PERFORMANCE OF SOCIAL MACHINE LEARNING UNDER LIMITED DATA

Ping Hu, Virginia Bordignon, Mert Kayaalp, Ali H. Sayed

School of Engineering, École Polytechnique Fédérale de Lausanne (EPFL)

ABSTRACT

This paper studies the non-asymptotic classification performance of the social machine learning strategy. This strategy involves an independent training phase followed by a cooperative inference phase to classify a growing number of samples. By considering instead a finite number of samples, we provide an upper bound for the probability of misclassification. This bound helps characterize the generalization ability of the social machine learning strategy, in terms of the statistical properties of the classification problem and the combination policy among the distributed classifiers. The analysis establishes the exponential decay of the probability of error with the number of samples when the training phase is consistent.

***Index Terms*—** Social machine learning, probability of error, non-asymptotic analysis.

1. INTRODUCTION

Social learning is a cooperative inference scheme where a collection of networked agents work together to infer the true state, out of a finite number of hypotheses, from local observations. Various social learning rules have been proposed in the literature [1–9], which rely largely on adaptation and combination steps. In the adaptation step, the Bayes rule is employed to update the agents' belief vectors using the new observations; while in the combination step, agents aggregate the information from their neighbors using either arithmetic or geometric averaging. When the social learning problem is globally identifiable, these algorithms have been shown to attain asymptotic truth learning, i.e., the belief on the true state converges to 1 when a growing number of observations are collected. In order to utilize the Bayes rule in the adaptation step, a key assumption in the works [1–9] is knowledge of the underlying likelihood models for the observations. However, these likelihood models are generally unavailable in real-world applications, which motivated the introduction of the social machine learning (SML) framework in [10, 11].

The SML strategy is a two-phase learning framework, as depicted in Fig. 1. In the *training* phase, each agent trains a classifier independently with a finite set of labeled samples. The purpose is to learn the discriminative information for distinguishing different hypotheses. The output of the trained classifier is used as the local decision statistic for inference [12], playing the role of the log-likelihood ratio when the likelihood models are known [1–9]. In the *prediction* phase, agents observe unlabeled samples and implement a social learning protocol based on the trained classifiers to infer the true state. This framework is *fully data-driven* and therefore the assumption of known likelihood models is avoided.

In [10, 11], the authors provide a rigorous theoretical analysis on the probability of consistent learning, i.e., the probability of asymptotic truth learning in the prediction phase when a large number of unlabeled samples are available, and illustrate the excellent classification performance of the SML strategy with experiments. This work supplements the theoretical analysis and examines the performance of the SML strategy when the amount of samples for inference, i.e., in the prediction phase, is *limited*. Collecting abundant samples is a time-consuming and expensive task, therefore learning with a limited number of samples is of practical interest. To this end, we conduct a *non-asymptotic* analysis of the classification error for the SML strategy. By characterizing a decision margin for the trained classifiers, we provide an upper bound on the instantaneous probability of error during the prediction process. Our result captures the influence of data heterogeneity and graph topology on the classification performance of the SML strategy. Numerical simulations with a network of feedforward neural networks are implemented to help illustrate the results.

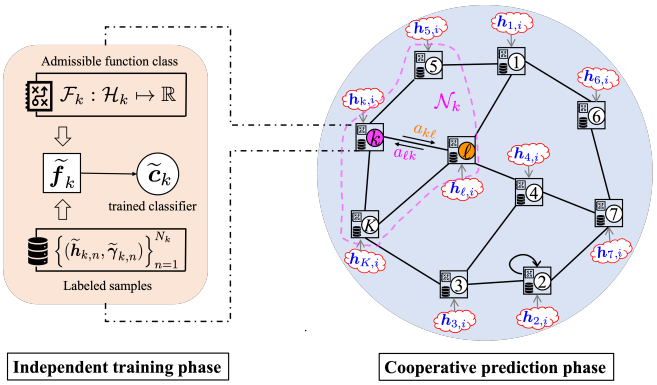


Fig. 1: SML architecture.

