DEEP LEARNING FROM CROWDSOURCED LABELS: COUPLED CROSS-ENTROPY MINIMIZATION, IDENTIFIABILITY, AND REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Using noisy crowdsourced labels from multiple annotators, a deep learning-based end-to-end (E2E) system aims to learn the label correction mechanism and the neural classifier simultaneously. To this end, many E2E systems concatenate the neural classifier with multiple annotator-specific "label confusion" layers and co-train the two parts in a parameter-coupled manner. The formulated *coupled cross-entropy minimization* (CCEM)-type criteria are intuitive and work well in practice. Nonetheless, theoretical understanding of the CCEM criterion has been limited. The contribution of this work is twofold: First, performance guarantees of the CCEM criterion are presented. Our analysis reveals for the first time that the CCEM can indeed correctly identify the annotators' confusion characteristics and the desired "ground-truth" neural classifier under realistic conditions, e.g., when only incomplete annotator labeling and finite samples are available. Second, based on the insights learned from our analysis, two regularized variants of the CCEM are proposed. The regularization terms provably enhance the identifiability of the target model parameters in various more challenging cases. A series of synthetic and real data experiments are presented to showcase the effectiveness of our approach.

1 INTRODUCTION

The success of deep learning has escalated the demand for labeled data to an unprecedented level. Some learning tasks can easily consume millions of labeled data (Najafabadi et al., 2015; Goodfellow et al., 2016). However, acquiring data labels is a nontrivial task—it often requires a pool of annotators with sufficient domain expertise to manually label the data items. For example, the popular Microsoft COCO dataset contains 2.5 million images and around 20,000 work hours aggregated from multiple annotators were used for its category labeling (Lin et al., 2014).

Crowdsourcing is considered an important working paradigm for data labeling. In crowdsourcing platforms, e.g., Amazon Mechanical Turk (Buhrmester et al., 2011), Crowdflower (Wazny, 2017), and ClickWork (Vakharia & Lease, 2013), data items are dispatched and labeled by many annotators; the annotations are then integrated to produce reliable labels. A notable challenge is that annotator-output labels are sometimes considerably noisy. Training machine learning models using noisy labels could seriously degrade the system performance (Arpit et al., 2017; Zhang et al., 2016a). In addition, the labels provided by individual annotators are often largely incomplete, as a dataset is often divided and dispatched to different annotators.

Early crowdsourcing methods often treat annotation integration and downstream operations, e.g., classification, as separate tasks; see, (Dawid & Skene, 1979; Karger et al., 2011a; Whitehill et al., 2009; Snow et al., 2008; Welinder et al., 2010; Liu et al., 2012; Zhang et al., 2016b). This pipeline estimates the annotators' confusion parameters (e.g., the confusion matrices under the Dawid & Skene (DS) model (Dawid & Skene, 1979)) in the first stage. Then, the corrected and integrated labels along with the data are used for training the downstream tasks' classifiers. However, simultaneously learning the annotators' confusions and the classifier in an end-to-end (E2E) manner has shown substantially improved performance in practice. (Raykar et al., 2010; Khetan et al., 2018; Tanno et al., 2019; Rodrigues & Pereira, 2018; Chu et al., 2021; Guan et al., 2018; Cao et al., 2019; Li et al., 2020; Wei et al., 2022; Chen et al., 2020).

To achieve the goal of E2E-based learning under crowdsourced labels, a class of methods concatenate a "confusion layer" for each annotator to the output of a neural classifier, and jointly learn these parts via a parameter-coupled manner. This gives rise to a *coupled cross entropy minimization* (CCEM) criterion (Rodrigues & Pereira, 2018; Tanno et al., 2019; Chen et al., 2020; Chu et al., 2021; Wei et al., 2022). In essence, the CCEM criterion models the observed labels as annotators' confused outputs from the *"ground-truth" predictor* (GTP)—i.e., the classifier as if being trained using the noiseless labels and with perfect generalization—which is natural and intuitive. A notable advantage of the CCEM criteria is that they often lead to computationally convenient optimization problems, as the confusion characteristics are modeled as just additional structured layers added to the neural networks. Nonetheless, the seemingly simple CCEM criterion and its variants have shown promising performance. For example, the *crowdlayer* method is a typical CCEM approach, which has served as a widely used benchmark for E2E crowdsourcing since its proposal(Rodrigues & Pereira, 2018). In (Tanno et al., 2019), a similar CCEM-type criterion is employed, with a trace regularization added. More CCEM-like learning criteria are seen in (Chen et al., 2020; Chu et al., 2021; Wei et al., 2022), with some additional constraints and considerations.

Challenges. Apart from its empirical success, understanding of the CCEM criterion has been limited. Particularly, it is often unclear if CCEM can correctly identify the annotators' confusion characteristics (that are often modeled as "confusion matrices") and the GTP under reasonable settings (Rodrigues & Pereira, 2018; Chu et al., 2021; Chen et al., 2020; Wei et al., 2022), while identifiability stands as the key for ensured performance. The only existing identifiability result on CCEM was derived under restricted conditions, e.g., the availability of infinite samples and the assumption that there exists annotators with diagonally-dominant confusion matrices (Tanno et al., 2019). To our best knowledge, model identifiability of CCEM has not been established under realistic conditions, e.g., in the presence of incomplete annotator labeling and non-experts under finite samples.

We should note that a couple of non-CCEM approaches proposed identifiability-guaranteed solutions for E2E crowdsourcing. The work in (Khetan et al., 2018) showed identifiability under their *expectation maximization* (EM)-based learning approach, but the result is only applicable to binary classification. An information-theoretic loss-based learning approach proposed in (Cao et al., 2019) presents some identifiability results, but under the presence of a group of independent expert annotators and infinite samples. These conditions are hard to meet or verify. In addition, the computation of these methods are often more complex relative to the CCEM-based methods.

Contributions. Our contributions are as follows:

• **Identifiability Characterizations of CCEM-based E2E Crowdsourcing.** In this work, we show that the CCEM criterion can indeed provably identify the annotators' confusion matrices and the GTP up to inconsequential ambiguities under mild conditions. Specifically, we show that, if the number of annotator-labeled items is sufficiently large and some other reasonable assumptions hold, the two parts can be recovered with bounded errors. Our result is the first finite-sample identifiability result for CCEM-based crowdsourcing. Moreover, our analysis reveals that the success of CCEM does *not* rely on conditional independence among the annotators. This is favorable (as annotator independence is a stringent requirement) and surprising, since conditional independence is often used to derive the CCEM criterion; see, e.g., (Tanno et al., 2019; Rodrigues & Pereira, 2018; Chu et al., 2021).

• **Regularization Design for CCEM With Provably Enhanced Identfiability.** Based on the key insights revealed in our analysis, we propose two types of regularizations that can provide enhanced identifiability guarantees under challenging scenarios. To be specific, the first regularization term ensures that the confusion matrices and the GTP can be identified without having any expert annotators if one has sufficiently large amount of data. The second regularization term ensures identifiability when class specialists are present among annotators. These identifiability-enhanced approaches demonstrate promising label integration performance in our experiments.

Notation. The notations are summarized in the supplementary material.

2 BACKGROUND

Crowdsourcing is a core working paradigm for labeling data, as individual annotators are often not reliable enough to produce quality labels. Consider a set of N items $\{\boldsymbol{x}_n\}_{n=1}^N$. Each item belongs

to one of the K classes. Here, $\boldsymbol{x}_n \in \mathbb{R}^D$ denote the feature vector of nth data item. Let $\{y_n\}_{n=1}^N$ denote the set of *ground-truth* labels, where $y_n \in [K],\ \forall n$. The ground-truth labels $\{y_n\}_{n=1}^N$ are *unknown*. We ask M annotators to provide their labels for their assigned items. Consequently, each item $\boldsymbol{x}_n$ is labeled by a subset of annotators indexed by $\mathcal{S}_n \subseteq [M]$. Let $\{\widehat{y}_n^{(m)}\}_{(m,n)\in\mathcal{S}}$ denote the set of labels provided by all annotators, where $\mathcal{S} = \{(m,n) \mid m \in \mathcal{S}_n, n \in [N]\}$ and $\widehat{y}_n^{(m)} \in [K]$. Note that $|\mathcal{S}| \ll NM$ usually holds and that the annotators may often *incorrectly* label the items, leading to an incomplete and noisy $\{\widehat{y}_n^{(m)}\}_{(m,n)\in\mathcal{S}}$. Here, "incomplete" means that $\mathcal{S}$ does not cover all possible combinations of (m,n)—i.e., not all annotators label all items. The goal of an E2E crowdsourcing system is to train a reliable machine learning model (often a classifier) using such noisy and incomplete annotations. This is normally done via blindly estimating the confusion characteristics of the annotators from the noisy labels.

Classic Approaches. Besides the naive majority voting method, one of the most influential crowdsourcing model is by Dawid & Skene (Dawid & Skene, 1979). Dawid & Skene models the confusions of the annotators using the conditional probabilities $\mathsf{Pr}(\widehat{y}_n^{(m)}|y_n)$. Consequently, annotator m's confusion is fully characterized by a matrix defined as follows: $\boldsymbol{A}_m(k,k') := \mathsf{Pr}(\widehat{y}_n^{(m)} = k|y_n = k'), \forall k,k' \in [K]$. Dawid & Skene also proposed an EM algorithm that estimate $\boldsymbol{A}_m$'s under a naive Bayes generative model. Many crowdsourcing methods follow Dawid & Skene's modeling ideas (Ghosh et al., 2011; Dalvi et al., 2013; Karger et al., 2013; Zhang et al., 2016b; Traganitis et al., 2018). Like (Dawid & Skene, 1979), many of these methods do not exploit the data features $\boldsymbol{x}_n$ while learning the confusion matrices of the annotators. They often treat confusion estimation and classifier training as separate and sequential tasks. Such two-stage approaches may suffer from error propagation.

Problem Setting of E2E Crowdsourcing. The E2E approaches in (Rodrigues & Pereira, 2018; Tanno et al., 2019; Khetan et al., 2018) model the probability of mth annotator's response to the data item $\boldsymbol{x}_n$ as follows:

$$\mathsf{Pr}(\widehat{y}_n^{(m)} = k|\boldsymbol{x}_n) = \sum_{k'=1}^{K} \mathsf{Pr}(\widehat{y}_n^{(m)} = k|y_n = k')\mathsf{Pr}(y_n = k'|\boldsymbol{x}_n),\ k \in [K], \tag{1}$$

where the assumption that the annotator confusion is data-independent has been used to derive the right-hand side. In the above, the distribution $\mathsf{Pr}(\widehat{y}_n^{(m)}|y_n)$ models the confusions in annotator responses given the true label y_n and $\mathsf{Pr}(y_n|\boldsymbol{x}_n)$ denotes the true label distribution of the data item $\boldsymbol{x}_n$. The distribution $\mathsf{Pr}(y_n|\boldsymbol{x}_n)$ can be represented using a mapping $\boldsymbol{f} : \mathbb{R}^D \to [0,1]^K,\ [\boldsymbol{f}(\boldsymbol{x}_n)]_k = \mathsf{Pr}(y_n = k|\boldsymbol{x}_n)$. Define a probability vector $\boldsymbol{p}_n^{(m)} \in [0,1]^K$ for every m and n, such that $[\boldsymbol{p}_n^{(m)}]_k = \mathsf{Pr}(\widehat{y}_n^{(m)} = k|\boldsymbol{x}_n)$. With these notations, the following model holds:

$$\boldsymbol{p}_n^{(m)} = \boldsymbol{A}_m \boldsymbol{f}(\boldsymbol{x}_n), \forall m, n. \tag{2}$$

In practice, one does not observe $\boldsymbol{p}_n^{(m)}$. Instead, the annotations $\widehat{y}_n^{(m)}$ are categorical realizations of the distribution $\boldsymbol{p}_n^{(m)}$; i.e.,

$$\widehat{y}_n^{(m)} \sim \mathsf{Cart}(\boldsymbol{p}_n^{(m)}). \tag{3}$$

The goal of E2E crowdsourcing is to identify $\boldsymbol{A}_m$ and $\boldsymbol{f}$ in (2) from noisy labels $\{\widehat{y}_n^{(m)}\}_{(m,n)\in\mathcal{S}}$ and data $\{\boldsymbol{x}_n\}_{n=1}^N$. Learning $\boldsymbol{f}$ not only identifies the true labels for the training data (data for which the annotations have been collected), but is useful to predict the true labels for unseen test data items. In addition, learning $\boldsymbol{A}_m$'s helps quantify how reliable each annotator is.

CCEM-based Approaches and Challenges. A number of existing E2E methods are inspired by the model in (2) to formulate CCEM-based learning losses.

• *Crowdlayer* (Rodrigues & Pereira, 2018): For each $(m,n) \in \mathcal{S}$, the generative model in (2) naturally leads to a cross-entropy based learning loss, i.e., $\mathsf{CE}(\boldsymbol{A}_m \boldsymbol{f}(\boldsymbol{x}_n), \widehat{y}_n^{(m)})$. As some annotators co-label items, the components $\boldsymbol{f}(\boldsymbol{x}_n)$ associated with the co-labeled items are shared by these annotators.

With such latent component coupling and putting the cross-entropy losses for all $(m, n) \in \mathcal{S}$ together, the work in (Rodrigues & Pereira, 2018) formulate the following CCEM loss:

$$\underset{\boldsymbol{f}\in\mathcal{F},\{\boldsymbol{A}_m\in\mathbb{R}^{K\times K}\}}{\text{minimize}} \ \frac{1}{|\mathcal{S}|}\sum_{(m,n)\in\mathcal{S}} \mathsf{CE}(\mathsf{softmax}(\boldsymbol{A}_m\boldsymbol{f}(\boldsymbol{x}_n)), \widehat{y}_n^{(m)}), \tag{4}$$

where $\mathcal{F} \subseteq \{\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^K \,|\, \boldsymbol{f}(\boldsymbol{x}) \in \boldsymbol{\Delta}_K, \ \forall \boldsymbol{x}\}$ is a function class and the softmax operator is applied as the output of $\boldsymbol{A}_m\boldsymbol{f}(\boldsymbol{x}_n)$ is supposed to be a probability mass function (PMF). The method based on the formulation in (4) was called *crowdlayer* (Rodrigues & Pereira, 2018). The name comes from the fact that the confusion matrices of the "crowd" $\boldsymbol{A}_m$'s can be understood as additional layers if $\boldsymbol{f}$ is represented by a neural network. The loss function in (4) is relatively easy to optimize, as any off-the-shelf neural network training algorithms can be directly applied. This seemingly natural and simple CCEM approach demonstrated substantial performance improvement upon classic methods. However, there is no performance characterization for (4).

• *TraceReg* (Tanno et al., 2019): The work in (Tanno et al., 2019) proposed a similar loss function—with a regularization term in order to establish identifiability of the $\boldsymbol{A}_m$'s and $\boldsymbol{f}$:

$$\underset{\boldsymbol{f}\in\mathcal{F},\{\boldsymbol{A}_m\in\mathcal{A}\}}{\text{minimize}} \ \frac{1}{|\mathcal{S}|}\sum_{(m,n)\in\mathcal{S}} \mathsf{CE}(\boldsymbol{A}_m\boldsymbol{f}(\boldsymbol{x}_n), \widehat{y}_n^{(m)}) + \lambda\sum_{m=1}^{M}\mathsf{trace}(\boldsymbol{A}_m), \tag{5}$$

where $\mathcal{A}$ is the constrained set $\{\boldsymbol{A} \in \mathbb{R}^{K\times K} | \boldsymbol{A} \geq \boldsymbol{0}, \boldsymbol{1}^\top\boldsymbol{A} = \boldsymbol{1}^\top\}$. It was shown that (i) if $\boldsymbol{p}_n^{(m)}$ instead of $\widehat{y}_n^{(m)}$ is observed for every m and n and (ii) if the mean of the $\boldsymbol{A}_m$'s are diagonally dominant, then solving (5) with $\mathsf{CE}(\boldsymbol{A}_m\boldsymbol{f}(\boldsymbol{x}_n), \widehat{y}_n^{(m)})$ replaced by enforcing $\boldsymbol{p}_n^{(m)} = \boldsymbol{A}_m\boldsymbol{f}(\boldsymbol{x}_n)$ ensures identifying $\boldsymbol{A}_m$ for all m, up to column permutations. Even though, the result provides some justification for using CCEM criterion under trace regularization, the result is clearly unsatisfactory, as it requires many stringent assumptions to establish identifiability.

• *SpeeLFC* (Chen et al., 2020), *Union-Net* (Wei et al., 2022), *CoNAL* (Chu et al., 2021): A number of other variants of the CCEM criterion are also proposed for E2E learning with crowdsourced labels. The work in (Chen et al., 2020) uses a similar criterion as in (4), but with the softmax operator applied on the columns of $\boldsymbol{A}_m$'s, instead on the output of $\boldsymbol{A}_m\boldsymbol{f}(\boldsymbol{x}_n)$. The work in (Wei et al., 2022) concatenates $\boldsymbol{A}_m$'s vertically and applies the softmax operator on the columns of such concatenated matrix, while formulating the CCEM criterion. The method in (Chu et al., 2021) models a common confusion matrix in addition to the annotator-specific confusion matrices $\boldsymbol{A}_m$'s and employs a CCEM-type criterion to learn both in a coupled fashion. Nonetheless, none of these approaches have successfully tackled the identifiability aspect. The work (Wei et al., 2022) claims theoretical support for their approach. However, their proof strategy is flawed, based on our analysis. In fact, understanding to the success of CCEM-based methods under practical E2E learning settings is still unclear.

Other Related Works. A couple of non-CCEM approaches have also considered the identifiability aspects (Khetan et al., 2018; Cao et al., 2019). The work in (Khetan et al., 2018) presents an EM-inspired alternating optimization strategy. Such procedure involves training the neural classifier multiple times, making it a computationally intensive approach. In addition, its identifiability guarantees are under certain strict assumptions, e.g., binary classification, identical confusions and conditional independence for annotators—which are hard to satisfy in practice. The approach in (Cao et al., 2019) employs an information-theoretic loss function to jointly train a predictor network and annotation aggregation network. The identifiability guarantees are again under restrictive assumptions, e.g., infinite data, no missing annotations and existence of mutually independent expert annotators.

We should also note that there exist works considering data-dependent annotator confusions as well, e.g., (Zhang et al., 2020) (multiple annotator case) and (Cheng et al., 2021; Xia et al., 2020; Zhu et al., 2022) (single annotator case)—leveraging more complex models and learning criteria. In our work, we focus on the criterion designed using a model in (1) as it is shown to be effective in practice.

3 Identifiability of CCEM-Based E2E Crowdsourcing

In this section, we offer identifiability analysis of the CCEM learning loss under realistic settings. We consider the following re-expressed objective function of CCEM:

$$\underset{\boldsymbol{f},\{\boldsymbol{A}_m\}}{\text{minimize}} \; -\frac{1}{|\mathcal{S}|}\sum_{(m,n)\in\mathcal{S}}\sum_{k=1}^{K}\mathbb{I}[\widehat{y}_n^{(m)}=k]\log[\boldsymbol{A}_m\boldsymbol{f}(\boldsymbol{x}_n)]_k \tag{6a}$$

$$\text{subject to} \;\; \boldsymbol{f}\in\mathcal{F},\; \boldsymbol{A}_m\in\mathcal{A},\; \forall m, \tag{6b}$$

where $\mathcal{F}\subseteq\{\boldsymbol{f}(\boldsymbol{x})\in\mathbb{R}^K|\,\boldsymbol{f}(\boldsymbol{x})\in\boldsymbol{\Delta}_K,\;\forall\boldsymbol{x}\}$ is a function class and $\mathcal{A}$ is the constrained set $\{\boldsymbol{A}\in\mathbb{R}^{K\times K}|\boldsymbol{A}\geq\boldsymbol{0},\boldsymbol{1}^\top\boldsymbol{A}=\boldsymbol{1}^\top\}$, since each column of $\boldsymbol{A}_m$'s are conditional probability distributions.

The rationale is that the confusion matrices $\boldsymbol{A}_m$'s act as correction terms for annotator's labeling noise to output a true label classifier $\boldsymbol{f}$. Intuitively, the objective in (6a) encourages the output estimates $\widehat{\boldsymbol{f}},\{\widehat{\boldsymbol{A}}_m\}$ to satisfy the relation $\boldsymbol{p}_n^{(m)}=\widehat{\boldsymbol{A}}_m\widehat{\boldsymbol{f}}(\boldsymbol{x}_n),\;\forall m,n$. One can easily note that the identifiability guarantees for $\boldsymbol{A}_m$ and $\boldsymbol{f}$ (i.e., when does the ground-truth $\boldsymbol{A}_m$ and the ground-truth $\boldsymbol{f}$ can be identified) are not so straightforward since there exists an infinite number of nonsingular matrices $\boldsymbol{Q}\in\mathbb{R}^{K\times K}$ such that $\boldsymbol{p}_n^{(m)}=(\widehat{\boldsymbol{A}}_m\boldsymbol{Q})(\boldsymbol{Q}^{-1}\widehat{\boldsymbol{f}}(\boldsymbol{x}_n))$.

To proceed, we make the following assumptions:

Assumption 1 *Each data item $\boldsymbol{x}_n,\; n\in[N]$ is drawn from a distribution $\mathcal{D}$ independently at random and has a bounded ℓ_2 norm. The observed index pairs (m,n) are included in $\mathcal{S}$ uniformly at random.*

Assumption 2 *The neural network function class $\mathcal{F}$ has a complexity measure denoted as $\mathscr{R}_{\mathcal{F}}$.*

Here, we use the *sensitive complexity* parameter introduced in (Lin & Zhang, 2019) as the complexity measure $\mathscr{R}_{\mathcal{F}}$. In essence, $\mathscr{R}_{\mathcal{F}}$ gets larger as the neural network function class $\mathcal{F}$ gets deeper and wider—also see supplementary material Sec. L for more discussion.

Assumption 3 *There exists a GTP $\boldsymbol{f}^\natural:\mathbb{R}^D\to[(1/\beta),1]^K$, for a certain $\beta>1$, such that $[\boldsymbol{f}^\natural(\boldsymbol{x}_n)]_k=\mathsf{Pr}(y_n=k|\boldsymbol{x}_n)$, where y_n denotes the true label of the data item $\boldsymbol{x}_n$. In addition, there exists $\widetilde{\boldsymbol{f}}\in\mathcal{F}$ such that $\|\widetilde{\boldsymbol{f}}(\boldsymbol{x}_n)-\boldsymbol{f}^\natural(\boldsymbol{x}_n)\|_2\leq\nu$ for all $\boldsymbol{x}_n\sim\mathcal{D}$, where $0\leq\nu<\infty$.*

Assumption 4 *(Near-Class Specialist Assumption) Let $\boldsymbol{A}_m^\natural$ denote the ground-truth confusion matrix for annotator m. For each class k, there exists a near-class specialist, indexed by m_k, such that $\|\boldsymbol{A}_{m_k}^\natural(k,:)-\boldsymbol{e}_k^\top\|_2\leq\xi_1$, where $0\leq\xi_1<\infty$.*

Assumption 5 *(Near-Anchor Point Assumption) For each class k, there exists a near-anchor point, $\boldsymbol{x}_{n_k}$, such that $\|\boldsymbol{f}^\natural(\boldsymbol{x}_{n_k})-\boldsymbol{e}_k\|_2\leq\xi_2)$, where $0\leq\xi_2<\infty$.*

The Near-Class Specialist Assumption (NCSA) implies that there exist annotators for each class $k\in[K]$ who can correctly distinguish items belonging to class k from those belonging to other classes. It is important to note that NCSA assumption is different from the commonly used assumptions such as the existence of expert annotators (Cao et al., 2019) or diagonal dominance of $\boldsymbol{A}_m^\natural$'s (Tanno et al., 2019). Such assumptions require the same annotators to be specialists with respect to all classes—also see Fig. 1 in the supplementary material in Sec. M. Indeed, NCSA assumption is more relaxed, by making use of the collective "wisdom of the crowd". The Near-Anchor Point Assumption (NAPA) is the relaxed version of the exact anchor point assumption commonly used in provable label noise learning methods (Xia et al., 2019; Li et al., 2021).

