Consistency in Models for Distributed Learning Under Communication Constraints

Joel B. Predd, *Student Member, IEEE*, Sanjeev R. Kulkarni, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—Motivated by sensor networks and other distributed settings, several models for distributed learning are presented. The models differ from classical works in statistical pattern recognition by allocating observations of an independent and identically distributed (i.i.d.) sampling process among members of a network of simple learning agents. The agents are limited in their ability to communicate to a central fusion center and thus, the amount of information available for use in classification or regression is constrained. For several basic communication models in both the binary classification and regression frameworks, we question the existence of agent decision rules and fusion rules that result in a universally consistent ensemble; the answers to this question present new issues to consider with regard to universal consistency. This paper addresses the issue of whether or not the guarantees provided by Stone's theorem in centralized environments hold in distributed settings.**

***Index Terms*—Classification, consistency, distributed learning, nonparametric, regression, sensor networks, statistical pattern recognition.**

I. Introduction

A. Models for Distributed Learning

CONSIDER the following learning model: Let X and Y be $\mathcal{X}$-valued and $\mathcal{Y}$-valued random variables, respectively, with a joint distribution denoted by $\boldsymbol{P}_{XY}$. $\mathcal{X}$ is known as the feature, input, or observation space; $\mathcal{Y}$ is known as the label, output, or target space. Throughout, we take $\mathcal{X} \subseteq \mathbb{R}^d$ and consider two cases corresponding to binary classification ($\mathcal{Y} = \{0,1\}$) and regression estimation ($\mathcal{Y} = \mathbb{R}$). Given a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the decision-theoretic problem is to design a decision rule $g : \mathcal{X} \to \mathcal{Y}$ that achieves the minimal expected loss $L^\star = \inf_g \boldsymbol{E}\{l(g(X), Y)\}$. Without prior knowledge of the distribution $\boldsymbol{P}_{XY}$, computing a loss minimizing decision rule is not possible. Instead, $D_n = \{(X_i, Y_i)\}_{i=1}^n$, an independent and identically distributed (i.i.d.) collection of training data with $(X_i, Y_i) \sim \boldsymbol{P}_{XY}$ for all $i \in \{1, \ldots, n\}$ is available; the learning problem is to use this data to infer decision rules with small expected loss.

This standard learning model invites one to consider numerous questions; however, in this work, we focus on the statistical property known as *universal consistency* [7], [12]. In traditional, centralized settings, D_n is provided to a single learning agent, and questions have been posed about the existence of classifiers or estimators that are universally consistent. The answers to such questions are well understood and are provided by results such as Stone's theorem [35], [7], [12], and numerous others in the literature.

Suppose, in contrast with the standard centralized setting, that for each $i \in \{1, \ldots, n\}$, the training datum (X_i, Y_i) is received by a distinct member of a network of n simple learning agents. When a central authority observes a new observation $X \sim \boldsymbol{P}_X$, it broadcasts the observation to the network in a request for information. At this time, each agent can respond with at most one bit. That is, each learning agent chooses whether or not to respond to the central authority's request for information; if it chooses to respond, an agent sends either a 1 or a 0 based on its local decision algorithm. Upon observing the response of the network, the central authority acts as a fusion center, combining the information to create an estimate of Y. As in the centralized setting, a key question arises: do there exist agent decision rules and a fusion rule that result in a universally consistent network in the limit as the number of agents increases without bound?

In what follows, we answer this question in the affirmative for both binary classification and regression estimation. In the binary classification setting, we demonstrate agent decision rules and a fusion rule that correspond nicely with classical kernel classifiers. With this connection to classical work, the universal Bayes-risk consistency of this ensemble then follows immediately from celebrated analyses such as Stone's theorem, etc. In the regression setting, we demonstrate that under regularity, randomized agent decision rules exist such that when the central authority applies a scaled average vote combination of the agents' responses, the resulting estimator is universally consistent under L_2-loss.

In this model, the agents convey slightly more information than is suggested by the mere one bit that we have allowed them to physically transmit to the fusion center. Indeed, each agent decides not between sending 1 or 0. Rather, each agent's decision rule can be viewed as a selection of one of *three* states: abstain, vote and send 0, and vote and send 1. With this observation, these results can be interpreted as follows: $\log_2(3)$ bits per agent per classification is sufficient for universal consistency to hold for both distributed classification and regression *with abstention*.

In this view, it is natural to ask whether these $\log_2(3)$ bits are necessary. Can consistency results be proven at lower bit rates? Consider a revised model, precisely the same as above, except that in response to the central authority's request for informa-

Manuscript received February 6, 2004; revised September 13, 2005. This work was supported in part by the Army Research Office under Grant DAAD19-00-1-0466, in part by Draper Laboratory under Grant IR&D 6002, in part by the National Science Foundation under Grant CCR-0312413, and in part by the Office of Naval Research under Grant N00014-03-1-0102.

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: jpredd@princeton.edu; kulkarni@princeton.edu; poor@princeton.edu).

Communicated by A. B. Nobel, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2005.860420

0018-9448/$20.00 © 2006 IEEE

tion, each agent must respond with 1 or 0; abstention is not an option and thus, each agent responds with exactly one bit per classification. Are there rules for which universal consistency results hold in distributed classification and regression *without abstention*?

Interestingly, we demonstrate that in the binary classification setting, randomized agent decision rules exist such that when a majority vote fusion rule is applied, universal Bayes-risk consistency holds. Next, we establish natural regularity conditions for candidate fusion rules and specify a reasonable class of agent decision rules. As an important negative result, we then demonstrate that for any agent decision rule within the class, there does not exist a regular fusion rule that is L_2 consistent for every distribution $\boldsymbol{P}_{XY}$. This result establishes the impossibility of universal consistency in this model for distributed regression without abstention for a restricted, but reasonable class of decision rules.