Notation: We use boldface fonts to denote random variables, and normal fonts for their realizations. $\mathbb{E}$ and $\mathbb{P}$ denote expectation and probability operators, respectively. Variables related to the training phase are topped with the symbol $\sim$.

2. THE SML STRATEGY

We consider a network of K connected agents or classifiers indexed by $k \in \mathcal{K} \triangleq \{1, 2, \ldots, K\}$, trying to solve a binary classification task. The agents are *heterogeneous* in that their observations may follow different statistical models even when they arise from the same class. For instance, in multi-view learning, each agent observes a different view of the same data [13]. In Fig. 1, we present a diagram of the SML architecture. We denote the binary hypotheses by $\Gamma = \{+1, -1\}$. Each agent k has a set of N_k labeled examples consisting of pairs $\{(\widetilde{\boldsymbol{h}}_{k,n}, \widetilde{\boldsymbol{\gamma}}_{k,n})\}_{n=1}^{N_k}$, where $\widetilde{\boldsymbol{h}}_{k,n} \in \mathcal{H}_k$ is the n-th feature vector and $\widetilde{\boldsymbol{\gamma}}_{k,n} \in \Gamma$ is the corresponding label. The pair

This work was supported in part by grant 205121-184999 from the Swiss National Science Foundation (SNSF). E-mails: ping.hu@epfl.ch, virginia.bordignon@epfl.ch, mert.kayaalp@epfl.ch, ali.sayed@epfl.ch.

$(\widetilde{\boldsymbol{h}}_{k,n}, \widetilde{\boldsymbol{\gamma}}_{k,n})$ is distributed according to

$$(\widetilde{\boldsymbol{h}}_{k,n}, \widetilde{\boldsymbol{\gamma}}_{k,n}) \sim \widetilde{p}_k(h,\gamma) = L_k(h|\gamma)\widetilde{p}_k(\gamma), \tag{1}$$

where $L_k(h|\gamma)$ is the *unknown* likelihood model and $\widetilde{p}_k(\gamma)$ is the class probability in the training set. We assume that the training set is balanced, i.e., $\widetilde{p}_k(\gamma) = \frac{1}{2}$. For the binary classification problem, the logit (i.e., log-ratio between posterior probabilities), defined as

$$c_k(h) \triangleq \log \frac{\widetilde{p}_k(+1|h)}{\widetilde{p}_k(-1|h)} \overset{\text{uniform prior}}{=} \log \frac{L_k(h|+1)}{L_k(h|-1)}, \tag{2}$$

is an important statistic for decision. The optimal Bayes classifier would assign $\gamma = +1$ to the feature vector h if $c_k(h)$ is positive and $\gamma = -1$ otherwise. However, $c_k(h)$ is not accessible due to the unknown likelihood models $L_k(h|\gamma)$. The two phases of the SML strategy operate as follows [11].

2.1. The training phase

In the training phase, each agent k approximates the logit function c_k defined in (2) by means of a function $\widetilde{\boldsymbol{f}}_k$ from an admissible class $\mathcal{F}_k : \mathcal{H}_k \mapsto \mathbb{R}$ by minimizing the following local empirical logistic risk:

$$\widetilde{\boldsymbol{f}}_k = \underset{f_k \in \mathcal{F}_k}{\arg\min}\, \widetilde{\boldsymbol{R}}_k(f_k) \triangleq \frac{1}{N_k} \sum_{n=1}^{N_k} \log\left(1 + e^{-\widetilde{\boldsymbol{\gamma}}_{k,n} f_k(\widetilde{\boldsymbol{h}}_{k,n})}\right). \tag{3}$$

