MV}$
   $T_{00}^{oMAP} - T_{00}^{MV} > (T_{11}^{MV} - T_{11}^{MAP})\left(\frac{1-\nu}{\nu}\right) \Rightarrow \mathbb{P}(y_{MAP} = y) > \mathbb{P}(y_{MV} = y)$;
3. If $T_{11}^{MAP} > T_{11}^{MV}$ and $T_{00}^{oMAP} < T_{00}^{MV} \Rightarrow T_{00}^{MV} - T_{00}^{MV} < (T_{11}^{oMAP} - T_{11}^{MV})\left(\frac{1-\nu}{\nu}\right)$;
4. Otherwise $\mathbb{P}(y_{oMAP} = y) < \mathbb{P}(y_{MV} = y)$;

In general if $x, y \in \mathbb{R}$ $x \leq y$ it's sufficient to have $\lceil x \rceil \leq \lceil y \rceil$. If one wants the strict inequality for the ceiling, a sufficient condition is $x \leq y - 1$.

From the previous sections we know that:

$$T_{cc}^{MV} = \sum_{i=\lceil \frac{H}{2} \rceil}^{H} \binom{H}{i} T_{cc}^{i}(1 - T_{cc})^{H-i} \quad \text{and} \quad T_{cc}^{oMAP} = \sum_{k=\lceil A_c \rceil}^{H} \binom{H}{k} T_{cc}^{k}(1 - T_{cc})^{H-k}.$$

Lemma A.1. *Let T be the workers' confusion matrix, and (ν_0, ν_1) the class distribution:*

$$T_{cc}^{oMAP} > T_{cc}^{MV} \iff \frac{\nu_c}{\nu_{1-c}} \geq \left(\frac{\delta_{1-c}}{\delta_c}\right)^{\frac{H}{2}} \sqrt{\rho} \tag{12}$$

$$T_{cc}^{oMAP} = T_{cc}^{MV} \iff \left(\frac{\delta_{1-c}}{\delta_c}\right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}} < \frac{\nu_c}{\nu_{1-c}} < \left(\frac{\delta_{1-c}}{\delta_c}\right)^{\frac{H}{2}} \sqrt{\rho} \tag{13}$$

Where:

$$\delta_c = \frac{T_{cc}}{1 - T_{1-c,1-c}} \quad \rho = \frac{T_{cc}T_{1-c,1-c}}{(1 - T_{cc})(1 - T_{1-c,1-c})} \tag{14}$$

Proof of statement in Eq. (13). The condition $T_{cc}^{oMAP} = T_{cc}^{MV}$ is equivalent to $\lceil A_c \rceil = \frac{H+1}{2}$, that is satisfied when:

$$\frac{H-1}{2} < \frac{-\log \frac{\nu_c}{\nu_{1-c}} + H \log \frac{T_{1-c,1-c}}{1-T_{cc}}}{\log \frac{T_{00}T_{11}}{(1-T_{00})(1-T_{11})}} \leq \frac{H+1}{2}$$

Using the notation from Eq. (14).

We remove the denominator, by multiplying both sides for it, notice we can do it without changing the direction of the inequality since $\rho > 1$ so $\log(\rho) > 1$. We can rewrite the equation above as:

$$\frac{H-1}{2} < \frac{-\log \frac{\nu_c}{\nu_{1-c}} + H \log \delta_{1-c}}{\log \rho} \leq \frac{H+1}{2}$$

$$\Updownarrow$$

$$H\left(\frac{\log \rho}{2} - \log \delta_{1-c}\right) - \frac{\log \rho}{2} < -\log \frac{\nu_c}{\nu_{1-c}} \leq H\left(\frac{\log \rho}{2} - \log \delta_{1-c}\right) + \frac{\log \rho}{2}$$

$$\Updownarrow$$

$$H\left(\frac{\log \rho}{2} - \log \delta_{1-c}\right) - \frac{\log \rho}{2} < -\log \frac{\nu_c}{\nu_{1-c}} \leq H\left(\frac{\log \rho}{2} - \log \delta_{1-c}\right) + \frac{\log \rho}{2}.$$

Noticing that:

$$\log \sqrt{\rho} - \log \delta_{1-c} = \log \sqrt{\frac{\delta_c}{\delta_{1-c}}},$$

It follows that:

$$\left(\frac{\delta_c}{\delta_{1-c}}\right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}} < \frac{\nu_{1-c}}{\nu_c} \leq \left(\frac{\delta_c}{\delta_{1-c}}\right)^{\frac{H}{2}} \sqrt{\rho}$$

This concludes the proof of statement in Eq. (13).

In case we wanted to derive a condition for the single class distributions ν_c we could derive:

$$\left(\frac{\delta_c}{\delta_{1-c}}\right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}} + 1 < \frac{1}{\nu_c} \leq \left(\frac{\delta_c}{\delta_{1-c}}\right)^{\frac{H}{2}} \sqrt{\rho} + 1$$

From which:

$$\frac{1}{\left(\frac{\delta_c}{\delta_{1-c}}\right)^{\frac{H}{2}} \sqrt{\rho} + 1} \leq \nu_c < \frac{1}{\left(\frac{\delta_c}{\delta_{1-c}}\right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}} + 1}$$

Proof of statement in Eq. (12)

$$\begin{aligned} T_{c,c}^{oMAP} > T_{c,c}^{MV} &\iff \lceil A_c \rceil < \lceil \frac{H}{2} \rceil = \frac{H+1}{2} \iff A_c \leq \frac{H-1}{2} \\ &\iff \frac{-\log \frac{\nu_c}{\nu_{1-c}} + H \log \frac{T_{1-c,1-c}}{1-T_{cc}}}{\log \frac{T_{00}T_{11}}{(1-T_{00})(1-T_{11})}} \leq \frac{H-1}{2} \end{aligned}$$

Using the notation from Eq. (14):

$$\begin{aligned} &-\log \frac{\nu_c}{\nu_{1-c}} + H \log \delta_{1-c} \leq \frac{H}{2} \log \rho - \frac{1}{2} \log \rho \Leftrightarrow -\log \frac{\nu_c}{\nu_{1-c}} \leq H \left[-\log \delta_{1-c} + \log(\sqrt{\rho}) \right] - \log(\sqrt{\rho}) \\ &\Leftrightarrow -\log \frac{\nu_c}{\nu_{1-c}} \leq \frac{H}{2} \left[\log \left(\frac{\delta_c}{\delta_{1-c}} \right) \right] - \log \sqrt{\rho} \Leftrightarrow \frac{\nu_{1-c}}{\nu_c} \leq \left(\frac{\delta_c}{\delta_{1-c}} \right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}} \Leftrightarrow \frac{\nu_c}{\nu_{1-c}} \geq \left(\frac{\delta_{1-c}}{\delta_c} \right)^{\frac{H}{2}} \sqrt{\rho} \end{aligned}$$

Remark. *Notice that from the Theorem above it follows directly that it can never happen that both $T_{00}^{MAP} > T_{00}^{MV}$ and $T_{11}^{MAP} > T_{11}^{MV}$ indeed we have that:*

$$\frac{\nu_0}{\nu_1} \geq \left(\frac{\delta_1}{\delta_0} \right)^{\frac{H}{2}} \sqrt{\rho} \quad \textit{and} \quad \frac{\nu_1}{\nu_0} \geq \left(\frac{\delta_0}{\delta_1} \right)^{\frac{H}{2}} \sqrt{\rho} \iff \sqrt{\rho} = \frac{1}{\sqrt{\rho}} \iff T_{cc} = 1 - T_{cc},$$

which is against our assumption.

We now state a sufficient condition to have that oMAP is equivalent to MV, precisely we have that if their confusion matrices are the same, the two methods are equivalent, the following Theorem states the condition under which this happens.

Theorem A.2. *Using the notation from Eq.* (14)*, we have that if:*

$$\left(\frac{\delta_0}{\delta_1} \right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}} < \frac{\nu_1}{\nu_0} < \left(\frac{\delta_0}{\delta_1} \right)^{\frac{H}{2}} \sqrt{\rho} \tag{15}$$

Then $T_{00}^{oMAP} = T_{00}^{MV}$ and that $T_{11}^{oMAP} = T_{11}^{MV}$ from which it follows that under the conditions in described in Eq. (15) *oMAP is equivalent to MV instance wise.*

Proof. From Lemma A.1 we obtained that have $T_{00}^{oMAP} = T_{00}^{MV}$ we need

$$\begin{cases} \left(\frac{\delta_0}{\delta_1} \right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}} < \frac{\nu_1}{\nu_0} \leq \left(\frac{\delta_0}{\delta_1} \right)^{\frac{H}{2}} \sqrt{\rho} \\ \left(\frac{\delta_1}{\delta_0} \right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}} < \frac{\nu_0}{\nu_1} \leq \left(\frac{\delta_1}{\delta_0} \right)^{\frac{H}{2}} \sqrt{\rho} \end{cases} \iff \left(\frac{\delta_0}{\delta_1} \right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}} < \frac{\nu_1}{\nu_0} < \left(\frac{\delta_0}{\delta_1} \right)^{\frac{H}{2}} \sqrt{\rho} \tag{16}$$

□

Notice that the same condition described by the theorem hold for $\frac{\nu_0}{\nu_1}$.

Corollary A.2.1. *In the case the noise rate of the two classes is symmetric, i.e.* $T_{00} = T_{11}$ *we have that if:*

$$1 - T_{00} < \nu_0 < T_{00}$$

it follows that $T_{00}^{oMAP} = T_{00}^{MV}$ *and* $T_{11}^{oMAP} = T_{11}^{MV}$.

Proof. The statement follows from Theorem A.2 noticing that in the case $T_{00} = T_{11}$ we have that $\frac{\delta_0}{\delta_1} = 1$ and $\sqrt{\rho} = \frac{T_{00}}{1-T_{00}}$. From this we obtain that:

$$\frac{1-T_{00}}{T_{00}} < \frac{\nu_1}{\nu_0} < \frac{T_{00}}{1-T_{00}} \iff \frac{1}{T_{00}} < \frac{1}{\nu_0} < \frac{1}{1-T_{00}}$$

□

Corollary A.2.2. *In the case the noise rate of the two classes is symmetric, i.e.* $T_{00} = T_{11}$ *we have that* $T_{cc}^{oMAP} > T_{cc}^{MV}$ *if and only if:*

$$\nu_c \geq T_{00}$$

Proof. The statement follows substituting $\frac{\delta_0}{\delta_1} = 1$ and $\sqrt{\rho} = \frac{T_{00}}{1-T_{00}}$ in Eq. (12). □

Let us see what happens in the case where we have that T_{cc}^{oMAP} is greater than T_{cc}^{MV}.

Theorem A.3. *If:*

$$\frac{\nu_1}{\nu_0} < \left(\frac{\delta_0}{\delta_1}\right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}} \qquad or \qquad \frac{\nu_1}{\nu_0} > \left(\frac{\delta_0}{\delta_1}\right)^{\frac{H}{2}} \sqrt{\rho}$$

oMAP is strictly better than MV.