Under Assumptions 1-5, we have the following result:

Theorem 1 *Assume that each $[\boldsymbol{A}_m\boldsymbol{f}(\boldsymbol{x}_n)]_k,\;\forall\boldsymbol{A}_m\in\mathcal{A},\forall\boldsymbol{f}\in\mathcal{F}$ is at least $(1/\beta)$. Also assume that $\sigma_{\max}(\boldsymbol{A}_m^\natural)\leq\sigma,\;\forall m$, for a certain $\sigma>0$. Then, with probability greater than $1-{}^{K}/_{N^{\frac{1}{8}}}$, any optimal solution $\widehat{\boldsymbol{A}}_m$ and $\widehat{\boldsymbol{f}}$ of the problem (6) satisfies the following relations:*

$$\min_{\boldsymbol{\Pi}}\|\widehat{\boldsymbol{A}}_m-\boldsymbol{A}_m^\natural\boldsymbol{\Pi}\|_{\mathrm{F}}^2=K\sigma^2(\eta+\xi_1+\xi_2),\;\forall m\in[M], \tag{7a}$$

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}\left[\min_{\boldsymbol{\Pi}}\|\widehat{\boldsymbol{f}}(\boldsymbol{x})-\boldsymbol{\Pi}^\top\boldsymbol{f}^\natural(\boldsymbol{x})\|_2^2\right]=K(\eta+\xi_1+\xi_2), \tag{7b}$$

where $\mathbf{\Pi} \in \{0,1\}^K$ *denotes a column permutation matrix,*

$$\eta^2 = \frac{\beta N^{\frac{1}{8}}}{\sqrt{S}} \left(\sqrt{M \log N} + (\|\boldsymbol{X}\|_{\mathrm{F}} \mathscr{R}_{\mathcal{F}})^{\frac{1}{4}} \right) + \beta \sqrt{K} N^{\frac{1}{8}} \nu,$$

and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$*, if conditions* $\xi_1, \xi_2 \leq 1/K$*,* $\nu \leq 1/\beta K^2 N^{\frac{1}{8}}$*, and* $S = |\mathcal{S}| = \Omega\left(\beta^2 N^{\frac{1}{4}} K^2 \max\left(M \log N, \sqrt{\|\boldsymbol{X}\|_{\mathrm{F}} \mathscr{R}_{\mathcal{F}}}\right)\right)$ *hold.*

The proof of Theorem 1 is relegated to the supplementary material in Sec. B.

4 Enhancing Identifiability via Regularization

Theorem 1 confirms that the CCEM is a provably effective criterion for identifying $\boldsymbol{A}_m^\natural$'s and $\boldsymbol{f}^\natural$. The caveats lie in Assumptions 4 and 5. Essentially, these assumptions require that some rows of the collection $\{\boldsymbol{A}_m^\natural\}_{m=1}^M$ and some vectors of the collection $\{\boldsymbol{f}^\natural(\boldsymbol{x}_n)\}_{n=1}^N$ are close to the canonical vectors, respectively. These are reasonable assumptions, but satisfying them simultaneously may not always be possible. A natural question is if these assumptions can be relaxed, at least partially.

In this section, we propose variants of the CCEM that admits enhanced identifiability. To this end, we introduce the following geometric condition that is often used in the literature of nonnegative matrix factorization (NMF):

Definition 1 *(SSC)(Gillis, 2020) Consider the second-order cone* $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^K | \sqrt{K-1}\|\boldsymbol{x}\|_2 \leq \mathbf{1}^\top \boldsymbol{x}\}$*. A nonnegative matrix* $\boldsymbol{Z} \in \mathbb{R}_+^{L \times K}$ *satisfies the sufficiently scattered condition (SSC) if (i)* $\mathcal{C} \subseteq \mathrm{cone}(\boldsymbol{Z}^\top)$ *and (ii)* $\mathrm{cone}\{\boldsymbol{Z}^\top\} \subseteq \mathrm{cone}\{\boldsymbol{Q}\}$ *does not hold for any orthogonal* $\boldsymbol{Q} \in \mathbb{R}^{K \times K}$ *except for the permutation matrices.*

Geometrically, the matrix $\boldsymbol{Z}$ satisfying the SSC means that its rows span a large "area" within the nonnegative orthant, so that the conic hull spanned by the rows contains the second order cone $\mathcal{C}$ as a subset. Note that the SSC condition subsumes the the NCSA (Assumption (4) where $\xi_1 = 0$) and the NAPA (Assumption (5) where $\xi_1 = 0$) as special cases. Using SSC, we first show that the CCEM criterion in (6) attains identifiability of $\boldsymbol{A}_m^\natural$'s and $\boldsymbol{f}^\natural$ without using the NCSA or the NAPA, when the problem size grows sufficiently large.

Theorem 2 *Suppose that the incomplete labeling paradigm in Assumption 1 holds with* $S = \Omega(t)$ *and* $N = O(t^3)$*, for a certain* $t > 0$*. Also, assume that there exists* $\mathcal{Z} = \{n_1, \ldots, n_T\}$ *such that* $\boldsymbol{F}_{\mathcal{Z}}^\natural = [\boldsymbol{f}^\natural(\boldsymbol{x}_{n_1}), \ldots, \boldsymbol{f}^\natural(\boldsymbol{x}_{n_T})]$ *satisfies the SSC and* $\boldsymbol{W}^\natural = [\boldsymbol{A}_1^{\natural\top}, \ldots, \boldsymbol{A}_M^{\natural\top}]$ *satisfies the SSC as well. Then, at the limit of* $t \to \infty$*, if* $\boldsymbol{f}^\natural \in \mathcal{F}$*, the optimal solutions* $(\{\widehat{\boldsymbol{A}}_m\}, \widehat{\boldsymbol{f}})$ *of* (6) *satisfies* $\widehat{\boldsymbol{A}}_m = \boldsymbol{A}_m^\natural \mathbf{\Pi}$ *for all* m *and* $\widehat{\boldsymbol{f}}(\boldsymbol{x}) = \mathbf{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x})$ *for all* $\boldsymbol{x} \sim \mathcal{D}$*, where* $\mathbf{\Pi} \in \{0,1\}^{K \times K}$ *is a column permutation matrix.*

The proof is via connecting the CCEM problem to the classic nonnegative matrix factorization (NMF) problem at the limit of $t \to \infty$; see Sec. D. It is also worth to note that Theorem 2 does not require complete observation of the annotations to guarantee identifiability. Another remark is that $\boldsymbol{W}^\natural$ and $\boldsymbol{F}^\natural$ satisfying the SSC is more relaxed compared to that they satisfying the NCSA and NAPA, respectively. For example, one may need class specialists with much higher expertise to satisfy NCSA than to satisfy SSC on $\boldsymbol{W}^\natural$—also see geometric illustration in Sec. M. Compared to the result in Theorem 1, Theorem 2 implies that when N is very large, the CCEM should work under less stringent conditions relative to the NCSA and NAPA. Nonetheless, the assumption that both the GTP and the annotators' confusion matrices should meet certain requirements may not always hold. In the following, we further show that the SSC on either $\boldsymbol{W}^\natural$ or $\boldsymbol{F}^\natural$ can be relaxed:

Theorem 3 *Suppose that the incomplete labeling paradigm in Assumption 1 holds with* $S = \Omega(t)$ *and* $N = O(t^3)$*, for a certain* $t > 0$*. Then, at the limit of* $t \to \infty$*, if* $\boldsymbol{f}^\natural \in \mathcal{F}$*, we have* $\boldsymbol{A}_m^* = \boldsymbol{A}_m^\natural \mathbf{\Pi}$ *for all* m *and* $\boldsymbol{f}^*(\boldsymbol{x}) = \mathbf{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x})$ *for all* $\boldsymbol{x} \sim \mathcal{D}$*, when any of the following conditions hold:*

(a) If there exists $\mathcal{Z} = \{n_1, \ldots, n_T\}$ *such that* $\boldsymbol{F}_{\mathcal{Z}}^{\natural} = [\boldsymbol{f}^{\natural}(\boldsymbol{x}_{n_1}), \ldots, \boldsymbol{f}^{\natural}(\boldsymbol{x}_{n_T})]$ *satisfies the SSC,* $\mathrm{rank}(\boldsymbol{W}^{\natural}) = K$*, and* $(\{\boldsymbol{A}_m^*\}, \boldsymbol{f}^*)$ *is the optimal solution of* (6) *with the maximal* $\log\det(\boldsymbol{F}^*\boldsymbol{F}^{*\top})$.

(b) If $\boldsymbol{W}^{\natural} = [\boldsymbol{A}^{\natural_1\top}, \ldots, \boldsymbol{A}_M^{\natural\top}]$ *satisfies the SSC,* $\mathrm{rank}(\boldsymbol{F}^{\natural}) = K$*, and* $(\boldsymbol{A}_m^*, \boldsymbol{f}^*)$ *is the optimal solution of* (6) *with the maximal* $\log\det((\boldsymbol{W}^*)^\top\boldsymbol{W}^*)$.

Theorem 3(a) shows that if the number of annotated data items is large and the GTP-outputs associated with $\boldsymbol{x}_n$ are diverse enough to satisfy the SSC, then $\boldsymbol{f}^{\natural}$ and $\boldsymbol{A}_m^{\natural}$ are identifiable even if there are no class specialists available. Compared to Theorem 2, the identifiability is clearly enhanced, as the conditions on the annotators have been relaxed. Theorem 3(b) implies that if $\boldsymbol{W}^{\natural}$ satisfies SSC, identifiability can be established irrespective of the geometry of $\boldsymbol{F}^{\natural}$. Intuitively, when there are more annotators available, $\boldsymbol{W}^{\natural}$ is more likely to satisfy SSC. Hence, in practice, Theorem 3(b) may offer ensured performance in cases where more annotators are available.

Similar to the proof in Theorem 2, the proof here connects the CCEM-based E2E crowdsourcing problem to the simplex-structured matrix factorization (SSMF) problem (see (Gillis, 2020)) at the limit of $t \to \infty$. The detailed proof is provided in Sec. E. We should mention that various connections between matrix factorization models and data labeling problems were also observed in related prior works, e.g, the classic approaches based on Dawid & Skene model (Dawid & Skene, 1979; Zhang et al., 2016b; Traganitis et al., 2018) and E2E learning under noisy labels (Li et al., 2021). Nonetheless, the former does not consider cross-entropy losses or involve any feature extractor, whereas, the latter does not consider multiple annotators or incomplete labeling.

Implementation via Regularization Theorem 3 naturally leads to regularized versions of CCEM. Correspondingly, we have the following two cases:

Following Theorem 3(a), when the GTP is believed to be diverse over different $\boldsymbol{x}_n$'s, we propose to employ the following regularized CCEM:

$$\underset{\boldsymbol{A}_m\in\mathcal{A},\forall m,\ \boldsymbol{f}\in\mathcal{F}}{\text{minimize}} \quad -\frac{1}{|\mathcal{S}|}\sum_{(m,n)\in\mathcal{S}}\sum_{k=1}^{K}\mathbb{I}[\widehat{y}_n^{(m)}=k]\log[\boldsymbol{A}_m\boldsymbol{f}(\boldsymbol{x}_n)]_k - \lambda\log\det\boldsymbol{F}\boldsymbol{F}^\top. \tag{8}$$

In similar spirit, following Theorem 3(b), we propose the following regularized CCEM for enhanced identifiability when M is large (so that $\boldsymbol{W}^{\natural}$ likely satisfies SSC):

$$\underset{\boldsymbol{A}_m\in\mathcal{A},\forall m,\ \boldsymbol{f}\in\mathcal{F}}{\text{minimize}} \quad -\frac{1}{|\mathcal{S}|}\sum_{(m,n)\in\mathcal{S}}\sum_{k=1}^{K}\mathbb{I}[\widehat{y}_n^{(m)}=k]\log[\boldsymbol{A}_m\boldsymbol{f}(\boldsymbol{x}_n)]_k - \lambda\log\det\boldsymbol{W}^\top\boldsymbol{W}. \tag{9}$$

The implementations for (8) and (9) are called as geometry-regularized crowdsourcing network, abbreviated as `GeoCrowdNet(F)` and `GeoCrowdNet(W)`, respectively.

Remark 1 Under big data settings, N is often large and the GTP outputs of $\boldsymbol{x}_n$'s are often reasonably diverse. Hence, (8) oftentimes works well, as will be seen in our experiments. Nonetheless, for cases where the geometry of $\boldsymbol{F}$ does not satisfy the SSC, the employment of more annotators can help via using (9)—which is also an intuitive advantage of crowdsourcing.

Remark 2 Many works derive the CCEM criterion using the assumption that the annotators are conditionally independent; see (Rodrigues & Pereira, 2018; Chen et al., 2020; Tanno et al., 2019). The CCEM is a maximum likelihood criterion if such independence holds. Interestingly, as we revealed in our analysis, the identifiability of $\boldsymbol{A}_m^{\natural}$ and $\boldsymbol{f}^{\natural}$ under the CCEM criterion does *not* rely on the annotators' independence with each other.

5 EXPERIMENTS

Baselines. The proposed methods are compared with a number of existing E2E crowdsourcing methods, namely, `TraceReg`(Tanno et al., 2019), `MBEM` (Khetan et al., 2018), `CrowdLayer` (Rodrigues & Pereira, 2018), `CoNAL` (Chu et al., 2021), and `Max-MIG` (Cao et al., 2019). In addition

Table 1: Average test accuracies of the proposed methods and the baselines on MNIST dataset ($K = 10$), labeled by machine annotators.

	Case 1		*Case 2*	
	$N = 1000$		$M = 5$	
Methods	$M = 15$	$M = 20$	$N = 5000$	$N = 10000$
`GeoCrowdNet(F)`	79.89 ± 3.08	82.18 ± 3.48	$\mathbf{85.92 \pm 2.73}$	$\mathbf{87.21 \pm 2.47}$
`GeoCrowdNet(W)`	$\mathbf{80.97 \pm 1.31}$	$\mathbf{83.69 \pm 2.37}$	77.79 ± 8.97	82.37 ± 9.18
`GeoCrowdNet(`$\lambda = 0$`)`	71.15 ± 6.73	69.17 ± 2.61	71.66 ± 4.48	60.29 ± 7.91
`TraceReg`	70.14 ± 6.93	78.09 ± 3.52	63.71 ± 10.76	64.45 ± 10.14
`Crowdlayer`	65.72 ± 4.81	72.90 ± 2.31	51.25 ± 6.24	53.12 ± 14.02
`MBEM`	37.24 ± 8.54	33.39 ± 5.72	28.14 ± 6.24	26.90 ± 2.01
`CoNAL`	51.33 ± 4.27	55.59 ± 5.32	44.79 ± 7.01	39.41 ± 9.21
`Max-MIG`	$\mathbf{81.32 \pm 1.31}$	$\mathbf{83.34 \pm 3.36}$	$\mathbf{83.71 \pm 1.23}$	81.05 ± 0.65
`NN-MV`	78.98 ± 3.21	82.78 ± 1.45	83.44 ± 1.98	$\mathbf{83.10 \pm 2.86}$
`NN-DSEM`	73.95 ± 4.90	77.57 ± 4.06	65.81 ± 2.74	69.09 ± 3.21

to these baselines, we also learn neural network classifiers using the labels aggregated from *majority voting* and the classic EM algorithm proposed by Dawid &S kene (Dawid & Skene, 1979), denoted as `NN-MV` and `NN-DSEM`, respectively.

Semi-Real Data Experiments. We consider CIFAR-10 (Krizhevsky, 2009) dataset for experiments with labels generated following the model in (3). Due to page limitations, we present the results in Sec. O. The results are used to validate our identifiability claims under ideal scenarios where no model mismatches exist.

Real-Data Experiments with Machine Annotations. The settings are as follows:

Dataset. We use the MNIST dataset (Deng, 2012) which consists of $70,000$ handwritten digit images of 0 to 9 (i.e., $K = 10$); see more details in Sec. O.

Noisy Label Generation. We train a number of machine classifiers to produce noisy labels. Five types of classifiers, including support vector machines (SVM), k-Nearest Neighbour (kNN), logistic regression, convolutional neural network (CNN), and fully connected neural network are considered to act as annotators. To create more annotators, we change some of their parameters (e.g., the number of nearest neighbour parameter k of kNN and the number of epochs for training the CNN) while training them. We consider two different cases and use the following strategies for training the machine annotators:

(i) Case 1: NCSA (Assumption 4) holding: Each annotator is chosen to be a specialist uniformly at random with probability 0.1 (Here, we choose the same annotator to be a specialist for all classes.) If a particular annotator is chosen to be a specialist, it is trained carefully so that its classification accuracy exceeds a certain threshold; see details in Sec. O.

(ii) Case 2: NCSA not holding: Each machine annotator is trained by randomly choosing a very small subset of the MNIST training data (size ranging between 100 to 500) so that the accuracy of these machine annotators are expected to be low.

Once the machine annotators are trained, we let them label unseen data items of size N. To evaluate the methods under incomplete labeling, we observe each label from an annotator uniformly at random with a probability of 0.1.

Settings. The neural network classifier architecture for MNIST dataset is chosen to be Lenet-5 (Lecun et al., 1998), which consists of two sets of convolutional and max pooling layers, followed by a flattening convolutional layer, two fully-connected layers and finally a softmax layer. Adam (Kingma & Ba, 2015) is used an optimizer with weight decay of 10^{-4} and batch size of 128. The regularization parameter λ and the initial learning rate of the Adam optimizer are chosen via grid search method using the validation set from $\{0.01, 0.001, 0.0001\}$ and $\{0.01, 0.001\}$, respectively. We choose the same neural network structures for all the baselines. The confusion matrices are initialized with identity matrices of size K for proposed methods and the baselines `TraceReg` and `CrowdLayer`.

Table 2: Average test accuracies of the proposed methods and the baselines on LabelMe ($M = 59, K = 8$) and Music ($M = 44, K = 10$) datasets.

Methods	LabelMe	Music
GeoCrowdNet(F)	$\mathbf{85.85 \pm 0.33}$	$\mathbf{67.13 \pm 1.27}$
GeoCrowdNet(W)	81.59 ± 0.73	66.46 ± 1.77
GeoCrowdNet (λ=0)	80.35 ± 0.40	65.93 ± 1.81
TraceReg	80.89 ± 0.73	66.40 ± 0.90
Crowdlayer	81.41 ± 0.90	65.53 ± 1.74
MBEM	80.05 ± 3.31	$\mathbf{69.00 \pm 2.15}$
CoNAL	81.68 ± 2.72	60.93 ± 0.61
Max-MIG	$\mathbf{85.03 \pm 0.44}$	64.33 ± 2.31
NN-MV	79.57 ± 1.24	63.40 ± 2.68
NN-DSEM	78.26 ± 1.18	64.66 ± 2.34

Results. Table 1 presents the average label prediction accuracies on the testing data of the MNIST over 5 random trials, for various cases. One can observe that GeoCrowdNet(F) performs the best when there are more number of annotated items, even when there are no class specialist annotators present (i.e., case 2 with NCSA not holding). On the other hand, GeoCrowdNet(W) starts showing its advantage over GeoCrowdNet(F) when there are more number of annotators and class specialists among them. Both are consistent with our theorems. Overall, GeoCrowdNet(F) exhibits consistently good performance in all the cases.

Real-Data Experiments with Human Annotators. The settings are as follows:

Datasets. We consider two different datasets, namely, LabelMe ($M = 59$) (Rodrigues et al., 2017; Russell et al., 2007) and Music ($M = 44$) (Rodrigues et al., 2014), both having noisy and incomplete labels provided by human workers from AMT.

Settings. For LabelMe, we employ the pretrained VGG-16 embeddings followed by a fully connected layer as used in (Rodrigues & Pereira, 2018). For Music, we choose a similar architecture, but with batch normalization layers. We use a batch size of 128 for LabelMe and a batch size of 100 for Music dataset. All other settings are the same as the machine annotator case. More details of the datasets and architecture can be seen in Sec. O.

Results. Table 2 shows the average label prediction accuracies on the test data for 5 random trials. One can see that GeoCrowdNet (λ=0) works reasonably well relative to the machine annotation case in the previous experiment. This may be because the number of annotators for both LabelMe and Music are reasonably large ($M = 59$ and $M = 44$, respectively), which makes $\boldsymbol{W}^\natural$ satisfying Assumption 4 become easier than before. This validates our claims in Theorem 1, but also shows the limitations of the plain-vanilla CCEM—that is, it may require a fairly large number of annotators to start showing promising results. Similar to the previous experiment, both GeoCrowdNet(F) and GeoCrowdNet(W) outperform the unregularized version GeoCrowdNet (λ=0), showing the advantages of the identifiability-enhancing regularization terms.

6 Conclusion

In this work, we revisited the CCEM criterion—one of the most popular E2E learning criteria for crowdsourced label integration. We provided the first finite-sample identifiability characterization of the confusion matrices and the neural classifier which are learned using the CCEM criterion. Compared to many exiting identifiability results, our guarantees are under more relaxed and more realistic settings. In particular, a take-home point revealed in our analysis is that CCEM can provably identify the desired model parameters even if the annotators are dependent, which is a surprising but favorable result. We also proposed two regularized variants of the CCEM, based on the insights learned from our identifiability analysis for the plain-vanilla CCEM. The regularized CCEM criteria provably enhance the identifiability of the confusion matrices and the neural classifier under more challenging scenarios. We evaluated the proposed approaches on various synthetic and real datasets. The results corroborate our theoretical claims.

REFERENCES

Devansh Arpit, Stanisław Jastrzundefinedbski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 233–242. JMLR.org, 2017.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Proc. NeurIPS*, volume 30, 2017.

Michael D. Buhrmester, T. Kwang, and S. Gosling. Amazon's mechanical turk. *Perspectives on Psychological Science*, 6:3–5, 2011.

Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. Max-mig: an information theoretic approach for joint learning from crowds. *CoRR*, abs/1905.13436, 2019.

Zhijun Chen, Huimin Wang, Hailong Sun, Pengpeng Chen, Tao Han, Xudong Liu, and Jie Yang. Structured probabilistic end-to-end learning from crowds. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 1512–1518, 7 2020.

Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *ICLR*, 2021.

Zhendong Chu, Jing Ma, and Hongning Wang. Learning from crowds by modeling common confusions. In *AAAI*, 2021.

Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. Aggregating crowdsourced binary ratings. In *Proceedings of International Conference on World Wide Web*, pp. 285–294, 2013.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pp. 20–28, 1979.

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

A.A. Fedotov, P. Harremoes, and F. Topsoe. Refinements of pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6):1491–1498, 2003.

X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos. Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain. *IEEE Trans. Signal Process.*, 63(9):2306–2320, May 2015.

Arpita Ghosh, Satyen Kale, and Preston McAfee. Who moderates the moderators?: crowdsourcing abuse detection in user-generated content. In *Proceedings of the ACM conference on Electronic commerce*, pp. 167–176, 2011.

Nicolas Gillis. *Nonnegative Matrix Factorization*. Society for Industrial and Applied Mathematics, 2020.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Melody Y. Guan, Varun Gulshan, Andrew M. Dai, and Geoffrey E. Hinton. Who said what: Modeling individual labelers improves classification. In *AAAI*, 2018.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Kejun Huang, Nicholas D. Sidiropoulos, Evangelos E. Papalexakis, Christos Faloutsos, Partha Pratim Talukdar, and Tom M. Mitchell. Principled neuro-functional connectivity discovery. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, jun 2015.

David Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in Neural Information Processing Systems*, volume 24, pp. 1953–1961, 2011a.

David R Karger, Sewoong Oh, and Devavrat Shah. Efficient crowdsourcing for multi-class labeling. *ACM Sigmetrics Performance Evaluation Review*, 41(1):81–92, 2013.

Ashish Khetan, Zachary Chase Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. *ArXiv*, abs/1712.04577, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Shikun Li, Shiming Ge, Yingying Hua, Chunhui Zhang, Hao Wen, Tengfei Liu, and Weiqiang Wang. Coupled-view deep classifier learning from multiple noisy annotators. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4667–4674, Apr. 2020. doi: 10.1609/aaai.v34i04.5898. URL `https://ojs.aaai.org/index.php/AAAI/article/view/5898`.

Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. In *ICML*, pp. 6403–6413, 2021.

Shan Lin and Jingwei Zhang. Generalization bounds for convolutional neural networks, 2019.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Qiang Liu, Jian Peng, and Alexander T Ihler. Variational inference for crowdsourcing. In *Advances in neural information processing systems*, volume 25, pp. 692–700, 2012.

Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 730–734, 2015. doi: 10.1109/ACPR.2015.7486599.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. *ArXiv*, abs/1605.00251, 2016.

Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):1–21, 2015.

Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr): 1297–1322, 2010.

Filipe Rodrigues and Francisco C. Pereira. Deep learning from crowds. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. Gaussian process classification and active learning with multiple annotators. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 433–441. PMLR, 22–24 Jun 2014.

Filipe Rodrigues, Mariana Lourenco, Bernardete Ribeiro, and Francisco C. Pereira. Learning supervised topic models for classification and regression from crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2409–2422, 2017. doi: 10.1109/TPAMI.2017.2648786.

Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, 2007.

Pinsker M S. The information stability of gaussian random variables and processes (in russian). *Dokl. Akad. Nauk SSSR*, 133:28–30, 1960.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. doi: 10.1017/CBO9781107298019.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, 2008.

Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11236–11245, 2019.

Panagiotis A Traganitis, Alba Pages-Zamora, and Georgios B Giannakis. Blind multiclass ensemble classification. *IEEE Trans. Signal Process.*, 66(18):4737–4752, 2018.

Donna Vakharia and Matthew Lease. Beyond AMT: an analysis of crowd work platforms. *Computing Research Repository*, 2013.

Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, pp. 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006.