The learned function $\widetilde{\boldsymbol{f}}_k$ is used to formulate the local decision statistic $\widetilde{\boldsymbol{c}}_k(h)$ of agent k for the unseen feature vector $h \in \mathcal{H}_k$:

$$\widetilde{\boldsymbol{c}}_k(h) = \widetilde{\boldsymbol{f}}_k(h) - \widetilde{\boldsymbol{\mu}}_k(\widetilde{\boldsymbol{f}}_k), \tag{4}$$

where $\widetilde{\boldsymbol{\mu}}_k(\widetilde{\boldsymbol{f}}_k)$ is the *empirical training mean* calculated via

$$\widetilde{\boldsymbol{\mu}}_k(f_k) = \frac{1}{N_k} \sum_{n=1}^{N_k} f_k(\widetilde{\boldsymbol{h}}_{k,n}), \quad \forall f_k \in \mathcal{F}_k. \tag{5}$$

Discounting the empirical training mean is suggested in [11] as a *debiasing* operation, i.e., to mitigate possible biased models resulting from the training process.

2.2. The prediction phase

In the prediction phase, the agents work jointly to solve the binary classification problem with the learned decision function $\widetilde{\boldsymbol{c}}_k$ in (4). In [11], the social learning problem (i.e., a classification problem with an infinite number of streaming observations) is considered. Specifically, we assume that at each time instant i, each agent k receives a new feature vector $\boldsymbol{h}_{k,i} \in \mathcal{H}_k$. That is, we consider that each agent k has access to a growing stream of observations

$$\boldsymbol{h}_{k,1}, \boldsymbol{h}_{k,2}, \ldots \tag{6}$$

generated from the unknown likelihood model $L_k(\cdot|\boldsymbol{\gamma}_0)$, where $\boldsymbol{\gamma}_0 \in \Gamma$ is the true state. The random variables $\{\boldsymbol{h}_{k,i} : k \in \mathcal{K}\}$ are mutually independent conditioned on the underlying truth $\boldsymbol{\gamma}_0$. We use the boldface font $\boldsymbol{\gamma}_0$ here to highlight that the true label is a random variable for the prediction phase. The class probability of $\boldsymbol{\gamma}_0$ is denoted by $P(\boldsymbol{\gamma}_0)$. To solve the social learning problem, the distributed learning rule based on the learned decision function $\widetilde{\boldsymbol{c}}_k$ is formed as [11]:

$$\boldsymbol{\lambda}_{k,i} = \sum_{\ell=1}^{K} a_{\ell k} \left(\boldsymbol{\lambda}_{\ell,i-1} + \widetilde{\boldsymbol{c}}_\ell(\boldsymbol{h}_{\ell,i})\right) \tag{7}$$

where $\boldsymbol{\lambda}_{k,i}$ denotes the *decision variable* for agent k at time instant i. That is, agent k prefers the label $+1$ (or -1) at time instant i when $\boldsymbol{\lambda}_{k,i}$ is positive (or negative). For the social learning problem with accurate likelihood models [1–9], $\boldsymbol{\lambda}_{k,i}$ represents the log-belief ratio between labels $+1$ and -1 for agent k at time instant i. The combination weight $a_{\ell k}$ that agent k assigns to its neighbor ℓ satisfies:

$$\sum_{\ell=1}^{K} a_{\ell k} = 1, \quad a_{\ell k} \geq 0, \quad a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k, \tag{8}$$

where $\mathcal{N}_k$ denotes the neighboring set of agent k (see Fig. 1 for an illustration). The communication network is assumed to be strongly connected, which means that the combination matrix $A = [a_{\ell k}]$ is primitive and its Perron eigenvector π satisfies [14]:

$$A\pi = \pi, \quad \sum_{k=1}^{K} \pi_k = 1, \quad \pi_k > 0, \ \forall k \in \mathcal{K}. \tag{9}$$