Proof. Without loss of generality we can consider the case in which $T_{00}^{oMAP} > T_{00}^{MV}$ and $T_{11}^{oMAP} < T_{11}^{MV}$, similar reasoning can be applied in the other case. MAP is better than MV if:

$$\nu T_{00}^{oMAP} + (1-\nu) T_{11}^{oMAP} > \nu T_{00}^{MV} + (1-\nu) T_{11}^{MV}$$

That is true if:

$$\sum_{k=\lceil A_0 \rceil}^{\frac{H-1}{2}} \binom{H}{k} T_{00}^k (1 - T_{00}^{H-k}) - \left(\frac{1-\nu}{\nu}\right) \sum_{k=\frac{H-1}{2}}^{\lceil A_1 \rceil} \binom{H}{k} T_{11}^k (1 - T_{11}^{H-k}) > 0$$

Denoting by $\eta = \left(\frac{1-\nu}{\nu}\right)$, that happens when:

$$\sum_{k=\lceil A_0 \rceil}^{\frac{H-1}{2}} \binom{H}{k} \left[T_{00}^k (1 - T_{00}^{H-k}) - \eta T_{11}^{H-k} (1 - T_{11}^k)\right] > 0$$

We know inspect the terms of the sum, if they are all positive than for sure the sum is positive. Namely, we want now to check if:

$$T_{00}^k (1 - T_{00}^{H-k}) - \eta T_{11}^{H-k} (1 - T_{11}^k) > 0.$$

$$\Updownarrow$$

$$\eta < \frac{T_{11}^{H-k}(1 - T_{11}^k)}{T_{00}^k (1 - T_{00}^{H-k})} = \frac{\delta_0^k}{\delta_1^{H-k}} = \frac{\rho^k}{\delta_1^H}$$

$$\Updownarrow$$

$$k > \frac{H \log \delta_1 + \log \eta}{\log \rho} = A_0.$$

The index k of the sum goes from $k = \lceil A_0 \rceil$ to $k = \frac{H-1}{2}$. Moreover we were assuming $\lceil A_0 \rceil \notin \mathbb{N}$, this means that for all terms of the sum it is satisfied that $k > A_0$.

□

Putting Theorem A.2 and Theorem A.3 together we are able to prove a necessary and sufficient condition to have oMAP equivalent to MV.

Theorem (Theorem 3.4 in the main paper. Necessary and sufficent condition for equivalence between MAP and MV)**.** *Using the notation in Eq.* (14)*, we have that oMAP is equivalent to MV instance wise if and only if:*

$$\left(\frac{\delta_0}{\delta_1}\right)^{\frac{H}{2}} \frac{1}{\sqrt{\rho}} < \frac{\nu_1}{\nu_0} < \left(\frac{\delta_0}{\delta_1}\right)^{\frac{H}{2}} \sqrt{\rho} \tag{17}$$

Proof. The Theorem is a corollary of Theorem A.3 and Theorem A.2.

□

Theorem 3.3 is a Corollary of the Theorem above and follows immediately considering that in that case $\delta_1 = \delta_0$ and $\sqrt{\rho} = \frac{T_{00}}{1-T_{00}}$.

A.3 Including estimation of T

Theorem (Gap with one-parameter $\widetilde{T}$ for binary tasks)**.** *Given the model assumptions as in Theorem 3.3 an an estimate $\widetilde{T}$ of T satisfying $\|\widetilde{T} - T\|_\infty < \epsilon$, we have with at least probability $(1-\delta)$ for any arbitrarily small δ:*

$$|\mathbb{P}(y_{eMAP} = y) - \mathbb{P}(y_{MV} = y)| \leq \frac{2\epsilon H(1+\epsilon)^{H-1}}{\nu} + \mathbb{P}(y_{oMAP} = y) - \mathbb{P}(y_{MV} = y). \tag{18}$$

Proof. Suppose we have that with probability higher than $1-\delta$ it happens that $||\widehat{T} - T||_\infty < \epsilon$. It follows that with probability at least $1-\delta$:

$$\mathbb{P}(y_{eMAP} = y) - \mathbb{P}(y_{MV} = y) = [\mathbb{P}(y_{eMAP} = y) - \mathbb{P}(y_{MAP} = y)] + [\mathbb{P}(y_{MAP} = y) - \mathbb{P}(y_{MV} = y)]$$

So:

$$\begin{aligned} |\mathbb{P}(y_{eMAP} = y) - \mathbb{P}(y_{MV} = y)| &\leq |\mathbb{P}(y_{eMAP} = y) - \mathbb{P}(y_{MAP} = y)| + |\mathbb{P}(y_{MAP} = y) - \mathbb{P}(y_{MV} = y)| \\ &\leq \frac{\epsilon}{\nu} + |\mathbb{P}(y_{MAP} = y) - \mathbb{P}(y_{MV} = y)| \end{aligned} \tag{19}$$

Indeed $\mathbb{P}(y_{MAP} = y) - \mathbb{P}(y_{eMAP} = y) = T_{00}^{MAP} - T_{11}^{MAP}\frac{1-\nu}{\nu} - T_{00}^{\widehat{MAP}} + T_{11}^{\widehat{MAP}}\frac{1-\nu}{\nu}$

It follows that:

$$|\mathbb{P}(y_{MAP} = y) - \mathbb{P}(y_{eMAP} = y)| \leq |T_{00}^{MAP} - T_{00}^{\widehat{MAP}}| + \left(\frac{1-\nu}{\nu}\right)|T_{11}^{MAP} - T_{11}^{\widehat{MAP}}|$$

We recall that:

$$T_{cc}^{\text{oMAP}} = \sum_{k=\lceil A_c \rceil}^{H} \binom{H}{k} T_{cc}^k (1-T_{cc})^{H-k} \text{ and } T_{cc}^{\text{eMAP}} = \sum_{k=\lceil \widehat{A_c} \rceil}^{H} \binom{H}{k} \widehat{T}_{cc}^k (1-\widehat{T}_{cc})^{H-k}$$

Let's first show that the error in the noise transition matrix ϵ, is small enough $\lceil A_c \rceil = \lceil \widehat{A_c} \rceil$.

$\chi = (\nu_0, 1-\nu_0)$ as the vector having as element class distributions, $||\rho|| = 1$. If we are able to recover the actual noise transition matrix up to an error of ϵ, as a consequence we are also able to recover the data distribution up to an ϵ, indeed, the class distribution after the noise introduction is $\chi' = T\chi$ and $\chi = T^{-1}\chi'$. So we can retrieve $\hat{\chi} = \hat{T}^{-1}\rho'$ and:

$$||\hat{\chi} - \chi|| = |T^{-1}\chi' - \hat{T}^{-1}\nu'| \leq ||T^{-1} - \hat{T}^{-1}|| \cdot ||\chi'|| \leq \epsilon \cdot 1.$$

If $|x-\hat{x}| \leq \varepsilon$ it follows that $1-\frac{\varepsilon}{\hat{x}} \leq \frac{x}{\hat{x}} \leq 1+\frac{\varepsilon}{\hat{x}}$, i.e. $|\frac{x}{\hat{x}}-1| \leq \frac{\varepsilon}{\hat{x}}$ We assume, wlog that $\log \frac{x}{1-x} > \log \frac{\hat{x}}{1-\hat{x}}$, namely $\frac{x(1-\hat{x})}{\hat{x}(1-x)} > 1$ if this is not the case we can swap the two terms:

$$\log \frac{x}{1-x} - \log \frac{\hat{x}}{1-\hat{x}} = \log \frac{x(1-\hat{x})}{\hat{x}(1-x)} \leq \log(1+\varepsilon) \leq \epsilon$$

Moreover:

$$\frac{T_{00}T_{11}}{(1-T_{11})(1-T_{00})} - \frac{\widehat{T}_{00}\widehat{T}_{11}}{(1-\widehat{T}_{11})(1-\widehat{T}_{00})} \leq \frac{[1-\min(T_{00},T_{11})]\varepsilon}{(1-\max(T_{11},T_{00}))^4-\varepsilon}$$

And:

$$\log \frac{T_{00}T_{11}}{(1-T_{11})(1-T_{00})} \log \frac{\widehat{T}_{00}\widehat{T}_{11}}{(1-\widehat{T}_{11})(1-\widehat{T}_{00})} \geq \left[\log \frac{T_{00}T_{11}}{(1-T_{11})(1-T_{00})}\right]^2 - \varepsilon \left[\log \frac{T_{00}T_{11}}{(1-T_{11})(1-T_{00})}\right]$$

It follows that:

$$|\hat{A}_c - A_c| \leq \frac{\log(1+\varepsilon) + \log\left(1+\varepsilon\frac{T_{1-c,1-c}}{(1-T_{cc})^2-\varepsilon}\right)}{\left[\log \frac{T_{00}T_{11}}{(1-T_{11})(1-T_{00})}\right]^2 - \varepsilon \left[\log \frac{T_{00}T_{11}}{(1-T_{11})(1-T_{00})}\right]}$$

Meaning that if ε is small enough, we can have that:

$\lceil \hat{A}_c \rceil - \lceil A_c \rceil$.

Now, using the fact that $|ab - \hat{a}\hat{b}| \leq b|a-\hat{a}| + \hat{a}|b-\hat{b}|$:

$$\begin{aligned} T_{cc}^{\text{oMAP}} - T_{cc}^{\text{eMAP}} &\leq \sum_{k=\lceil A_c \rceil}^{H} \binom{H}{k} \{(1-\widehat{T}_{cc})|T_{cc}^k - \widehat{T}_{cc}^k| + T_{cc}[(1-T_{cc})^{H-k} - (1-\widehat{T}_{cc})^{H-k}]\} \\ &= \underbrace{\sum_{k=\lceil A_c \rceil}^{H} \binom{H}{k} (1-\widehat{T}_{cc})^{H-k}|T_{cc}^k - \widehat{T}_{cc}^k|}_{\text{Term 1}} + \underbrace{\sum_{k=\lceil A_c \rceil}^{H} \binom{H}{k} T_{cc}^k[(1-T_{cc})^{H-k} - (1-\widehat{T}_{cc})^{H-k}]}_{\text{Term 2}} \end{aligned} \tag{20}$$

Now using the fact that $(x^n - y^n) = (x-y)(x^{n-1} + x^{n-2}y + x^{n-3}y^2 \dots x^2y^{n-3} + xy^{n-2} + y^{n-1})$ we can obtain that if $|T_{cc} - \widehat{T}^{cc}| < \epsilon$ it follows that:

$$||T_{cc}^k - \widehat{T_{cc}}^k| \leq |T_{cc} - \widehat{T}_{cc}| \max(T_{cc}, \widehat{T}_{cc})^{k-1}(k-1)$$

So:

$$\begin{aligned} ||T_{cc}^k - \widehat{T_{cc}}^k| &\leq |T_{cc} - \widehat{T}^{cc}| \max(T_{cc}, \widehat{T}_{cc})^{k-1}(k-1) \\ &\leq \epsilon(\widehat{T}_{cc} + \epsilon)^{k-1}(k-1) \\ &\leq \epsilon(\widehat{T}_{cc} + \epsilon)^{k-1}k \end{aligned} \tag{21}$$

And:

$$\begin{aligned} |(1-T_{cc})^{H-k} - (1-\widehat{T_{cc}})^{H-k}| &\leq |T_{cc} - \widehat{T}_{cc}| \max(1-T_{cc}, 1-\widehat{T}_{cc})^{H-k-1}(H-k-1) \\ &\leq \epsilon(H-k-1)(1-\widehat{T}_{cc} + \epsilon)^{H-k-1} \\ &\leq \epsilon(H-k)(1-\widehat{T}_{cc} + \epsilon)^{H-k-1} \end{aligned} \tag{22}$$