Kerri Wazny. "crowdsourcing" ten years in: A review. *Journal of Global Health*, 7:020602, 2017.

Hongxin Wei, Renchunzi Xie, Lei Feng, Bo Han, and Bo An. Deep learning from multiple noisy annotators as a union. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2022. doi: 10.1109/TNNLS.2022.3168696.

Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, pp. 2424–2432, 2010.

Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, volume 22, pp. 2035–2043. 2009.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning?, 2019.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7597–7610, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016a. URL `http://arxiv.org/abs/1611.03530`.

Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Cicarrelli, Frederik Barkhof, and Daniel Alexander. Disentangling human error from ground truth in segmentation of medical images. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15750–15762, 2020.

Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(102): 1–44, 2016b.

Zhaowei Zhu, Jialu Wang, and Yang Liu. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27633–27653, 17–23 Jul 2022.

Supplementary Material of " Deep Learning From Crowdsourced Labels: Coupled Cross-entropy Minimization, Identifiability, and Regularization"

A NOTATION

x, $\boldsymbol{x}$, and $\boldsymbol{X}$ represent a scalar, a vector, and a matrix, respectively. $[\boldsymbol{x}]_i$ and $\boldsymbol{x}(i)$ both denote the ith entry of the vector $\boldsymbol{x}$. $\boldsymbol{X}(:,j)$ denotes the jth column vector of $\boldsymbol{X}$ and $\boldsymbol{X}(i,:)$ denotes the ith row vector of $\boldsymbol{X}$. $[\boldsymbol{X}]_{i,j}$ and $\boldsymbol{X}(i,j)$ both mean the (i,j)th entry of $\boldsymbol{X}$. $\|\boldsymbol{x}\|_2$ and $\|\boldsymbol{X}\|_{\mathrm{F}}$ mean the Euclidean (Frobenius) norm of the augment. $[I]$ means an integer set $\{1,2,\ldots,I\}$. $^\top$ denote transpose. $\boldsymbol{\Delta}_K = \{\boldsymbol{x} \in \mathbb{R}^K : \sum_i \boldsymbol{x}(i) = 1, \boldsymbol{x} \geq \boldsymbol{0}\}$ denotes the probability simplex. $\boldsymbol{X} \geq \boldsymbol{0}$ implies that all the entries of the matrix $\boldsymbol{X}$ are nonnegative. $\mathbb{I}[A]$ denotes an indicator function for the event A such that $\mathbb{I}[A] = 1$ if the event A happens, otherwise $\mathbb{I}[A] = 0$. $\mathsf{CE}(\boldsymbol{x}, y) = -\sum_{k=1}^K \mathbb{I}[y=k]\log(\boldsymbol{x}(k))$ denotes the cross entropy function. $\boldsymbol{I}_K$ denotes an identity matrix of size $K \times K$. $\sigma_{\max}(\boldsymbol{X})$ and $\sigma_{\min}(\boldsymbol{X})$ denote the largest and the smallest singular values of the matrix $\boldsymbol{X}$, respectively. $\mathsf{cone}(\boldsymbol{X})$ denotes the conic hull formed by the columns of the matrix $\boldsymbol{X}$. $\det(\boldsymbol{X})$ represents the determinant of $\boldsymbol{X}$. $\mathsf{vec}(\boldsymbol{X})$ denotes the vectorization operation that concatenates the columns of $\boldsymbol{X}$. $\mathsf{trace}(\boldsymbol{X})$ outputs the trace (the sum of the diagonal elements) of $\boldsymbol{X}$.

B PROOF OF THEOREM 1

If the vectors $\boldsymbol{p}_n^{(m)}$ for all m, n are available, then one can represent the model in (2) as a *nonnegative matrix factorization* (NMF) model as follows:

$$\underbrace{\begin{bmatrix} \boldsymbol{p}_1^{(1)} & \ldots & \boldsymbol{p}_N^{(1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{p}_1^{(M)} & \ldots & \boldsymbol{p}_N^{(M)} \end{bmatrix}}_{\boldsymbol{P} \in \mathbb{R}^{MK \times N}} = \underbrace{\begin{bmatrix} \boldsymbol{A}_1 \\ \vdots \\ \boldsymbol{A}_M \end{bmatrix}}_{\boldsymbol{W} \in \mathbb{R}^{MK \times K}} \underbrace{[\boldsymbol{f}(\boldsymbol{x}_1) \quad \ldots \quad \boldsymbol{f}(\boldsymbol{x}_N)]}_{\boldsymbol{F} \in \mathbb{R}^{K \times N}}, \tag{10}$$

where the factors $\boldsymbol{W}$ and $\boldsymbol{F}$ are both nonnegative per their physical meaning.

Let $\boldsymbol{P}^\natural$ denote the ground-truth and $\boldsymbol{p}_n^{\natural(m)}$ denote the (m,n)th block in $\boldsymbol{P}^\natural$, following the representation in (10). Note that we do not observe the entire $\boldsymbol{P}^\natural$ but only the $\widehat{y}_n^{(m)}$ sampled from $\boldsymbol{p}_n^{\natural(m)}$ for $(m,n) \in \mathcal{S}$. We first show that the CCEM criterion in (6) implicitly estimate $\boldsymbol{P}^\natural$ from incomplete labeling represented by $\mathcal{S}$. To this end, we consider the objective function in (6a) by choosing $\mathsf{CE}(\boldsymbol{x}, y) = -\sum_{k=1}^K \mathbb{I}[y=k]\log(\boldsymbol{x}(k))$:

$$D_{\mathcal{S}}(\boldsymbol{P}; \widehat{\boldsymbol{Y}}) \triangleq -\frac{1}{S} \sum_{(m,n)\in\mathcal{S}} \sum_{k=1}^K \mathbb{I}[\widehat{y}_n^{(m)} = k] \log \boldsymbol{P}((m-1)K + k, n), \tag{11}$$

where $\widehat{\boldsymbol{Y}}$ denotes the collection of all $\widehat{y}_n^{(m)}, (m,n) \in \mathcal{S}$. Let $\{\widehat{\boldsymbol{A}}_m\}_{m=1}^M$ and $\widehat{\boldsymbol{f}}$ denote the estimates of the learning criterion in (6). Then, the criterion (6) can be re-written as:

$$\widehat{\boldsymbol{P}} \triangleq \arg\min_{\boldsymbol{P} \in \mathcal{P}} D_{\mathcal{S}}(\boldsymbol{P}; \widehat{\boldsymbol{Y}}), \tag{12}$$

where $\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{W}} \left[\widehat{\boldsymbol{f}}(\boldsymbol{x}_1) \quad \ldots \quad \widehat{\boldsymbol{f}}(\boldsymbol{x}_N)\right]$, $\widehat{\boldsymbol{W}} = [\widehat{\boldsymbol{A}}_1^\top, \ldots, \widehat{\boldsymbol{A}}_M^\top]^\top$ and

$$\begin{aligned} \mathcal{P} &\triangleq \{\boldsymbol{P} \in \mathbb{R}^{MK \times N} \mid \boldsymbol{P} = \boldsymbol{W}\left[\boldsymbol{f}(\boldsymbol{x}_1) \quad \ldots \quad \boldsymbol{f}(\boldsymbol{x}_N)\right], \boldsymbol{W} \in \mathcal{W},\ \boldsymbol{f} \in \mathcal{F}\}, \\ \mathcal{W} &\triangleq \{\boldsymbol{W} = [\boldsymbol{A}_1^\top, \ldots, \boldsymbol{A}_M^\top]^\top \in \mathbb{R}^{MK \times K} \mid \boldsymbol{1}^\top \boldsymbol{A}_m = \boldsymbol{1}^\top, \boldsymbol{A}_m \geq \boldsymbol{0},\ \forall m\}, \\ \mathcal{F} &\subseteq \{\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^K \mid \boldsymbol{f}(\boldsymbol{x}) \in \boldsymbol{\Delta}_K,\ \forall \boldsymbol{x} \in \mathbb{R}^D\}. \end{aligned}$$

Our main goal is to characterize the estimation errors of $\{\widehat{\boldsymbol{A}}_m\}_{m=1}^M$ and $\widehat{\boldsymbol{f}}$. This can be done via first bounding the estimation errors $\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2$, where $\boldsymbol{p}_n^{\natural(m)}$ is given by:

$$\boldsymbol{p}_n^{\natural(m)} = \boldsymbol{A}_m^\natural \boldsymbol{f}^\natural(\boldsymbol{x}_n),\ \forall m, n$$

in which $\boldsymbol{A}_m^\natural$ and $\boldsymbol{f}^\natural$ denote the mth ground-truth confusion matrix and the GTP, respectively and $\widehat{\boldsymbol{p}}_n^{(m)}$ denote the (m,n)th block in $\widehat{\boldsymbol{P}}$.

B.1 ESTIMATING $\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2$

We have the following proposition:

Proposition 1 *Under the assumptions given in Theorem 1, the following results hold:*

1. *With probability greater than* $1-\delta$*:*

$$\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2 \leq 8\sqrt{2}K\beta\mathfrak{R} + 10\log(\beta)\frac{\sqrt{2\log(4/\delta)}}{\sqrt{S}} + \beta\sqrt{K}\nu,$$

where $\mathfrak{R} = \frac{16}{\sqrt{S}}\sqrt{MK\log\left(4K\sqrt{NM}\right) + (2\|\boldsymbol{X}\|_{\mathrm{F}}\mathscr{R}_{\mathcal{F}})^{\frac{1}{2}}}$.

2. *With probability greater than* $1-\delta-\frac{1}{N^{\frac{1}{8}}}$*, for any* $n\in[N]$*:*

$$\sum_{m=1}^{M}\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2 \leq N^{\frac{1}{8}}\left(\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2\right) + N^{\frac{1}{8}}\sqrt{\frac{\log\beta\log(2/\delta)}{S}}.$$

The proof is provided in Sec. C.

B.2 ESTIMATION ERRORS FOR $\widehat{\boldsymbol{A}}_m$ AND $\widehat{\boldsymbol{f}}$

In this section, we characterize the estimation errors for the confusion matrices and the GTP. Before proceeding, we re-express the results in Proposition 1:

$$\|\boldsymbol{P}^{\natural} - \widehat{\boldsymbol{P}}\|_{\mathrm{F}} \leq \zeta, \tag{13}$$

$$\sum_{m=1}^{M}\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2 \leq \varphi, \forall n. \tag{14}$$

with probability greater than $1-\delta$ and $1-\delta-\frac{1}{N^{\frac{1}{8}}}$, respectively, for certain $\zeta,\varphi>0$, where $\boldsymbol{p}_n^{\natural(m)}$ denotes (m,n)th block in the matrix $\boldsymbol{P}^{\natural}$ and $\widehat{\boldsymbol{p}}_n^{(m)}$ denotes (m,n)th block in the matrix $\widehat{\boldsymbol{P}}$—see (10). Assumptions 4-5 imply that there exists index sets $\varLambda=\{\tilde{m}_1,\ldots,\tilde{m}_K\}$ and $\varPsi=\{\tilde{n}_1,\ldots,\tilde{n}_K\}$ such that:

$$\boldsymbol{W}^{\natural}(\varLambda,:) = \boldsymbol{I}_K + \boldsymbol{N}_W$$
$$\boldsymbol{F}^{\natural}(:,\varPsi) = \boldsymbol{I}_K + \boldsymbol{N}_F,$$

where $\|\boldsymbol{N}_W\|_{\mathrm{F}} \leq \sqrt{K}\xi_1$, and $\|\boldsymbol{N}_F\|_{\mathrm{F}} \leq \sqrt{K}\xi_2$. Note that the conditions do not require any confusion matrices to be near identity, i.e., we do not require any annotators to be all-class specialists. Nonetheless, for notation simplicity, we let $\varLambda=\{1,\ldots,K\}$ and $\varPsi=\{1,\ldots,K\}$, without loss of generality. Hence, we have the following:

$$\boldsymbol{W}^{\natural} = \begin{bmatrix}\boldsymbol{A}_1^{\natural}\\ \boldsymbol{A}_2^{\natural}\\ \vdots\\ \boldsymbol{A}_M^{\natural}\end{bmatrix},\ \boldsymbol{F}^{\natural} = \begin{bmatrix}\boldsymbol{f}^{\natural}(\boldsymbol{x}_1) & \ldots & \boldsymbol{f}^{\natural}(\boldsymbol{x}_N)\end{bmatrix} = \begin{bmatrix}\boldsymbol{F}_1^{\natural} & \boldsymbol{F}_2^{\natural}\end{bmatrix},$$

where $\boldsymbol{A}_1^{\natural} = \boldsymbol{I}_K + \boldsymbol{N}_W$, $\boldsymbol{F}_1^{\natural} = \boldsymbol{I}_K + \boldsymbol{N}_F$, and $\boldsymbol{F}_2^{\natural} \in \mathbb{R}^{K\times(N-K)}$, $\|\boldsymbol{N}_W\|_{\mathrm{F}} \leq \sqrt{K}\xi_1$, and $\|\boldsymbol{N}_F\|_{\mathrm{F}} \leq \sqrt{K}\xi_2$.

Note that letting $\varLambda=\{1,\ldots,K\}$ and $\varPsi=\{1,\ldots,K\}$ for our analysis is without loss of any generality since we will be analyzing the error term $\|\boldsymbol{A}_1^{\natural}\boldsymbol{F}_1^{\natural} - \widehat{\boldsymbol{A}}_1\widehat{\boldsymbol{F}}_1\|_{\mathrm{F}}^2$ and it satisfies

$$\|\boldsymbol{A}_1^{\natural}\boldsymbol{F}_1^{\natural} - \widehat{\boldsymbol{A}}_1\widehat{\boldsymbol{F}}_1\|_{\mathrm{F}}^2 = \|\boldsymbol{\Pi}_1\boldsymbol{A}_1^{\natural}\boldsymbol{F}_1^{\natural}\boldsymbol{\Pi}_2 - \boldsymbol{\Pi}_1\widehat{\boldsymbol{A}}_1\widehat{\boldsymbol{F}}_1\boldsymbol{\Pi}_2\|_{\mathrm{F}}^2$$

for any row permutation matrix $\boldsymbol{\Pi}_1$ and column permutation matrix $\boldsymbol{\Pi}_2$.

Let $\widehat{\boldsymbol{A}}_m$ and $\widehat{\boldsymbol{f}}$ denote the estimated values of $\boldsymbol{A}_m^\natural$ and $\boldsymbol{f}^\natural$ using the learning criterion (6) such that

$$\widehat{\boldsymbol{W}} = \begin{bmatrix} \widehat{\boldsymbol{A}}_1 \\ \widehat{\boldsymbol{A}}_2 \\ \vdots \\ \widehat{\boldsymbol{A}}_M \end{bmatrix}, \ \widehat{\boldsymbol{F}} = \begin{bmatrix} \widehat{\boldsymbol{f}}(\boldsymbol{x}_1) & \dots & \widehat{\boldsymbol{f}}(\boldsymbol{x}_N) \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{F}}_1 & \widehat{\boldsymbol{F}}_2 \end{bmatrix},$$

where $\widehat{\boldsymbol{A}}_m \in \mathbb{R}^{K\times K}$, $\widehat{\boldsymbol{F}}_1 \in \mathbb{R}^{K\times K}$, and $\widehat{\boldsymbol{F}}_2 \in \mathbb{R}^{K\times(N-K)}$. Then, we have:

$$\begin{aligned} \|\boldsymbol{P}^\natural - \widehat{\boldsymbol{P}}\|_{\mathrm{F}}^2 &= \|\boldsymbol{W}^\natural \boldsymbol{F}^\natural - \widehat{\boldsymbol{W}}\widehat{\boldsymbol{F}}\|_{\mathrm{F}}^2 \\ &= \|\boldsymbol{I}_K + \boldsymbol{N}_W + \boldsymbol{N}_F + \boldsymbol{N}_W\boldsymbol{N}_F - \widehat{\boldsymbol{A}}_1\widehat{\boldsymbol{F}}_1\|_{\mathrm{F}}^2 + \|\boldsymbol{F}_2^\natural + \boldsymbol{N}_W\boldsymbol{F}_2^\natural - \widehat{\boldsymbol{A}}_1\widehat{\boldsymbol{F}}_2\|_{\mathrm{F}}^2 \\ &\quad + \sum_{m=2}^{M} \|\boldsymbol{A}_m^\natural + \boldsymbol{A}_m^\natural\boldsymbol{N}_F - \widehat{\boldsymbol{A}}_m\widehat{\boldsymbol{F}}_1\|_{\mathrm{F}}^2 + \sum_{m=2}^{M} \|\boldsymbol{A}_m^\natural\boldsymbol{F}_2^\natural - \widehat{\boldsymbol{A}}_m\widehat{\boldsymbol{F}}_2\|_{\mathrm{F}}^2. \end{aligned} \tag{15}$$

Let us define the error matrices as below:

$$\boldsymbol{E}_A^{(1)} \triangleq \widehat{\boldsymbol{A}}_1 - (\boldsymbol{I}_K + \boldsymbol{N}_W)\boldsymbol{\Pi},\ \boldsymbol{E}_A^{(m)} \triangleq \widehat{\boldsymbol{A}}_m - \boldsymbol{A}_m^\natural\boldsymbol{\Pi},\ \forall m > 1, \tag{16}$$

$$\boldsymbol{E}_F^{(1)} \triangleq \widehat{\boldsymbol{F}}_1 - \boldsymbol{\Pi}^\top(\boldsymbol{I}_K + \boldsymbol{N}_F),\ \boldsymbol{E}_F^{(2)} \triangleq \widehat{\boldsymbol{F}}_2 - \boldsymbol{\Pi}^\top\boldsymbol{F}_2^\natural, \tag{17}$$

where $\boldsymbol{\Pi} \in [0,1]^{K\times K}$ is a column permutation matrix and is the same across all the error blocks. We hope to characterize the norm of these error matrices.

Upper bound for $\|\boldsymbol{E}_A^{(1)}\|_{\mathrm{F}}$ and $\|\boldsymbol{E}_F^{(1)}\|_{\mathrm{F}}$. We start by considering the first term on the R.H.S. of (15), i.e., $\|\boldsymbol{I}_K + \boldsymbol{N}_W + \boldsymbol{N}_F + \boldsymbol{N}_W\boldsymbol{N}_F - \widehat{\boldsymbol{A}}_1\widehat{\boldsymbol{F}}_1\|_{\mathrm{F}}^2$. From (14), we have the following with probability greater than $1 - K\delta - \frac{K}{N^{\frac{1}{8}}}$:

$$\begin{aligned} \varphi^2 &\geq \|\boldsymbol{I}_K + \boldsymbol{N}_W + \boldsymbol{N}_F + \boldsymbol{N}_W\boldsymbol{N}_F - \widehat{\boldsymbol{A}}_1\widehat{\boldsymbol{F}}_1\|_{\mathrm{F}}^2 \\ &= \sum_{k=1}^{K} |1 + \widetilde{\boldsymbol{N}}(k,k) - \widehat{\boldsymbol{A}}_1(k,:)\widehat{\boldsymbol{F}}_1(:,k)|^2 + \sum_{j=1}^{K}\sum_{k\neq j} |\widetilde{\boldsymbol{N}}(j,k) - \widehat{\boldsymbol{A}}_1(j,:)\widehat{\boldsymbol{F}}_1(:,k)|^2, \end{aligned} \tag{18}$$

where $\widetilde{\boldsymbol{N}} = \boldsymbol{N}_W + \boldsymbol{N}_F + \boldsymbol{N}_W\boldsymbol{N}_F$. The absolute value of the largest entry in $\widetilde{\boldsymbol{N}}$ can be bounded by $\xi_1 + \xi_2 + \sqrt{K}\xi_1\xi_2$. Let us denote $\kappa = \varphi + \xi_1 + \xi_2 + \sqrt{K}\xi_1\xi_2$ for conciseness. Then, from (18), we get the following conditions:

$$1 - \kappa \leq |\widehat{\boldsymbol{A}}_1(k,:)\widehat{\boldsymbol{F}}_1(:,k)| \leq 1 + \kappa,\ \forall k \in [K] \tag{19a}$$

$$|\widehat{\boldsymbol{A}}_1(k,:)\widehat{\boldsymbol{F}}_1(:,j)| \leq \kappa, \forall k \neq j. \tag{19b}$$

From the conditions (19a) and (19b), we have the following result:

Lemma 1 *Assume that the conditions in* (19a) *and* (19b) *hold and that* $\kappa \leq \frac{1}{K+1}$. *Then, we have the following relations satisfied:*

$$\textit{arg max}_\ell \widehat{\boldsymbol{A}}_1(k,\ell) \neq \textit{arg max}_\ell \widehat{\boldsymbol{A}}_1(j,\ell),\ \forall k \neq j \tag{20a}$$

$$\textit{arg max}_q \widehat{\boldsymbol{F}}_1(q,k) \neq \textit{arg max}_q \widehat{\boldsymbol{F}}_1(q,j),\ \forall k \neq j \tag{20b}$$

$$\textit{arg max}_\ell \widehat{\boldsymbol{A}}_1(k,\ell) = \textit{arg max}_q \widehat{\boldsymbol{F}}_1(q,k),\ \forall k. \tag{20c}$$

The proof of the lemma is given in Sec. F.

From the lower bound condition in (19a), we have the following result:

$$\begin{aligned}\sum_{k=1}^{K}(1-\kappa) &\leq \sum_{k=1}^{K}\widehat{\boldsymbol{A}}_1(k,1)\widehat{\boldsymbol{F}}_1(1,k)+\cdots+\widehat{\boldsymbol{A}}_1(k,K)\widehat{\boldsymbol{F}}_1(K,k)\\ &\leq \sum_{k=1}^{K}\max_{\ell\in[K]}\widehat{\boldsymbol{A}}_1(k,\ell)(\widehat{\boldsymbol{F}}_1(1,k)+\cdots+\widehat{\boldsymbol{F}}_1(K,k))\\ &= \sum_{k=1}^{K}\max_{\ell\in[K]}\widehat{\boldsymbol{A}}_1(k,\ell) \qquad (21)\end{aligned}$$

where the last equality is obtained due to the probability simplex constraints on the columns of $\widehat{\boldsymbol{F}}_1$.

From the lower bound condition in (19a), we have the below result as well:

$$\begin{aligned}\sum_{k=1}^{K}(1-\kappa) &\leq \sum_{k=1}^{K}\Big(\widehat{\boldsymbol{A}}_1(k,1)\widehat{\boldsymbol{F}}_1(1,k)+\cdots+\widehat{\boldsymbol{A}}_1(k,K)\widehat{\boldsymbol{F}}_1(K,k)\Big)\\ &\leq \sum_{k=1}^{K}\max_{\ell}\widehat{\boldsymbol{F}}_1(k,\ell)(\widehat{\boldsymbol{A}}_1(1,k)+\cdots+\widehat{\boldsymbol{A}}_1(K,k))\\ &= \sum_{k=1}^{K}\max_{\ell}\widehat{\boldsymbol{F}}_1(k,\ell). \qquad (22)\end{aligned}$$

where the last equality is obtained due to the probability simplex constraints on all the columns of $\widehat{\boldsymbol{A}}_1$.

Next, we characterize the term $\|\boldsymbol{E}_A^{(1)}\|_{\mathrm{F}}^2$. To achieve this, by employing the result (20a) in Lemma 1, we fix the column permutation $\boldsymbol{\Pi}$ as follows:

$$\boldsymbol{\Pi}(j,k)=\begin{cases}1, & j=\arg\max_{\ell}\widehat{\boldsymbol{A}}_1(k,\ell)\\ 0, & \text{otherwise.}\end{cases} \qquad (23)$$

We get the following set of relations:

$$\begin{aligned}\|\widehat{\boldsymbol{A}}_1-\boldsymbol{I}_K\boldsymbol{\Pi}\|_{\mathrm{F}}^2 &= \sum_{k=1}^{K}\|\widehat{\boldsymbol{A}}_1(k,:)-\boldsymbol{e}_k^\top\boldsymbol{\Pi}\|_2^2\\ &= \sum_{k=1}^{K}\|\widehat{\boldsymbol{A}}_1(k,:)\|_2^2+K-2\sum_{k=1}^{K}\widehat{\boldsymbol{A}}_1(k,:)\boldsymbol{\Pi}^\top\boldsymbol{e}_k\\ &= \|\widehat{\boldsymbol{A}}_1\|_{\mathrm{F}}^2+K-2\sum_{k=1}^{K}\widehat{\boldsymbol{A}}_1(k,:)\boldsymbol{\Pi}^\top\boldsymbol{e}_k\\ &\leq 2K-2\sum_{k=1}^{K}\widehat{\boldsymbol{A}}_1(k,:)\boldsymbol{\Pi}^\top\boldsymbol{e}_k\\ &\overset{(a)}{=} 2K-2\sum_{k=1}^{K}\max_{\ell}\widehat{\boldsymbol{A}}_1(k,\ell)\\ &\overset{(b)}{\leq} 2K-2\sum_{k=1}^{K}(1-\kappa)=2K\kappa,\end{aligned}$$

where the last inequality (b) is by (21) and the relation (a) is obtained by choosing $\boldsymbol{\Pi}$ as defined in (23).