B. Motivation and Background

Motivation for studying distributed learning in general and the current models in particular arise from wireless sensor networks and distributed databases, applications that have attracted considerable attention in recent years [1]. Research in wireless sensor networks has focused on two separate aspects: networking issues, such as capacity, delay, and routing strategies; and applications issues. This paper is concerned with the second of these aspects, and in particular with the problem of distributed inference. Wireless sensor networks are *a fortiori* designed for the purpose of making inferences about the environments that they are sensing, and they are typically characterized by limited communications capabilities due to tight energy and bandwidth limitations, as well as the typically *ad hoc* nature of wireless networks. Thus, distributed inference is a major issue in the study of wireless sensor networks.

In problems of distributed databases, there is a collection of training data that is massive in both the dimension of the feature space and quantity of data. For political, economic, social, or technological reasons, this database is distributed geographically or in such a way that it is infeasible for any single agent to access the entire database. Multiple agents may be deployed to make inferences from various segments of the database, but communication constraints arising from privacy or security concerns highlight distributed inference as a key issue in this setting as well. Recent research has studied inference in the distributed databases setting from an algorithmic point of view; for example, [22] proposed a distributed boosting algorithm and studied its performance empirically.

Distributed detection and estimation is a well-developed field with a rich history. Much of the work in this area has focused on either parametric problems, in which strong statistical assumptions are made [36], [37], [3], [38], [23], [21], [6], [17], [8], or on traditional nonparametric formalisms, such as constant-false-alarm-rate detection [2]. Recently, [34] advocated a learning theoretic approach to wireless sensor networks and [26], in the context of kernel methods commonly used in machine learning, considered the classical model for decentralized detection [36] in a nonparametric setting.

In this paper, we consider an alternative nonparametric approach to the study of distributed inference that is most closely aligned with models considered in nonparametric statistics and the study of kernel estimators and other Stone-type rules. Extensive work has been done related to the consistency of Stone-type rules under various sampling processes; for example, [7], [12], and references therein; [5], [11], [18], [19], [20], [25], [27], [28], [29], [33], [35], [39], [40]. These models focus on various dependency structures within the training data and assume that a single processor has access to the entire data stream.

The nature of the work considered in this paper is to consider similar questions of universal consistency in models that capture some of the structure in a distributed environment. As motivated earlier, agents in distributed scenarios have constrained communication capabilities and moreover, each may have access to distinct data streams that differ in distribution and may depend on parameters such as the state of a sensor network or location of a database. We consider the question: for a given model of communication among agents, each of whom has been allocated a small portion of a larger learning problem, can enough information be exchanged to allow for a universally consistent ensemble? In this work, the learning problem is divided among agents by allocating each a unique observation of an i.i.d. sampling process. As explained earlier, we consider simple communication models with and without abstention. Insofar as these models present a useful picture of distributed scenarios, this paper addresses the issue of whether or not the guarantees provided by Stone's theorem in centralized environments hold in distributed settings. Notably, the models under consideration will be similar in spirit to their classical counterparts; indeed, similar techniques can be applied to prove results.

Note that [30] studies a similar model for distributed learning under communication constraints. Whereas [30] allocates regions of feature space among agents, here we allocate observations of an i.i.d. sampling process. Moreover, here we study a richer class of communication constraints. A related area of research lies in the study of ensemble methods in machine learning; examples of these techniques include bagging, boosting, mixtures of experts, and others [13], [4], [9], [10], [15]. These techniques are similar to the problem of interest here in that they aggregate many individually trained classifiers. However, the focus of these works is on the statistical and algorithmic advantages of learning with an ensemble and not on the nature of learning under communication constraints. Notably, [14] considered a probably approximately correct (PAC)-like model for learning with many individually trained hypotheses in a distribution-specific (i.e., parametric) framework.

Numerous other works in the literature are relevant to the research presented here. However, different points need to be made depending on whether we consider regression or classification with or without abstention. Lacking such context here, we will save such discussion of these results for the appropriate sections in the paper.

C. Organization

The remainder of this paper is organized as follows. In Section II, the notation and technical assumptions relevant to the remainder of the paper are introduced. In Sections III, and IV,

we study the models for binary classification in communication with and without abstention, respectively. In Sections V and VI, we study the models for regression estimation with and without abstention in turn. In each section, we present the main results, discuss important connections to other work in nonparametric statistics, and then proceed with a proof that further emphasizes differences from classical analyses like Stone's theorem. In Section VII, we conclude with a discussion of future work. Technical lemmas that are readily apparent from the literature are left to the Appendix.

II. Preliminaries

In this section, we introduce notation and technical assumptions relevant to the remainder of the paper.

As stated earlier, let X and Y be $\mathcal{X}$-valued and $\mathcal{Y}$-valued random variables, respectively, with a joint distribution denoted by $\boldsymbol{P}_{XY}$. $\mathcal{X}$ is known as the feature, input, or observation space; $\mathcal{Y}$ is known as the label, output, or target space. Throughout, we will take $\mathcal{X} \subseteq \mathbb{R}^d$ and consider two cases corresponding to binary classification ($\mathcal{Y} = \{0,1\}$) and regression estimation ($\mathcal{Y} = \mathbb{R}$). Let $D_n = \{(X_i, Y_i)\}_{i=1}^n$ denote an i.i.d. collection of training data with $(X_i, Y_i) \sim \boldsymbol{P}_{XY}$ for all $i \in \{1, \ldots, n\}$.