We can now use the Eqs. (21) and (22) To bound Term 1 and Term 2 of Eq. (20) it follows that:

$$\begin{aligned}\text{Term 1} &\leq \epsilon \sum_{k=\lceil A_c \rceil}^{H} \binom{H}{k} (1-\widehat{T}_{cc})^{H-k} (\widehat{T}_{cc}+\epsilon)^{k-1} k \\ &= \epsilon H \sum_{k=\lceil A_c \rceil - 1}^{H-1} \binom{H-1}{h} (1-\widehat{T}_{cc})^{H-1-h} (\widehat{T}_{cc}+\epsilon)^{h} \\ &\leq \epsilon H (1+\epsilon)^{H-1}\end{aligned}$$

$$\begin{aligned}\text{Term 2} &\leq \epsilon \sum_{k=\lceil A_c \rceil}^{H} \binom{H}{k} (H-k) T_{cc}^{k} (1-\widehat{T}_{cc}+\epsilon)^{H-k-1} \\ &= \sum_{k=\lceil A_c \rceil}^{H-1} \binom{H}{k} (H-k) T_{cc}^{k} (1-\widehat{T}_{cc}+\epsilon)^{H-k-1} \text{ [we are using that } H-H=0] \\ &= \epsilon H \sum_{k=\lceil A_c \rceil}^{H-1} \binom{H-1}{k} T_{cc}^{k} (1-\widehat{T}_{cc}+\epsilon)^{H-1-k} \\ &\leq \epsilon H (1+\epsilon)^{H-1}\end{aligned}$$

As a consequence $T_{cc}^{\text{oMAP}} - T_{cc}^{\text{eMAP}} \leq 2\epsilon H (1+\epsilon)^{H-1}$

From this it follows that:

$$\begin{aligned}|\mathbb{P}(y_{oMAP}=y) - \mathbb{P}(y_{eMAP}=y)| &\leq 2\epsilon H (1+\epsilon)^{H-1} + \left(\frac{1-\nu}{\nu}\right) |2\epsilon H (1+\epsilon)^{H-1} \\ &= \frac{2\epsilon H (1+\epsilon)^{H-1}}{\nu}\end{aligned}$$

□

A.4 Proof of Theorem 3.5

Lemma A.4. *Given N noisy samples and an approximation of the noise transition matrix $\tilde{T}$, such that the inequality $||T-\tilde{T}||_2 \leq \epsilon$ holds with probability at least $1-\gamma$, we define $\tilde{\nu} = \tilde{T}^{-1}\hat{\nu}_{noisy}$, where $\hat{\nu}_{noisy}$ is an approximation of the noisy label distribution. Under these conditions, it follows that, with probability $1-\alpha$, the following bound holds:*

$$||\nu - \tilde{\nu}||_2 \leq \frac{\epsilon}{\lambda_{min}(\tilde{T})} \left[\frac{1}{\lambda_{min}(T)-\epsilon} + \sqrt{C}\right].$$

where $\alpha = 1 - 2e^{-2\epsilon^2 N} - \gamma$.

Proof. If with probability at least $1-\gamma(\epsilon)$ is true that $||T-\tilde{T}|| < \epsilon$. From the definition of noise transition matrix, we have that $\nu_{\text{noise}} = T\nu$.

Now ν_{noisy} is something we can approximate looking at the distribution of the noisy data.

We call this approximation $\hat{\nu}_{\text{noisy}}$ and it is true that (we can use Dvoretzky-Kiefer-Wolfowitz inequality (Dvoretzky et al., 1956)):

$$\mathbb{P}(||\hat{\nu}_{\text{noisy}} - \nu_{\text{noisy}}||_\infty > \epsilon) \leq 2e^{-2\epsilon^2 n}.$$

Let us consider $\tilde{T}$ the approximation of T and $\hat{\nu}$ the approximation of ν, we define $\tilde{\nu} = \tilde{T}^{-1}\hat{\nu}_{\text{noisy}}$.

Since the distribution of the classes must sum up to 1, $\sum_{i=1}^{C} \nu^i_{\text{noisy}} = 1$ so $||\nu_{\text{noisy}}||_2 \leq 1$.

$$\begin{aligned}
\nu - \tilde{\nu} &= T^{-1}\nu_{\text{noisy}} - \tilde{T}^{-1}\hat{\nu}_{\text{noisy}} \\
&= T^{-1}\nu_{\text{noisy}} - \tilde{T}^{-1}\nu_{\text{noisy}} + \tilde{T}^{-1}\nu_{\text{noisy}} - \tilde{T}^{-1}\hat{\nu}_{\text{noisy}} \\
&= (T^{-1} - \tilde{T}^{-1})\nu_{\text{noisy}} + \tilde{T}^{-1}(\nu_{\text{noisy}} - \hat{\nu}_{\text{noisy}})
\end{aligned}$$

So:

$$||\nu - \tilde{\nu}||_2 \leq ||T^{-1} - \tilde{T}^{-1}||_2||\nu_{\text{noisy}}||_2 + ||\tilde{T}^{-1}||_2||\nu_{\text{noisy}} - \hat{\nu}_{\text{noisy}}||_2$$

Notice that ν_{noisy} and $\hat{\nu}_{\text{noisy}}$ are vectors in $\mathbb{R}^C$. It follows that $||\nu_{\text{noisy}} - \hat{\nu}_{\text{noisy}}||_2 \leq \sqrt{C}||\nu_{\text{noisy}} - \hat{\nu}_{\text{noisy}}||_\infty$.

Moreover, from the definition of noise transition matrix, $\tilde{T}$ is symmetric so $\tilde{T}^{-1}$ symmetric, it follows that:

$$\begin{aligned}
||\tilde{T}^{-1}||_2 = \lambda_{\max}(\tilde{T}^{-1}) &= \frac{1}{\lambda_{\min}(T)} \\
T^{-1} - \tilde{T}^{-1} &= T^{-1}(\tilde{T} - T)\tilde{T}^{-1} \\
||T^{-1} - \tilde{T}^{-1}||_2 &\leq \frac{\epsilon}{\lambda_{\min}(T)\lambda_{\min}(\tilde{T})}.
\end{aligned}$$

Using Weyl's inequality (Weyl, 1912) on perturbed eigenvalues:

$$\lambda_{\min}(\tilde{T}) \leq \lambda_{\min}(T) + \lambda_{\max}(T - \tilde{T}) = \lambda_{\min}(T) + ||T - \tilde{T}||_2$$

It follows that $\lambda_{\min}(T) > \lambda_{\min}(\tilde{T}) - \epsilon$.

Using Boole's inequality Boole (1847) we have that with probability $1 - 2e^{-2\epsilon^2 N} - \delta$:

$$||\nu - \tilde{\nu}||_2 \leq \frac{\epsilon}{\lambda_{\min}(\tilde{T})[\lambda_{\min}(\tilde{T}) - \epsilon)]} + \epsilon\frac{\sqrt{C}}{\lambda_{\min}(\tilde{T})} = \frac{\epsilon}{\lambda_{\min}(\tilde{T})}\left[\frac{1}{\lambda_{\min}(\tilde{T}) - \epsilon} + \sqrt{C}\right]$$

□

Lemma A.5. *Let the function $A(x, y)$ be defined as $A(x, y) = x^a(1 - x)^b y^c(1 - y)^d$.*
In the interval $\frac{1}{2} < x < 1 - \xi$ and $\frac{1}{2} < y < 1 - \xi$, where ξ is a positive constant smaller than 1, the function $A(x, y)$ is Lipschitz continuous with Lipschitz constant
$(1 - \xi)^{a+c-1}(\frac{1}{2})^{b+d-1}\sqrt{\max(|\frac{a-b}{2}|, |-b + \xi(a + b)|) + \max(|\frac{c-d}{2}|, |-d + \xi(c + d)|)}$.

Proof. in the interval $\frac{1}{2} < x < 1 - \xi$, $\frac{1}{2} < y < 1 - \xi$ is differentiable and its gradient is

$$\nabla A(x, y) = \begin{pmatrix} (a - ax - bx)\, x^{a-1}(1 - x)^{b-1}y^c(1 - y)^d \\ (c - cy - dy)\, x^a(1 - x)^b y^{c-1}(1 - y)^{d-1} \end{pmatrix}.$$

Now using that $\frac{1}{2} < x < 1-\xi$ and $\frac{1}{2} < y < 1-\xi$, it follows that $\xi < 1-x < \frac{1}{2}$ and $\xi < 1-y < \frac{1}{2}$. As a consequence:

$$\begin{aligned}
|x| < 1 - \xi \text{ and } |1 - x| &< \frac{1}{2} \\
-b + \xi(a + b) < a - (a + b)x &< \frac{a - b}{2} \\
-d + \xi(c + d) < c - (c + d)y &< \frac{c - d}{2} \\
|a - (a + b)x| &< \max(|\frac{a - b}{2}|, |-b + \xi(a + b)|) \\
|c - (c + d)y| &< \max(|\frac{c - d}{2}|, |-d + \xi(c + d)|)
\end{aligned}$$

It follows that:

$$|\nabla A(x,y)_x| < \max(|\frac{a-b}{2}|, |-b+\xi(a+b)|)(1-\xi)^{a+c-1}(\frac{1}{2})^{b+d-1}$$
$$|\nabla A(x,y)_y| = \max(|\frac{c-d}{2}|, |-d+\xi(c+d)|)(1-\xi)^{a+c-1}(\frac{1}{2})^{b+d-1}$$

Finally:

$$\|\nabla A(x,y)\|_2 \leq (1-\xi)^{a+c-1}(\frac{1}{2})^{b+d-1}\sqrt{\max(|\frac{a-b}{2}|, |-b+\xi(a+b)|) + \max(|\frac{c-d}{2}|, |-d+\xi(c+d)|)}$$

□

Lemma A.6. *Let the function $A(x,y)$ be defined as $A(x,y) = x^{\frac{H-1}{2}}(1-x)^{\frac{H+1}{2}}y^{-\frac{H+1}{2}}(1-y)^{-\frac{H-1}{2}}$.*

In the interval $\frac{1}{2} < x < 1-\xi$ and $\frac{1}{2} < y < 1-\xi$, where ξ is a positive constant smaller than 1, the function $A(x,y)$ is Lipschitz continuous with Lipschitz constant $(1-\xi)^{-2}(\frac{1}{2})^1\sqrt{\max(\frac{1}{2}, \frac{H+1}{2} - H\xi)^2 + \max(\frac{1}{2}, \frac{H-1}{2} - \xi H)^2}$.

The proof follows straightforwardly from Lemma A.5.

Lemma A.7. *Let the function $B(x,y)$ be defined as $B(x,y) = x^{\frac{H+1}{2}}(1-x)^{\frac{H-1}{2}}y^{-\frac{H-1}{2}}(1-y)^{-\frac{H+1}{2}}$ In the interval $\frac{1}{2} < x < 1-\xi$ and $\frac{1}{2} < y < 1-\xi$, where ξ is a positive constant smaller than 1, the function $B(x,y)$ is Lipschitz continuous with Lipschitz constant $4(1-\xi)\sqrt{\max(\frac{1}{2}, \frac{H-1}{2} - H\xi)^2 + \max(\frac{1}{2}, \frac{H+1}{2} - \xi H)^2}$.*

The proof follows straightforwardly from Lemma A.5.