An important feature of (7) is that the information from the observations is aggregated over both space (through ℓ) and time (through i), which strengthens the decision-making capabilities of the agents. The SML strategy is called *consistent* if asymptotic truth learning is attained, i.e., the true state $\boldsymbol{\gamma}_0$ is learned by all agents when the number of observations goes to infinity. For all $k \in \mathcal{K}$, we introduce the following *asymptotic decision statistic* $\widehat{\boldsymbol{\lambda}}_{k,\infty}$:

$$\widehat{\boldsymbol{\lambda}}_{k,\infty} \triangleq \lim_{i \to \infty} \frac{1}{i} \boldsymbol{\lambda}_{k,i}, \tag{10}$$

then the SML strategy is consistent if $\boldsymbol{\gamma}_0 \widehat{\boldsymbol{\lambda}}_{k,\infty} > 0$. That is, $\widehat{\boldsymbol{\lambda}}_{k,\infty} > 0$ if $\boldsymbol{\gamma}_0 = +1$, and $\widehat{\boldsymbol{\lambda}}_{k,\infty} < 0$ otherwise. We recall here an important conclusion of the social learning rule (7) from [7, 11]:

$$\widehat{\boldsymbol{\lambda}}_{k,\infty} \overset{\text{a.s.}}{=} \sum_{k=1}^{K} \pi_k \mathbb{E}_{\boldsymbol{h}_{k,i} \sim L_k(\cdot|\boldsymbol{\gamma}_0)} \widetilde{\boldsymbol{c}}_k(\boldsymbol{h}_{k,i}), \tag{11}$$

where "a.s." means almost sure convergence.

2.3. Consistent learning

Based on the learning rule (7) and its convergence property in (11), the authors in [11] established the following condition for consistent learning:

$$\mu^+(\widetilde{\boldsymbol{f}}) > \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) \quad \text{and} \quad \mu^-(\widetilde{\boldsymbol{f}}) < \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}), \tag{12}$$

where $\widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) = \sum_{k=1}^{K} \pi_k \widetilde{\boldsymbol{\mu}}_k(\widetilde{\boldsymbol{f}}_k)$ is the network average of the empirical training mean (5) and

$$\mu^+(\widetilde{\boldsymbol{f}}) = \sum_{k=1}^{K} \pi_k \mathbb{E}_{\boldsymbol{h}_{k,i} \sim L_k(\cdot|+1)} \widetilde{\boldsymbol{f}}_k(\boldsymbol{h}_{k,i}), \tag{13}$$

$$\mu^-(\widetilde{\boldsymbol{f}}) = \sum_{k=1}^{K} \pi_k \mathbb{E}_{\boldsymbol{h}_{k,i} \sim L_k(\cdot|-1)} \widetilde{\boldsymbol{f}}_k(\boldsymbol{h}_{k,i}) \tag{14}$$

are the network average of the conditional mean for the two hypotheses. Let P_c denote the probability of consistent learning:

$$P_c \triangleq \mathbb{P}\left(\mu^+(\widetilde{\boldsymbol{f}}) > \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}), \mu^-(\widetilde{\boldsymbol{f}}) < \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}})\right). \tag{15}$$

Under some technical assumptions, it is shown in [11] that if the *network target risk* $\mathsf{R}^o < \log 2$ and the *network Rademarcher complexity* $\rho < \mathcal{E}(\mathsf{R}^o)$, then P_c is lower bounded by

$$P_c \geq 1 - 2\exp\left\{-\frac{8N_{\max}}{\alpha^2\beta^2}\left(\mathcal{E}(\mathsf{R}^o) - \rho\right)^2\right\} \tag{16}$$

where $N_{\max} \triangleq \max_k N_k$, $\alpha \triangleq \sum_{k=1}^K \pi_k N_{\max}/N_k$, and β is a uniform bound on the functions f_k, i.e.,

$$|f_k(h)| \leq \beta,\ \forall h \in \mathcal{H}_k, f_k \in \mathcal{F}_k \text{ and } k \in \mathcal{K}. \tag{17}$$

The formal definitions of R^o, ρ, and $\mathcal{E}(\mathsf{R}^o)$ can be found in [11], and are omitted here due to space limitation.