Throughout this paper, we will use δ_{ni} to denote the randomized response of the ith learning agent in an ensemble of n agents. For each $i \in \{1, \ldots, n\}$, δ_{ni} is an $\mathcal{S}$-valued random variable, where $\mathcal{S}$ is the decision space for the agent; in models *with abstention* we take $\mathcal{S} = \{\text{abstain}, 1, 0\}$ and in models *without abstention* we take $\mathcal{S} = \{1, 0\}$. As an important consequence of the assumed lack of interagent communication and the assumption that D_n is i.i.d., we have the following observation which will be fundamental to the subsequent analysis.

(A) The ith agent's response δ_{ni} may be dependent on X, X_i, and Y_i, but is statistically independent of $\{(X_j, Y_j)\}_{j \neq i}$ and conditionally independent of $\{\delta_{nj}\}_{j \neq i}$ given X.

Thus, to specify δ_{ni} and thereby design agent decision rules, it suffices to define the conditional distribution $\boldsymbol{P}\{\delta_{ni} \,|X, X_i, Y_i\}$ for all $(X, X_i, Y_i) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y}$. In each of the subsequent sections, we will find it convenient to do so by specifying a function

$$\bar{\delta}_n(x) : \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \to \{\text{abstain}\} \cup [0,1].$$

In particular, we define

$$\begin{aligned}
\boldsymbol{P}\{\delta_{ni} = \text{abstain} \,|X, X_i, Y_i\} &= \begin{cases} 1, & \text{if } \bar{\delta}_n(X, X_i, Y_i) = \text{abstain} \\ 0, & \text{otherwise} \end{cases} \\
\boldsymbol{P}\{\delta_{ni} = 1 \,|X, X_i, Y_i\} &= \begin{cases} 0, & \text{if } \bar{\delta}_n(X, X_i, Y_i) = \text{abstain} \\ \bar{\delta}_n(X, X_i, Y_i), & \text{otherwise} \end{cases} \\
\boldsymbol{P}\{\delta_{ni} = 0 \,|X, X_i, Y_i\} &= \begin{cases} 0, & \text{if } \bar{\delta}_n(X, X_i, Y_i) = \text{abstain} \\ 1 - \bar{\delta}_n(X, X_i, Y_i), & \text{otherwise.} \end{cases}
\end{aligned} \tag{1}$$

It is straightforward to verify that (1) is a valid probability distribution for every $(X, X_i, Y_i) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y}$. Therefore, together with (A), δ_{ni} is clearly specified by $\bar{\delta}_{ni}(x)$ and (1).

Note that this formalism serves merely as a technical convenience and should not mask the simplicity of the agent decision rules. In words, an agent will abstain from voting if $\bar{\delta}_n(X, X_i, Y_i) = \text{abstain}$; else, the agent flips a biased coin to send 1 or 0, with the bias determined by $\bar{\delta}_n(X, X_i, Y_i)$. Though this formalism may appear restrictive since rules of this form do not allow randomized decisions to abstain, the results in this paper do not rely on this flexibility.

To emphasize, note that communication is constrained between the agents and the fusion center via the limited decision space $\mathcal{S}$ and as above, communication between agents is not allowed (the latter is a necessary precondition for observation (A)). Consistent with the notation, we assume that the agents have knowledge of n, the number of agents in the ensemble. Moreover, we assume that for each n, every agent has the same local decision rule; i.e., the ensemble is homogenous in this sense. An underlying assumption is that each agent is able to generate random numbers, independent of the rest of the network.

Consistent with convention, we use

$$g_n(x) = g_n(x, \{\delta_{ni}\}_{i=1}^n) : \mathcal{X} \times \mathcal{S}^n \to \{0,1\}$$

to denote the central authority's fusion rule in the binary classification frameworks and similarly, we use

$$\hat{\eta}_n(x) = \hat{\eta}_n(x, \{\delta_{ni}\}_{i=1}^n\}) : \mathcal{X} \times \mathcal{S}^n \to \mathbb{R}$$

to denote its fusion rule in the regression frameworks. In defining fusion rules throughout the remainder of the paper, it will be convenient to denote the random set

$$I_V = I_V(X, D_n) \triangleq \{i \in \{1, \ldots, n\} : \delta_{ni} \neq \text{abstain}\}$$

as the set of agents that vote and, hence, do not abstain. To emphasize the central authority's primary role of aggregating the response of the network, we shall henceforth refer to this agent as a *fusion center*.

Defining a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, we seek a decision rule $f : \mathcal{X} \to \mathcal{Y}$ (i.e., agent decision rules and a fusion rule) that minimizes the expected loss $\boldsymbol{E}\{l(f(X), Y)\}$. In the binary classification setting, the criterion of interest is the probability of misclassification; we let $l(y, y') = 1_{\{y \neq y'\}}$, the well-known zero–one loss. The structure of the risk minimizing maximum *a posteriori* probability (MAP) decision rule is well understood [7]; let $\delta_B : \mathcal{X} \to \{0,1\}$ denote this Bayes decision rule. In regression settings, we consider the squared error criterion; we let $l(y, y') = |y - y'|^2$. It is well known that the regression function

$$\eta(x) = \boldsymbol{E}\{Y \,|X = x\} \tag{2}$$

achieves the minimal expected loss in this case. Throughout the remainder of the paper, we let $L^\star = \inf_f \boldsymbol{E}\{l(f(X), Y)\}$ denote the minimal expected loss. Depending on whether we find ourselves in the binary classification or regression setting, it will be clear from the context whether $L^\star$ refers to the optimal (binary) Bayes risk or minimal mean squared error.

In this work, we focus on the statistical property known as *universal consistency* [7], [12], defined as follows.

Definition 1: Let $L_n = \boldsymbol{E}\{l(f_n(X, D_n), Y)\,|D_n\}$. $\{f_n\}_{n=1}^{\infty}$ is said to be *universally consistent* if $\bar{\boldsymbol{E}}\{L_n\} \to L^\star$ for *all* distributions $\boldsymbol{P}_{XY}$.