Assume we are able to verify the condition of Theorem 3.4, given the definition of $f(T), g(\nu), h(T)$ from the same Theorem. When dealing with real-world data and estimated quantities experiments quantities $\tilde{T}$ and $\tilde{\nu}$, a question arises: under what conditions does the following implication hold?

$$f(\tilde{T}) < g(\tilde{\nu}) < h(\tilde{T}) \Rightarrow f(T) < g(\nu) < h(T)$$

From previous Lemmas, there exist bounds ϵ_1, ϵ_2, and ϵ_3, such that:

$$|g(\nu) - g(\tilde{\nu})| < \epsilon_2, \quad |f(T) - f(\tilde{T})| < \epsilon_1, \quad \text{and} \quad |h(T) - h(\tilde{T})| < \epsilon_3.$$

with probability at least $1-\beta$ for some β we can derive.

Therefore, if the approximations $\tilde{T}$ and $\tilde{\nu}$ are sufficiently accurate, the inequalities of the estimated conditions can be used to infer the true conditions.

If the following inequalities hold:

$$\begin{cases} g(\tilde{\nu}) - f(\tilde{T}) > \epsilon_2 + \epsilon_1 \\ h(\tilde{T}) - g(\tilde{\nu}) > \epsilon_3 + \epsilon_2 \end{cases}$$

then it follows that:

$$f(T) < g(\nu) < h(T).$$

Given the function $g_0(\nu_0) = \frac{1-\nu_0}{\nu_0}$, defined on the interval $\eta < \nu_0 < 1-\eta$, where $0 < \eta < 1$. The function is differentiable, with its derivative given by $g_0'(\nu_0) = -\frac{1}{\nu_0^2}$. Furthermore, the magnitude of the derivative is bounded as $|g_0'(\nu_0)| \leq \frac{1}{\min(\eta, 1-\eta)^2}$. Thus, if $||\tilde{\nu} - \nu||_2 \leq \epsilon_2$, it follows that $|g_0(\nu_0) - g_0(\tilde{\nu}_0)| < \frac{\epsilon_2}{\min(\eta, 1-\eta)^2}$.

By applying Lemma A.4, we obtain that, with probability at least $1 - \gamma - 2e^{-2\epsilon^2 N}$, the following inequality holds:

$$|g_0(\nu_0) - g_0(\tilde{\nu}_0)| < \frac{\epsilon}{\lambda_{\min}(\tilde{T})}\left[\frac{1}{\lambda_{\min}(\tilde{T}) - \epsilon} + \sqrt{C}\right]\frac{1}{\min(\eta, 1-\eta)^2}$$

Assuming that $0.5 < T_{cc} \leq 1-\xi$ for some positive $\xi < 1$ and that we have an approximation of the noise transition matrix $\tilde{T}$ so that with probability $1-\gamma$, $||T-\tilde{T}||_2 < \epsilon$, from Lemma A.6 and Lemma A.7 it follows that with probability $1-\gamma$ holds that:

$$|f_0(T) - f_0(\tilde{T})| \leq \epsilon(1-\xi)^{-2}\frac{1}{2}\sqrt{\max(\frac{1}{2}, \frac{H+1}{2} - H\xi)^2 + \max(\frac{1}{2}, \frac{H-1}{2} - \xi H)^2}$$

$$|h_0(T) - h_0(\tilde{T})| \leq 4\epsilon\sqrt{\max(\frac{1}{2}, \frac{H-1}{2} - H\xi)^2 + \max(\frac{1}{2}, \frac{H+1}{2} - \xi H)^2}$$

B EXPERIMENTS

B.1 Experimental Setup

All experiments were performed on an Amazon Web Services (AWS) Elastic Compute Cloud (EC2) instance with an AMD EPYC 7R32 CPU with 8 cores and 16 threads and no GPU.

All the code was written in Python 3.10.9. To implement aggregation methods as Dawid-Skene, GLAD and MACE was used the code from the Toloka library[3]. To implement BWA, EBCC, IWMV, and LA$^{\text{twopass}}$ the code was inspired by Yang et al. (2024b). For all the methods, we use the parameters presented in their respective papers.

Regarding the datasets, all of them, except D-Product, can be downloaded from the ActiveCrowdToolkit[4] page. The description of data they were aggregated can be found in Venanzi et al. (2015). The D-Product dataset was downloaded from the Crowdsourcing datasets repository[5].

B.2 Optimality conditions with estimated quantities

Table 3: Different sets of annotations with N samples are synthetically generated of which only n_e fraction are used to estimate $(\tilde{\nu}, \widetilde{T})$. $\mathbb{I}_o(\cdot)$ indicates if optimality condition in Theorem 3.4 is satisfied using real/estimated parameters. $\mathcal{A}_\varphi$ and $\mathcal{T}_\varphi$ report accuracy and run time, in seconds, for aggregation method φ. $\mathcal{T}_o$ is time to verify $\mathbb{I}_o(\tilde{\nu}, \widetilde{T})$.

N	n_e	Method	$\mathbb{I}_o(\nu,T)$	$\mathbb{I}_o(\tilde{\nu},\widetilde{T})$	$\mathcal{A}_{\text{oMAP}}$	$\mathcal{A}_{\text{MV}}$	$\mathcal{A}_{\text{EBCC}}$	$\mathcal{A}_{\text{LA}}$	$\mathcal{A}_{\text{IWMV}}$	$\mathcal{T}_{\text{MV}}$	$\mathcal{T}_{\text{EBCC}}$	$\mathcal{T}_{\text{LA}}$	$\mathcal{T}_{\text{IWMV}}$	$\mathcal{T}_{\text{o}}$
5000	0.05	IWMV	✗	✓	0.776	0.695	0.695	0.695	0.681	**0.011**	3.251	0.028	0.034	0.010
10000	0.05	IAA	✓	✓	0.810	0.810	0.810	0.810	0.810	**0.024**	6.549	0.058	0.069	0.026
20000	0.1	IAA	✗	✗	0.901	0.720	0.774	0.720	0.716	**0.047**	36.203	0.114	0.149	0.057
50000	0.1	EBCC	✓	✓	0.735	0.735	0.735	0.735	0.735	**0.121**	37.834	0.294	0.375	8.122
65000	0.1	EBCC	✓	✗	0.783	0.783	0.783	0.783	0.783	**0.158**	101.664	0.382	0.489	42.972
100000	0.05	IWMV	✓	✓	0.701	0.701	0.701	0.701	0.701	**0.239**	118.871	0.626	0.761	0.177

Table 3 shows all four cases of true/false-positive/negative from three different estimation methods (IWMV, IAA and EBCC). As expected, when both $\mathbb{I}_o(\nu,T)$ and $\mathbb{I}_o(\tilde{\nu},\tilde{T})$ are ✓, then all the methods behave as good as oMAP. When $\mathbb{I}_o(\nu,T)$ is ✗, then oMAP will always be the best solution. For reference, we also report the computational time for each aggregation algorithm and MV always remains the fastest, with the second place for Yang et al. (2024b). Additional statistics about the data can be found in Table 4, where are shown the synthetically generated noise transition matrices (T) and class distributions (ν) and their corresponding estimates ($\tilde{T}$ and $\tilde{\nu}$). From Fig. 3f from the main paper the empirical approach is working perfectly in all cases. However, this approach does not always work perfectly, as also shown in Table 3.

Fig. 5 shows the confusion matrices of the the empirical approach with the IAA and EBCC methods, respectively. It can be seen how the quality of the estimation influences the performance of the estimation method. When using the EBCC methods, the empirical approach is making really a small number of mistakes. This is also confirmed by the quality of the estimated matrices, which are really close to the original ones.

Fig. 6 shows with non-red bars the proportion of experiments where the verification of Theorem 3.4 using the estimated parameters from the candidate methods yields the same conclusion as the verification of Theorem 3.4 using the true (ν, T) values. This is calculated for the cases where the theorem is verified as true with the actual

[3] `https://crowd-kit.readthedocs.io/en/latest/`

[4] `https://orchidproject.github.io/active-crowd-toolkit/`

[5] `https://dbgroup.cs.tsinghua.edu.cn/ligl/crowddata/`

Table 4: Statistics of data from Table 3. This Table shows the noise transition matrix (T) and the distribution of classes (ν) real and estimated for each experiment by using the methods shown in Table 3.

T	$\tilde{T}$	ν	$\tilde{\nu}$
$\begin{bmatrix} 0.55 & 0.45 \\ 0.2 & 0.8 \end{bmatrix}$	$\begin{bmatrix} 0.71 & 0.29 \\ 0.29 & 0.71 \end{bmatrix}$	$\begin{bmatrix} 0.65 & 0.35 \end{bmatrix}$	$\begin{bmatrix} 0.54 & 0.46 \end{bmatrix}$
$\begin{bmatrix} 0.75 & 0.25 \\ 0.35 & 0.65 \end{bmatrix}$	$\begin{bmatrix} 0.68 & 0.32 \\ 0.32 & 0.68 \end{bmatrix}$	$\begin{bmatrix} 0.7 & 0.3 \end{bmatrix}$	$\begin{bmatrix} 0.67 & 0.33 \end{bmatrix}$
$\begin{bmatrix} 0.65 & 0.35 \\ 0.35 & 0.65 \end{bmatrix}$	$\begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.77 \end{bmatrix}$	$\begin{bmatrix} 0.9 & 0.1 \end{bmatrix}$	$\begin{bmatrix} 0.98 & 0.02 \end{bmatrix}$
$\begin{bmatrix} 0.55 & 0.45 \\ 0.20 & 0.80 \end{bmatrix}$	$\begin{bmatrix} 0.67 & 0.33 \\ 0.33 & 0.67 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.37 & 0.63 \end{bmatrix}$
$\begin{bmatrix} 0.70 & 0.30 \\ 0.30 & 0.70 \end{bmatrix}$	$\begin{bmatrix} 0.72 & 0.28 \\ 0.28 & 0.72 \end{bmatrix}$	$\begin{bmatrix} 0.68 & 0.32 \end{bmatrix}$	$\begin{bmatrix} 0.78 & 0.23 \end{bmatrix}$
$\begin{bmatrix} 0.70 & 0.30 \\ 0.45 & 0.55 \end{bmatrix}$	$\begin{bmatrix} 0.77 & 0.23 \\ 0.23 & 0.77 \end{bmatrix}$	$\begin{bmatrix} 0.68 & 0.32 \end{bmatrix}$	$\begin{bmatrix} 0.64 & 0.36 \end{bmatrix}$

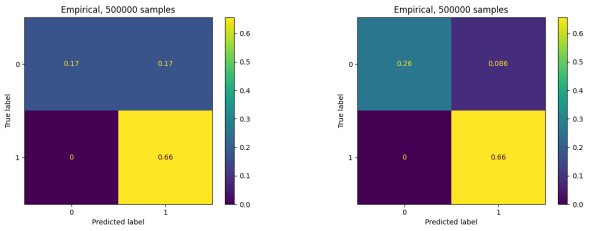
Table 5: Noise transition matrices T computed using anchor points for SP, SP_amt, ZenCrowd_in, D-Product, that are the binary classes real-world dataset we are using in our experiments.