Hence, we have the following with probability greater than $1-K\delta-\frac{K}{N^{\frac{1}{8}}}$:

$$\|\boldsymbol{E}_A^{(1)}\|_{\mathrm{F}}\leq\|\widehat{\boldsymbol{A}}_1-\boldsymbol{I}_K\boldsymbol{\Pi}\|_{\mathrm{F}}+\|\boldsymbol{N}_W\|_{\mathrm{F}}\leq\sqrt{2K\kappa}+\sqrt{K}\xi_1. \qquad (24)$$

We proceed to consider the error term $\|\boldsymbol{E}_F^{(1)}\|_{\mathrm{F}}^2$. To achieve this, consider the following:

$$\begin{aligned}
\|\widehat{\boldsymbol{F}}_1 - \boldsymbol{\Pi}^\top \boldsymbol{I}_K\|_{\mathrm{F}}^2 &= \sum_{k=1}^{K} \|\widehat{\boldsymbol{F}}_1(k,:) - \boldsymbol{e}_k^\top \boldsymbol{\Pi}^\top\|_2^2 \\
&= \sum_{k=1}^{K} \|\widehat{\boldsymbol{F}}_1(k,:)\|_2^2 + K - 2\sum_{k=1}^{K} \widehat{\boldsymbol{F}}_1(k,:)\boldsymbol{\Pi}^\top \boldsymbol{e}_k \\
&= \|\widehat{\boldsymbol{F}}_1\|_{\mathrm{F}}^2 + K - 2\sum_{k=1}^{K} \widehat{\boldsymbol{F}}_1(k,:)\boldsymbol{\Pi}^\top \boldsymbol{e}_k \\
&\leq 2K - 2\sum_{k=1}^{K} \widehat{\boldsymbol{F}}_1(k,:)\boldsymbol{\Pi}^\top \boldsymbol{e}_k \\
&\overset{(a)}{=} 2K - 2\sum_{k=1}^{K} \max_{\ell} \widehat{\boldsymbol{F}}_1(k,\ell) \\
&\overset{(b)}{\leq} 2K - 2\sum_{k=1}^{K} (1-\kappa) = 2K\kappa,
\end{aligned}$$

where the last inequality (b) is by (22) and the relation (a) is by combining the definition of $\boldsymbol{\Pi}$ in (23) and (20c) in Lemma 1. Hence, we get the following with probability greater than $1 - K\delta - \frac{K}{N^{\frac{1}{8}}}$:

$$\|\boldsymbol{E}_F^{(1)}\|_{\mathrm{F}} \leq \|\widehat{\boldsymbol{F}}_1 - \boldsymbol{I}_K\boldsymbol{\Pi}\|_{\mathrm{F}} + \|\boldsymbol{N}_F\|_{\mathrm{F}} \leq \sqrt{2K\kappa} + \sqrt{K}\xi_2. \tag{25}$$

Upper bound for $\|\boldsymbol{E}_F^{(2)}\|_{\mathrm{F}}$. Next, we consider the second term in (15), i.e., $\|\boldsymbol{F}_2^\natural + \boldsymbol{N}_W\boldsymbol{F}_2^\natural - \widehat{\boldsymbol{A}}_1\widehat{\boldsymbol{F}}_2\|_{\mathrm{F}}$. From (13), we have the following with probability greater than $1-\delta$:

$$\begin{aligned}
\zeta \geq \|\boldsymbol{F}_2^\natural + \boldsymbol{N}_W\boldsymbol{F}_2^\natural - \widehat{\boldsymbol{A}}_1\widehat{\boldsymbol{F}}_2\|_{\mathrm{F}} &= \|\boldsymbol{F}_2^\natural + \boldsymbol{N}_W\boldsymbol{F}_2^\natural - (\boldsymbol{I}_K\boldsymbol{\Pi} + \boldsymbol{N}_W\boldsymbol{\Pi} + \boldsymbol{E}_A^{(1)})(\boldsymbol{E}_F^{(2)} + \boldsymbol{\Pi}^\top\boldsymbol{F}_2^\natural)\|_{\mathrm{F}} \\
&= \|(\boldsymbol{I}_K\boldsymbol{\Pi} + \boldsymbol{N}_W\boldsymbol{\Pi} + \boldsymbol{E}_A^{(1)})\boldsymbol{E}_F^{(2)} + \boldsymbol{E}_A^{(1)}\boldsymbol{\Pi}^\top\boldsymbol{F}_2^\natural\|_{\mathrm{F}} \\
&= \|\widehat{\boldsymbol{A}}_1\boldsymbol{E}_F^{(2)} + \boldsymbol{E}_A^{(1)}\boldsymbol{\Pi}^\top\boldsymbol{F}_2^\natural\|_{\mathrm{F}} \\
&\overset{(a)}{\geq} \left|\|\widehat{\boldsymbol{A}}_1\boldsymbol{E}_F^{(2)}\|_{\mathrm{F}} - \|\boldsymbol{E}_A^{(1)}\boldsymbol{\Pi}^\top\boldsymbol{F}_2^\natural\|_{\mathrm{F}}\right| \\
&\overset{(b)}{\geq} \left|\sigma_{\min}(\widehat{\boldsymbol{A}}_1)\|\boldsymbol{E}_F^{(2)}\|_{\mathrm{F}} - \|\boldsymbol{E}_A^{(1)}\|_{\mathrm{F}}\|\boldsymbol{F}_2^\natural\|_{\mathrm{F}}\right| \\
&\overset{(c)}{\geq} \left|\sigma_{\min}(\widehat{\boldsymbol{A}}_1)\|\boldsymbol{E}_F^{(2)}\|_{\mathrm{F}} - (\sqrt{2K\kappa} + \sqrt{K}\xi_1)\|\boldsymbol{F}_2^\natural\|_{\mathrm{F}}\right|
\end{aligned} \tag{26}$$

where the inequality (a) is by using the triangle inequality, (b) is obtained by applying the following two relations for any two matrices $\boldsymbol{A} \in \mathbb{R}^{K\times K}, \boldsymbol{B} \in \mathbb{R}^{K\times L},\ L \geq K$:

$$\|\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}} \geq \sigma_{\min}(\boldsymbol{A})\|\boldsymbol{B}\|_{\mathrm{F}} \tag{27}$$

$$\|\boldsymbol{A}\boldsymbol{B}\|_{\mathrm{F}} \leq \|\boldsymbol{A}\|_{\mathrm{F}}\|\boldsymbol{B}\|_{\mathrm{F}}. \tag{28}$$

The inequality (c) is obtained by applying (24).

Hence, (26) combined with the fact that $\|\boldsymbol{F}_2\|_{\mathrm{F}} \leq \|\boldsymbol{F}^\natural\|_{\mathrm{F}}$ gives the below:

$$\|\boldsymbol{E}_F^{(2)}\|_{\mathrm{F}} \leq \frac{\zeta + (\sqrt{2K\kappa} + \sqrt{K}\xi_1)\|\boldsymbol{F}^\natural\|_{\mathrm{F}}}{\sigma_{\min}(\widehat{\boldsymbol{A}}_1)}. \tag{29}$$

Next, we use the following lemma to characterize $\sigma_{\min}(\widehat{\boldsymbol{A}}_1)$:

Lemma 2 *Suppose that the matrix $\boldsymbol{X} \in \mathbb{R}^{K\times K}$ takes the form $\boldsymbol{X} = \boldsymbol{I}_K + \boldsymbol{E}_1 + \boldsymbol{E}_2$. Assume that $\|\boldsymbol{E}_1\|_F \leq \upsilon_1$ and $\|\boldsymbol{E}_2\|_F \leq \upsilon_2$, for a certain $\upsilon_1, \upsilon_2 > 0$. Then, we have*

$$\sigma_{\min}(\boldsymbol{X}) \geq |1 - \upsilon_1 - \upsilon_2|.$$

The proof is relegated to Sec. G. Applying Lemma 2, we get the following with probability greater than $1-(K+1)\delta-\frac{K}{N^{\frac{1}{8}}}$:

$$\|\boldsymbol{E}_F^{(2)}\|_{\mathrm{F}} \leq \frac{\zeta+(\sqrt{2K\kappa}+\sqrt{K}\xi_1)\|\boldsymbol{F}^{\natural}\|_{\mathrm{F}}}{|1-\sqrt{2K\kappa}-2\sqrt{K}\xi_1|} \leq \frac{c_1(\zeta+\sqrt{K\kappa}\|\boldsymbol{F}^{\natural}\|_{\mathrm{F}})}{|1-\sqrt{K\kappa}|} \tag{30}$$

for certain constant $c_1>0$. where we used the fact that $\xi_1 \leq \sqrt{\kappa}$ since $\kappa=\varphi+\xi_1+\xi_2+\sqrt{K}\xi_1\xi_2$ and and $\kappa \leq 1$.

Upper bound for $\|\boldsymbol{E}_A^{(m)}\|_{\mathrm{F}},\ m>1$. We consider the third term on the R.H.S. of (15). From (14), for each m, we have the following with probability greater than $1-K\delta-\frac{K}{N^{\frac{1}{8}}}$:

$$\begin{aligned}
\sqrt{K}\varphi \geq \|\boldsymbol{A}_m^{\natural}+\boldsymbol{A}_m^{\natural}\boldsymbol{N}_F-\widehat{\boldsymbol{A}}_m\widehat{\boldsymbol{F}}_1\|_{\mathrm{F}} &= \|\boldsymbol{A}_m^{\natural}+\boldsymbol{A}_m^{\natural}\boldsymbol{N}_F-(\boldsymbol{A}_m^{\natural}\boldsymbol{\Pi}+\boldsymbol{E}_A^{(m)})(\boldsymbol{\Pi}^\top\boldsymbol{I}_K+\boldsymbol{\Pi}^\top\boldsymbol{N}_F+\boldsymbol{E}_F^{(1)})\|_{\mathrm{F}} \\
&= \|\boldsymbol{E}_A^{(m)}(\boldsymbol{\Pi}^\top\boldsymbol{I}_K+\boldsymbol{\Pi}^\top\boldsymbol{N}_F+\boldsymbol{E}_F^{(1)})+\boldsymbol{A}_m^{\natural}\boldsymbol{\Pi}\boldsymbol{E}_F^{(1)}\|_{\mathrm{F}} \\
&= \|\boldsymbol{E}_A^{(m)}\widehat{\boldsymbol{F}}_1+\boldsymbol{A}_m^{\natural}\boldsymbol{\Pi}\boldsymbol{E}_F^{(1)}\|_{\mathrm{F}} \\
&\overset{(a)}{\geq} \left|\|\boldsymbol{E}_A^{(m)}\widehat{\boldsymbol{F}}_1\|_{\mathrm{F}}-\|\boldsymbol{A}_m^{\natural}\boldsymbol{\Pi}\boldsymbol{E}_F^{(1)}\|_{\mathrm{F}}\right| \\
&\overset{(b)}{\geq} \left|\sigma_{\min}(\widehat{\boldsymbol{F}}_1)\|\boldsymbol{E}_A^{(m)}\|_{\mathrm{F}}-\sigma_{\max}(\boldsymbol{A}_m^{\natural})\|\boldsymbol{E}_F^{(1)}\|_{\mathrm{F}}\right| \\
&\overset{(c)}{\geq} \left|\sigma_{\min}(\widehat{\boldsymbol{F}}_1)\|\boldsymbol{E}_A^{(m)}\|_{\mathrm{F}}-(\sqrt{2K\kappa}+\sqrt{K}\xi_2)\sigma_{\max}(\boldsymbol{A}_m^{\natural})\right|,
\end{aligned}$$

where the relation (a) by the triangle inequality, (b) is by applying (27) and (28), and (c) is via (25).

Hence, $\forall m>1$, we get the following with probability greater than $1-2K\delta-\frac{2K}{N^{\frac{1}{8}}}$:

$$\|\boldsymbol{E}_A^{(m)}\|_{\mathrm{F}} \leq \frac{\sqrt{K}\varphi+(\sqrt{2K\kappa}+\sqrt{K}\xi_2)\sigma_{\max}(\boldsymbol{A}_m^{\natural})}{|1-\sqrt{2K\kappa}-2\sqrt{K}\xi_2|} \leq \frac{c_2\sqrt{K\kappa}\sigma_{\max}(\boldsymbol{A}_m^{\natural})}{|1-\sqrt{K\kappa}|} \tag{31}$$

for certain constant $c_2>0$, where we used the fact that $\varphi \leq \sqrt{\kappa}$ and $\xi_2 \leq \sqrt{\kappa}$ since $\kappa=\varphi+\xi_1+\xi_2+\sqrt{K}\xi_1\xi_2$ and $\kappa \leq 1$. In the above, we have also applied $\sigma_{\min}(\widehat{\boldsymbol{F}}_1) \geq |1-\sqrt{2K\kappa}-2\sqrt{K}\xi_2|$ following Lemma 2 and (25).

Putting Together. By applying the upper bounds obtained in (24) and (31), we have the following with probability greater than $1-2K\delta-\frac{2K}{N^{\frac{1}{8}}}$:

$$\|\boldsymbol{E}_A^{(m)}\|_{\mathrm{F}}^2 = \|\widehat{\boldsymbol{A}}_m-\boldsymbol{A}_m^{\natural}\boldsymbol{\Pi}\|_{\mathrm{F}}^2 \leq \frac{c_2^2K^2\kappa}{|1-\sqrt{K\kappa}|^2},\ \forall m.$$

where we have used the fact that $\sigma_{\max}(\boldsymbol{A}_m^{\natural}) \leq \|\boldsymbol{A}_m^{\natural}\|_{\mathrm{F}} \leq \sqrt{K}$.

Similarly, by combining (25) and (30), we have the following with probability greater than $1-(2K+1)\delta-\frac{2K}{N^{\frac{1}{8}}}$

$$\begin{aligned}
\|\widehat{\boldsymbol{F}}-\boldsymbol{\Pi}^\top\boldsymbol{F}^{\natural}\|_{\mathrm{F}}^2 &= \|\boldsymbol{E}_F^{(1)}\|_{\mathrm{F}}^2+\|\boldsymbol{E}_F^{(2)}\|_{\mathrm{F}}^2 \\
&\leq (\sqrt{2K\kappa}+\sqrt{K}\xi_2)^2+\frac{c_1^2(\zeta+\sqrt{K\kappa}\|\boldsymbol{F}^{\natural}\|_{\mathrm{F}})^2}{|1-\sqrt{K\kappa}|^2} \\
&\leq \frac{c_3(\zeta+\sqrt{NK\kappa})^2}{|1-\sqrt{K\kappa}|^2}
\end{aligned} \tag{32}$$

for certain constant $c_3>0$, where we have used the fact that $\|\boldsymbol{F}^{\natural}\|_{\mathrm{F}} \leq \sqrt{N}$.

We hope to characterize the generalization performance of the predicted function $\widehat{\boldsymbol{f}}$ from (32). Towards this, we have the following result:

Lemma 3 *Under Assumptions 1 and 2, the following holds with probability greater than* $1-\delta$*:*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\widehat{\boldsymbol{f}}(\boldsymbol{x})-\boldsymbol{\Pi}^\top\boldsymbol{f}^\natural(\boldsymbol{x})\|_2^2\right]\leq\frac{1}{N}\|\widehat{\boldsymbol{F}}-\boldsymbol{\Pi}^\top\boldsymbol{F}^\natural\|_{\mathrm{F}}^2+64N^{-5/8}\left(2\|\boldsymbol{X}\|_{\mathrm{F}}\mathscr{R}_{\mathcal{F}}\right)^{\frac{1}{4}}+16\sqrt{\frac{2\log(4/\delta)}{N}}. \tag{33}$$

The proof is given in Sec. H. Hence, combining Lemma 3 with (32), we have the following with probability greater than $1-2(K+1)\delta-\frac{2K}{N^{\frac{1}{8}}}$

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\widehat{\boldsymbol{f}}(\boldsymbol{x})-\boldsymbol{\Pi}^\top\boldsymbol{f}^\natural(\boldsymbol{x})\|_2^2\right]\leq\frac{c_3(\zeta+\sqrt{NK\kappa})^2}{N|1-\sqrt{K\kappa}|^2}+64N^{-5/8}\left(2\|\boldsymbol{X}\|_{\mathrm{F}}\mathscr{R}_{\mathcal{F}}\right)^{\frac{1}{4}}+16\sqrt{\frac{2\log(4/\delta)}{N}},$$

where

$$\begin{aligned}
\kappa &\leq \varphi+2\sqrt{K}(\xi_1+\xi_2),\\
\zeta^2 &\leq 8\sqrt{2}NMK\beta\mathfrak{R}+10NM\log(\beta)\frac{\sqrt{2\log(4/\delta)}}{\sqrt{S}}+\beta NM\sqrt{K}\nu\\
\varphi^2 &\leq 8\sqrt{2}N^{1/8}\beta\mathfrak{R}+10\log(\beta)N^{1/8}\frac{\sqrt{2\log(4/\delta)}}{\sqrt{S}}\\
&\quad+\beta N^{1/8}\sqrt{K}\nu+N^{\frac{1}{8}}\sqrt{\frac{\log\beta\log(2/\delta)}{S}},\\
\mathfrak{R} &= \frac{16}{\sqrt{S}}\sqrt{MK\log\left(4K\sqrt{NM}\right)+\left(2\|\boldsymbol{X}\|_{\mathrm{F}}\mathscr{R}_{\mathcal{F}}\right)^{\frac{1}{2}}}.
\end{aligned}$$

C Proof of Proposition 1

Let us consider the following notation:

$$\boldsymbol{P}(\omega)=\boldsymbol{p}_n^{(m)}, \tag{34}$$

where $\omega=(m,n)\in[M]\times[N]$; i.e., $\boldsymbol{P}(\omega)$ "reads out" the (m,n)th block in (10). Similarly, we define $\widehat{\boldsymbol{Y}}(\omega)=\widehat{y}_n^{(m)}$. With this notation, we have

$$D_{\mathcal{S}}(\boldsymbol{P};\widehat{\boldsymbol{Y}})=\frac{1}{S}\sum_{s=1}^{S}\mathsf{CE}(\boldsymbol{P}(\omega_s),\widehat{\boldsymbol{Y}}(\omega_s)),$$

where $\mathsf{CE}(\boldsymbol{x},y)=-\sum_{k=1}^{K}\mathbb{I}[y=k]\log(\boldsymbol{x}(k))$, $S=|\mathcal{S}|$, and $\omega_s=(m_s,n_s)\in\mathcal{S}$. Under Assumption 1, we can define:

$$D_{\Pi}(\boldsymbol{P},\widehat{\boldsymbol{Y}})\triangleq\mathbb{E}_{\mathcal{S}\sim\Pi}[\mathsf{CE}(\boldsymbol{P}(\omega),\widehat{\boldsymbol{Y}}(\omega))]=\sum_{m=1}^{M}\sum_{n=1}^{N}\pi_n^{(m)}\mathsf{CE}(\boldsymbol{p}_n^{(m)},\widehat{y}_n^{(m)}),$$

where Π denotes the uniform distribution and $\pi_n^{(m)}=\frac{1}{NM}$ denotes the probability of observing the annotation for the index pair (m,n).

Eq. (12) implies

$$D_{\mathcal{S}}(\widehat{\boldsymbol{P}};\widehat{\boldsymbol{Y}})\leq D_{\mathcal{S}}(\widetilde{\boldsymbol{P}};\widehat{\boldsymbol{Y}}), \tag{35}$$

where $\widetilde{\boldsymbol{P}}$ is given by:

$$\widetilde{\boldsymbol{P}}=\boldsymbol{W}^\natural\left[\widetilde{\boldsymbol{f}}(\boldsymbol{x}_1)\quad\ldots\quad\widetilde{\boldsymbol{f}}(\boldsymbol{x}_N)\right],$$

and $\widetilde{\boldsymbol{f}}$ is a learning function constructed under Assumption 3. Hence, by taking expectation w.r.t. $\widehat{\boldsymbol{Y}}$ we have

$$\begin{aligned}
\mathbb{E}[D_\Pi(\widehat{\boldsymbol{P}};\widehat{\boldsymbol{Y}}) - D_\Pi(\boldsymbol{P}^\natural;\widehat{\boldsymbol{Y}})] &= \mathbb{E}[D_\Pi(\widehat{\boldsymbol{P}};\widehat{\boldsymbol{Y}})] - D_{\mathcal{S}}(\widehat{\boldsymbol{P}};\widehat{\boldsymbol{Y}}) + D_{\mathcal{S}}(\boldsymbol{P}^\natural;\widehat{\boldsymbol{Y}}) - \mathbb{E}[D_\Pi(\boldsymbol{P}^\natural;\widehat{\boldsymbol{Y}})] \\
&\quad + D_{\mathcal{S}}(\widehat{\boldsymbol{P}};\widehat{\boldsymbol{Y}}) - D_{\mathcal{S}}(\widetilde{\boldsymbol{P}};\widehat{\boldsymbol{Y}}) + D_{\mathcal{S}}(\widetilde{\boldsymbol{P}};\widehat{\boldsymbol{Y}}) - D_{\mathcal{S}}(\boldsymbol{P}^\natural;\widehat{\boldsymbol{Y}}) \\
&\leq \mathbb{E}[D_\Pi(\widehat{\boldsymbol{P}};\widehat{\boldsymbol{Y}})] - D_{\mathcal{S}}(\widehat{\boldsymbol{P}};\widehat{\boldsymbol{Y}}) + D_{\mathcal{S}}(\boldsymbol{P}^\natural;\widehat{\boldsymbol{Y}})] - \mathbb{E}[D_\Pi(\boldsymbol{P}^\natural;\widehat{\boldsymbol{Y}})] \\
&\quad + D_{\mathcal{S}}(\widetilde{\boldsymbol{P}};\widehat{\boldsymbol{Y}}) - D_{\mathcal{S}}(\boldsymbol{P}^\natural;\widehat{\boldsymbol{Y}}) \\
&\leq \sup_{\boldsymbol{P}\in\mathcal{P}} \left| D_{\mathcal{S}}(\boldsymbol{P};\widehat{\boldsymbol{Y}}) - \mathbb{E}[D_\Pi(\boldsymbol{P};\widehat{\boldsymbol{Y}})] \right| + \left| D_{\mathcal{S}}(\boldsymbol{P}^\natural;\widehat{\boldsymbol{Y}}) - \mathbb{E}[D_\Pi(\boldsymbol{P}^\natural;\widehat{\boldsymbol{Y}})] \right| \\
&\quad + \left| D_{\mathcal{S}}(\widetilde{\boldsymbol{P}};\widehat{\boldsymbol{Y}}) - D_{\mathcal{S}}(\boldsymbol{P}^\natural;\widehat{\boldsymbol{Y}}) \right|, \qquad (36)
\end{aligned}$$

where the first inequality is obtained from (35).

Let us consider the L.H.S. of (36):

$$\begin{aligned}
&\mathbb{E}\left[D_\Pi(\widehat{\boldsymbol{P}};\widehat{\boldsymbol{Y}}) - D_\Pi(\boldsymbol{P}^\natural;\boldsymbol{Y})\right] \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M}\left(\pi_n^{(m)}\sum_{k=1}^{K} -\boldsymbol{P}^\natural((m-1)K+k,n)\log\widehat{\boldsymbol{P}}((m-1)K+k,n)\right. \\
&\qquad \left. +\boldsymbol{P}^\natural((m-1)K+k,n)\log\boldsymbol{P}^\natural((m-1)K+k,n)\right) \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M}\pi_n^{(m)}\sum_{k=1}^{K}\boldsymbol{P}^\natural((m-1)K+k,n)\log\frac{\boldsymbol{P}^\natural((m-1)K+k,n)}{\widehat{\boldsymbol{P}}((m-1)K+k,n)} \\
&= \mathsf{D}_{\mathsf{KL}}\left(\boldsymbol{P}^\natural,\widehat{\boldsymbol{P}}\right), \qquad (37)
\end{aligned}$$

where expectation is taken w.r.t. $\widehat{\boldsymbol{Y}}$ (while taking expectation, we used the uniform probability $\pi_n^{(m)} = \frac{1}{NM}, \forall, n, m$ for observing each annotation $\widehat{y}_n^{(m)}$ and used the ground-truth probability $\boldsymbol{P}^\natural((m-1)K+k,n)$ for each event $\mathbb{I}[\widehat{y}_n^{(m)} = k]$) and $\mathsf{D}_{\mathsf{KL}}(\boldsymbol{P}^\natural,\widehat{\boldsymbol{P}})$ is the average Kullback–Leibler (KL) divergence between the entries of the matrices $\boldsymbol{P}^\natural$ and $\widehat{\boldsymbol{P}} \in \mathbb{R}^{MK\times N}$, which is given by

$$\mathsf{D}_{\mathsf{KL}}\left(\boldsymbol{P}^\natural,\widehat{\boldsymbol{P}}\right) = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\mathsf{D}_{\mathsf{KL}}\left(\boldsymbol{p}_n^{\natural(m)},\widehat{\boldsymbol{p}}_n^{(m)}\right).$$

Upper-bounding the first term on the R.H.S of (36)**.** Next, we characterize the first term on the R.H.S. of (36). To achieve this, we invoke the following theorem (Theorem 26.5 in (Shalev-Shwartz & Ben-David, 2014)) :

Theorem 4 *(Shalev-Shwartz & Ben-David, 2014, Theorem 26.5) Assume that for all y and for all $\boldsymbol{x}$, we have $|\mathsf{CE}(\boldsymbol{x};y)| \leq z_{\max}$. Then for any $\boldsymbol{P} \in \mathcal{P}$, the following holds with probability greater than $1-\delta$:*

$$\left| D_{\mathcal{S}}(\boldsymbol{P};\widehat{\boldsymbol{Y}}) - \mathbb{E}[D_\Pi(\boldsymbol{P};\widehat{\boldsymbol{Y}})] \right| \leq 2\mathfrak{R}_{\mathcal{S}}(\ell\circ\mathcal{P}\circ\mathcal{S}) + 4z_{\max}\sqrt{\frac{2\log(4/\delta)}{S}}, \qquad (38)$$

where $\ell\circ\mathcal{P}\circ\mathcal{S}$ denotes the set

$$\ell\circ\mathcal{P}\circ\mathcal{S} \triangleq \left\{\left(\mathsf{CE}(\boldsymbol{P}(\omega_1);\widehat{\boldsymbol{Y}}(\omega_1)),\ldots,\mathsf{CE}(\boldsymbol{P}(\omega_S);\widehat{\boldsymbol{Y}}(\omega_S))\right) \mid \boldsymbol{P}\in\mathcal{P}\right\}$$

and $\mathfrak{R}_{\mathcal{S}}(\mathcal{X})$ denotes the empirical Rademacher complexity of the set $\mathcal{X}$.