The *probability of error* achieved by the SML strategy is defined as the probability of inconsistent learning:

$$P_e \triangleq \mathbb{P}\left(\boldsymbol{\gamma}_0 \widehat{\boldsymbol{\lambda}}_{k,\infty} \leq 0\right) \tag{18}$$

where the randomness stems from both the training phase (i.e., the training set) and the prediction phase (i.e., the true label $\boldsymbol{\gamma}_0$). We next show that an upper bound for P_e can be obtained from the probability of consistent learning P_c. Combining the convergence result (11) with the definitions (4), (5), (13), and (14), we have

$$\widehat{\boldsymbol{\lambda}}_{k,\infty} \overset{\text{a.s.}}{=} \begin{cases} \mu^+(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}), & \boldsymbol{\gamma}_0 = +1, \\ \mu^-(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}), & \boldsymbol{\gamma}_0 = -1. \end{cases} \tag{19}$$

This yields

$$\begin{aligned} P_e &\overset{(18)}{=} \mathbb{P}(\boldsymbol{\gamma}_0 = +1)\mathbb{P}\left(\widehat{\boldsymbol{\lambda}}_{k,\infty} \leq 0 | \boldsymbol{\gamma}_0 = +1\right) \\ &\quad + \mathbb{P}(\boldsymbol{\gamma}_0 = -1)\mathbb{P}\left(\widehat{\boldsymbol{\lambda}}_{k,\infty} \geq 0 | \boldsymbol{\gamma}_0 = -1\right) \\ &\overset{(19)}{=} P(+1)\mathbb{P}\left(\mu^+(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) \leq 0\right) \\ &\quad + P(-1)\mathbb{P}\left(\mu^-(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) \geq 0\right) \\ &\leq P(+1)(1 - P_c) + P(-1)(1 - P_c) \\ &= 1 - P_c. \end{aligned} \tag{20}$$

Therefore, when the number of observations for inference is large enough in the prediction phase, the probability of error achieved by the SML strategy is upper bounded by $1 - P_c$, which can be further bounded using (16).

3. NON-ASYMPTOTIC PERFORMANCE

In this section, we analyze the probability of error for the binary classification task with a *finite number* of observations. In this setting, the agents try to identify the true label $\boldsymbol{\gamma}_0$ given a sequence of streaming feature vectors

$$\boldsymbol{h}_{k,1}, \boldsymbol{h}_{k,2}, \ldots, \boldsymbol{h}_{k,S} \tag{21}$$

where S is the size of the unlabeled samples in the prediction phase. It is noted that when S tends to infinity, we recover the classical social learning problem [11], whose probability of error can be upper bounded by $1 - P_c$ as shown in (20). Since S is finite, the non-asymptotic performance analysis on the prediction phase is required. To this end, we characterize the *instantaneous* probability of error for each agent k, whose decision at time instant i is denoted by $\boldsymbol{\gamma}_{k,i}$. Without loss of generality, we assume an uninformed initial condition in the subsequent analysis, i.e., $\boldsymbol{\lambda}_{k,0} = 0, \forall k \in \mathcal{K}$.

Since the agents make a decision according to the sign of their decision variables, i.e.,

$$\boldsymbol{\gamma}_{k,i} \triangleq \operatorname{sign}(\boldsymbol{\lambda}_{k,i}), \tag{22}$$

an error occurs at agent k if $\boldsymbol{\lambda}_{k,i}$ and $\boldsymbol{\gamma}_0$ have different signs. Let $P_{k,i}^e$ denote the instantaneous probability of error associated with agent k at time instant i:

$$P_{k,i}^e \triangleq \mathbb{P}\left(\boldsymbol{\gamma}_0 \boldsymbol{\lambda}_{k,i} \leq 0\right). \tag{23}$$

Before analyzing the classification error (23), we first introduce a δ-*margin* consistent learning condition:

$$\mu^+(\widetilde{\boldsymbol{f}}) > \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) + \delta \quad \text{and} \quad \mu^-(\widetilde{\boldsymbol{f}}) < \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) - \delta \tag{24}$$

where $\delta \geq 0$ describes the expected distance between the asymptotic decision statistic $\widehat{\boldsymbol{\lambda}}_{k,\infty}$ and the decision boundary 0, which we will refer to as *decision margin*.