This definition requires convergence in expectation and according to convention, defines *weak* universal consistency. This notion is contrasted with *strong* universal consistency where $L_n \to L^\star$ almost surely. Extending results of weak universal consistency to the strong sense has generally required the theory of large deviations, in particular McDiarmid's inequality [7]. Though the focus in this paper is on the weaker sense, the results in this paper might be extended to strong universal consistency using similar techniques. In particular, note that since consistency in distributed classification *with abstention* can be reduced to Stone's theorem, the extension to strong universal consistency follows immediately from standard results. Further, the negative result for distributed regression *without abstention* automatically precludes consistency in the strong sense. An extension for distributed classification without abstention and distributed regression with abstention may be possible under a refined analysis; the authors leave such analysis for future research.

III. Distributed Classification With Abstention: Stone's Theorem

In this section, we show that the universal consistency of distributed classification with abstention follows immediately from Stone's theorem and the classical analysis of naive kernel classifiers. To start, let us briefly recap the model. Since we are in the classification framework, $\mathcal{Y} = \{0, 1\}$. Suppose that for each $i \in \{1, \ldots, n\}$, the training datum $(X_i, Y_i) \in D_n$ is received by a distinct member of a network of n learning agents. When the fusion center observes a new observation $X \sim \boldsymbol{P}_X$, it broadcasts the observation to the network in a request for information. At this time, each of the learning agents can respond with at most one bit. That is, each learning agent chooses whether or not to respond to the fusion center's request for information; and if an agent chooses to respond, it sends either a 1 or a 0 based on a local decision algorithm. Upon receiving the agents' responses, the fusion center combines the information to create an estimate of Y.

To answer the question of whether agent decision rules and fusion rules exist that result in a universally consistent ensemble, let us construct one natural choice. With

$$B_{r_n}(x) = \{x' \in \mathbb{R}^d : \| x - x' \|_2 \leq r_n\}$$

let

$$\bar{\delta}_n(x, X_i, Y_i) = \begin{cases} Y_i, & \text{if } X_i \in B_{r_n}(x) \\ \text{abstain}, & \text{otherwise} \end{cases} \tag{3}$$

and

$$g_n(x) = \begin{cases} 1, & \text{if } \sum_{i \in I_V} \delta_{ni} \geq \frac{1}{2}|I_V| \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

so that $g_n(x)$ amounts to a majority vote fusion rule. Recall from (1) that the agents' randomized responses are defined by $\bar{\delta}_n(\cdot)$. In words, agents respond according to their training data label as long as the new observation X is sufficiently close to their training observation X_i; else, they abstain. In this model with abstention, note that δ_{ni} is $\{\text{abstain}, 1, 0\}$-valued since Y_i is binary-valued and thus, the communications constraints are obeyed.

With this choice, it is straightforward to see that the net decision rule is equivalent to the plug-in kernel classifier rule with the naive kernel. Indeed

$$g_n(x) = \begin{cases} 1, & \text{if } \frac{\sum_{i=1}^n Y_i 1_{B_{r_n}(x)}(X_i)}{\sum_{i=1}^n 1_{B_{r_n}(x)}(X_i)} \geq \frac{1}{2} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

With this equivalence,[1] the universal consistency of the ensemble follows from Stone's theorem applied to naive kernel classifiers. With $L_n = \boldsymbol{P}\{g_n(X) \neq Y\,|D_n\}$, the probability of error of the ensemble conditioned on the random training data, we state this known result without proof as Theorem 1.

Theorem 1: ([7]) If $r_n \to 0$ and $(r_n)^d n \to \infty$ as $n \to \infty$, then $\boldsymbol{E}\{L_n\} \to L^*$ for all distributions $\boldsymbol{P}_{XY}$.

The kernel classifier with the naive kernel is somewhat unique among other frequently analyzed universally consistent classifiers in its relevance to the current model. More general kernels (for instance, a Gaussian kernel) are not easily applicable as the real-valued weights do not naturally form a randomized decision rule. Furthermore, nearest neighbor rules do not apply as a given agent's decision rule would then need to depend on the data observed by the other agents; such interagent communication is not allowed in the current model.

IV. Distributed Classification Without Abstention

As noted in the Introduction, given the result of the previous section, it is natural to ask whether the communication constraints can be tightened. Let us consider the second model in which the agents cannot choose to abstain. In effect, each agent communicates one bit per decision. Again, we consider the binary classification framework but as a technical convenience, adjust our notation so that $\mathcal{Y} = \{+1, -1\}$ instead of the usual $\{0, 1\}$; also, agents now decide between sending ± 1. The formalism introduced in Section II can be extended naturally to allow this slight modification; we allow δ_{ni} to be specified so that $\boldsymbol{P}\{\delta_{ni} = +1\,|X, X_i, Y_i\} = \bar{\delta}_{ni}(x, X_i, Y_i)$. We again consider whether universally Bayes-risk consistent schemes exist for the ensemble.

Consider the randomized agent decision rule specified as follows:

$$\bar{\delta}_{ni}(x, X_i, Y_i) = \begin{cases} \frac{1}{2}Y_i + \frac{1}{2}, & \text{if } X_i \in B_{r_n}(x) \\ \frac{1}{2}, & \text{otherwise.} \end{cases} \tag{6}$$

Recall from (1) that the agents' randomized responses are defined by $\bar{\delta}_n(\cdot)$. Note that $\boldsymbol{P}\{\delta_{ni} = Y_i\,|X_i \in B_{r_n}(x)\} = 1$, and thus, the agents respond according to their training data label if x is sufficiently close to X_i. Else, they simply "guess," flipping an unbiased coin. In this model without abstention, it is readily verified that each agent transmits one bit per decision as δ_{ni} is $\{\pm 1\}$-valued since $\boldsymbol{P}\{\delta_{ni} = \text{abstain}\} = 0$; thus, the communication constraints are obeyed.