To apply Theorem 4, we will characterize $\mathfrak{R}_{\mathcal{S}}(\ell\circ\mathcal{P}\circ\mathcal{S})$ which is defined as follows (Shalev-Shwartz & Ben-David, 2014):

$$\mathfrak{R}_{\mathcal{S}}(\ell\circ\mathcal{P}\circ\mathcal{S}) \triangleq \frac{1}{S}\mathbb{E}\left[\sup_{\boldsymbol{P}\in\mathcal{P}}\sum_{s=1}^{S}\sigma_s\mathsf{CE}(\boldsymbol{P}(\omega_s);\widehat{\boldsymbol{Y}}(\omega_s))\right], \qquad (39)$$

where expectation is w.r.t. the independent Rademacher random variables $\sigma_s \in \{-1,1\}$. Note that $\boldsymbol{P}(\omega)$ is a vector-valued [see Eq. (34)]. Hence, we invoke the following contraction result to upper bound $\mathfrak{R}_{\mathcal{S}}(\ell\circ\mathcal{P}\circ\mathcal{S})$:

Lemma 4 *(Maurer, 2016) Let $\mathcal{P}$ be a class of mappings $\{\boldsymbol{P} : \mathcal{X} \to \mathbb{R}^K\}$, where $\mathcal{X}$ be any set and $(\omega_1, \ldots, \omega_S) \in \mathcal{X}^S$. Also assume that $\ell : \mathbb{R}^K \to \mathbb{R}$ has the Lipschitz constant L. Then*

$$\mathbb{E}\left[\sup_{\boldsymbol{P}\in\mathcal{P}} \sum_{s=1}^{S} \sigma_s \mathsf{CE}(\boldsymbol{P}(\omega_s))\right] \leq \sqrt{2}L\mathbb{E}\left[\sup_{\boldsymbol{P}\in\mathcal{P}} \sum_{s,k} \sigma_{sk}\boldsymbol{P}_k(w_s)\right]$$

where $\boldsymbol{P}_k(w)$ denotes the kth component of $\boldsymbol{P}(\omega)$ and σ_{sk} is an independent (doubly indexed) Rademacher random variable.

Let us define a vector $\boldsymbol{z} \triangleq (\boldsymbol{P}_1(\omega_1), \boldsymbol{P}_2(\omega_1), \ldots, \boldsymbol{P}_{K-1}(\omega_s), \boldsymbol{P}_K(\omega_s)) \in \mathbb{R}^{SK}$ and the set $\mathcal{Z} \triangleq \{\boldsymbol{z} = (\boldsymbol{P}_1(\omega_1), \ldots, \boldsymbol{P}_K(\omega_s)) \mid \boldsymbol{P} \in \mathcal{P}\}$. With these definitions, we apply Lemma 4 in (39) and obtain

$$\begin{aligned}
\mathfrak{R}_{\mathcal{S}}(\ell \circ \mathcal{P} \circ \mathcal{S}) &\leq \frac{\sqrt{2}\beta}{S}\mathbb{E}\left[\sup_{\boldsymbol{P}\in\mathcal{P}} \sum_{s,k} \sigma_{s,k}\boldsymbol{P}_k(w_s)\right] \\
&= \frac{\sqrt{2}\beta}{S}\mathbb{E}\left[\sup_{\boldsymbol{z}\in\mathcal{Z}} \sum_{i} \sigma_i \boldsymbol{z}(i)\right] \\
&= \sqrt{2}\beta K \frac{1}{SK}\mathbb{E}\left[\sup_{\boldsymbol{z}\in\mathcal{Z}} \sum_{i} \sigma_i \boldsymbol{z}(i)\right] = \sqrt{2}\beta K \mathfrak{R}_{\mathcal{S}}(\mathcal{Z}),
\end{aligned} \tag{40}$$

where β is an upper bound of the Lipschitz constant of the cross entropy loss function $\mathsf{CE}(\boldsymbol{x}; y) = -\sum_{k=1}^{K} \mathbb{I}[y = k] \log \boldsymbol{x}(k)$ when $\boldsymbol{x} \in \boldsymbol{\Delta}_K$ with $\boldsymbol{x}(k) > (1/\beta)$. $\forall k$.

Next, we will characterize $\mathfrak{R}_{\mathcal{S}}(\mathcal{Z})$ using the covering number of the set $\mathcal{Z}$.

Definition 2 *(Vershynin, 2012) The ϵ-net covering of the set $\mathcal{Z}$ (denoted as $\overline{\mathcal{Z}}$) is a finite subset of $\mathcal{Z}$ (i.e.,$\overline{\mathcal{Z}} \subseteq \mathcal{Z}$) such that for any $\boldsymbol{z} \in \mathcal{Z}$, there exists an $\overline{\boldsymbol{z}} \in \overline{\mathcal{Z}}$ satisfying $\|\boldsymbol{z} - \overline{\boldsymbol{z}}\|_2^2 \leq \epsilon$. The smallest cardinality of the ϵ-nets of $\overline{\mathcal{Z}}$ is known as the covering number of $\mathcal{Z}$, which is denoted as $\overline{N}(\epsilon, \mathcal{Z})$.*

Let us consider a pair of vectors $\boldsymbol{z}, \overline{\boldsymbol{z}} \in \mathcal{Z}$ as below:

$$\begin{aligned}
\boldsymbol{z} &= (\boldsymbol{P}_1(\omega_1), \ldots, \boldsymbol{P}_K(\omega_s)) \in \mathbb{R}^{SK},\ \boldsymbol{P}(\omega_s) = \boldsymbol{A}_{m_s}\boldsymbol{F}(:, n_s), \\
\overline{\boldsymbol{z}} &= \left(\overline{\boldsymbol{P}}_1(\omega_1), \ldots, \overline{\boldsymbol{P}}_K(\omega_s)\right) \in \mathbb{R}^{SK},\ \overline{\boldsymbol{P}}(\omega_s) = \overline{\boldsymbol{A}}_{m_s}\overline{\boldsymbol{F}}(:, n_s),
\end{aligned}$$

where all $\boldsymbol{A}_m, \overline{\boldsymbol{A}}_m, \boldsymbol{F}$, and $\overline{\boldsymbol{F}}$ satisfy the nonnegativity constraints and have unit ℓ_1-norm on the columns. Then, we have

$$\begin{aligned}
\|\boldsymbol{z} - \overline{\boldsymbol{z}}\| &\leq \|\boldsymbol{P} - \overline{\boldsymbol{P}}\|_{\mathrm{F}} \\
&= \|\boldsymbol{W}\boldsymbol{F} - \overline{\boldsymbol{W}\boldsymbol{F}}\|_2 \\
&= \|\boldsymbol{W}\boldsymbol{F} - \boldsymbol{W}\overline{\boldsymbol{F}} + \boldsymbol{W}\overline{\boldsymbol{F}} - \overline{\boldsymbol{W}\boldsymbol{F}}\|_{\mathrm{F}} \\
&\leq \|\boldsymbol{W}\boldsymbol{F} - \boldsymbol{W}\overline{\boldsymbol{F}}\|_{\mathrm{F}} + \|\boldsymbol{W}\overline{\boldsymbol{F}} - \overline{\boldsymbol{W}\boldsymbol{F}}\|_{\mathrm{F}} \\
&\leq \|\boldsymbol{W}\|_{\mathrm{F}}\|\boldsymbol{F} - \overline{\boldsymbol{F}}\|_{\mathrm{F}} + \|\boldsymbol{W} - \overline{\boldsymbol{W}}\|_{\mathrm{F}}\|\overline{\boldsymbol{F}}\|_{\mathrm{F}} \\
&\leq \sqrt{MK}\|\boldsymbol{F} - \overline{\boldsymbol{F}}\|_{\mathrm{F}} + \sqrt{N}\|\boldsymbol{W} - \overline{\boldsymbol{W}}\|_{\mathrm{F}},
\end{aligned}$$

where the last inequality uses the fact the Frobenius norm of $\boldsymbol{W}$ is bounded by $\sqrt{MK}$ and the Frobenius norm of $\overline{\boldsymbol{F}}$ is bounded by $\sqrt{N}$. Hence, to obtain an ε-net covering for the set $\mathcal{Z}$ (i.e., $\|\boldsymbol{z} - \overline{\boldsymbol{z}}\| \leq \varepsilon$), we only need to show that there exists a $\frac{\varepsilon}{2\sqrt{MK}}$-net covering for $\boldsymbol{F}$ and a $\frac{\varepsilon}{2\sqrt{N}}$-net covering for each $\boldsymbol{W}$ since

$$\sqrt{MK}\|\boldsymbol{F} - \overline{\boldsymbol{F}}\|_{\mathrm{F}} + \sqrt{N}\|\boldsymbol{W} - \overline{\boldsymbol{W}}\|_{\mathrm{F}} \leq \sqrt{MK}\left(\frac{\varepsilon}{2\sqrt{MK}}\right) + \sqrt{N}\left(\frac{\varepsilon}{2\sqrt{N}}\right) = \varepsilon.$$

Note that the full rank matrix $\boldsymbol{W} \in \mathbb{R}^{MK\times K}$ can be represented as an MK^2-dimensional vector whose Euclidean norm is bounded by $\sqrt{MK}$. Hence, the cardinality of the $\frac{\varepsilon}{2\sqrt{N}}$-net covering

for $\boldsymbol{W} \in \mathbb{R}^{MK \times K}$ is at most $\left(\frac{4MK\sqrt{NK}}{\varepsilon}\right)^{MK^2}$ (Shalev-Shwartz & Ben-David, 2014). We have $\boldsymbol{F}(:, n) = \boldsymbol{f}(\boldsymbol{x}_n)$, where $\boldsymbol{f}$ belongs to a function class $\mathcal{F}$. Hence, we use the covering number corresponding to the neural network function class $\mathcal{F}$. Using Lemma 14 of (Lin & Zhang, 2019), we get the cardinality of the $\frac{\varepsilon}{2\sqrt{MK}}$-net covering for $\boldsymbol{F} = [\boldsymbol{f}(\boldsymbol{x}_1), \ldots, \boldsymbol{f}(\boldsymbol{x}_N)],\ \boldsymbol{f} \in \mathcal{F}$ as below:

$$\overline{N}\left(\frac{\varepsilon}{2\sqrt{MK}}, \mathcal{F}\right) \leq \exp\left(\sqrt{\frac{2\sqrt{MK}\|\boldsymbol{X}\|_{\mathrm{F}}\mathscr{R}_{\mathcal{F}}}{\varepsilon}}\right),$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{d \times N}$ and the parameter $\mathscr{R}_{\mathcal{F}}$ is from Assumption 2.

Using the covering number results, the cardinality of the ε-net covering of set $\mathcal{Z}$ is bounded by the following:

$$\overline{N}(\varepsilon, \mathcal{Z}) \leq \left(\frac{4MK\sqrt{NK}}{\varepsilon}\right)^{MK^2} \times \exp\left(\sqrt{\frac{2\sqrt{MK}\|\boldsymbol{X}\|_{\mathrm{F}}\mathscr{R}_{\mathcal{F}}}{\varepsilon}}\right). \tag{41}$$

Now that we have characterized $\overline{N}(\epsilon, \mathcal{Z})$, we invoke the below lemma to obtain the Rademacher complexity $\mathfrak{R}_{\mathcal{S}}(\mathcal{Z})$:

Lemma 5 *(Bartlett et al., 2017, Lemma A.5) The empirical Rademacher complexity of the set $\mathcal{Z}$ with respect to the observed set $\mathcal{S}$ having size SK is upper bounded as follows:*

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{Z}) \leq \inf_{a>0}\left(\frac{4a}{\sqrt{SK}} + \frac{12}{SK}\int_a^{\sqrt{SK}} \sqrt{\log \overline{N}(\mu, \mathcal{Z})}d\mu\right). \tag{42}$$

We apply (41) in Lemma 5. and obtain

$$\begin{aligned}
\mathfrak{R}_{\mathcal{S}}(\mathcal{Z}) &\overset{(a)}{\leq} \inf_{a>0}\left(\frac{4a}{\sqrt{SK}} + \frac{12}{SK}\sqrt{SK}\sqrt{\log \overline{N}(a, \mathcal{Z})}\right) \\
&\overset{(b)}{\leq} \inf_{a>0}\left(\frac{4a}{\sqrt{SK}} + \frac{12}{\sqrt{SK}}\sqrt{MK^2 \log\left(\frac{4MK\sqrt{NK}}{a}\right) + \left(\frac{2\sqrt{MK}\|\boldsymbol{X}\|_{\mathrm{F}}\mathscr{R}_{\mathcal{F}}}{a}\right)^{\frac{1}{2}}}\right) \\
&\overset{(c)}{\leq} \frac{4\sqrt{MK}}{\sqrt{SK}} + \frac{12}{\sqrt{SK}}\sqrt{MK^2 \log\left(4K\sqrt{NM}\right) + (2\|\boldsymbol{X}\|_{\mathrm{F}}\mathscr{R}_{\mathcal{F}})^{\frac{1}{2}}} \\
&\leq \frac{16}{\sqrt{S}}\sqrt{MK \log\left(4K\sqrt{NM}\right) + (2\|\boldsymbol{X}\|_{\mathrm{F}}\mathscr{R}_{\mathcal{F}})^{\frac{1}{2}}}.
\end{aligned} \tag{43}$$

In the above, the first inequality (a) is obtained by using the relation

$$\int_a^{\sqrt{SK}} \sqrt{\log \overline{N}(\mu, \mathcal{Z})}d\mu \ \leq\ \sqrt{SK}\sqrt{\log \overline{N}(a, \mathcal{Z})},$$

which holds because $\sqrt{\log \overline{N}(\mu, \mathcal{Z})}$ decreases monotonically as μ increases. The inequality (b) is obtained by applying (41), and (d) is obtained by fixing $a = \sqrt{MK}$ which is smaller than $\sqrt{SK}$.

Combining the upperbound of $\mathfrak{R}_{\mathcal{S}}(\mathcal{Z})$ given by (43) with the upper bound of $\mathfrak{R}_{\mathcal{S}}(\ell \circ \mathcal{P} \circ \mathcal{S})$ as given by (40) with the result in (38), we get that with probability greater than $1 - \delta$,

$$\left|D_{\mathcal{S}}(\boldsymbol{P}; \widehat{\boldsymbol{Y}}) - \mathbb{E}[D_{\Pi}(\boldsymbol{P}; \widehat{\boldsymbol{Y}})]\right| \leq 2\sqrt{2}\beta K\mathfrak{R}_{\mathcal{S}}(\mathcal{Z}) + 4z_{\max}\sqrt{\frac{2\log(4/\delta)}{S}}, \tag{44}$$

where $\mathfrak{R}_{\mathcal{S}}(\mathcal{Z})$ is upper bounded by (43) and $z_{\max}$ is the upperbound of the value of the function $\mathsf{CE}(\boldsymbol{x}, y)$ which can be characterized as below:

$$z_{\max} = \max_{\substack{\boldsymbol{x}(k) > \frac{1}{\beta} \\ y \in [K]}} \mathsf{CE}(\boldsymbol{x}, y) \leq \max_{\substack{\boldsymbol{x}(k) > \frac{1}{\beta} \\ y \in [K]}} -\sum_{k=1}^{K} \mathbb{I}[y = k] \log \boldsymbol{x}(k) \leq \max_{u > \frac{1}{\beta}} -\log u = \log(\beta). \tag{45}$$

Upper-bounding the second term on the R.H.S of (36)**.** Next, we proceed to upper bound the second term on the R.H.S of (36). Let us consider the Hoeffding's inequality

Lemma 6 *Let $Z_1, \ldots, Z_S$ be independent bounded random variables with $Z_s \in [z_{\min}, z_{\max}]$ for all s where $-\infty < z_{\min} \leq z_{\max} < \infty$. Then for all $t \geq 0$,*

$$\mathsf{Pr}\left(\frac{1}{S}\sum_{s=1}^{S}(Z_s - \mathbb{E}[Z_s]) \geq t\right) \leq \exp\left(-\frac{2St^2}{(z_{\max} - z_{\min})^2}\right).$$

To use Lemma 6, let us define the random variable $Z_n^{(m)}$ as follows:

$$Z_n^{(m)} \triangleq \mathsf{CE}(\boldsymbol{p}_n^{\natural(m)}, \widehat{y}_n^{(m)}),$$

where $\boldsymbol{p}_n^{\natural(m)} = \boldsymbol{A}_m^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_n)$. The maximum and minimum values of $Z_n^{(m)}$ are $z_{\max} = \log(\beta)$ (see (45)) and $z_{\min} = 0$, respectively. Then, invoking Lemma 6, one can obtain

$$\mathsf{Pr}\left(D_{\mathcal{S}}(\boldsymbol{P}^{\natural}; \widehat{\boldsymbol{Y}})] - \mathbb{E}[D_{\Pi}(\boldsymbol{P}^{\natural}; \widehat{\boldsymbol{Y}})] \geq t\right) \leq \exp\left(-\frac{2St^2}{(\log(\beta))^2}\right). \tag{46}$$

Hence, by substituting $t = \log(\beta)\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2S}}$, where $\delta \in (0,1)$ in (46), we get that with probability greater than $1-\delta$

$$D_{\mathcal{S}}(\boldsymbol{P}^{\natural}; \widehat{\boldsymbol{Y}})] - \mathbb{E}[D_{\Pi}(\boldsymbol{P}^{\natural}; \widehat{\boldsymbol{Y}})] \leq \log(\beta)\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2S}}. \tag{47}$$

Upper-bounding the third term on the R.H.S of (36)**.**

$$\begin{aligned}
\left|D_{\mathcal{S}}(\widetilde{\boldsymbol{P}}; \widehat{\boldsymbol{Y}}) - D_{\mathcal{S}}(\boldsymbol{P}^{\natural}; \widehat{\boldsymbol{Y}})\right| &= \Bigg|-\frac{1}{S}\sum_{(m,n)\in\mathcal{S}}\sum_{k=1}^{K}\mathbb{I}[\widehat{y}_n^{(m)} = k]\log\widetilde{\boldsymbol{P}}((m-1)K+k, n) \\
&\quad + \frac{1}{S}\sum_{(m,n)\in\mathcal{S}}\sum_{k=1}^{K}\mathbb{I}[\widehat{y}_n^{(m)} = k]\log\boldsymbol{P}^{\natural}((m-1)K+k, n)\Bigg| \\
&\leq \frac{1}{S}\sum_{(m,n)\in\mathcal{S}}\sum_{k=1}^{K}\mathbb{I}[\widehat{y}_n^{(m)} = k]\left|\log[\boldsymbol{A}_m^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_n)]_k - \log[\boldsymbol{A}_m^{\natural}\widetilde{\boldsymbol{f}}(\boldsymbol{x}_n)]_k\right| \\
&\leq \frac{1}{S}\sum_{(m,n)\in\mathcal{S}}\sum_{k=1}^{K}\mathbb{I}[\widehat{y}_n^{(m)} = k]\beta\left|[\boldsymbol{A}_m^{\natural}\boldsymbol{f}^{\natural}(\boldsymbol{x}_n)]_k - [\boldsymbol{A}_m^{\natural}\widetilde{\boldsymbol{f}}(\boldsymbol{x}_n)]_k\right| \\
&\leq \frac{1}{S}\sum_{(m,n)\in\mathcal{S}}\sum_{k=1}^{K}\mathbb{I}[\widehat{y}_n^{(m)} = k]\beta\|\boldsymbol{A}_m^{\natural}(k,:)\|_2\|\boldsymbol{f}^{\natural}(\boldsymbol{x}_n) - \widetilde{\boldsymbol{f}}(\boldsymbol{x}_n)\|_2 \\
&\leq \beta\sqrt{K}\nu,
\end{aligned} \tag{48}$$

where the first inequality uses the triangle inequality, the second inequality uses the concavity of the $\log$ function, the third inequality is via Cauchy Schwartz inequality, and the last inequality employs Assumption 3.

Putting Together. Hence, by combining the result in (44) with (37), (36), (47), and (48), we get that with probability greater than $1-2\delta$,

$$\mathsf{D}_{\mathsf{KL}}\left(\boldsymbol{P}^{\natural}, \widehat{\boldsymbol{P}}\right) \leq 4\sqrt{2}\beta K\mathfrak{R}_{\mathcal{S}}(\mathcal{Z}) + 8\log(\beta)\sqrt{\frac{2\log(4/\delta)}{S}} + \log(\beta)\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2S}} + \beta\sqrt{K}\nu. \tag{49}$$

Using Pinsker's inequality (S, 1960; Fedotov et al., 2003), we get

$$\mathsf{D}_{\mathsf{KL}}\left(\boldsymbol{P}^{\natural}, \widehat{\boldsymbol{P}}\right) = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\mathsf{D}_{\mathsf{KL}}\left(\boldsymbol{p}_n^{\natural(m)}, \widehat{\boldsymbol{p}}_n^{(m)}\right) \geq \frac{1}{2NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_1^2$$
$$\geq \frac{1}{2NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2$$

where the last inequality uses the fact that $\|\boldsymbol{x}\|_1 \geq \|\boldsymbol{x}\|_2$. The above relation combined with (49), implies that with probability greater than $1-2\delta$:

$$\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2 \leq 8\sqrt{2}K\beta\mathfrak{R}_{\mathcal{S}}(\mathcal{Z}) + 10\log(\beta)\frac{\sqrt{2\log(4/\delta)}}{\sqrt{S}} + \beta\sqrt{K}\nu, \quad (50)$$

where $\mathfrak{R}_{\mathcal{S}}(\mathcal{Z})$ is upper-bounded in (43)

Next, we aim to bound each term $\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2$. Towards this, we have the below result:

Lemma 7 *Suppose that* $\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2$ *is upper-bounded. Then, with probability greater than* $1-\delta-\frac{1}{N^{\frac{1}{8}}}$*, we have for any* m, n*,*

$$\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2 \leq N^{\frac{1}{8}}\left(\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2\right) + N^{\frac{1}{8}}\sqrt{\frac{\log\beta\log(1/\delta)}{S}}. \quad (51)$$

D Proof of Theorem 2

The proof of Theorem 2 utilizes some key results from the proof of Theorem 1. We start by employing the result in Proposition 1. It implies that when $\boldsymbol{f}^{\natural} \in \mathcal{F}$, i.e., $\nu = 0$, $S \geq C_1 t$, $N \leq C_2 t^3$, for certain constants $C_1, C_2 > 0$ and at the limit of $t \to \infty$, we get the following:

$$\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2 = 0, \forall m, n$$

with probability 1. Since all the entries of the matrices $\boldsymbol{P}^{\natural}$ and $\widehat{\boldsymbol{P}}$ are nonnegative, it implies that $\widehat{\boldsymbol{P}} = \boldsymbol{P}^{\natural}$, i.e.,

$$\widehat{\boldsymbol{P}} = \boldsymbol{W}^{\natural}\boldsymbol{F}^{\natural}. \quad (52)$$

On the other hand, the matrix $\widehat{\boldsymbol{P}}$ can be constructed using the estimates of the CCEM criterion (6) via

$$\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{W}}\widehat{\boldsymbol{F}}. \quad (53)$$

Hence, in order to identify $\boldsymbol{W}^{\natural}$ and $\boldsymbol{F}^{\natural}$ from the NMF model in (52), we invoke the following result:

Lemma 8 *(Gillis, 2020) Consider the matrix factorization model* $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{Y}$*, where* $\boldsymbol{X} \in \mathbb{R}^{I\times K}$*,* $\boldsymbol{Y} \in \mathbb{R}^{K\times J}$ *and* $\mathsf{rank}(\boldsymbol{X}) = \mathsf{rank}(\boldsymbol{Y}) = K$*. If* $\boldsymbol{X}, \boldsymbol{Y} \geq \boldsymbol{0}$ *and both* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *satisfy SSC, then any* $\widehat{\boldsymbol{X}}$ *and* $\widehat{\boldsymbol{Y}}$ *that satisfy* $\boldsymbol{Z} = \widehat{\boldsymbol{X}}\widehat{\boldsymbol{Y}}$ *must have the following form:*

$$\widehat{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{\Pi}\boldsymbol{\Sigma},\ \widehat{\boldsymbol{Y}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Pi}^{\top}\boldsymbol{Y},$$

where $\boldsymbol{\Pi}$ *is a column permutation matrix and* $\boldsymbol{\Sigma}$ *is a diagonal nonnegative and scaling matrix.*

Next, we show that Lemma 8 can be applied given the conditions in Theorem 2. The matrix $\boldsymbol{F}_{\mathcal{Z}}^{\natural} = [\boldsymbol{f}^{\natural}(\boldsymbol{x}_{n_1}), \ldots, \boldsymbol{f}^{\natural}(\boldsymbol{x}_{n_T})]$ includes a subset of columns of $\boldsymbol{F}^{\natural}$, i.e., $\mathrm{cone}(\boldsymbol{F}_{\mathcal{Z}}^{\natural}) \subseteq \mathrm{cone}(\boldsymbol{F}^{\natural})$. Hence, we have

$$\mathcal{C} \subseteq \mathrm{cone}(\boldsymbol{F}_{\mathcal{Z}}^{\natural}) \implies \mathcal{C} \subseteq \mathrm{cone}(\boldsymbol{F}^{\natural}), \quad (54)$$

where $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^K | \sqrt{K-1}\|\boldsymbol{x}\|_2 \leq \mathbf{1}^\top \boldsymbol{x}\}$ is the second-order cone.