Dataset name	anchor Map T matrix
SP	$\begin{bmatrix} 0.57 & 0.43 \\ 0.35 & 0.65 \end{bmatrix}$
SP_amt	$\begin{bmatrix} 0.64 & 0.36 \\ 0.36 & 0.64 \end{bmatrix}$
ZC_in	$\begin{bmatrix} 0.58 & 0.42 \\ 0.49 & 0.51 \end{bmatrix}$
D-Product	$\begin{bmatrix} 0.72 & 0.28 \\ 0.45 & 0.55 \end{bmatrix}$

parameters. The red bars indicate the cases where Theorem 3.5 and Theorem 3.4 with true parameters lead to the same conclusion. The experiments were conducted with synthetic data of varying sample sizes N and with values of ν_0 ranging from 0.6 to 0.9. The case with $\nu_0 = 0.5$ is in Fig. 4 from the main paper. The percentage of data used for estimation is always equal to 10%.

B.3 Real-data Noise Transition Matrices

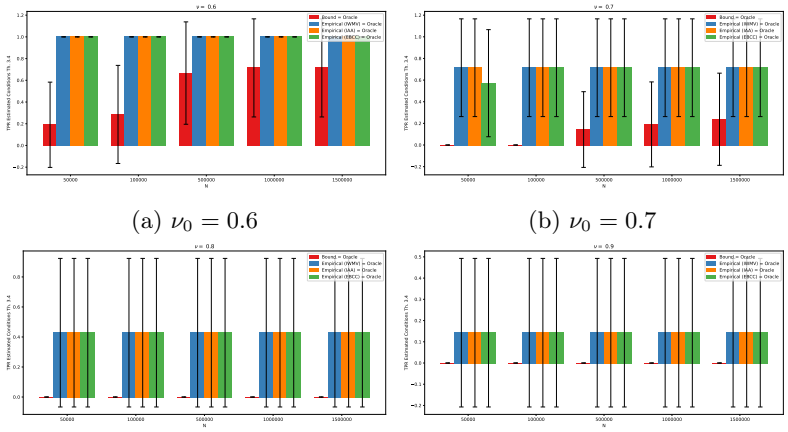
Anchor Map T matrix computed using anchor points for datasets with 2 classes (SP, SP_amt, ZenCrowd_in, D-Product) are reported in Table 5.



(a) IAA (b) EBCC

(c) IAA (d) EBCC

Figure 5: Confusion matrices describing the performance of the empirical method comparing it with the oracle results. Different T matrices and class distributions ν are used to perform the experiments. These results are based on $H = 3$ and $N = 10^6$. With this value of N, the empirical approach has good performance with the EBCC method, which slightly decrease with the IAA estimation approach.



(a) $\nu_0 = 0.6$ (b) $\nu_0 = 0.7$

(c) $\nu_0 = 0.8$ (d) $\nu_0 = 0.9$

Figure 6: Non-red bars show the fraction of experiments where verification of Theorem 3.4 with estimated parameters from the candidate methods aligns with that of Theorem 3.4 using the true (ν, T), considering cases where the theorem is verified with true parameters. Red bars indicate cases where Theorem 3.5 aligns with Theorem 3.4 using true parameters. Synthetic data have various sample sizes N, and the average True Positive Rate is plotted.

Anchor Map T matrix computed using anchor points for MS real-world dataset is reported in Eq. (23):

$$\widetilde{T}_{MS} = \begin{bmatrix} 0.41 & 0.05 & 0.05 & 0.13 & 0.07 & 0.05 & 0.09 & 0.06 & 0.05 & 0.04 \\ 0.05 & 0.41 & 0.05 & 0.04 & 0.07 & 0.05 & 0.12 & 0.08 & 0.05 & 0.07 \\ 0.08 & 0.05 & 0.31 & 0.21 & 0.05 & 0.03 & 0.09 & 0.05 & 0.05 & 0.07 \\ 0.06 & 0.07 & 0.06 & 0.39 & 0.08 & 0.08 & 0.07 & 0.03 & 0.08 & 0.08 \\ 0.04 & 0.08 & 0.06 & 0.08 & 0.41 & 0.06 & 0.07 & 0.06 & 0.07 & 0.08 \\ 0.06 & 0.06 & 0.08 & 0.23 & 0.08 & 0.26 & 0.11 & 0.06 & 0.05 & 0.04 \\ 0.15 & 0.08 & 0.05 & 0.09 & 0.09 & 0.06 & 0.31 & 0.06 & 0.05 & 0.05 \\ 0.06 & 0.05 & 0.07 & 0.07 & 0.07 & 0.03 & 0.05 & 0.49 & 0.09 & 0.03 \\ 0.05 & 0.06 & 0.04 & 0.06 & 0.07 & 0.28 & 0.08 & 0.05 & 0.28 & 0.03 \\ 0.06 & 0.06 & 0.05 & 0.10 & 0.10 & 0.08 & 0.05 & 0.03 & 0.07 & 0.41 \end{bmatrix} \tag{23}$$

Anchor Map T matrix computed using anchor points for CF_amt real-world dataset is reported in Eq. (24):

$$\widetilde{T}_{CF\_amt} = \begin{bmatrix} 0.33 & 0.21 & 0.24 & 0.08 & 0.14 \\ 0.14 & 0.31 & 0.22 & 0.15 & 0.18 \\ 0.12 & 0.20 & 0.49 & 0.10 & 0.09 \\ 0.14 & 0.22 & 0.20 & 0.29 & 0.15 \\ 0.18 & 0.22 & 0.21 & 0.12 & 0.27 \end{bmatrix} \tag{24}$$

B.4 Heatmap visualization of Theorem 3.4

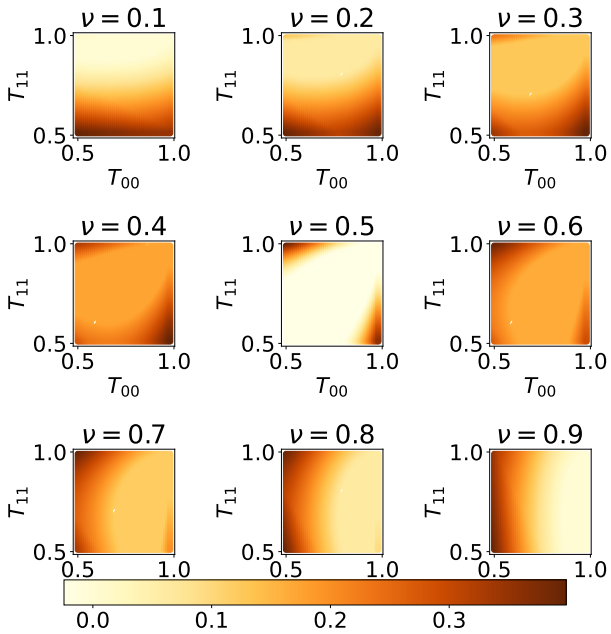


Figure 7: Heatmap visualization of the gap between oMAP and MV in the two-coin case.

The heatmap in Fig. 7 shows the gap between MV and oMAP. The experiment is obtained via simulations with different ν values ranging from 0.1 to 0.9. The plot obtained through simulations accurately reflects the theorem's conditions.

C BEYOND EQUAL RELIABILITY ASSUMPTION

C.1 Uniformly perturbed T

In this setting we imagine that instead of being fixed, the annotators noise transition matrix are sampled form a distribution so that:

$$T_h = \begin{bmatrix} T_{00} - \sigma_h & T_{01} + \sigma_h \\ T_{10} + \sigma_h & T_{11} - \sigma_h \end{bmatrix} \quad \text{with } \sigma_h \sim \text{Unif}[-\sigma, \sigma] \tag{25}$$

For a given σ.

Suppose we want the quantity to not depend of the particular pool of annotators we choose but simply on their number. What we can do in this case, having the knowledge annotators are sampled from that distribution is to use them to answer the question:

"*Given annotators sampled around that distribution, not having the need of actually them being exactly equal, neither the necessity of knowing the exact reliability of each annotator, what is the expectation, on annotator distribution, of the probability of MV match the theoretical optima upper bound of the estimation given by oMAP?*" In this case we're trying to answer the above question:

$$\mathbb{E}_\sigma[\mathbb{P}(y^{MV} = y)|\sigma_1, \dots, \sigma_h)] = \mathbb{E}_\sigma[\mathbb{P}(y^{oMAP} = y|\sigma_1, \dots, \sigma_h)]$$

Lemma C.1. *Let H be the number of annotators and T_h their noise transition matrix, samples from the distribution defined in Eq.* (25)*. For $c \in \{0, 1\}$, it holds that:*

$$\mathbb{E}[T_{cc}^{MV}] = \sum_{s=\lceil \frac{H}{2} \rceil}^{H} \binom{H}{s} T_{cc}^s (1 - T_{cc})^{H-s}$$

Proof. Denoting by $[H] = \{1, \dots, H\}$, in the setting we just described we have that:

$$\begin{aligned} T_{cc}^{MV} &= \mathbb{P}(y_{MV} = c|y = c) \\ &= \mathbb{P}(\text{at least } \lceil \frac{H}{2} \rceil \text{annotators vote class } c|y = c) \\ &= \sum_{s=\lceil \frac{H}{2} \rceil}^{H} \mathbb{P}(\text{exactly } s \text{ annotators vote class } c|y = c) \\ &= \sum_{s=\frac{H+1}{2}}^{H} \sum_{\substack{A \subseteq [H] \\ \text{s.t. } |A|=s}} \prod_{h \in A} (T_h)_{cc} \prod_{k \in H \setminus A} (T_k)_{c1-c} \end{aligned}$$

And the expected value of:

$$\mathbb{E}[T_{cc}^{MV}] = \mathbb{E}\left[\sum_{s=\frac{H+1}{2}}^{H} \sum_{\substack{A \subseteq [H] \\ \text{s.t. } |A|=s}} \prod_{h \in A} (T_h)_{cc} \prod_{k \in [H] \setminus A} (T_k)_{c1-c} \right] \tag{26}$$

The random variables T^h are independent since we assume the σ^h to be independent. So we have that:

$$\mathbb{E}[T_{cc}^{MV}] = \sum_{s=\frac{H+1}{2}}^{H} \sum_{\substack{A\subseteq[H] \\ \text{s.t. } |A|=s}} \prod_{h\in A} \mathbb{E}\left[(T_h)_{cc}\right] \prod_{k\in[H]\setminus A} \mathbb{E}\left[(T_k)_{c1-c}\right] \tag{27}$$

$$= \sum_{s=\frac{H+1}{2}}^{H} \sum_{\substack{A\subseteq[H] \\ \text{s.t. } |A|=s}} \prod_{h\in A} T_{cc} \prod_{k\in[H]\setminus A} T_{c1-c} \tag{28}$$

$$= \sum_{s=\lceil\frac{H+1}{2}\rceil}^{H} \binom{H}{s} T_{cc}^{s}(1-T_{cc})^{H-s} \tag{29}$$

□

Let us now derive the formulation for oMAP. The following lemma hold.