It is clear that the δ-margin consistent learning condition (24) is stronger than the consistent learning condition given in (12). In fact, expression (24) specifies the condition of asymptotic truth learning when the agents are conservative in making decisions. That is, they remain uncertain about the underlying label when the decision threshold of $\widehat{\boldsymbol{\lambda}}_{k,\infty}$ is not larger than δ. Let $P_{c,\delta}$ denote the probability of δ-margin consistent learning:

$$P_{c,\delta} \triangleq \mathbb{P}\left(\mu^+(\widetilde{\boldsymbol{f}}) > \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) + \delta, \mu^-(\widetilde{\boldsymbol{f}}) < \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) - \delta\right) \tag{25}$$

We have $P_{c,\delta} \leq P_c$ and $P_{c,0} = P_c$. A lower bound on $P_{c,\delta}$ is obtained as follows.

Lemma 1 (Probability of δ-margin consistent learning). *Assume that $0 \leq \delta < \frac{\mathsf{R}^o}{2}$ and $\rho < \mathcal{E}(\mathsf{R}^o - 2\delta) - \frac{\delta}{4}$, then we have the following bound for the probability of δ-margin consistent learning:*

$$P_{c,\delta} \geq 1 - 2\exp\left\{-\frac{8N_{\max}}{\alpha^2\beta^2}\left(\mathcal{E}(\mathsf{R}^o - 2\delta) - \frac{\delta}{4} - \rho\right)^2\right\} \tag{26}$$

Sketch of proof. The key to the proof is to rewrite $1 - P_{c,\delta}$ as

$$\begin{aligned} 1 - P_{c,\delta} &= \mathbb{P}\left(\left\{\mu^+(\widetilde{\boldsymbol{f}}) \leq \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) + \delta\right\} \cup \left\{\mu^-(\widetilde{\boldsymbol{f}}) \geq \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) - \delta\right\}\right) \\ &= \mathbb{P}\left(\left|\widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) - \mu(\widetilde{\boldsymbol{f}})\right| \geq \frac{\mu^+(\widetilde{\boldsymbol{f}}) - \mu^-(\widetilde{\boldsymbol{f}})}{2} - \delta\right) \end{aligned}$$

with $\mu(\widetilde{\boldsymbol{f}}) = \frac{\mu^+(\widetilde{\boldsymbol{f}}) + \mu^-(\widetilde{\boldsymbol{f}})}{2}$, and to upper bound this probability using the techniques introduced in [11]. □

Let $\mathcal{A}$ denote the event of wrong classification of agent k at time instant i and let $\mathcal{B}$ denote the event of δ-margin consistent learning:

$$\mathcal{A} \triangleq \left\{\boldsymbol{\gamma}_0 \boldsymbol{\lambda}_{k,i} \leq 0\right\}, \tag{27}$$

$$\mathcal{B} \triangleq \left\{\mu^+(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) > \delta, \mu^-(\widetilde{\boldsymbol{f}}) - \widetilde{\boldsymbol{\mu}}(\widetilde{\boldsymbol{f}}) < -\delta\right\}. \tag{28}$$

According to the law of total probability, the following inequality holds:

$$P_{k,i}^e \overset{(23)}{=} \mathbb{P}(\mathcal{A} \cap \mathcal{B}) + \mathbb{P}(\mathcal{A} \cap \overline{\mathcal{B}}) \leq \mathbb{P}\left(\mathcal{A}|\mathcal{B}\right) + \mathbb{P}[\overline{\mathcal{B}}] \tag{29}$$

where $\overline{\mathcal{B}}$ denotes the complement of event $\mathcal{B}$. Therefore, an upper bound for $P_{k,i}^e$ can be obtained by characterizing the conditional probability $\mathbb{P}(\mathcal{A}|\mathcal{B})$. This is established in the following theorem.