Also for any orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{K \times K}$ except for the permutation matrices, the below holds:

$$\text{cone}(\boldsymbol{F}_{\mathcal{Z}}^\natural) \not\subset \text{cone}\{\boldsymbol{Q}\} \implies \text{cone}(\boldsymbol{F}^\natural) \not\subset \text{cone}\{\boldsymbol{Q}\}. \tag{55}$$

Eqs (54) and (55) imply that $\boldsymbol{F}^\natural$ satisfies SSC, given the assumption that $\boldsymbol{F}_{\mathcal{Z}}^\natural$ satisfies SSC. In addition, the column scaling does not affect the conic hull, i.e,

$$\text{cone}(\boldsymbol{F}^\natural) = \text{cone}(\boldsymbol{\Sigma}^{-1}\boldsymbol{F}^\natural),$$
$$\text{cone}(\boldsymbol{W}^{\natural\top}) = \text{cone}(\boldsymbol{\Sigma}\boldsymbol{W}^{\natural\top}).$$

Since both $\boldsymbol{W}^\natural$ and $\boldsymbol{F}^\natural$ satisfy SSC, $\mathsf{rank}(\boldsymbol{F}^\natural) = \mathsf{rank}(\boldsymbol{W}^\natural) = K$ hold (see (Gillis, 2020)). Hence, the conditions in Lemma 8 holds for (52). Comparing (52) and (53) and invoking Lemma 8, we get

$$\widehat{\boldsymbol{W}} = \boldsymbol{W}^\natural \boldsymbol{\Pi}, \tag{56a}$$
$$\widehat{\boldsymbol{F}} = \boldsymbol{\Pi}^\top \boldsymbol{F}^\natural, \tag{56b}$$

where the scaling $\boldsymbol{\Sigma}$ is automatically removed due to the sum-to-one constraints on the columns of $\widehat{\boldsymbol{A}}_m$'s and $\widehat{\boldsymbol{F}}$.

The first result (56a) implies that

$$\widehat{\boldsymbol{A}}_m = \boldsymbol{A}_m^\natural \boldsymbol{\Pi}, \forall m.$$

Applying the second result (56b) in Lemma 3 along with the assumption that $N \to \infty$, we get

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\widehat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x})\|_2^2\right] = 0. \tag{57}$$

Fact 1 Let X be a nonnegative random variable with $\mathbb{E}[X] = 0$. Then, X is zero almost surely, i.e.,

$$\mathsf{Pr}(X = 0) = 1.$$

Employing Fact 1 in (57), we get that

$$\|\widehat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x})\|_2^2 = 0, \forall \boldsymbol{x} \sim \mathcal{D}, \implies \widehat{\boldsymbol{f}}(\boldsymbol{x}) = \boldsymbol{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x}), \forall \boldsymbol{x} \sim \mathcal{D},$$

due to the nonnegativity of $\boldsymbol{f}^\natural$ and $\widehat{\boldsymbol{f}}$.

E Proof of Theorem 3

To prove Theorem 3, let us first consider the CCEM criterion in (6). Let $\widehat{\boldsymbol{W}}$ and $\widehat{\boldsymbol{F}}$ be any optimal solution of (6) and

$$\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{W}}\widehat{\boldsymbol{F}}. \tag{58}$$

From Proposition 1, using the conditions in the statement of Theorem 3, i.e., $\boldsymbol{f}^\natural \in \mathcal{F}$ implying $\nu = 0$, $S \geq C_1 t$, $N \leq C_2 t^3$, for certain constants $C_1, C_2 > 0$, we get the following at the limit of $t \to \infty$:

$$\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2 = 0, \forall m, n,$$

with probability 1.

Since all the entries of the matrices $\boldsymbol{P}^\natural$ and $\widehat{\boldsymbol{P}}$ are nonnegative, it implies that $\widehat{\boldsymbol{P}} = \boldsymbol{P}^\natural$, i.e.,

$$\widehat{\boldsymbol{P}} = \boldsymbol{W}^\natural \boldsymbol{F}^\natural. \tag{59}$$

Proof of Theorem 3(a): We consider the following result which is distilled and summarized from the proof of Theorem 1 in (Fu et al., 2015):

Lemma 9 *Suppose a matrix $\boldsymbol{Y} \in \mathbb{R}^{K \times J}$ satisfies $\boldsymbol{Y} \geq \boldsymbol{0}$, $\boldsymbol{1}^\top \boldsymbol{Y} = \boldsymbol{1}^\top$, $\mathsf{rank}(\boldsymbol{Y}) = K$, and SSC. Then, for any $\widehat{\boldsymbol{Y}} = \boldsymbol{Q}\boldsymbol{Y}$ satisfying $\widehat{\boldsymbol{Y}} \geq \boldsymbol{0}$, $\boldsymbol{1}^\top \widehat{\boldsymbol{Y}} = \boldsymbol{1}^\top$, the following holds:*

$$|\det(\boldsymbol{Q})| \leq 1,$$

The equality holds only if $\boldsymbol{Q}$ is a permutation matrix.

Let us start by considering the following criterion:

$$\underset{\boldsymbol{W},\boldsymbol{F}}{\text{maximize}} \ \det(\boldsymbol{F}\boldsymbol{F}^\top) \tag{60a}$$

$$\text{s.t. } \widehat{\boldsymbol{P}} = \boldsymbol{W}\boldsymbol{F} \tag{60b}$$

$$\boldsymbol{F} \geq \boldsymbol{0}, \boldsymbol{1}^\top \boldsymbol{F} = \boldsymbol{1}^\top \tag{60c}$$

where $\widehat{\boldsymbol{P}}$ is the optimal solution from the CCEM criterion (6) and hence, as shown before, $\widehat{\boldsymbol{P}}$ satisfies (59).

Let $\boldsymbol{W}^*$ and $\boldsymbol{F}^*$ be optimal solutions of (60), then the following holds:

$$\det(\boldsymbol{F}\boldsymbol{F}^{*\top}) \geq \det(\boldsymbol{F}^\natural \boldsymbol{F}^{\natural\top}). \tag{61}$$

From (59), we observe that $\boldsymbol{W}^*$ and $\boldsymbol{F}^*$ satisfies $\boldsymbol{W}^* = \boldsymbol{W}^\natural \boldsymbol{Q}^{-1}$, $\boldsymbol{F}^* = \boldsymbol{Q}\boldsymbol{F}^\natural$, for a certain invertible matrix $\boldsymbol{Q}$. Let us assume that $\boldsymbol{Q}$ is not a permutation matrix, Then, we have

$$\begin{aligned}\det(\boldsymbol{F}^*\boldsymbol{F}^{*\top}) &= \det(\boldsymbol{Q}^\top \boldsymbol{F}^\natural \boldsymbol{F}^{\natural\top} \boldsymbol{Q}) \\ &= |\det(\boldsymbol{Q})|^2 \det(\boldsymbol{F}^\natural \boldsymbol{F}^{\natural\top}) \\ &< \det(\boldsymbol{F}^\natural \boldsymbol{F}^{\natural\top})\end{aligned}$$

where the last inequality is from Lemma 9 using the SSC condition on $\boldsymbol{F}^\natural$. Note that the result is a contradiction from (61). Hence, $\boldsymbol{Q}$ must be a permutation matrix. This implies that the optimal solution $\boldsymbol{W}^*$ and $\boldsymbol{F}^*$ of Problem 60 satisfies the following:

$$\boldsymbol{W}^* = \boldsymbol{W}^\natural \boldsymbol{\Pi}, \tag{62a}$$

$$\boldsymbol{F}^* = \boldsymbol{\Pi}^\top \boldsymbol{F}^\natural, \tag{62b}$$

Following the last part of Theorem 2, by using Lemma 3 and Fact 1, we have

$$\boldsymbol{A}_m^* = \boldsymbol{A}_m^\natural \boldsymbol{\Pi}, \forall m.$$
$$\boldsymbol{f}^*(\boldsymbol{x}) = \boldsymbol{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x}), \forall \boldsymbol{x} \sim \mathcal{D}.$$

Note that since $\log$ is a monotonically increasing function, by using $\log\det(\boldsymbol{F}\boldsymbol{F}^\top)$ in (60) in place of $\det(\boldsymbol{F}\boldsymbol{F})^\top$ does not change the optimal solution, yet it keeps the objective function in differentiable domain (this is because, $\det(\boldsymbol{X})$ becomes zero if $\boldsymbol{X}$ is singular). Hence, we can conclude that the following criterion

$$\underset{\boldsymbol{W},\boldsymbol{F}}{\text{maximize}} \ \log\det(\boldsymbol{F}\boldsymbol{F}^\top) \tag{63a}$$

$$\text{s.t. } \widehat{\boldsymbol{P}} = \boldsymbol{W}\boldsymbol{F} \tag{63b}$$

$$\boldsymbol{F} \geq \boldsymbol{0}, \tag{63c}$$

also results in the same optimal solutions $\boldsymbol{W}^*$ and $\boldsymbol{F}^*$ satisfying (62).

Proof of Theorem 3(b): We consider the following result:

Lemma 10 *(Huang et al., 2015) Suppose a matrix $\boldsymbol{X} \in \mathbb{R}^{I \times K}$ satisfies $\boldsymbol{X} \geq \boldsymbol{0}$, $\boldsymbol{1}^\top \boldsymbol{X} = \rho \boldsymbol{1}^\top$, $\rho > 0$, $\mathsf{rank}(\boldsymbol{X}) = K$, and SSC. Then, for any $\widehat{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{Q}$ satisfying $\widehat{\boldsymbol{X}} \geq \boldsymbol{0}$, $\boldsymbol{1}^\top \widehat{\boldsymbol{X}} = \rho \boldsymbol{1}^\top$, the following holds:*

$$|\det(\boldsymbol{Q})| \leq 1,$$

The equality holds only if $\boldsymbol{Q}$ is a permutation matrix.

Note that, in Lemma 10, we have $\mathbf{1}^\top \boldsymbol{X} = \rho \mathbf{1}^\top$. Then, we have

$$\mathbf{1}^\top \boldsymbol{X}\boldsymbol{Q} = \rho \mathbf{1}^\top \boldsymbol{Q}$$

Since $\mathbf{1}^\top \boldsymbol{X}\boldsymbol{Q} = \rho \mathbf{1}^\top$ holds by the assumption, we get

$$\rho \mathbf{1}^\top \boldsymbol{Q} = \rho \mathbf{1}^\top \implies \mathbf{1}^\top \boldsymbol{Q} = \mathbf{1}^\top. \tag{64}$$

The result in (64) is used in order to conclude Lemma 10 from the results in (Huang et al., 2015).

We consider the following criterion:

$$\underset{\boldsymbol{W},\boldsymbol{F}}{\text{maximize}} \ \det(\boldsymbol{W}^\top \boldsymbol{W}) \tag{65a}$$

$$\text{s.t. } \widehat{\boldsymbol{P}} = \boldsymbol{W}\boldsymbol{F} \tag{65b}$$

$$\boldsymbol{W} \geq \mathbf{0}, \mathbf{1}^\top \boldsymbol{W} = M\mathbf{1}^\top, \tag{65c}$$

where we used the constraint $\mathbf{1}^\top \boldsymbol{W} = M\mathbf{1}^\top$ since $\mathbf{1}^\top \boldsymbol{A}_m = \mathbf{1}^\top$ for all m.

Let $\boldsymbol{W}^*$ and $\boldsymbol{F}^*$ be optimal solutions of (65), then we have:

$$\det(\boldsymbol{W}^{*\top} \boldsymbol{W}^*) \geq \det(\boldsymbol{W}^{\natural\top} \boldsymbol{W}^\natural). \tag{66}$$

Applying (59), $\boldsymbol{W}^*$ and $\boldsymbol{F}^*$ satisfies $\boldsymbol{W}^* = \boldsymbol{W}^\natural \boldsymbol{Q},\ \boldsymbol{F}^* = \boldsymbol{Q}^{-1}\boldsymbol{F}^\natural$, for a certain invertible matrix $\boldsymbol{Q}$. Let us assume that $\boldsymbol{Q}$ is not a permutation matrix, Then, we have

$$\begin{aligned}\det(\boldsymbol{W}^{*\top} \boldsymbol{W}^*) &= \det(\boldsymbol{Q}^\top \boldsymbol{W}^{\natural\top} \boldsymbol{W}^\natural \boldsymbol{Q}) \\ &= |\det(\boldsymbol{Q})|^2 \det(\boldsymbol{W}^{\natural\top} \boldsymbol{W}^\natural) \\ &< \det(\boldsymbol{W}^{\natural\top} \boldsymbol{W}^\natural)\end{aligned}$$

where the last inequality is from Lemma 10 using the SSC condition on $\boldsymbol{W}^\natural$ which is a contradiction from (66). Hence, $\boldsymbol{Q}$ must be a permutation matrix. This implies that the optimal solution $\boldsymbol{W}^*$ and $\boldsymbol{F}^*$ of Problem 65 satisfies the following:

$$\boldsymbol{W}^* = \boldsymbol{W}^\natural \boldsymbol{\Pi}, \tag{67a}$$

$$\boldsymbol{F}^* = \boldsymbol{\Pi}^\top \boldsymbol{F}^\natural, \tag{67b}$$

Similar to the last part of Theorem 2, by using Lemma 3 and Fact 1, we have

$$\boldsymbol{A}_m^* = \boldsymbol{A}_m^\natural \boldsymbol{\Pi}, \forall m.$$
$$\boldsymbol{f}^*(\boldsymbol{x}) = \boldsymbol{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x}), \forall \boldsymbol{x} \sim \mathcal{D}.$$

In this case as well, the optimal solutions do not change if we employ $\log\det(\boldsymbol{W}^\top \boldsymbol{W})$ in (65) in place of $\det(\boldsymbol{W}^\top \boldsymbol{W})$.

Remark 3 Geometrically, Theorem 3 seeks maximum volume solutions w.r.t. the conic hulls of $\boldsymbol{F}$ and $\boldsymbol{W}^\top$, respectively. One can also note that in both cases of Theorem 3, the corresponding optimal solutions do not change if we minimize $\log\det(\boldsymbol{W}^\top \boldsymbol{W})$ in (60) and minimize $\log\det(\boldsymbol{F}\boldsymbol{F}^\top)$ in (65). However, relying on the SSC of $\boldsymbol{F}^\natural$ ($\boldsymbol{W}^\natural$) and minimizing the volume of conic hull of $\boldsymbol{W}^\top$ ($\boldsymbol{F}$) may be inefficient since $\boldsymbol{P}^\natural = \boldsymbol{W}^\natural \boldsymbol{F}^\natural$ does not hold exactly in practice.

F Proof of Lemma 1

Let us define some notations:

$$\ell_k^* = \arg\max_{\ell \in [K]} \widehat{\boldsymbol{A}}_1(k, \ell)$$
$$q_k^* = \arg\max_{q \in [K]} \widehat{\boldsymbol{F}}_1(q, k).$$

Consider the following conditions given in Lemma 1

$$1 - \kappa \leq |\widehat{\boldsymbol{A}}_1(k,:)\widehat{\boldsymbol{F}}_1(:,k)| \leq 1 + \kappa,\ \forall k \in [K] \tag{68a}$$

$$|\widehat{\boldsymbol{A}}_1(k,:)\widehat{\boldsymbol{F}}_1(:,j)| \leq \kappa, \forall k \neq j. \tag{68b}$$

F.1 PROVING $\ell_k^* \neq q_j^*,\ \forall k \neq j$

We begin by noticing the below relation from (68a):

$$\begin{aligned}(1-\kappa) &\leq \widehat{\boldsymbol{A}}_1(k,1)\widehat{\boldsymbol{F}}_1(1,k) + \cdots + \widehat{\boldsymbol{A}}_1(k,K)\widehat{\boldsymbol{F}}_1(K,k) \\ &\leq \max_{\ell\in[K]} \widehat{\boldsymbol{A}}_1(k,\ell)(\widehat{\boldsymbol{F}}_1(1,k) + \cdots + \widehat{\boldsymbol{F}}_1(K,k)) \\ &= \max_{\ell\in[K]} \widehat{\boldsymbol{A}}_1(k,\ell) \\ &= \widehat{\boldsymbol{A}}_1(k,\ell_k^*),\ \forall k, \end{aligned} \tag{69}$$

where the first equality employs the probability simplex constraints on the columns of $\widehat{\boldsymbol{F}}$.

We further proceed to prove by using contradiction. First, assume the below for certain $k \neq j$,

$$\ell_k^* = \arg\max_\ell \widehat{\boldsymbol{A}}_1(k,\ell) = \arg\max_q \widehat{\boldsymbol{F}}_1(q,j) = q_j^*. \tag{70}$$

Under the assumption (70), consider (68b). For (68b) to hold, we need to satisfy the below for a certain $k \neq j$:

$$\begin{aligned}&\widehat{\boldsymbol{A}}_1(k,\ell_k^*)\widehat{\boldsymbol{F}}_1(q_j^*,j) < \kappa \\ &\implies \widehat{\boldsymbol{F}}_1(q_j^*,j) < \frac{\kappa}{1-\kappa}\end{aligned}$$

where the last relation is obtained from (69). This also implies that

$$\begin{aligned}1 = \sum_{q=1}^{K} \widehat{\boldsymbol{F}}_1(q,k) &< \frac{K\kappa}{1-\kappa} \\ \implies \kappa &> \frac{1}{K+1}.\end{aligned}$$

However, Lemma 1 assumes that $\kappa \leq \frac{1}{K+1}$, Hence, the assumption (70) does not hold and we get

$$\arg\max_\ell \widehat{\boldsymbol{A}}_1(k,\ell) \neq \arg\max_q \widehat{\boldsymbol{F}}_1(q,j),\ k \neq j. \tag{71}$$

F.2 PROVING $\ell_k^* = q_k^*,\ \forall k$

From the result in (71), we consider (68b) for a certain $k \neq j$.

$$\begin{aligned}&\widehat{\boldsymbol{A}}_1(k,\ell_k^*)\widehat{\boldsymbol{F}}_1(q,j) \leq \kappa \\ &\implies \widehat{\boldsymbol{F}}_1(q,j) \leq \frac{\kappa}{1-\kappa},\ q \neq q_j^*.\end{aligned} \tag{72}$$

where the last relation is obtained from (69).

Since $\sum_q \widehat{\boldsymbol{F}}_1(q,j) = 1$, we get

$$\widehat{\boldsymbol{F}}_1(q_j^*,j) \geq \frac{1-K\kappa}{1-\kappa}, \forall j. \tag{73}$$

From the result in (71) and the condition in (68b), we also have the following for a certain $k \neq j$

$$\begin{aligned}&\widehat{\boldsymbol{A}}_1(k,\ell)\widehat{\boldsymbol{F}}_1(q_j^*,j) \leq \kappa \\ &\implies \widehat{\boldsymbol{A}}_1(k,\ell) \leq \frac{\kappa(1-\kappa)}{1-K\kappa},\ \ell \neq \ell_k^*,\end{aligned} \tag{74}$$

where the last relation by (73).

Next, we proceed to prove by contradiction. First, assume the below for certain k

$$\ell_k^* \neq q_k^*. \tag{75}$$

Hence, for each k, we have

$$\begin{aligned}|\widehat{\boldsymbol{A}}_1(k,:)\widehat{\boldsymbol{F}}_1(:,k)| &= \widehat{\boldsymbol{A}}_1(k,1)\widehat{\boldsymbol{F}}_1(1,k) + \cdots + \widehat{\boldsymbol{A}}_1(k,K)\widehat{\boldsymbol{F}}_1(K,k) \\ &\leq \sum_{q\neq q_k^*} \widehat{\boldsymbol{F}}_1(q,k) + \widehat{\boldsymbol{A}}_1(k,\ell),\ \ell \neq \ell_k^* \\ &\leq \frac{(K-1)\kappa}{1-\kappa} + \frac{\kappa(1-\kappa)}{1-K\kappa} \end{aligned} \tag{76}$$

where we have used the fact that all the entries of $\widehat{\boldsymbol{A}}_1$ and $\widehat{\boldsymbol{F}}_1$ are smaller than 1, in the first inequality and have applied (72) and (74) in the last inequality.

The lower-bound condition in (68a) gives that for each k,

$$|\widehat{\boldsymbol{A}}_1(k,:)\widehat{\boldsymbol{F}}_1(:,k)| \geq 1-\kappa. \tag{77}$$

Then, comparing both (76) and (77), we hope to have

$$\begin{aligned}&\frac{(K-1)\kappa}{1-\kappa} \geq 1-\kappa \\ &\Rightarrow \kappa \geq \frac{K+1-\sqrt{(K+1)^2-4}}{2}\end{aligned}$$

However, $\frac{K+1-\sqrt{(K+1)^2-4}}{2} \geq \frac{1}{K+1}$, which leads to a contradiction to our assumption that $\kappa \leq \frac{1}{K+1}$. Hence, the assumption (75) does not hold and we get

$$\text{arg max}_\ell \widehat{\boldsymbol{A}}_1(k,\ell) = \text{arg max}_q \widehat{\boldsymbol{F}}_1(q,k),\ \forall k. \tag{78}$$

Combining both (78) and (78), we get

$$\begin{aligned}\text{arg max}_\ell \widehat{\boldsymbol{A}}_1(k,\ell) &\neq \text{arg max}_\ell \widehat{\boldsymbol{A}}_1(j,\ell),\ \forall k \neq j \\ \text{arg max}_q \widehat{\boldsymbol{F}}_1(q,k) &\neq \text{arg max}_q \widehat{\boldsymbol{F}}_1(q,j),\ \forall k \neq j.\end{aligned}$$

G Proof of Lemma 2

We have

$$\begin{aligned}\sigma_{\min}(\boldsymbol{X}) = \sigma_{\min}(\boldsymbol{I}_K + \boldsymbol{E}_1 + \boldsymbol{E}_2) &= \min_{\|\boldsymbol{x}\|_2=1} \|(\boldsymbol{I}_K + \boldsymbol{E}_1 + \boldsymbol{E}_2)\boldsymbol{x}\|_2 \\ &= \min_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{I}_K\boldsymbol{x} + (\boldsymbol{E}_1+\boldsymbol{E}_2)\boldsymbol{x}\|_2 \overset{(a)}{\geq} \min_{\|\boldsymbol{x}\|_2=1} |\|\boldsymbol{I}_K\boldsymbol{x}\|_2 - \|(\boldsymbol{E}_1+\boldsymbol{E}_2)\boldsymbol{x}\|_2| \\ &\geq \left|\min_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{I}_K\boldsymbol{x}\|_2 - \max_{\|\boldsymbol{x}\|_2=1} \|(\boldsymbol{E}_1+\boldsymbol{E}_2)\boldsymbol{x}\|_2\right| = \left|1 - \max_{\|\boldsymbol{x}\|_2=1} \|(\boldsymbol{E}_1+\boldsymbol{E}_2)\boldsymbol{x}\|_2\right| \\ &= |1 - \|\boldsymbol{E}_1+\boldsymbol{E}_2\|_2| \overset{(b)}{\geq} |1 - \|\boldsymbol{E}_1\|_{\mathrm{F}} - \|\boldsymbol{E}_2\|_{\mathrm{F}}| \\ &\overset{(c)}{\geq} |1 - \upsilon_1 - \upsilon_2|\end{aligned} \tag{79}$$

where the inequality (a) employs the triangle inequality, (b) is obtained via the matrix norm equivalence relation $\|\boldsymbol{E}_1+\boldsymbol{E}_2\|_2 \leq \|\boldsymbol{E}_1+\boldsymbol{E}_2\|_{\mathrm{F}} \leq \|\boldsymbol{E}_1\|_{\mathrm{F}} + \|\boldsymbol{E}_2\|_{\mathrm{F}}$, (c) is obtained by the assumption that $\|\boldsymbol{E}_1\|_F \leq \upsilon_1$ and $\|\boldsymbol{E}_2\|_F \leq \upsilon_2$.