[1]Strictly speaking, this equality holds almost surely (a.s.), since the agents' responses are random variables.

A natural fusion rule is the majority vote. That is, the fusion center decides according to

$$g_n(x) = \begin{cases} 1, & \text{if } \sum_{i=1}^n \delta_{ni} > 0 \\ -1, & \text{otherwise.} \end{cases} \tag{7}$$

As before, the natural performance metric for the ensemble is the probability of misclassification. Modifying our convention slightly, let $D_n = \{(X_i, Y_i, \delta_{ni})\}_{i=1}^n$ and define

$$L_n = \boldsymbol{P}\{g_n(X) \neq Y \,|D_n\}. \tag{8}$$

That is, L_n is the conditional probability of error of the majority vote fusion rule conditioned on the randomness in agent training and agent decision rules.

A. Main Result and Comments

Theorem 2 specifies sufficient conditions for consistency for an ensemble using the described decision rules.

Theorem 2: If $r_n \to 0$ and $(r_n)^d\sqrt{n} \to \infty$ as $n \to \infty$, then $\boldsymbol{E}\{L_n\} \to L^*$.

Yet again, the conditions of the theorem strike a similarity with consistency results for kernel classifiers using the naive kernel. Indeed, $r_n \to 0$ ensures that the bias of the classifier decays to zero. However, $\{r_n\}_{n=1}^\infty$ must not decay too rapidly. As the number of agents in the ensemble grows large, many, indeed most, of the agents will be "guessing" for any given classification; in general, only a decaying fraction of the agents will respond with useful information. In order to ensure that these informative bits can be heard through the noise introduced by the guessing agents, $(r_n)^d\sqrt{n} \to \infty$. Note the difference between this result and that for naive kernel classifiers where $(r_n)^d n \to \infty$ assures a sufficient rate of convergence for $\{r_n\}_{n=1}^\infty$.

Notably, to prove this result, we show directly that the expected probability of misclassification converges to the Bayes rate. This is unlike techniques commonly used to demonstrate the consistency of kernel classifiers, etc., which are so-called "plug-in" classification rules. These rules estimate the *a posteriori* probabilities $\boldsymbol{P}\{Y = i \,|X\}, i = \pm 1$ and construct classifiers based on thresholding the estimate. In this setting, it suffices to show that these estimates converge to the true probabilities in $L^p(\boldsymbol{P}_X)$. However, for this model, we cannot estimate the *a posteriori* probabilities and must resort to another proof technique; this observation foreshadows the negative result of Section VI.

With our choice of "coin flipping" agent decision rules, one may be tempted to model the observations made by the fusion center as noise-corrupted labels from the training set and to thereby recover Theorem 2 from the literature on learning with noisy data. However, note that since the fusion center does not have access to the agents' feature observations (i.e., $\{X_i\}_{i=1}^n$), the fusion rule cannot in general be modeled as a "plug-in" classication rule as analyzed, for instance, in [24]. Moreover, in contrast to the noise models considered in [24], the agent decision rules here are statistically dependent on X and are also dependent on X_i in an atypical way: the noise statistics depend on n and for particular $\boldsymbol{P}_{XY}$, one can show that as n increases without bound, the probability that an agent guesses (a label is noisy) grows toward 1. These differences distinguish Theorem 2 from results in the literature on learning with noisy data.

B. Proof of Theorem 2

Proof: Fix an arbitrary $\epsilon > 0$. We will show that $\boldsymbol{E}\{L_n\} - L^*$ is less than ϵ for all sufficiently large n. Using the notation in (2), we write

$$\begin{aligned}\eta(x) &= \boldsymbol{E}\{Y \,|X = x\} \\ &= \boldsymbol{P}\{Y = +1 \,|X = x\} - \boldsymbol{P}\{Y = -1 \,|X = x\}\end{aligned}$$

and define $A_\epsilon = \{x : |\eta(x)| > \frac{\epsilon}{2}\}$. It follows that

$$\begin{aligned}\boldsymbol{E}\{L_n\} - L^* &= \boldsymbol{E}\left\{\boldsymbol{P}\{g_n(X) \neq Y \,|D_n\}\right\} - \boldsymbol{P}\{\delta_B(X) \neq Y\} \\ &= \boldsymbol{E}\left\{(\boldsymbol{P}\{g_n(X) \neq Y \,|D_n, X\} \right. \\ &\qquad \left. - \boldsymbol{P}\{\delta_B(X) \neq Y \,|X\}) \cdot (1_{A_\epsilon}(X) + 1_{\bar{A}_\epsilon}(X))\right\}\end{aligned} \tag{9}$$

with the expectation in (9) being taken with respect to X and D_n. Note that for all $x \in \bar{A}_\epsilon$

$$\boldsymbol{P}\{\delta_B(X) \neq Y \,|X = x\} = \frac{1}{2} - \frac{|\eta(x)|}{2} \geq \frac{1}{2} - \frac{\epsilon}{4}$$

and therefore,

$$\boldsymbol{P}\{g_n(X) \neq Y \,|D_n, X\} \leq 1 - \boldsymbol{P}\{\delta_B(X) \neq Y \,|X = x\} \leq \frac{1}{2} + \frac{\epsilon}{4}.$$