Theorem 1 (**Classification error**). *Let σ denote the second largest-magnitude eigenvalue of combination matrix A. Suppose that agents perform the social learning protocol* (7)*, then under the δ-margin consistent learning condition, we have*

$$\mathbb{P}(\mathcal{A}|\mathcal{B}) \leq \exp\left\{-\frac{(\delta i-\kappa)^2}{2\beta^2 i}\right\} \tag{30}$$

for all $i \geq \frac{\kappa}{\delta}$, where

$$\kappa \triangleq \frac{8\beta \log K}{1-\sigma}. \tag{31}$$

Hence for any sequence of observations with size $S \geq \frac{\kappa}{\delta}$, the probability of classification error $P_{k,S}^e$ is upper bounded by

$$P_{k,S}^e \leq 2\exp\left\{-\frac{8N_{\max}}{\alpha^2\beta^2}\left(\mathcal{E}(\mathsf{R}^o-2\delta)-\frac{\delta}{4}-\rho\right)^2\right\} + \exp\left\{-\frac{(\delta S-\kappa)^2}{2\beta^2 S}\right\}. \tag{32}$$

Sketch of proof. The proof is based on the convergence property of the combination matrix A and McDiarmid's inequality [15, 16] as well as Lemma 1. □

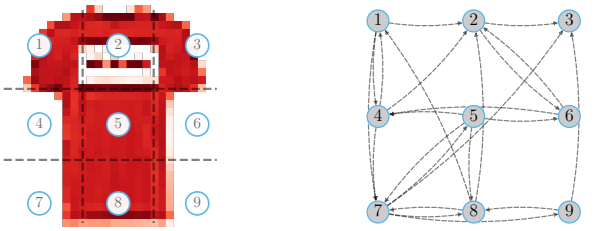
As the number of samples S grows, $P_{k,S}^e$ approaches the first term of (32), which is an upper bound for $1-P_{c,\delta}$. Since $P_{c,\delta} \leq P_c$, then in view of (20), $1-P_{c,\delta}$ is an upper bound on the probability of error for the social learning problem (i.e., $S=\infty$). By letting $\delta \to 0$, we recover the upper bound $1-P_c$ established in (20) [11]. According to (32), it is expected that the decay rate of $P_{k,S}^e$ with respect to S will be larger when the decision margin δ increases.

4. NUMERICAL SIMULATIONS

In the simulations, we consider the FashionMNIST dataset [17] and build a binary classification problem to distinguish 'T-Shirt' (labeled with $+1$) from 'Trouser' (labeled with -1). Each image of this dataset contains 784 pixels. We employ a network of 9 spatially distributed agents, where each agent observes a part of the image (see Fig. 2(a)) and they are connected through a strongly-connected communication network with the topology depicted in Fig. 2(b). We also assume a self-loop for each agent (not shown in Fig. 2(b)). A uniform averaging rule is employed for constructing the combination policy A [14].

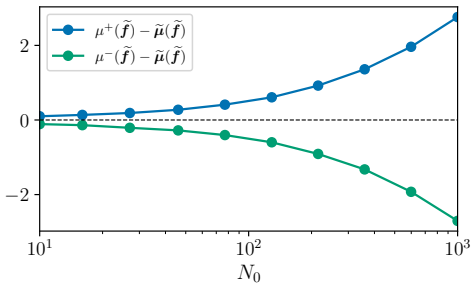
In the prediction phase, each agent trains its own classifier, which is a feedforward neural network with one hidden layer of 15 neurons and activation function $\tanh$. This simple structure is employed to better visualize the probability of error curves. To illustrate the δ-margin consistent learning condition, we consider different sizes of training sets. For simplicity, we assume an identical training size for all agents, i.e., $N_k = N_0, \forall k \in \mathcal{K}$. Given the value of N_0, a balanced training set is generated by randomly sampling from the FashionMNIST dataset. For each selected training set, the training is running using mini-batch iterates of 10 samples, over 30 epochs. We employ the Adam optimizer [16] with learning rate 0.0001 in the simulations.