H Proof of Lemma 3

First, let us define the following w.r.t the loss function $g(\boldsymbol{f},\boldsymbol{y}) = \|\boldsymbol{f}-\boldsymbol{y}\|_2^2,\ \forall \boldsymbol{f} \in \mathcal{F}, \boldsymbol{y} \in \mathbb{R}^K$ and the input data $\boldsymbol{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)$ where each $\boldsymbol{x}_n$ is sampled i.i.d. from the distribution $\mathcal{D}$ under Assumption 1:

$$L_{\boldsymbol{X}}(\widehat{\boldsymbol{f}}) \triangleq \frac{1}{N}\sum_{n=1}^N g(\widehat{\boldsymbol{f}}(\boldsymbol{x}_n), \boldsymbol{f}^\natural(\boldsymbol{x}_n)) = \frac{1}{N}\sum_{n=1}^N \|\widehat{\boldsymbol{f}}(\boldsymbol{x}_n) - \boldsymbol{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x}_n)\|_2^2 = \frac{1}{N}\|\widehat{\boldsymbol{F}} - \boldsymbol{\Pi}^\top \boldsymbol{F}^\natural\|_{\mathrm{F}}^2, \tag{80a}$$

$$L_{\mathcal{D}}(\widehat{\boldsymbol{f}}) \triangleq \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[g(\widehat{\boldsymbol{f}}(\boldsymbol{x}), \boldsymbol{f}^\natural(\boldsymbol{x}))\right] = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\widehat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x})\|_2^2\right]. \tag{80b}$$

We invoke Theorem 26.5 in (Shalev-Shwartz & Ben-David, 2014) and get that with probability greater than $1-\delta$:

$$L_{\mathcal{D}}(\widehat{\boldsymbol{f}}) \leq L_{\boldsymbol{X}}(\widehat{\boldsymbol{f}}) + 2\mathfrak{R}_N(g \circ \mathcal{F}) + 4\bar{c}\sqrt{\frac{2\log(4/\delta)}{N}}$$

$$\implies \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\|\widehat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x})\|_2^2\right] \leq \frac{1}{N}\|\widehat{\boldsymbol{F}} - \boldsymbol{\Pi}^\top \boldsymbol{F}^\natural\|_{\mathrm{F}}^2 + 4\mathfrak{R}_N(\mathcal{F}) + 16\sqrt{\frac{2\log(4/\delta)}{N}} \quad (81)$$

where the last inequality utilizes the definitions in (80), the contraction lemma (Lemma 26.9 from (Shalev-Shwartz & Ben-David, 2014)) and also applied $\bar{c} = 4$ since $|g(\boldsymbol{f}, \boldsymbol{y})| \leq 4$ in our case. The term $\mathfrak{R}_N(\mathcal{F})$ denotes the empirical Rademacher complexity of the neural network function class $\mathcal{F}$ which is upperbounded via the sensitive complexity parameter $\mathscr{R}_{\mathcal{F}}$ as follows (Lin & Zhang, 2019):

$$\mathfrak{R}_N(\mathcal{F}) \leq 16N^{-5/8}\left(2\|\boldsymbol{X}\|_{\mathrm{F}}\mathscr{R}_{\mathcal{F}}\right)^{\frac{1}{4}}.$$

I Proof of Lemma 7

Our goal is to bound $\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2$ for each m, n. To achieve this, we consider the below lemma:

Lemma 11 *(McDiarmid's inequality) Let $\mathcal{U}$ be some set and let $v : \mathcal{U}^N \to \mathbb{R}$ be a function of N variables such that for some $c > 0$, for all $i \in [N]$ and for all $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N, \boldsymbol{u}_i' \in \mathcal{U}$ we have*

$$|v(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N) - v(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{i-1}, \boldsymbol{u}_i', \boldsymbol{u}_{i+1}, \ldots, \boldsymbol{u}_N)| \leq c.$$

Let $U_1, \ldots, U_N$ be N independent random variables taking values in $\mathcal{U}$. Then, with probability greater than $1-\delta$,

$$|v(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N) - \mathbb{E}[v(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N)]| \leq c\sqrt{\frac{N\log(2/\delta)}{2}},$$

where expectation is taken w.r.t. $U_1, \ldots, U_N$.

In our settings, let us define $\boldsymbol{u}_n = \{\boldsymbol{x}_n, \{\widehat{\boldsymbol{y}}_n^{(m)}\}_{m\in\mathcal{S}_n}, \mathcal{S}_n\}$ where $\mathcal{S}_n$ denotes the indices of the annotators who labeled the data item $\boldsymbol{x}_n$ and let U_n be the random variable associated with u_n, One can note that $U_1, \ldots, U_N$ are independent in this case. Now, let us define the following:

$$\begin{aligned} v(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N) &\triangleq \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2 \\ &= \frac{1}{NM}\|\boldsymbol{P}^\natural - \widehat{\boldsymbol{P}}\|_{\mathrm{F}}^2, \end{aligned} \quad (82)$$

where $\widehat{\boldsymbol{P}}$ is determined by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$.

Next, we aim to characterize the constant c in Lemma 11 under the definition in (82). Towards this, we first characterize $\|\widehat{\boldsymbol{P}}' - \widehat{\boldsymbol{P}}\|_{\mathrm{F}}^2$, where $\widehat{\boldsymbol{P}}'$ is the optimal solution obtained by using data $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{i-1}, \boldsymbol{u}_i', \boldsymbol{u}_{i+1}, \ldots, \boldsymbol{u}_N$.

Lemma 12 *Let $\widehat{\boldsymbol{P}}$ is the optimal solution of the CCM criterion 6 using the data $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$ and let $\widehat{\boldsymbol{P}}'$ is the optimal solution of the CCM criterion 6 using the data $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{i-1}, \boldsymbol{u}_i', \boldsymbol{u}_{i+1}, \ldots, \boldsymbol{u}_N$. Assume that each $\widehat{\boldsymbol{P}}$ and $\widehat{\boldsymbol{P}}'$ are at least $(1/\beta)$, for a certain $\beta > 1$. Then, the following holds*

$$\|\widehat{\boldsymbol{P}}' - \widehat{\boldsymbol{P}}\|_{\mathrm{F}}^2 \leq \frac{2\bar{c}M\log\beta}{S},$$

where $\bar{c} > 0$ is a constant

Hence, we have

$$\begin{aligned}
&|v(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N)-v(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{i-1},\boldsymbol{u}',\boldsymbol{u}_{i+1},\ldots,\boldsymbol{u}_N)| \\
&=\frac{1}{NM}\left|\|\boldsymbol{P}^\natural-\widehat{\boldsymbol{P}}\|_{\mathrm{F}}^2-\|\boldsymbol{P}^\natural-\widehat{\boldsymbol{P}}'\|_{\mathrm{F}}^2\right| \\
&=\frac{1}{NM}\left|\|\boldsymbol{P}^\natural-\widehat{\boldsymbol{P}}'+\widehat{\boldsymbol{P}}'-\widehat{\boldsymbol{P}}\|_{\mathrm{F}}^2-\|\boldsymbol{P}^\natural-\widehat{\boldsymbol{P}}'\|_{\mathrm{F}}^2\right| \\
&\le\frac{1}{NM}\left|\|\widehat{\boldsymbol{P}}'-\widehat{\boldsymbol{P}}\|_{\mathrm{F}}^2+2\|\boldsymbol{P}^\natural-\widehat{\boldsymbol{P}}'\|_{\mathrm{F}}\|\widehat{\boldsymbol{P}}'-\widehat{\boldsymbol{P}}\|_{\mathrm{F}}\right| \\
&\le\frac{2\bar{c}M\log\beta}{NMS}+2\sqrt{2}\sqrt{\frac{2\bar{c}M\log\beta}{NMS}} \\
&\le\bar{c}_1\frac{\sqrt{\log\beta}}{\sqrt{NS}},
\end{aligned}\tag{83}$$

where $\bar{c}_1>0$ is a constant. where the first inequality is by triangle inequality, the second inequality is by using Lemma 12 and by the relation $\|\boldsymbol{P}_1-\boldsymbol{P}_2\|_{\mathrm{F}}\le\sqrt{2NM}$, for any $\boldsymbol{P}_1,\boldsymbol{P}_2\in\mathcal{P}$.

Hence, using the result in (83) and Lemma 11, we get the following with probability greater than $1-\delta$:

$$|v(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N)-\mathbb{E}[v(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N)]|\le\bar{c}_1\sqrt{\frac{\log\beta\log(2/\delta)}{S}}.\tag{84}$$

To proceed, we have the following lemma:

Lemma 13 *Let* $v_n(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N)\triangleq\sum_{m=1}^M\|\boldsymbol{p}_n^{\natural(m)}-\widehat{\boldsymbol{p}}_n^{(m)}\|_2^2$. *Then*

$$\mathbb{E}[v_n(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N)]=\mathbb{E}[v_{n'}(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N)],\ \forall,n,n'.$$

The proof is provided in Sec. K.

From Lemma 13, we get

$$\mathbb{E}[v(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N)]=\mathbb{E}[v_n(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N)],\ \forall n.$$

Hence, combining (84) with Lemma 13, we get the following with probability $1-\delta$,

$$\mathbb{E}\left[\sum_{m=1}^M\|\boldsymbol{p}_n^{\natural(m)}-\widehat{\boldsymbol{p}}_n^{(m)}\|_2^2\right]\le\frac{1}{NM}\sum_{n=1}^N\sum_{m=1}^M\|\boldsymbol{p}_n^{\natural(m)}-\widehat{\boldsymbol{p}}_n^{(m)}\|_2^2+\bar{c}_1\sqrt{\frac{\log\beta\log(2/\delta)}{S}}.\tag{85}$$

Applying Markov's inequality,

$$\mathsf{Pr}(\sum_{m=1}^M\|\boldsymbol{p}_n^{\natural(m)}-\widehat{\boldsymbol{p}}_n^{(m)}\|_2^2<t)\ge1-\frac{\mathbb{E}\left[\sum_{m=1}^M\|\boldsymbol{p}_n^{\natural(m)}-\widehat{\boldsymbol{p}}_n^{(m)}\|_2^2\right]}{t}.$$

Let $t=N^{\frac{1}{8}}\mathbb{E}\left[\sum_{m=1}^M\|\boldsymbol{p}_n^{\natural(m)}-\widehat{\boldsymbol{p}}_n^{(m)}\|_2^2\right]$, then with probability greater than $1-\frac{1}{N^{\frac{1}{8}}}$, we have

$$\sum_{m=1}^M\|\boldsymbol{p}_n^{\natural(m)}-\widehat{\boldsymbol{p}}_n^{(m)}\|_2^2\le N^{\frac{1}{8}}\mathbb{E}\left[\sum_{m=1}^M\|\boldsymbol{p}_n^{\natural(m)}-\widehat{\boldsymbol{p}}_n^{(m)}\|_2^2\right].\tag{86}$$

Combining (85) and (86), we get that with probability greater than $1-\frac{1}{N^{\frac{1}{8}}}-\delta$, for any n,

$$\sum_{m=1}^M\|\boldsymbol{p}_n^{\natural(m)}-\widehat{\boldsymbol{p}}_n^{(m)}\|_2^2\le N^{\frac{1}{8}}\left(\frac{1}{NM}\sum_{n=1}^N\sum_{m=1}^M\|\boldsymbol{p}_n^{\natural(m)}-\widehat{\boldsymbol{p}}_n^{(m)}\|_2^2\right)+N^{\frac{1}{8}}\bar{c}_1\sqrt{\frac{\log\beta\log(2/\delta)}{S}}.$$

J PROOF OF LEMMA 12

To prove Lemma 12, we follow the stability analysis presented in Chapter 13 of (Shalev-Shwartz & Ben-David, 2014). Let $\boldsymbol{u}_n = \{\boldsymbol{x}_n, \widehat{y}_n^{(m)}, m \in \mathcal{S}_n\}$ and let $\mathcal{T}$ collects all $\boldsymbol{u}_n$'s which is used for estimating $\widehat{\boldsymbol{P}}$, i.e.,

$$\mathcal{T} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$$

Let us also define the set $\mathcal{T}'$ as the following:

$$\mathcal{T}' = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{i-1}, \boldsymbol{u}', \boldsymbol{u}_{i+1}, \ldots, \boldsymbol{u}_N\}.$$

where $\boldsymbol{u}' \neq \boldsymbol{u}_i$ hold.The goal is to characterize the change in the optimal solutions of the CCEM criterion in (6) when $\mathcal{T}'$ is given as input in place of $\mathcal{T}$.

Let us concisely denote the objective function in (6) as below:

$$\mathcal{L}_{\mathcal{T}}(\boldsymbol{P}) = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{u}_n \in \mathcal{T}} \ell(\boldsymbol{u}_n; \boldsymbol{P})$$
$$\mathcal{L}_{\mathcal{T}'}(\boldsymbol{P}) = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{u}_n \in \mathcal{T}'} \ell(\boldsymbol{u}_n; \boldsymbol{P})$$

where $\ell(\boldsymbol{u}_n; \boldsymbol{P}) = \sum_{m \in \mathcal{S}_n} \mathsf{CE}(\boldsymbol{p}_n^{(m)}, \widehat{y}_n^{(m)})$. We have the corresponding optimal solution defined as follows:

$$\widehat{\boldsymbol{P}} = \arg\min_{\boldsymbol{P} \in \mathcal{P}} \mathcal{L}_{\mathcal{T}}(\boldsymbol{P}),$$
$$\widehat{\boldsymbol{P}}' = \arg\min_{\boldsymbol{P} \in \mathcal{P}} \mathcal{L}_{\mathcal{T}'}(\boldsymbol{P}).$$

Then, for any $\boldsymbol{P}_1, \boldsymbol{P}_2 \in \mathcal{P}$,

$$\mathcal{L}_{\mathcal{T}}(\boldsymbol{P}_1) - \mathcal{L}_{\mathcal{T}}(\boldsymbol{P}_2) = \mathcal{L}_{\mathcal{T}'}(\boldsymbol{P}_1) - \mathcal{L}_{\mathcal{T}'}(\boldsymbol{P}_2) + \frac{\ell(\boldsymbol{u}_i; \boldsymbol{P}_1) - \ell(\boldsymbol{u}_i; \boldsymbol{P}_2)}{|\mathcal{S}|} + \frac{\ell(\boldsymbol{u}'; \boldsymbol{P}_2) - \ell(\boldsymbol{u}'; \boldsymbol{P}_1)}{|\mathcal{S}|}$$

The above relation applies for $\widehat{\boldsymbol{P}}$ and $\widehat{\boldsymbol{P}}'$ as well and hence, we have

$$\mathcal{L}_{\mathcal{T}}(\widehat{\boldsymbol{P}}') - \mathcal{L}_{\mathcal{T}}(\widehat{\boldsymbol{P}}) \leq \frac{\ell(\boldsymbol{u}_i; \widehat{\boldsymbol{P}}') - \ell(\boldsymbol{u}_i; \widehat{\boldsymbol{P}})}{|\mathcal{S}|} + \frac{\ell(\boldsymbol{u}'; \widehat{\boldsymbol{P}}) - \ell(\boldsymbol{u}'; \widehat{\boldsymbol{P}}')}{|\mathcal{S}|}$$

since $\mathcal{L}_{\mathcal{T}'}(\widehat{\boldsymbol{P}}') - \mathcal{L}_{\mathcal{T}}(\widehat{\boldsymbol{P}}') \leq 0$ as $\widehat{\boldsymbol{P}}'$ is the minimizer of $\mathcal{L}_{\mathcal{T}'}(\boldsymbol{P})$.

We use the fact that the function $\mathcal{L}_{\mathcal{T}}(\boldsymbol{P})$ is strongly convex w.r.t. $\boldsymbol{P}$, with a parameter $\mu > 0$ since each entry of the $\boldsymbol{P}$ matrix is at least $1/\beta$, where $\beta > 1$.

Since $\widehat{\boldsymbol{P}}$ is the minimizer of $\mathcal{L}_{\mathcal{T}}(\boldsymbol{P})$ and using the strong convexity of the function $\mathcal{L}_{\mathcal{T}}(\boldsymbol{P})$, we have

$$\mathcal{L}_{\mathcal{T}}(\widehat{\boldsymbol{P}}') - \mathcal{L}_{\mathcal{T}}(\widehat{\boldsymbol{P}}) \geq \frac{\mu}{2} \|\widehat{\boldsymbol{P}}' - \widehat{\boldsymbol{P}}\|_{\mathrm{F}}^2$$

Hence, we get

$$\begin{aligned} \|\widehat{\boldsymbol{P}}' - \widehat{\boldsymbol{P}}\|_{\mathrm{F}}^2 &\leq \frac{2}{|\mu\mathcal{S}|} (\ell(\boldsymbol{z}_i; \widehat{\boldsymbol{P}}') - \ell(\boldsymbol{z}_i; \widehat{\boldsymbol{P}}) + \ell(\boldsymbol{z}'; \widehat{\boldsymbol{P}}) - \ell(\boldsymbol{z}'; \widehat{\boldsymbol{P}}')) \\ &\leq \frac{2M \log \beta}{\mu |\mathcal{S}|}, \end{aligned}$$

where $\mu > 0$.

K PROOF OF LEMMA 13

We have

$$v_n(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n,\ldots,\boldsymbol{u}_{n'},\ldots,\boldsymbol{u}_N) \triangleq \sum_{m=1}^{M} \|\boldsymbol{p}_n^{\natural(m)} - \widehat{\boldsymbol{p}}_n^{(m)}\|_2^2$$
$$= \sum_{m=1}^{M} \|\boldsymbol{A}_m^{\natural}\boldsymbol{f}(\boldsymbol{x}_n) - \widehat{\boldsymbol{A}}_m\widehat{\boldsymbol{f}}(\boldsymbol{x}_n)\|_2^2$$

where $\boldsymbol{u}_n = \{\boldsymbol{x}_n, \{\widehat{\boldsymbol{y}}_n^{(m)}\}_{m\in\mathcal{S}_n}, \mathcal{S}_n\}$ and $\widehat{\boldsymbol{f}}$ is the estimate obtained using the data $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n,\ldots,\boldsymbol{u}_{n'},\ldots,\boldsymbol{u}_N$. Note that $\boldsymbol{f}$ does not depend on the data $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n,\ldots,\boldsymbol{u}_{n'},\ldots,\boldsymbol{u}_N$.

Consider $n' \neq n$, then

$$v_{n'}(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{n'},\ldots,\boldsymbol{u}_n,\ldots,\boldsymbol{u}_N) \triangleq \sum_{m=1}^{M} \|\boldsymbol{p}_{n'}^{\natural(m)} - \widehat{\boldsymbol{p}}_{n'}^{(m)}\|_2^2$$
$$= \sum_{m=1}^{M} \|\boldsymbol{A}_m^{\natural}\boldsymbol{f}(\boldsymbol{x}_{n'}) - \widehat{\boldsymbol{A}}_m\widehat{\boldsymbol{f}}'(\boldsymbol{x}_{n'})\|_2^2$$
$$= \sum_{m=1}^{M} \|\boldsymbol{A}_m^{\natural}\boldsymbol{f}(\boldsymbol{x}_{n'}) - \widehat{\boldsymbol{A}}_m\widehat{\boldsymbol{f}}(\boldsymbol{x}_{n'})\|_2^2$$

where $\boldsymbol{u}_{n'} = \{\boldsymbol{x}_{n'}, \{\widehat{\boldsymbol{y}}_{n'}^{(m)}\}_{m\in\mathcal{S}_{n'}}, \mathcal{S}_{n'}\}$ where $\widehat{\boldsymbol{f}}'$ is the estimate obtained using $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{n'},\ldots,\boldsymbol{u}_n,\ldots,\boldsymbol{u}_N$ and is the same as $\widehat{\boldsymbol{f}}$ since optimal solutions do not depend on the order of the input. Hence, we can use the following notation:

$$g_{\boldsymbol{U}}(\boldsymbol{x}_n) = v_n(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n,\ldots,\boldsymbol{u}_{n'},\ldots,\boldsymbol{u}_N)$$
$$g_{\boldsymbol{U}}(\boldsymbol{x}_{n'}) = v_{n'}(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{n'},\ldots,\boldsymbol{u}_n,\ldots,\boldsymbol{u}_N)$$

where g is a certain real valued function and $\boldsymbol{U}$ captures the dependency of the function g w.r.t all $\boldsymbol{u}_n$'s.

Next, we consider taking expectation w.r.t. the random variables $U_1,\ldots,U_N$:

$$\mathbb{E}[g_{\boldsymbol{U}}(\boldsymbol{x}_n)] = \int_{\boldsymbol{U}} q(\boldsymbol{U}) g_{\boldsymbol{U}}(\boldsymbol{x}_n) d\boldsymbol{U}$$
$$\mathbb{E}[g_{\boldsymbol{U}}(\boldsymbol{x}_{n'})] = \int_{\boldsymbol{U}} q(\boldsymbol{U}) g_{\boldsymbol{U}}(\boldsymbol{x}_{n'}) d\boldsymbol{U}$$

where $q(\boldsymbol{U})$ denotes the joint probability of the random variables $U_1,\ldots,U_N$.

Since $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$ are identically distributed, we get

$$\mathbb{E}[g_{\boldsymbol{U}}(\boldsymbol{x}_n)] = \mathbb{E}[g_{\boldsymbol{U}}(\boldsymbol{x}_{n'})].$$

L MORE DETAILS ON THE COMPLEXITY MEASURE $\mathscr{R}_{\mathcal{F}}$

The work in (Lin & Zhang, 2019) introduces a complexity measure for measuring the expressiveness of different family of neural network classes. For example, consider a general convolution neural network (CNN) architecture with L layers such that

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{\sigma}_L(\gamma_L(\boldsymbol{H}_L)\boldsymbol{\sigma}_{L-1}(\gamma_{L-1}(\boldsymbol{H}_{L-1})\ldots\boldsymbol{\sigma}_1(\gamma_1(\boldsymbol{H}_1)\boldsymbol{x})\ldots)),\ \forall \boldsymbol{f}\in\mathcal{F},$$

where $\boldsymbol{\sigma}_i(\cdot) = [\sigma_i(\cdot),\ldots,\sigma_i(\cdot)] \in \mathbb{R}^{d_i}$ represents an element-wise activation function at ith layer satisfying $\sigma_i(0) = 0$ and σ_i being ρ-Lipschitz continuous for all i. Let $\mathcal{G}_{\rm F}$ denote the set of layer indices with fully connected layers and $\mathcal{G}_{\rm C}$ denote the layer indices with convolutional layers. If $i \in \mathcal{G}_{\rm F}$, we have $\gamma_i(\boldsymbol{H}_i) = \boldsymbol{H}_i \in \mathbb{R}^{d_i\times d_{i-1}}$. Suppose $i \in \mathcal{G}_{\rm C}$, then we have $\gamma_i(\boldsymbol{H}_i)\boldsymbol{z}_{i-1} = \boldsymbol{H}_i \circledast \boldsymbol{z}_{i-1}$,

$$\boldsymbol{W}^\natural = \begin{bmatrix} \vdots \\ \boldsymbol{A}^\natural_{m_1}(1,:) \\ \vdots \\ \boldsymbol{A}^\natural_{m_2}(2,:) \\ \vdots \\ \boldsymbol{A}^\natural_{m_3}(3,:) \\ \vdots \end{bmatrix} \begin{matrix} \\ \approx \ \boldsymbol{e}_1^\top \\ \\ \approx \ \boldsymbol{e}_2^\top \\ \\ \approx \ \boldsymbol{e}_3^\top \\ \\ \end{matrix} \qquad \boldsymbol{W}^\natural = \begin{bmatrix} \vdots \\ \vdots \\ \vdots \\ \boldsymbol{A}^\natural_{m_*} \\ \vdots \\ \vdots \\ \vdots \end{bmatrix} \approx \boldsymbol{I}_3$$

Figure 1: (left) $\boldsymbol{W}^\natural = [\boldsymbol{A}_1^{\natural\top}, \ldots, \boldsymbol{A}_M^{\natural\top}]^\top$ satisfying NCSA with $K = 3$, meaning that there exists specialists for each class $k \in [K]$ and (right) expert annotator m^* among M annotators, meaning $\boldsymbol{A}^\natural_{m_*} \approx \boldsymbol{I}_K$.