Thus,

$$\begin{aligned}\boldsymbol{E}\{L_n\} - L^* &\leq \boldsymbol{E}\left\{(\boldsymbol{P}\{g_n(X) \neq Y \,|D_n, X\} \right. \\ &\qquad \left. - \boldsymbol{P}\{\delta_B(X) \neq Y \,|X\})1_{A_\epsilon}(X) + \frac{\epsilon}{2}\right\} \\ &\leq \boldsymbol{P}\{g_n(X) \neq \delta_B(X) \,|X \in A_\epsilon\}\boldsymbol{P}\{A_\epsilon\} + \frac{\epsilon}{2}.\end{aligned}$$

Note that if $\boldsymbol{P}\{A_\epsilon\} = 0$, then the proof is complete. Let us proceed assuming $\boldsymbol{P}\{A_\epsilon\} > 0$. Clearly, it suffices to show that $\lim_{n\to\infty} \boldsymbol{P}\{g_n(X) \neq \delta_B(X) \,\big|\, X \in A_\epsilon\} \leq \frac{\epsilon}{2}$. Let us define the quantities

$$\begin{aligned}m_n(x) &= \boldsymbol{E}\{\eta(X)\delta_{ni} \,|X = x\} \\ \sigma_n^2(x) &= \boldsymbol{E}\{|\eta(X)\delta_{ni} - m_n(X)|^2 \,|X = x\}\end{aligned}$$

with the expectation being taken over the random training data and the randomness introduced by the agent decision rules. Respectively, $m_n(x)$ and $\sigma_n^2(x)$ can be interpreted as the mean and variance of the "margin" of the agent response δ_{ni}, conditioned on the observation X. For large positive $m_n(x)$, the agents can be expected to respond "confidently" (with large margin) according to the Bayes rule when asked to classify an object x. For large $\sigma_n^2(x)$, the fusion center can expect to observe a large variance among the individual agent responses to x.

Fix any integer $k > 0$. Consider the sequence of sets indexed by n

$$B_{n,k} = \{x \in \mathcal{X} : m_n(x)n > k\sqrt{n}\sigma_n(x)\}$$

so that $x \in B_{n,k}$ if and only if $\frac{m_n(x)\sqrt{n}}{\sigma_n(x)} > k$. We can interpret $B_{n,k}$ as the set of observations for which informed agents

have a sufficiently strong signal compared with the noise of the guessing agents. Then

$$\begin{aligned}
&\boldsymbol{P}\{g_n(X) \neq \delta_B(X) \,|\, X \in A_\epsilon\} \\
&\quad = \boldsymbol{P}\left\{ \eta(X) \sum_{i=1}^{n} \delta_{ni} < 0 \,\middle|\, X \in A_\epsilon \right\} \qquad (10)\\
&\quad = \boldsymbol{P}\left\{ \eta(X) \sum_{i=1}^{n} \delta_{ni} < 0 \,\middle|\, X \in A_\epsilon \cap B_{n,k} \right\} \\
&\qquad \cdot \boldsymbol{P}\{X \in B_{n,k} \,|X \in A_\epsilon\} \\
&\qquad + \boldsymbol{P}\left\{ \eta(X) \sum_{i=1}^{n} \delta_{ni} < 0 \,\middle|\, X \in A_\epsilon \cap \bar{B}_{n,k} \right\} \\
&\qquad \cdot \boldsymbol{P}\{X \in \bar{B}_{n,k} \,|X \in A_\epsilon\}. \qquad (11)
\end{aligned}$$

Note that conditioned on $X, \eta(X)\sum_{i=1}^{n}\delta_{ni}$ is a sum of i.i.d. random variables with mean $m_n(X)$ and variance $\sigma_n^2(X)$. Further, for $x \in B_{n,k}, \eta(x)\sum_{i=1}^{n}\delta_{ni} < 0$ implies

$$\left| \eta(x) \sum_{i=1}^{n} \delta_{ni} - m_n(x) n \right| > k\sqrt{n}\sigma_n^2(x).$$

Thus, it is straightforward to see that,

$$\begin{aligned}
&\boldsymbol{P}\left\{ \eta(X) \sum_{i=1}^{n} \delta_{ni} < 0 \,\middle|\, X \in A_\epsilon \cap B_{n,k} \right\} \\
&\quad = \boldsymbol{E}\left\{ \boldsymbol{P}\left\{ \eta(X) \sum_{i=1}^{n} \delta_{ni} < 0 \,\middle|\, X \right\} \,\middle|\, X \in A_\epsilon \cap B_{n,k} \right\} \\
&\quad \leq \boldsymbol{E}\left\{ \boldsymbol{P}\left\{ \left| \eta(X) \sum_{i=1}^{n} \delta_{ni} - m_n(X) n \right| > k\sqrt{n}\sigma_n(X) \,\middle|\, X \right\} \,\middle|\, X \in A_\epsilon \cap B_{n,k} \right\} \\
&\quad \leq \frac{1}{k^2}.
\end{aligned}$$

Here, the last statement follows from Markov's inequality. Choosing k sufficiently large and returning to (11)

$$\boldsymbol{P}\{g_n(X) \neq \delta_B(X) \,|X \in A_\epsilon\} \leq \frac{\epsilon}{2} + \boldsymbol{P}\{X \in \bar{B}_{n,k} \,|X \in A_\epsilon\}.$$