Lemma C.2. *Let H be the number of annotators and T_h their noise transition matrix, samples from the distribution defined in Eq.* (25)*, with :*

$$\sigma \leq \frac{\log\rho}{H}\left[\frac{2}{T_{1-c1-c}} + \frac{2}{T_{c1-c}} + \frac{1}{T_{cc}} + \frac{1}{T_{1-cc}}\right]^{-1} \min(A_c - \lfloor A_c \rfloor, 1 - A_c - \lfloor A_c \rfloor) \tag{30}$$

Where:

$$A_c = [\log\rho]^{-1}\left[\log\left(\frac{\nu_{1-c}}{\nu_c}\right) + H\log\delta_{1-c}\right]$$

for $c \in \{0,1\}$. Then, it holds that:

$$\mathbb{E}[T_{cc}^{oMAP}] = \sum_{k=\lceil A_c\rceil}^{H} \binom{H}{k} T_{cc}^{k}(1-T_{cc})^{H-k}.$$

Proof. Let us denote by C_c the set of annotators that correctly voted annotators c when the true label was c and by W_c the set of annotators that incorrectly voted class $1-c$ when the true class was c. Notice that $W_c = [H] \setminus C_c$ We recall that the element i-th of the posterior probabilities vector, $p_{i|\tilde{c}} = \nu_i \prod_{h\in C_{\tilde{c}}}(T_h)_{ic} \prod_{k\in W_{\tilde{c}}}(T_k)_{i1-c}$.
The $\text{argmax}_{i\in\{1,0\}}$ of p_i is c if $p_{c|c} > p_{1-c|c}$ i.e.:

$$
\begin{aligned}
&\nu_c \prod_{h\in C_c} (T_h)_{cc} \prod_{k\in W_c} (T_k)_{c1-c} > \nu_{1-c} \prod_{h\in C_c} (T_h)_{1-cc} \prod_{k\in[H]\setminus C_c} (T_k)_{1-c1-c} \\
&\Updownarrow \\
&\prod_{h\in C_c} \frac{(T_h)_{cc}}{(T_h)_{1-cc}} \prod_{k\in[H]\setminus C_c} \frac{(T_k)_{c1-c}}{(T_k)_{1-c1-c}} > \frac{\nu_{1-c}}{\nu_c} \\
&\Updownarrow \\
&\sum_{h\in C_c} \log\left(\frac{(T_h)_{cc}}{(T_h)_{1-cc}}\right) + \sum_{k\in[H]\setminus C_c} \log\left(\frac{(T_k)_{c1-c}}{(T_k)_{1-c1-c}}\right) > \log\left(\frac{\nu_{1-c}}{\nu_c}\right) \qquad (31) \\
&\Updownarrow \\
&\sum_{h\in H} m_h \log\left(\frac{(T_h)_{cc}}{(T_h)_{1-cc}}\right) + (1-m_h)\log\left(\frac{(T_h)_{c1-c}}{(T_h)_{1-c1-c}}\right) > \log\left(\frac{\nu_{1-c}}{\nu_c}\right) \text{ with } m_h = \mathbb{1}_{h\in C_c} \\
&\Updownarrow \\
&\sum_{h\in H} m_h \left[\log\left(\frac{(T_h)_{cc}}{(T_h)_{1-cc}}\right) - \log\left(\frac{(T_h)_{c1-c}}{(T_h)_{1-c1-c}}\right)\right] > \log\left(\frac{\nu_{1-c}}{\nu_c}\right) + \sum_{h\in H} \log\left(\frac{(T_h)_{1-c1-c}}{(T_h)_{c1-c}}\right) \\
&\Updownarrow \\
&\sum_{h\in H} m_h \left[\log\left(\frac{(T_h)_{cc}(T_h)_{1-c1-c}}{(T_h)_{1-cc}(T_h)_{c1-c}}\right)\right] > \log\left(\frac{\nu_{1-c}}{\nu_c}\right) + \sum_{h\in H} \log\left(\frac{(T_h)_{1-c1-c}}{(T_h)_{c1-c}}\right) \\
&\Updownarrow \\
&\sum_{h\in H} m_h \log\left(\rho_h\right) > \log\left(\frac{\nu_{1-c}}{\nu_c}\right) - \sum_{h\in H} \log\left((\delta_h)_c\right) \text{ with } m_h = \mathbb{1}_{h\in C_c}
\end{aligned}
\qquad (32)
$$

Denoting by $\rho_h = \frac{(T_h)_{cc}(T_h)_{1-c1-c}}{(T_h)_{1-cc}(T_h)_{c1-c}}$ and $(\delta_h)_c = \frac{(T_h)_{1-c1-c}}{(T_h)_{c1-c}}$.

The only random variable in equation (31) is the set C_c, with this writing we transferred the randomness to the variable m_h and m_h is a Bernoulli with parameter $(T_h)_{cc}$.

Let us denote by $\alpha_c = \log\left(\frac{\nu_{1-c}}{\nu_c}\right) - \sum_{h\in H} \log\left((\delta_h)_c\right)$ with $m_h = \mathbb{1}_{h\in C_c}$ and by $\beta_h = \log(\rho_h)$. We are so saying that the probability that:

$$\mathbb{P}(y^{oMAP} = y | y = c) = \mathbb{P}\left(\sum_{h\in H} m_h\beta_h > \alpha_c\right)$$

with $m_h \sim \text{Bern}((T_h)_{cc})$.

Now, we have discrete set of possible values for this sum $\sum_{h\in H} m_h\beta_h$, indeed it can take values in this set $V = \left\{\sum_{h\in S}\beta_h \;\middle|\; S\subseteq H\right\}$ and the probability of:

$$\mathbb{P}\left(\sum_{h\in H} m_h\beta_h = \sum_{h\in S}\beta_h\right) = \prod_{h\in S}(T_h)_{cc} \prod_{k\in[H]\setminus S}(1-(T_k)_{cc})$$

It follows that:

$$T_{cc}^{oMAP} = \mathbb{P}\left(\sum_{h\in H} m_h\beta_h > \alpha_c\right) = \sum_{\substack{S\subseteq[H] \\ \text{s.t. } \sum_{s\in S}\beta_s > \alpha_c}} \prod_{h\in S}(T_h)_{cc} \prod_{k\in H\setminus S}(1-(T_k)_{cc})$$

Without loss of generality, we can consider $c = 0$, we have that:

$$\beta_h = \log\left(\frac{(T_{00} - \sigma_h)(T_{11} - \sigma_h)}{(T_{10} + \sigma_h)(T_{01} + \sigma_h)}\right)$$

Substitute $T_{ij} \pm \sigma_h$ and rewrite:

$$\begin{aligned}\beta_h &= \log\left(\frac{T_{00}T_{11}}{T_{10}T_{01}} \cdot \frac{1 - \frac{\sigma_h}{T_{00}}}{1 + \frac{\sigma_h}{T_{10}}} \cdot \frac{1 - \frac{\sigma_h}{T_{11}}}{1 + \frac{\sigma_h}{T_{01}}}\right) \\ &= \log\left(\frac{T_{00}T_{11}}{T_{10}T_{01}}\right) + \log\left(\frac{1 - \frac{\sigma_h}{T_{00}}}{1 + \frac{\sigma_h}{T_{10}}} \cdot \frac{1 - \frac{\sigma_h}{T_{11}}}{1 + \frac{\sigma_h}{T_{01}}}\right).\end{aligned}$$

Similarly:

$$\log\left(\frac{T_{11} + \sigma_h}{T_{01} - \sigma_h}\right) = \log\left(\frac{T_{11}}{T_{01}}\right) + \log\left(\frac{1 + \frac{\sigma_h}{T_{11}}}{1 - \frac{\sigma_h}{T_{01}}}\right)$$

So:

$$\alpha_0 = \log\left(\frac{\nu_1}{\nu_0}\right) + H\log\left(\frac{T_{11}}{T_{01}}\right) + \sum_{h\in[H]} \log\left(\frac{1 + \frac{\sigma_h}{T_{11}}}{1 - \frac{\sigma_h}{T_{01}}}\right)$$

Substituting, we have that:

$$\sum_{sinS} \beta_s > \alpha_0$$

$$\Updownarrow$$

$$|S|\log\left(\frac{T_{00}T_{11}}{T_{10}T_{01}}\right) + \sum_{h\in S} \log\left(\frac{1 - \frac{\sigma_h}{T_{00}}}{1 + \frac{\sigma_h}{T_{10}}} \cdot \frac{1 - \frac{\sigma_h}{T_{11}}}{1 + \frac{\sigma_h}{T_{01}}}\right) > \log\left(\frac{\nu_1}{\nu_0}\right) + H\log\left(\frac{T_{11}}{T_{01}}\right) + \sum_{h\in[H]} \log\left(\frac{1 + \frac{\sigma_h}{T_{11}}}{1 - \frac{\sigma_h}{T_{01}}}\right)$$

$$\Updownarrow$$

$$|S|\log\left(\frac{T_{00}T_{11}}{T_{10}T_{01}}\right) > \log\left(\frac{\nu_1}{\nu_0}\right) + H\log\left(\frac{T_{11}}{T_{01}}\right) + \sum_{h\in[H]} \log\left(\frac{1 + \frac{\sigma_h}{T_{11}}}{1 - \frac{\sigma_h}{T_{01}}}\right) - \sum_{h\in S} \log\left(\frac{1 - \frac{\sigma_h}{T_{00}}}{1 + \frac{\sigma_h}{T_{10}}} \cdot \frac{1 - \frac{\sigma_h}{T_{11}}}{1 + \frac{\sigma_h}{T_{01}}}\right)$$

$$\Updownarrow$$

$$|S| > \log\left(\frac{T_{10}T_{01}}{T_{00}T_{11}}\right)\left[\log\left(\frac{\nu_1}{\nu_0}\right) + H\log\left(\frac{T_{11}}{T_{01}}\right)\right] + \log\left(\frac{T_{10}T_{01}}{T_{00}T_{11}}\right)\left[\sum_{h\in[H]} \log\left(\frac{1 + \frac{\sigma_h}{T_{11}}}{1 - \frac{\sigma_h}{T_{01}}}\right) - \sum_{h\in S} \log\left(\frac{1 - \frac{\sigma_h}{T_{00}}}{1 + \frac{\sigma_h}{T_{10}}} \cdot \frac{1 - \frac{\sigma_h}{T_{11}}}{1 + \frac{\sigma_h}{T_{01}}}\right)\right]$$

Let us denote by:

$$\xi = +\log\left(\frac{T_{10}T_{01}}{T_{00}T_{11}}\right)\left[\sum_{h\in[H]} \log\left(\frac{1 + \frac{\sigma_h}{T_{11}}}{1 - \frac{\sigma_h}{T_{01}}}\right) - \sum_{h\in S} \log\left(\frac{1 - \frac{\sigma_h}{T_{00}}}{1 + \frac{\sigma_h}{T_{10}}} \cdot \frac{1 - \frac{\sigma_h}{T_{11}}}{1 + \frac{\sigma_h}{T_{01}}}\right)\right]$$

We already defined in the previous section A_0 as:

$$A_0 = \left[\log\left(\frac{T_{00}T_{11}}{T_{10}T_{01}}\right)\right]^{-1}\left[\log\left(\frac{\nu_1}{\nu_0}\right) + H\log\left(\frac{T_{11}}{T_{01}}\right)\right]$$

Where $|S|$ is an integer.