In Fig. 3(a), we plot the decision margins achieved under different N_0, where the results are averaged over 100 different randomly generated training sets for each N_0. It can be seen that the decision margin increases as N_0 grows. This indicates a better learning condition for the prediction phase. In Fig. 3(b), the evolution of the instantaneous probability of error of agent 1 for $i \in [0, 30]$ under

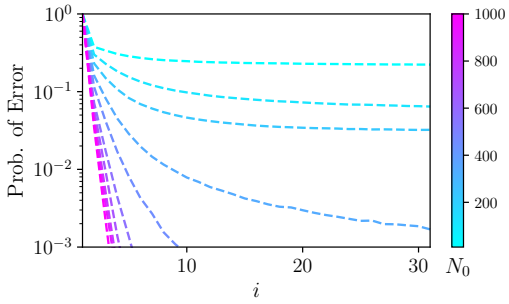


(a) Map of observations (b) Topology of the network

Fig. 2: Setting of distributed classifiers.



(a) Decision margin



(b) Classification error

Fig. 3: Performance of the SML strategy.

different N_0 is presented. The underlying class for all observations is set to be 'T-Shirt'. For each training set considered in Fig. 3(a), we conduct 2000 Monte Carlo runs of binary classification with the trained classifiers based on this dataset and obtain the average result. The simulation result for a specified size N_0 is then estimated empirically from the associated 100 training sets. We can see from Fig. 3(b) that for all training sizes, the instantaneous probability of error decreases over time i. For larger N_0, the decaying is almost exponential and the decay rate is positively correlated to the decision margin, which is consistent with (32).

5. CONCLUDING REMARKS

This paper studies the learning performance of the social machine learning strategy from [11] when the amount of samples for inference is finite in the prediction phase. An upper bound for the probability of error was presented. Our results extend the analysis in [11], which investigated the classification performance when the number of observations grows. An interesting future extension would be to consider the classification performance using a single sample.

6. REFERENCES

[1] V. Bordignon, V. Matta, and A. H. Sayed, "Adaptive social learning," *IEEE Transactions on Information Theory*, vol. 67, no. 9, pp. 6053–6081, 2021.

[2] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-Bayesian social learning," *Games and Economic Behavior*, vol. 76, no. 1, pp. 210–225, 2012.

[3] X. Zhao and A. H. Sayed, "Learning over social networks via diffusion adaptation," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, USA, 2012, pp. 709–713.

[4] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Distributed detection: Finite-time analysis and impact of network topology," *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3256–3268, 2016.

[5] A. Nedic, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-Bayesian learning," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5538–5553, 2017.

[6] H. Salami, B. Ying, and A. H. Sayed, "Social learning over weakly connected graphs," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 2, pp. 222–238, 2017.

[7] A. Lalitha, T. Javidi, and A. D. Sarwate, "Social learning and distributed hypothesis testing," *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6161–6179, 2018.

[8] V. Matta, V. Bordignon, A. Santos, and A. H. Sayed, "Interplay between topology and social learning over weak graphs," *IEEE Open Journal of Signal Processing*, vol. 1, pp. 99–119, 2020.

[9] M. Kayaalp, Y. Inan, E. Telatar, and A. H. Sayed, "On the arithmetic and geometric fusion of beliefs for distributed inference," *arXiv:2204.13741*, 2022.

[10] V. Bordignon, S. Vlaski, V. Matta, and A. H. Sayed, "Network classifiers based on social learning," in *Proc. IEEE ICASSP*, Toronto, ON, Canada, 2021, pp. 5185–5189.

[11] V. Bordignon, S. Vlaski, V. Matta, and A. H. Sayed, "Learning from heterogeneous data based on social interactions over graphs," to appear in *IEEE Transactions on Information Theory*, 2023. Also available at arXiv:2112.09483, 2022.

[12] V. Matta and A. H. Sayed, "Estimation and detection over adaptive networks," in *Cooperative and Graph Signal Processing*, P. M. Djurić and C. Richard, Eds. Amsterdam, The Netherlands: Elsevier, 2018, pp. 69–106.

[13] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[14] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

[15] C. McDiarmid, "On the method of bounded differences," *Surveys in combinatorics*, vol. 141, no. 1, pp. 148–188, 1989.

[16] A. H. Sayed, *Inference and Learning from Data*. Cambridge University Press, 2022.

[17] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms," *arXiv:1708.07747*, 2017.