Now let us determine specific expressions for $m_n(x)$ and $\sigma_n^2(x)$, as dictated by our choice of agent decision rules. Clearly

$$\begin{aligned}
m_n(x) &= \eta(x)\boldsymbol{E}\{\delta_{ni} \,|X = x\} \\
&= \eta(x)\boldsymbol{E}\{\boldsymbol{E}\{2\bar{\delta}_{ni}(X, X_i, Y_i) - 1 \,|X, X_i, Y_i\} \,|X = x\} \\
&= \eta(x)\left(0 \cdot \boldsymbol{P}\left\{X_i \in \bar{B}_{r_n}(x)\right\}\right. \\
&\quad \left. + \eta_n(x) \cdot \boldsymbol{P}\left\{X_i \in B_{r_n}(x)\right\}\right) \\
&= \eta(x)\eta_n(x) \int 1_{B_{r_n}(x)}(y) P_X(dy)
\end{aligned}$$

with $\eta_n(x) = \boldsymbol{E}\{\eta(X) \,|X \in B_{r_n}(x)\}$. Also,

$$\begin{aligned}
\sigma_n^2(x) &= \eta^2(x)\boldsymbol{E}\{|\delta_{ni} - \boldsymbol{E}\{\delta_{ni} \,|X = x\}|^2 \,|X = x\} \\
&= \eta^2(x)(1 - \boldsymbol{E}\{\delta_{ni} \,|X = x\}^2).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\boldsymbol{P}\{X \in \bar{B}_{n,k} \,|X \in A_\epsilon\} \\
&= \boldsymbol{P}\{m_n(X)n < k\sqrt{n}\sigma_n(X) \,|X \in A_\epsilon\} \\
&= \boldsymbol{P}\left\{ \frac{\eta(X)\eta_n(X) \int 1_{B_{r_n}(X)}(y) P_X(dy)\sqrt{n}}{|\eta(X)|\sqrt{1 - \boldsymbol{E}\{\delta_{ni} \,|X\}^2}} < k \,\middle|\, X \in A_\epsilon \right\} \\
&= \boldsymbol{P}\left\{ (\mathrm{sgn}(\eta(X))\eta_n(X)) \right. \\
&\qquad \left. \cdot \left(\frac{\sqrt{n} \int 1_{B_{r_n}(X)}(y) P_X(dy)}{\sqrt{1 - \boldsymbol{E}\{\delta_{ni} \,|X\}^2}} \right) < k \,\middle|\, X \in A_\epsilon \right\}.
\end{aligned}$$

For any $1 \geq \gamma > 0$, we have

$$\begin{aligned}
&\boldsymbol{P}\{X \in \bar{B}_{n,k} \,|X \in A_\epsilon\} \\
&\quad \leq \boldsymbol{P}\left\{ \frac{\sqrt{n}}{\sqrt{1 - \boldsymbol{E}\{\delta_{ni} \,|X\}^2}} \int 1_{B_{r_n}(X)}(y) P_X(dy) \right. \\
&\qquad \left. < k \,\middle|\, X \in A_\epsilon, \mathrm{sgn}(\eta(X))\eta_n(X) > \gamma \right\} \\
&\qquad + \boldsymbol{P}\{\mathrm{sgn}(\eta(X))\eta_n(X) \leq \gamma \,|X \in A_\epsilon\}. \qquad (12)
\end{aligned}$$

First, consider the second term. With $\gamma = \frac{\epsilon}{4}$, it follows from our choice of A_ϵ that $\{\mathrm{sgn}(\eta(X))\eta_n(X) \leq \frac{\epsilon}{4}\}$ implies $\{|\eta(X) - \eta_n(X)| > \frac{\epsilon}{4}\}$. Thus,

$$\begin{aligned}
&\boldsymbol{P}\left\{ \mathrm{sgn}(\eta(X))\eta_n(X) \leq \frac{\epsilon}{4} \,\middle|\, X \in A_\epsilon \right\} \\
&\qquad \leq \boldsymbol{P}\left\{ |\eta(X) - \eta_n(X)| > \frac{\epsilon}{4} \,\middle|\, X \in A_\epsilon \right\}.
\end{aligned}$$

Since by technical Lemma 2 (see the Appendix), $\eta_n(X) \to \eta(X)$ in probability and by assumption $\boldsymbol{P}\{A_\epsilon\} > 0$, it follows from technical Lemma 1 in the Appendix that

$$\boldsymbol{P}\{\mathrm{sgn}(\eta(X))\eta_n(X) \leq \frac{\epsilon}{4} \,|X \in A_\epsilon\} \to 0.$$

Returning to (12) with $\gamma = (\frac{\epsilon}{4})$, note that we have just demonstrated that $\lim_{n\to\infty} \boldsymbol{P}\{\mathrm{sgn}(\eta(X))\eta_n(X) > \frac{\epsilon}{4}\} = 1$. Thus, to show that the first term converges to zero, by technical Lemma 1, it suffices to show that

$$\frac{\sqrt{n}}{\sqrt{1 - \boldsymbol{E}\{\delta_{ni} \,|X\}^2}} \int 1_{B_{r_n}(X)}(y) P_X(dy) \to \infty \text{ i.p.} \qquad (13)$$

Since $\frac{1}{\sqrt{1-\boldsymbol{E}\{\delta_{ni} \,|X\}^2}} \geq 1$, this follows from technical Lemma 3 in the Appendix and the fact that $(r_n)^d\sqrt{n} \to \infty$. This completes the proof. □

V. Distributed Regression With Abstention

We now turn our attention to distributed regression. As in Section III, the model remains the same except that now $\mathcal{Y} = \mathbb{R}$; that is, Y is an $\mathbb{R}$-valued random variable and likewise, agents receive real-valued training data labels Y_i. In this section, we consider communication with abstention. With the aim of determining whether universally consistent ensembles can be constructed, let us devise candidate rules.