so $|S| > A_0 + \xi \iff |S| \geq \lceil A_0 + \xi \rceil$

If ξ is so small that $\lceil A_c + \xi \rceil = \lceil A_c \rceil$ we have that:

$$\lceil A_0 + \xi \rceil = \lceil A_c \rceil \iff -(A_0 - \lfloor A_0 \rfloor) \leq \xi \leq 1 - (A_0 - \lfloor A_0 \rfloor) \tag{33}$$

We can always seletc σ_h so that the maximum value of the σ_h is small so that the condition in Eq. (33) is met. We assumed, $|\sigma_h| < \sigma$.

$$\begin{aligned}
|\xi| &\leq \left[\log\left(\frac{T_{00}T_{11}}{T_{01}T_{01}}\right)\right]^{-1}\left[\left|\sum_{h\in[H]}\log\left(\frac{1+\frac{\sigma_h}{T_{11}}}{1-\frac{\sigma_h}{T_{01}}}\right)\right| + \left|\sum_{h\in S}\log\left(\frac{1-\frac{\sigma_h}{T_{00}}}{1+\frac{\sigma_h}{T_{10}}}\cdot\frac{1-\frac{\sigma_h}{T_{11}}}{1+\frac{\sigma_h}{T_{01}}}\right)\right|\right] \\
&\leq H\left[\log\left(\frac{T_{00}T_{11}}{T_{01}T_{01}}\right)\right]^{-1}\left[\log\left(\frac{1+\frac{\sigma}{T_{11}}}{1-\frac{\sigma}{T_{01}}}\right) + \log\left(\frac{1+\frac{\sigma}{T_{00}}}{1-\frac{\sigma}{T_{10}}}\cdot\frac{1+\frac{\sigma}{T_{11}}}{1-\frac{\sigma}{T_{01}}}\right)\right] \\
&\leq H\left[\log\left(\frac{T_{00}T_{11}}{T_{01}T_{01}}\right)\right]^{-1}\left[2\log\left(\frac{1+\frac{\sigma}{T_{11}}}{1-\frac{\sigma}{T_{01}}}\right) + \log\left(\frac{1+\frac{\sigma}{T_{00}}}{1-\frac{\sigma}{T_{10}}}\right)\right] \\
&= H\left[\log\left(\frac{T_{00}T_{11}}{T_{01}T_{01}}\right)\right]^{-1}\left[2\log\left(1+\frac{\sigma}{T_{11}}\right) - 2\log\left(1-\frac{\sigma}{T_{01}}\right) + \log\left(1+\frac{\sigma}{T_{00}}\right) - \log\left(1-\frac{\sigma}{T_{10}}\right)\right]
\end{aligned}$$

$$H\left[\log\left(\frac{T_{00}T_{11}}{T_{01}T_{01}}\right)\right]^{-1}$$

We can use that $\frac{x}{1+x} \leq \ln(1+x) \leq x$ for all $x > -1$ to obtain that:

$$\begin{aligned}
&2\log\left(1+\frac{\sigma}{T_{11}}\right) - 2\log\left(1-\frac{\sigma}{T_{01}}\right) + \log\left(1+\frac{\sigma}{T_{00}}\right) - \log\left(1-\frac{\sigma}{T_{10}}\right) \\
&\leq 2\frac{\sigma}{T_{11}} + 2\frac{\sigma}{\sigma+T_{01}} + \frac{\sigma}{T_{00}} + \frac{\sigma}{\sigma+T_{10}} \\
&\leq 2\frac{\sigma}{T_{11}} + 2\frac{\sigma}{T_{01}} + \frac{\sigma}{T_{00}} + \frac{\sigma}{T_{10}} \\
&= \sigma\left[\frac{2}{T_{11}} + \frac{2}{T_{01}} + \frac{1}{T_{00}} + \frac{1}{T_{10}}\right]
\end{aligned}$$

Putting everything together:

$$|\xi| \leq \sigma H\left[\log\left(\frac{T_{00}T_{11}}{T_{01}T_{01}}\right)\right]^{-1}\left[\frac{2}{T_{11}} + \frac{2}{T_{01}} + \frac{1}{T_{00}} + \frac{1}{T_{10}}\right] \leq \sigma H\left[\log\left(\frac{T_{00}T_{11}}{T_{01}T_{01}}\right)\right]^{-1}$$

The following conditions implies Eq. (33):

$$\begin{aligned}
\sigma H\left[\log\left(\frac{T_{00}T_{11}}{T_{01}T_{01}}\right)\right]^{-1}\left[\frac{2}{T_{11}} + \frac{2}{T_{01}} + \frac{1}{T_{00}} + \frac{1}{T_{10}}\right] &\leq \min(A_0 - \lfloor A_0 \rfloor, 1 - A_0 - \lfloor A_0 \rfloor) \\
\sigma \leq \frac{1}{H}\left[\log\left(\frac{T_{00}T_{11}}{T_{01}T_{01}}\right)\right]\left[\frac{2}{T_{11}} + \frac{2}{T_{01}} + \frac{1}{T_{00}} + \frac{1}{T_{10}}\right]^{-1} &\min(A_0 - \lfloor A_0 \rfloor, 1 - A_0 - \lfloor A_0 \rfloor) \\
\sigma \leq \frac{\log(\rho_0)}{H}\left[\frac{2}{T_{11}} + \frac{2}{T_{01}} + \frac{1}{T_{00}} + \frac{1}{T_{10}}\right]^{-1} &\min(A_0 - \lfloor A_0 \rfloor, 1 - A_0 - \lfloor A_0 \rfloor)
\end{aligned}$$

Now, assuming condition Eq. (33) is true:

$$T_{cc}^{oMAP} = \sum_{\substack{S\subseteq[H] \\ \text{s.t. } \sum_{l\in S}\beta_l > \alpha_c}} \prod_{h\in S}(T_h)_{cc} \prod_{k\in H\setminus S}(1-(T_k)_{cc}) \tag{34}$$

$$T_{cc}^{oMAP} = \sum_{\substack{S\subseteq[H] \\ \text{s.t. } |S| > \lceil a \rceil}} \prod_{h\in S}(T_h)_{cc} \prod_{k\in H\setminus S}(1-(T_k)_{cc}) \tag{35}$$

Again, using that the T_h are independent, we obtain that:

$$\mathbb{E}[T_{cc}^{oMAP}] = \sum_{\substack{S\subseteq[H]\\ \text{s.t. } |S|>\lceil a\rceil}} \prod_{h\in S} \mathbb{E}\left[(T_h)_{cc}\right] \prod_{k\in H\setminus S} \left(1-\mathbb{E}\left[(T_k)_{cc}\right]\right) \tag{36}$$

$$= \sum_{\substack{S\subseteq[H]\\ \text{s.t. } |S|>\lceil a\rceil}} T_{cc}^{|S|}\left(1-(T_k)_{cc}\right)^{H-|S|} \tag{37}$$

$$= \sum_{s=\lceil A_c\rceil}^{H} \binom{H}{s} T_{cc}^{s}(1-T_{cc})^{H-s} \tag{38}$$

□

Thanks to the previous Lemmas we are ready to prove the following Theorem.

Theorem C.3. *Using the notation from Eq.* (14)*. Let H be the number of annotators and T_h their noise transition matrix, samples from the distribution defined in Eq.* (25)*, with:*

$$\sigma \leq \frac{\log\rho}{H}\left[\frac{2}{T_{1-c1-c}}+\frac{2}{T_{c1-c}}+\frac{1}{T_{cc}}+\frac{1}{T_{1-cc}}\right]^{-1}\min(A_c-\lfloor A_c\rfloor, 1-A_c-\lfloor A_c\rfloor) \tag{39}$$

for $c\in\{0,1\}$ and with

$$A_c = [\log\rho]^{-1}\left[\log\left(\frac{\nu_{1-c}}{\nu_c}\right)+H\log\delta_{1-c}\right].$$

we have that if:

$$\left(\frac{\delta_0}{\delta_1}\right)^{\frac{H}{2}}\frac{1}{\sqrt{\rho}} < \frac{\nu_1}{\nu_0} < \left(\frac{\delta_0}{\delta_1}\right)^{\frac{H}{2}}\sqrt{\rho} \tag{40}$$

Then $\mathbb{E}[T_{cc}^{oMAP}] = \mathbb{E}[T_{cc}^{MV}]$ and in expectations, over the distribution of the annotators oMAP is equivalent to MV instance wise.

Proof.

$$\begin{array}{c} \mathbb{E}[\mathbb{P}(y_{MV}=y)] = \mathbb{E}[\mathbb{P}(y_{oMV}=y)] \\ \Updownarrow \\ \mathbb{E}[\mathbb{P}(y_{MV}=y|y=0)\mathbb{P}(y=0)+P(y_{MV}=y|y=1)\mathbb{P}(y=1)] \\ = \mathbb{E}[\mathbb{P}(y_{oMAP}=y|y=0)\mathbb{P}(y=0)+\mathbb{P}(y_{oMAP}=y|y=1)\mathbb{P}(y=1)] \qquad \Updownarrow \\ \mathbb{E}\left[T_{00}^{\text{MV}}+T_{11}^{\text{MV}}\frac{1-\nu}{\nu}\right] = \mathbb{E}\left[T_{00}^{\text{oMAP}}+T_{11}^{\text{oMAP}}\frac{1-\nu}{\nu}\right] \end{array}$$

Using the two Lemmas proved before we can just rely on the proof of Theorem A.2 to conclude this proof. □

C.1.1 Experiments

Table 6 confirms the theoretical results obtained in Section 3.4 and demonstrated in this section. This table uses $H=3$ and $N=10000$. To obtain the results each annotator h has a different T_h matrix, obtained via a perturbation of the original T (shown in the Table). The condition on Eq. (7) is checked and then the Accuracy of oMAP and MV is presented. As expected, all the times there is the ✓ symbol the accuracy of oMAP matches the one of MV, showing the correctness of the theoretical results. The experiments are averaged across multiple runs with different seed values.

Table 6: Each annotator h has a different T_h matrix, obtained via a perturbation of the original T. The condition on Eq. (7) is checked and then the Accuracy of oMAP and MV is presented. This table uses $H = 3$ and $N = 10000$.