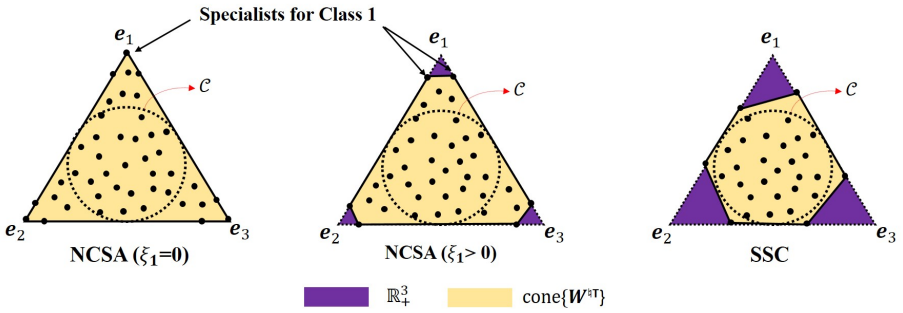


Figure 2: The dots denote the rows of $\boldsymbol{W}^\natural$ and the circle denote the second-order cone $\mathcal{C}$.

where the matrix $\boldsymbol{H}_i \in \mathbb{R}^{c_i \times r_i}$ contains c_i convolutional filters each of which having dimension r_i, $\boldsymbol{z}_{i-1}$ denotes the output of the $(i-1)$th layer, and $\circledast$ denotes the convolution operation. Here $d_0 = d$ is the dimension of the of the input data items $\boldsymbol{x}_n$'s. Then, its sensitive complexity $\mathscr{R}_{\mathcal{F}}$ is defined as follows:

$$\mathscr{R}_{\mathcal{F}} = 2\rho^L L^2 \left(\sum_{i \in \mathcal{G}_{\mathrm{F}}} d_i^2 d_{i-1}^2 + \sum_{i \in \mathcal{G}_{\mathrm{C}}} c_i^2 r_i^2 \sqrt{d_i / c_i} \right).$$

Note that such as a general CNN architecture covers many popular neural networks, e.g., fully connected neural networks, CNNs such as Lenet-5 (Lecun et al., 1998), VGG-16 (Liu & Deng, 2015), and so on.

M Near-Class Specialist Assumption and SSC

NCSA is a relaxed condition compared to having expert annotators or having diagonally dominant annotators–see Fig. 1. It can be understood that NCSA assumption does not require any single annotator to be specialists with respect to all classes. Nonetheless, having an expert annotator m_* satisfies NCSA, but not vice versa.

The SSC (Definition 1) is a relaxed condition compared to the NCSA (Assumption 4). To illustrate this, we consider the geometry of the matrix $\boldsymbol{W}^\natural$ satisfying the NCSA and the SSC as shown in Fig. 2.

N Algorithm Description

In this section, we detail the implementation of the regularized criteria in (8) and (9). The neural network predictor function $\boldsymbol{f} \in \mathcal{F}$ is parameterized using $\boldsymbol{\theta}$ and can be denoted as $\boldsymbol{f}_\theta : \mathbb{R}^d \to \boldsymbol{\Delta}$. Let $\boldsymbol{\psi} = [\mathsf{vec}(\boldsymbol{A}_1)^\top, \ldots, \mathsf{vec}(\boldsymbol{A}_M)^\top, \mathsf{vec}(\boldsymbol{\theta})^\top]^\top$. Let us also define the following for certain batch

$\mathcal{B} \subset [N]$:

$$\mathcal{L}_F(\boldsymbol{\psi}, \mathcal{B}) = -\frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \sum_{m \in \mathcal{S}_n} \sum_{k=1}^{K} \mathbb{I}[\widehat{y}_n^{(m)} = k] \log[\boldsymbol{A}_m \boldsymbol{f}_\theta(\boldsymbol{x}_n)]_k - \lambda \log\det \widetilde{\boldsymbol{F}} \widetilde{\boldsymbol{F}}^\top,$$

$$\mathcal{L}_W(\boldsymbol{\psi}, \mathcal{B}) = -\frac{1}{|\mathcal{B}|} \sum_{n \in \mathcal{B}} \sum_{m \in \mathcal{S}_n} \sum_{k=1}^{K} \mathbb{I}[\widehat{y}_n^{(m)} = k] \log[\boldsymbol{A}_m \boldsymbol{f}_\theta(\boldsymbol{x}_n)]_k - \lambda \log\det \boldsymbol{W}^\top \boldsymbol{W},$$

where $\widetilde{\boldsymbol{F}} = [\boldsymbol{f}(\boldsymbol{x}_{n_1}), \ldots, \boldsymbol{f}(\boldsymbol{x}_{n_{|\mathcal{B}|}})]$, $n_1, \ldots, n_{|\mathcal{B}|} \in \mathcal{B}$ and $\boldsymbol{W} = [\boldsymbol{A}_1^\top, \ldots, \boldsymbol{A}_M^\top]^\top$. Let $\nabla_{\boldsymbol{\psi}} \mathcal{L}_F(\boldsymbol{\psi}, \mathcal{B})$ denote the stochastic gradient of $\mathcal{L}_F(\boldsymbol{\psi}, \mathcal{B})$ w.r.t. $\boldsymbol{\psi}$ and $\nabla_{\boldsymbol{\psi}} \mathcal{L}_W(\boldsymbol{\psi}, \mathcal{B})$ denote the stochastic gradient of $\mathcal{L}_W(\boldsymbol{\psi}, \mathcal{B})$ w.r.t. $\boldsymbol{\psi}$. Under these notations, the proposed methods `GeoCrowdNet(F)` and `GeoCrowdNet(W)` are described in Algorithm 1 and 2, respectively.

Algorithm 1: `GeoCrowdNet(F)`

input : Data $\{\boldsymbol{x}_n, \widehat{y}_n^{(m)}, m \in \mathcal{S}_n\}_{n=1}^N$
Initialize $\boldsymbol{A}_m$'s to identity matrices $\boldsymbol{I}_K$;
Initialize the parameters $\boldsymbol{\theta}$ of the neural network function class $\mathcal{F}$;
$\boldsymbol{\psi} = [\mathsf{vec}(\boldsymbol{A}_1)^\top, \ldots, \mathsf{vec}(\boldsymbol{A}_M)^\top, \mathsf{vec}(\boldsymbol{\theta})^\top]^\top$;
while *stopping criterion is not reached* **do**
  while *stopping criterion is not reached* **do**
    Draw a random batch $\mathcal{B}$;
    Compute $\nabla_{\boldsymbol{\psi}} \mathcal{L}_F(\boldsymbol{\psi}, \mathcal{B})$;
    $\boldsymbol{\psi} \leftarrow \mathsf{AdamOptimizer}(\boldsymbol{\psi}, \nabla_{\boldsymbol{\psi}} \mathcal{L}_F(\boldsymbol{\psi}, \mathcal{B}))$;
    for $m = 1$ **to** M **do**
      $\boldsymbol{A}_m \leftarrow \mathsf{softmax}(\boldsymbol{A}_m)$ ▷ softmax operation on the columns of $\boldsymbol{A}_m$
    end
  end
end
output : Estimates $\widehat{\boldsymbol{A}}_m, \forall m, \widehat{\boldsymbol{\theta}}$

Algorithm 2: `GeoCrowdNet(W)`

input : Data $\{\boldsymbol{x}_n, \widehat{y}_n^{(m)}, m \in \mathcal{S}_n\}_{n=1}^N$
Initialize $\boldsymbol{A}_m$'s to identity matrices $\boldsymbol{I}_K$;
Initialize the parameters $\boldsymbol{\theta}$ of the neural network function class $\mathcal{F}$;
$\boldsymbol{\psi} = [\mathsf{vec}(\boldsymbol{A}_1)^\top, \ldots, \mathsf{vec}(\boldsymbol{A}_M)^\top, \mathsf{vec}(\boldsymbol{\theta})^\top]^\top$;
while *stopping criterion is not reached* **do**
  while *stopping criterion is not reached* **do**
    Draw a random batch $\mathcal{B}$;
    Compute $\nabla_{\boldsymbol{\psi}} \mathcal{L}_W(\boldsymbol{\psi}, \mathcal{B})$;
    $\boldsymbol{\psi} \leftarrow \mathsf{AdamOptimizer}(\boldsymbol{\psi}, \nabla_{\boldsymbol{\psi}} \mathcal{L}_W(\boldsymbol{\psi}, \mathcal{B}))$;
    for $m = 1$ **to** M **do**
      $\boldsymbol{A}_m \leftarrow \mathsf{softmax}(\boldsymbol{A}_m)$ ▷ softmax operation on the columns of $\boldsymbol{A}_m$
    end
  end
  output : Estimates $\widehat{\boldsymbol{A}}_m, \forall m, \widehat{\boldsymbol{\theta}}$
end

O Additional Experiments and Implementation Details

O.1 Synthetic Annotators

Dataset. We consider CIFAR-10 (Krizhevsky, 2009) dataset for synthetic noisy label experiments. CIFAR-10 dataset comprises of $60,000$ labeled color images of animals, vehicles and so on, each having a size of 32×32 and belonging $K = 10$ different classes. We use $45,000$ images for training, $5,000$ images for validation, and $10,000$ images for testing.

Noisy Label Generation. In order to produce noisy annotations for the images of the training data, we simulate $M = 5$ synthetic annotators. We randomly choose an annotator m^* among M annotators

Table 3: Average test accuracies of the proposed methods and the baselines on CIFAR-10 dataset ($K = 10$), labeled by $M = 5$ synthetic annotators.

Methods	$\gamma = 0.01$	$\gamma = 0.2$	$\gamma = 0.3$
`GeoCrowdNet(F)`	$\mathbf{84.82 \pm 0.43}$	$\mathbf{71.66 \pm 1.08}$	$\mathbf{67.97 \pm 1.44}$
`GeoCrowdNet(W)`	82.76 ± 0.46	69.20 ± 0.59	63.86 ± 1.64
`GeoCrowdNet` (λ=0)	82.50 ± 0.38	69.14 ± 1.29	63.73 ± 2.64
`TraceReg`	82.54 ± 0.34	70.01 ± 1.16	65.01 ± 2.30
`Crowdlayer`	$\mathbf{83.01 \pm 0.23}$	69.91 ± 1.48	$\mathbf{65.70 \pm 2.32}$
`MBEM`	81.01 ± 0.32	69.45 ± 1.56	64.15 ± 2.21
`CoNAL`	81.41 ± 0.43	68.56 ± 1.78	63.21 ± 1.98
`Max-MIG`	82.01 ± 0.78	$\mathbf{69.96 \pm 1.23}$	62.01 ± 2.46
`NN-MV`	50.96 ± 0.83	37.90 ± 1.19	33.24 ± 1.17
`NN-DSEM`	52.02 ± 0.80	37.95 ± 1.41	33.64 ± 0.69

and its confusion matrix is generated by $\boldsymbol{A}^{\natural}_{m*} = \boldsymbol{I}_K + \gamma \texttt{rand}(K, K)$, followed by normalization w.r.t. ℓ_1 norm on the columns. Here $\boldsymbol{I}_K$ denotes the identity matrix of size K and $\texttt{rand}(K, K)$ denotes the $K \times K$ matrix with its entries randomly chosen from a uniform distribution between 0 and 1. The parameter γ controls how well the annotator m^* correctly identifies the ground-truth labels. The confusion matrices of the remaining $M - 1$ annotators are generated such that they provide labels uniformly at random. This type of modeling for the confusion matrices resembles the *hammer-spammer* model as employed in the works (Rodrigues & Pereira, 2018; Tanno et al., 2019). Using the ground-truth label y_n's provided by the dataset and the generated confusion matrices, the noisy annotations $\widehat{y}_n^{(m)}$'s are produced. Each annotation $\widehat{y}_n^{(m)}$ is observed independently at random such that only 20% of the total annotations are available for training. Note that, the true labels are not accessible by any of the methods.

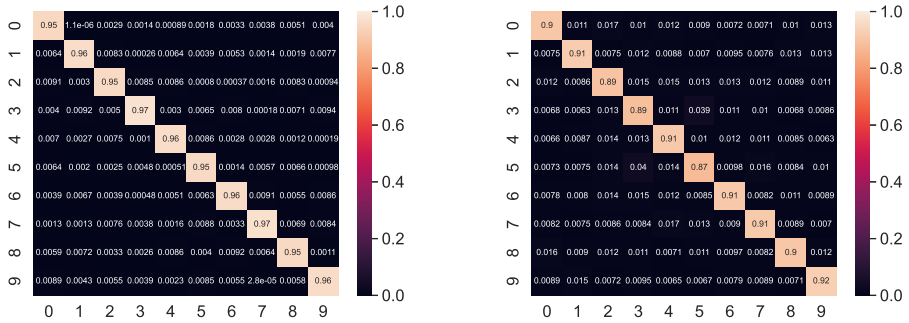
Neural Network Architectures and Settings. For CIFAR-10 dataset, we choose ResNet-9 architecture (He et al., 2016). Adam (Kingma & Ba, 2015) is used as the optimizer with weight decay of 10^{-4} and batch size of 128. The regularization parameter λ and the initial learning rate of the Adam optimizer are chosen via grid search method using the validation set from $\{0.01, 0.001, 0.0001\}$ and $\{0.01, 0.001\}$, respectively. We choose the same neural network structures for all the baselines. The confusion matrices are initialized with identity matrices of size K for the proposed method and the baselines `TraceReg` and `CrowdLayer`.

Results. Table 3 presents the results for various values of γ—when γ is smaller, the expertise of annotator m^* in correctly labeling the data items is better. One can note that the proposed approach, notably `GeoCrowdNet(F)`, outperforms the baselines in all these challenging scenarios. When annotator m^*'s labeling accuracy drops drastically as well, i.e., γ becomes larger, `GeoCrowdNet(F)` performs the best, This implies that even if there are no class specialists, `GeoCrowdNet(F)` thrives better for this large N scenario. Figs. 3, 4 and 5 show the ground-truth confusion matrices and the corresponding estimates when $\gamma = 0.01$ for `GeoCrowdNet` methods. One can see all methods estimate the confusion matrices reasonably well, corroborating our identifiability claims. The `GeoCrowdNet(F)` exhibits lower average mean squared error (MSE) for the estimated confusion matrices.

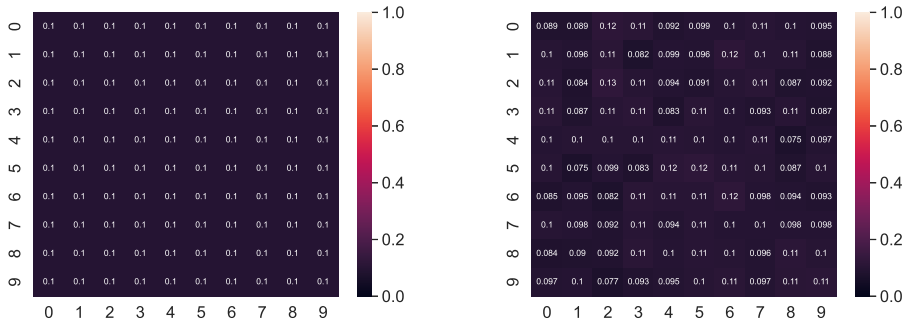
O.2 Real Data Experiments - Machine Annotations

Here, we provide additional details of the real data experiments. For experiments with machine annotations, we consider the MNIST dataset. Each image in the dataset is of size 28×28 in grey-scale format. We consider $57,000$ images as training data, 3000 images for validation, and $10,000$ images for testing.

For machine annotations, we train a number of machine classifiers to act as annotators. In Table 1, under case 2, we choose $M = 5$ annotators. Specifically, Linear SVM, logistic regression, k-NN with $k = 5$, CNN with two convolution layers followed by a max pooling layer, a fully connected layer and a softmax layer, fully connected neural network (FCNN) with 1 hidden layer and 128 hidden units are trained. The CNN is trained for 5 epochs and the FCNN is trained for 10 epochs.



The ground-truth $\boldsymbol{A}^{\natural}_{m*}$ (left) and the estimate $\widehat{\boldsymbol{A}}_{m*}$ (right)



The ground-truth $\boldsymbol{A}^{\natural}_{m}$ (left) and the estimate $\widehat{\boldsymbol{A}}_{m}$ (right), $m \neq m^*$

Figure 3: The illustration of the ground-truth confusion matrices and the estimated confusion matrices by the proposed approach `GeoCrowdNet (F)` for CIFAR-10 dataset, with $M = 5$ machine annotators and $\gamma = 0.01$. The average mean squared error (MSE) of the confusion matrices is 0.088.

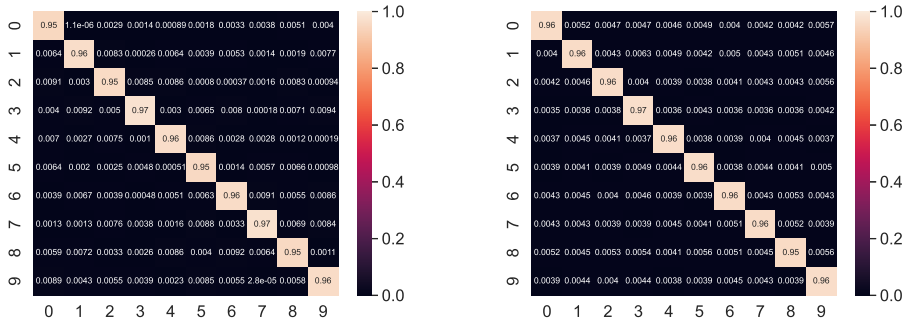
The individual label prediction accuracy of these annotators on unseen data items of size N ranges from 15.88% to 82.23% (averaged over 5 random trials) and 2 annotators are having less than 50% accuracy on average.

For case 2, we trained more annotators by changing some parameters of the above mentioned machine classifiers. Specifically, we use linear SVM, polynomial kernel-based SVM, and Gaussian SVM as different variants of SVM. For k-NN, we choose k from $\{3, 5, 7, 10\}$. FCNN, CNN, and logistic regression are trained with different number of epochs $\{10, 15, 20, 25\}$ and with random initializations for each case. Under case 1, if an annotator is chosen to be a specialist, it is trained with more number of samples ($10,000$ samples) with more number of epochs such that its label prediction accuracy is higher than 90%. Under this strategy, in Table 1 for the case $M = 15$, the individual label prediction accuracy of the annotators on unseen data items of size N ranges from 3.15% to 95.27% (2 annotators with more than 90% accuracy) . For the case with $M = 20$, the individual label prediction accuracy of the annotators on unseen data items of size N ranges from 3.24% to 95.73% (3 annotators with more than 90% accuracy).

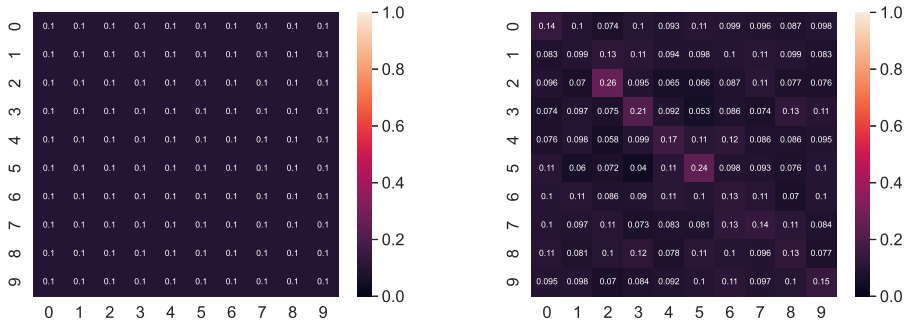
O.3 Real Data Experiments - Human Annotators

For experiments with annotations collected from AMT, we use LabelMe dataset (Rodrigues et al., 2017; Russell et al., 2007) and Music dataset (Rodrigues et al., 2014).

LabelMe dataset (Rodrigues et al., 2017; Russell et al., 2007) is an image classification dataset consisting of 2688 images from $K = 8$ different classes, namely, highway, inside city, tall building,



The ground-truth $\boldsymbol{A}^{\natural}_{m*}$ (left) and the estimate $\widehat{\boldsymbol{A}}_{m*}$ (right)



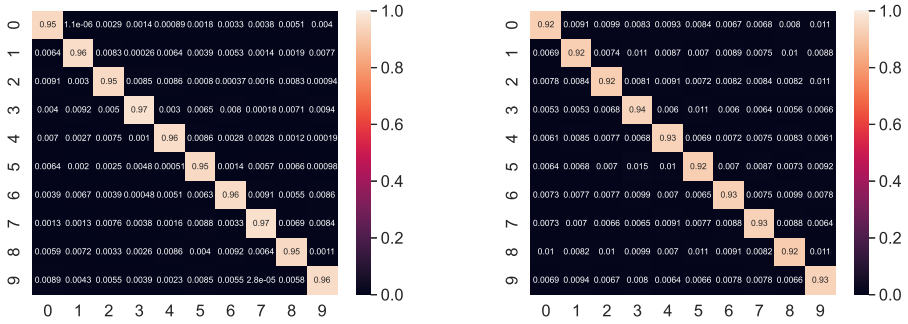
The ground-truth $\boldsymbol{A}^{\natural}_{m}$ (left) and the estimate $\widehat{\boldsymbol{A}}_{m}$ (right), $m \neq m^*$

Figure 4: The illustration of the ground-truth confusion matrices and the estimated confusion matrices by the proposed approach `GeoCrowdNet (W)` for CIFAR-10 dataset, with $M = 5$ machine annotators and $\gamma = 0.01$. The average mean squared error (MSE) of the confusion matrices is 0.102.

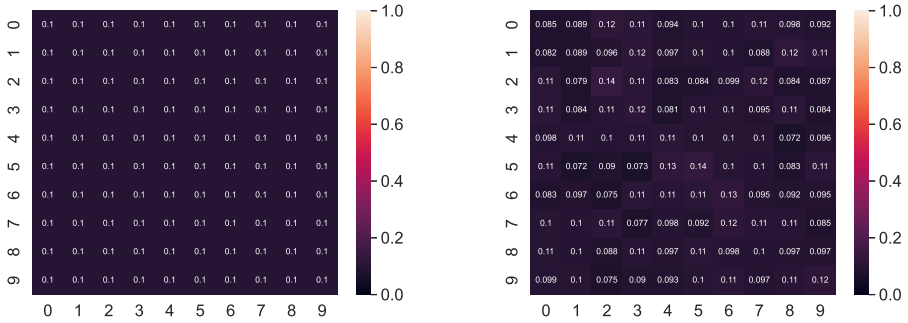
street, forest, coast, mountain, and open country. From the available images, 1000 images are annotated by $M = 59$ AMT workers. In total, about 2547 image annotations are obtained by the AMT workers whose labeling accuracy ranges from 0% to 100% with a mean accuracy of 69.2%. In order to enrich the training dataset, standard augmentation techniques such as rescaling, cropping, horizontal flips etc. are employed and accordingly, the training dataset consists of $10,000$ images annotated by 59 workers. The validation set consists of 500 images and the remaining 1188 images are used for testing.

Music dataset (Rodrigues et al., 2014) consists of audio samples of 10 different genres of music such as classical, country, disco, hiphop, jazz, rock, blues, reggae, pop, and metal. The dataset has about 1000 samples of songs each having a duration of 30 seconds. About 700 samples are annotated by 44 AMT workers (overall, about 2946 annotations are observed) and the remaining 300 are allocated for testing. Out of the annotated 700 samples, we consider 595 samples for training, and the remaining 105 samples are used for validation.

For LabelMe dataset, we employ the settings used in (Rodrigues & Pereira, 2018). The pretrained VGG-16 embeddings for the images are given as inputs to a fully connected (FC) neural network with one hidden layer having 128 hidden units and ReLU activation functions. A dropout layer with rate 50% is also used, followed by a softmax layer. For Music dataset, we choose the same FC layer and the softmax layer as employed for LabelMe, but with batch normalization layers before each of these layers.



The ground-truth $\boldsymbol{A}^{\natural}_{m*}$ (left) and the estimate $\widehat{\boldsymbol{A}}_{m*}$ (right)



The ground-truth $\boldsymbol{A}^{\natural}_{m}$ (left) and the estimate $\widehat{\boldsymbol{A}}_{m}$ (right), $m \neq m^*$

Figure 5: The illustration of the ground-truth confusion matrices and the estimated confusion matrices by `GeoCrowdNet` ($\lambda = 0$) for CIFAR-10 dataset, with $M = 5$ machine annotators and $\gamma = 0.01$. The average mean squared error (MSE) of the confusion matrices is 0.097.