T_{00}	T_{11}	ν_0	Eq. (7) satisfied	oMAP Acc.	MV Acc.
0.7	0.8	0.6	✓	0.8328	0.8328
0.7	0.8	0.9	✗	0.9261	0.7961
0.55	0.55	0.6	✗	0.6099	0.5677
0.55	0.55	0.9	✗	0.8992	0.5785
0.9	0.7	0.6	✓	0.8974	0.8974
0.9	0.7	0.9	✓	0.9521	0.9521
0.6	0.6	0.6	✓	0.6499	0.6499
0.6	0.6	0.9	✓	0.9030	0.6485

C.2 Annotators of two categories

Let us now consider the case in which we have to group of annotators with different reliabilities. Specifically we have group A, with noise transition matrix T_A and group B with noise tranistion matrix T_B. Let us derive what are T_{cc}^{MV} and T_{cc}^{oMAP} in this case. Wlog we can assume $|A| < |B|$ and $|A| < \lceil \frac{H}{2} \rceil$.

$$\begin{aligned} T_{cc}^{MV} &= \mathbb{P}(\text{at least } \lceil \frac{H}{2} \rceil \text{annotators vote class } c | y = c) \\ &= \sum_{k=1}^{|A|} \sum_{l=\lceil \frac{H}{2} \rceil - k}^{|B|} \binom{|A|}{k} \binom{|B|}{l} (T_A)_{cc}^k (1-(T_A)_{cc})^{|A|-k} (T_B)_{cc}^l (1-(T_B)_{cc})^{|B|-l}. \end{aligned}$$

We define:

$\alpha_c = \log\left(\frac{\nu_{1-c}}{\nu_c}\right) + |B|\log(\delta_B)_{1-c} + |A|\log(\delta_A)_{1-c} = \log\left(\frac{\nu_{1-c}}{\nu_c}\right) + H\log(\delta_B)_{1-c} + |A|\log\frac{(\delta_A)_{1-c}}{(\delta_B)_{1-c}}$, for oMAP:

$$T_{cc}^{oMAP} = \sum_{k=1}^{|A|} \sum_{l=\lceil \frac{\alpha}{\log\rho_B} - k\frac{\log\rho_A}{\log\rho_B} \rceil}^{|B|} \binom{|A|}{k} \binom{|B|}{l} (T_A)_{cc}^k (1-(T_A)_{cc})^{|A|-k} (T_B)_{cc}^l (1-(T_B)_{cc})^{|B|-l}.$$

Theorem C.4. *Let us assume the annotators belong to two groups A and B, with different reliabilities. Specifically group A has noise transition matrix T_A and group B has noise transition matrix T_B. Denoting by $\rho_A = \frac{(T_A)_{cc}(T_A)_{1-c,1-c}}{(1-(T_A)_{cc})(1-(T_A)_{1-c,1-c})}$ and $(\delta_A)_c = \frac{(T_A)_{cc}}{1-(T_A)_{1-c,1-c}}$, if $\rho_A = \rho_B$ we have that if*

$$\left(\frac{(\delta_B)_c}{(\delta_B)_{1-c}}\right)^{\frac{H}{2}} \sqrt{\frac{1}{\rho_B}} \left(\frac{(\delta_B)_{1-c}}{(\delta_A)_{1-c}}\right)^{|A|} < \frac{\nu_{1-c}}{\nu_c} \leq \left(\frac{(\delta_B)_c}{(\delta_B)_{1-c}}\right)^{\frac{H}{2}} \sqrt{\rho_B} \left(\frac{(\delta_B)_{1-c}}{(\delta_A)_{1-c}}\right)^{|A|} \tag{41}$$

Then $T_{00}^{oMAP} = T_{00}^{MV}$ and that $T_{11}^{oMAP} = T_{11}^{MV}$ from which it follows that under the conditions in described in Eq. (15) oMAP is equivalent to MV instance wise.

Proof. Now, $T_{cc}^{MV} = T_{cc}^{oMAP} \iff \forall k \in \{1,\dots,|A|\} \lceil \frac{H}{2} \rceil - k = \lceil \frac{\alpha_c}{\log\rho_B} - k\frac{\log\rho_A}{\log\rho_B} \rceil$

$\iff \forall k \in \{1,\dots,|A|\} \frac{H+1}{2} - k - 1 < \frac{\alpha_c}{\log\rho_B} - k\frac{\log\rho_A}{\log\rho_B} \leq \frac{H+1}{2} - k$

$\iff \forall k \in \{1,\dots,|A|\} \frac{H+1}{2} - k\left(1 - \frac{\log\rho_A}{\log\rho_B}\right) - 1 < \frac{\alpha_c}{\log\rho_B} \leq \frac{H+1}{2} - k\left(1 - \frac{\log\rho_A}{\log\rho_B}\right)$

$$\iff \frac{H+1}{2} - k\left(1 - \frac{\log\rho_A}{\log\rho_B}\right) - 1 < \frac{\log\left(\frac{\nu_{1-c}}{\nu_c}\right) + H\log(\delta_B)_{1-c} + |A|\log\frac{(\delta_A)_{1-c}}{(\delta_B)_{1-c}}}{\log\rho_B} \leq \frac{H+1}{2} - k\left(1 - \frac{\log\rho_A}{\log\rho_B}\right)$$

$$\frac{(H-1)\log\rho_B}{2} - k\left(\log\frac{\rho_B}{\rho_A}\right) < \log\left(\frac{\nu_{1-c}}{\nu_c}\right) + H\log(\delta_B)_{1-c} + |A|\log\frac{(\delta_A)_{1-c}}{(\delta_B)_{1-c}} \leq \frac{(H+1)\log\rho_B}{2} - k\left(\log\frac{\rho_B}{\rho_A}\right)$$

$$\Updownarrow$$

$$\frac{(H-1)\log\rho_B}{2} - k\left(\log\frac{\rho_B}{\rho_A}\right) - H\log(\delta_B)_{1-c} - |A|\log\frac{(\delta_A)_{1-c}}{(\delta_B)_{1-c}}$$
$$< \log\left(\frac{\nu_{1-c}}{\nu_c}\right)$$
$$\leq \frac{(H+1)\log\rho_B}{2} - k\left(\log\frac{\rho_B}{\rho_A}\right) - H\log(\delta_B)_{1-c} - |A|\log\frac{(\delta_A)_{1-c}}{(\delta_B)_{1-c}}$$

$$\Updownarrow$$

$$\frac{(H-1)\log\rho_B}{2} - k\left(\log\frac{\rho_B}{\rho_A}\right) - H\log(\delta_B)_{1-c} - |A|\log\frac{(\delta_A)_{1-c}}{(\delta_B)_{1-c}}$$
$$< \log\left(\frac{\nu_{1-c}}{\nu_c}\right)$$
$$\leq \frac{(H+1)\log\rho_B}{2} - k\left(\log\frac{\rho_B}{\rho_A}\right) - H\log(\delta_B)_{1-c} - |A|\log\frac{(\delta_A)_{1-c}}{(\delta_B)_{1-c}}$$

Noticing that:

$$\log\sqrt{\rho} - \log\delta_{1-c} = \log\sqrt{\frac{\delta_c}{\delta_{1-c}}},$$

$$\Updownarrow$$

$$\frac{H}{2}\log\left(\frac{(\delta_B)_c}{(\delta_B)_{1-c}}\right) - \frac{1}{2}\log\rho_B - k\left(\log\frac{\rho_B}{\rho_A}\right) - |A|\log\frac{(\delta_A)_{1-c}}{(\delta_B)_{1-c}}$$
$$< \log\left(\frac{\nu_{1-c}}{\nu_c}\right)$$
$$\leq \frac{H}{2}\log\left(\frac{(\delta_B)_c}{(\delta_B)_{1-c}}\right) + \frac{1}{2}\log\rho_B - k\left(\log\frac{\rho_B}{\rho_A}\right) - |A|\log\frac{(\delta_A)_{1-c}}{(\delta_B)_{1-c}}$$

$$\Updownarrow$$

$$\left(\frac{(\delta_B)_c}{(\delta_B)_{1-c}}\right)^{\frac{H}{2}}\sqrt{\frac{1}{\rho_B}}\left(\frac{\rho_A}{\rho_B}\right)^k\left(\frac{(\delta_B)_{1-c}}{(\delta_A)_{1-c}}\right)^{|A|}$$
$$< \frac{\nu_{1-c}}{\nu_c}$$
$$\leq \left(\frac{(\delta_B)_c}{(\delta_B)_{1-c}}\right)^{\frac{H}{2}}\sqrt{\rho_B}\left(\frac{\rho_A}{\rho_B}\right)^k\left(\frac{(\delta_B)_{1-c}}{(\delta_A)_{1-c}}\right)^{|A|}$$

$$\Updownarrow$$

$$\left(\frac{(\delta_B)_c}{(\delta_B)_{1-c}}\right)^{\frac{H}{2}}\sqrt{\frac{1}{\rho_B}}\left(\frac{\rho_A}{\rho_B}\right)^k\left(\frac{(\delta_B)_{1-c}}{(\delta_A)_{1-c}}\right)^{|A|} < \frac{\nu_{1-c}}{\nu_c} \leq \left(\frac{(\delta_B)_c}{(\delta_B)_{1-c}}\right)^{\frac{H}{2}}\sqrt{\rho_B}\left(\frac{\rho_A}{\rho_B}\right)^k\left(\frac{(\delta_B)_{1-c}}{(\delta_A)_{1-c}}\right)^{|A|}$$

$$\Updownarrow$$

$$\left(\frac{(\delta_B)_c}{(\delta_B)_{1-c}}\right)^{\frac{H}{2}}\sqrt{\frac{1}{\rho_B}}\left(\frac{\rho_A}{\rho_B}\right)^k\left(\frac{(\delta_B)_{1-c}}{(\delta_A)_{1-c}}\right)^{|A|} < \frac{\nu_{1-c}}{\nu_c} \leq \left(\frac{(\delta_B)_c}{(\delta_B)_{1-c}}\right)^{\frac{H}{2}}\sqrt{\rho_B}\left(\frac{\rho_A}{\rho_B}\right)^k\left(\frac{(\delta_B)_{1-c}}{(\delta_A)_{1-c}}\right)^{|A|}$$

Under the hypothesis $\rho_A = \rho_B$:

$$\left(\frac{(\delta_B)_c}{(\delta_B)_{1-c}}\right)^{\frac{H}{2}} \sqrt{\frac{1}{\rho_B}} \left(\frac{(\delta_B)_{1-c}}{(\delta_A)_{1-c}}\right)^{|A|} < \frac{\nu_{1-c}}{\nu_c} \leq \left(\frac{(\delta_B)_c}{(\delta_B)_{1-c}}\right)^{\frac{H}{2}} \sqrt{\rho_B} \left(\frac{(\delta_B)_{1-c}}{(\delta_A)_{1-c}}\right)^{|A|} \tag{42}$$

□

C.2.1 Experiments

Table 7: Experiments on synthetic data to check the correctness of Eq. (42). Two class of annotators A and B (with $|A| = 3$ and $|B| = 4$) have two noise transition matrices T_A and T_B. As expected, when there is a ✓ the accuracy of MV matches the one of oMAP, showing the empirical correctness of the proposed formulation.

T_{00}^A	T_{11}^A	T_{00}^B	T_{11}^B	ν_0	Eq. (42) Satisfied	oMAP Acc.	MV Acc.
0.58	0.8	0.8	0.58	0.55	✓	0.8686	0.8686
0.58	0.8	0.8	0.58	0.65	✓	0.8567	0.8567
0.58	0.8	0.8	0.58	0.95	✗	0.9644	0.8438
0.78	0.65	0.65	0.78	0.55	✓	0.8993	0.8993
0.78	0.65	0.65	0.78	0.65	✓	0.8950	0.8950
0.78	0.65	0.65	0.78	0.95	✗	0.9658	0.9058

To verify the correctness of Eq. (42) we perform the following experiment: two classes of annotators A and B are presented, each with its own noise transition matrix T_A and T_B (diagonal values presented in the Table). Then Eq. (42) is used to verify if, with the current data, MV is the optimal solution or not. To check if the results from the condition are confirmed we present the accuracy of MV and oMAP in Table 7. When the condition is satisfied (✓) the accuracy of the two aggregation methods is the same, confirming the correctness of the proposed solution. Since one condition for this theorem is to have $\rho_A = \rho_B$, for sake of simplicity, we simply inverted the rows of the noise transition matrices A and B.