For some as yet unspecified sequence of functions $T_n : \mathbb{R} \to [0,1]$ and a sequence of real numbers $\{r_n\}_{n=1}^{\infty}$, consider the randomized agent decision rules specified as follows:

$$\bar{\delta}_{ni}(x) = \begin{cases} T_n(Y_i) & \text{if } X_i \in B_{r_n}(x) \\ \text{abstain}, & \text{otherwise} \end{cases} \tag{14}$$

for $i = 1, \ldots, n$. In words, the agents choose to vote only if X_i is close enough to X; to vote, they flip a biased coin, with the bias determined by the size of the ensemble n and Y_i, via the function $T_n(\cdot)$. In this model with abstention, note that δ_{ni} is $\{\text{abstain}, 1, 0\}$-valued and thus, the communication constraints are obeyed.

It is intuitively clear that $T_n(\cdot)$ should be designed so that the realization of random bit $\delta_{n,i}$ reveals information about the real-valued label Y_i to the fusion center. In particular, it is natural to ask whether any continuous bijective mapping $\mathbb{R}$ to the interval $(0,1)$ would suffice in biasing the coin in a manner that is informative enough to provide universal consistency. For example, one might chose $T_n(y) = T(y) = \frac{1}{1+e^{-y}}$ and consider agent decision rules of the form (14) in conjunction with a fusion rule like

$$\hat{\eta}_n(x) = T^{-1}\left(\frac{\sum_{i \in I_V} \delta_{ni}}{|I_V|}\right). \tag{15}$$

Since agents have the flexibility to abstain, the fusion center can accurately estimate the average bias chosen by nonabstaining agents; the hope, then, is to determine the corresponding average label by inverting $T(\cdot)$. As will be observed in the proof of Theorem 3, such a choice is not possible in general, since $T(\cdot)$ is nonlinear; such an approach introduces a systematic bias to the estimator and thereby prevents consistency.

If, however, $|Y| \leq B$ a.s. for some known $B > 0$, it suffices to choose $T_n(\cdot)$ as the linear function mapping $[-B, B]$ to $[0,1]$. Since in this case, $T_n^{-1}(\boldsymbol{E}\{\delta_{n,i} \,|X, X_i\}) = \boldsymbol{E}\{Y_i \,|X_i\}$, universal consistency then follows with trivial modifications to the proof of Stone's theorem.

This intuition leads us to a rule that captures consistency in the general case. Though choices abound, we can choose T_n to be piecewise linear. In particular, let $\{c_n\}_{n=1}^{\infty}$ be an arbitrary sequence of real numbers such that $c_n \to \infty$ as $n \to \infty$ and choose

$$T_n(Y_i) = \begin{cases} \frac{1}{2c_n} Y_i + \frac{1}{2} & |Y_i| \leq c_n \\ \frac{1}{2}, & \text{otherwise} \end{cases} \tag{16}$$

and specify the fusion rule as

$$\hat{\eta}_n(x) = 2c_n\left(\frac{\sum_{i \in I_V} \delta_{ni}}{|I_V|} - \frac{1}{2}\right). \tag{17}$$

In words, the fusion center shifts and scales the average vote. For appropriately chosen sequences $\{c_n\}_{n=1}^{\infty}$ and $\{r_n\}_{n=1}^{\infty}$, this ensemble is universally consistent, as proved by Theorem 3.

In particular, let $D_n = \{(X_i, Y_i, \delta_{ni})\}_{t=1}^{n}$ and define $L_n = \boldsymbol{E}\{|\hat{\eta}(X) - Y|^2 \,|D_n\}$. That is, L_n is the conditional mean squared error, conditioned on the randomness in the training data and in the agent decision rules.

A. Main Result and Comments

Assuming an ensemble using the described decision rules, Theorem 3 specifies sufficient conditions for consistency.

Theorem 3: Suppose $\boldsymbol{P}_{XY}$ is such that $\boldsymbol{P}_X$ is compactly supported and $\boldsymbol{E}\{Y^2\} < \infty$. If, as $n \to \infty$

1) $c_n \to \infty$,
2) $r_n \to 0$, and
3) $(\frac{c_n^2}{n r_n^d}) \to 0$,

then $\boldsymbol{E}\{L_n\} \to L^{\star}$.

More generally, the constraint regarding the compactness of $\boldsymbol{P}_X$ can be weakened. As will be observed in the proof later, $\boldsymbol{P}_X$ must be such that when coupled with a bounded random variable Y, there is a known convergence rate of the variance term of the naive kernel classifier (under a standard i.i.d. sampling model). $\{c_n\}_{n=1}^{\infty}$ should be chosen so that it grows at a rate slower than the rate at which the variance term decays. Notably, to select $\{c_n\}_{n=1}^{\infty}$, one does not need to understand the convergence rate of the bias term, and this is why continuity conditions are not required; the bias term will converge to zero universally as long as $c_n \to \infty$ and $r_n \to 0$ as $n \to \infty$.

In observing the response of the network, the fusion center sees δ_{ni} from those agents who have not abstained. Since these random variables can be viewed as random quantizations or transformations of the labels in the training data, it is natural to ask whether the consistency of these rules follows as a special case of models for learning with noisy data. In this case, the underlying noise model would transform the label Y_i to the set $\{0,1\}$ in a manner that would be statistically dependent on X, X_i, Y_i itself and n. Though it is possible to view the current question in this framework, to our knowledge such a highly structured noise model has not been considered in the literature.

Finally, those familiar with the classical statistical pattern recognition literature will find the style of proof very familiar; special care must be taken to demonstrate that the variance of the estimate does not decrease too slowly compared to $\{c_n\}_{n=1}^{\infty}$ and to show that the bias introduced by the "clipped" agent decision rules converges to zero.

B. Proof of Theorem 3

Proof: By standard orthogonality arguments [12], it suffices to show that $\boldsymbol{E}\{|\hat{\eta}_n(X) - \eta(X)|^2\} \to 0$ as $n \to 0$.

Define $\bar{\eta}_n(x) \triangleq \boldsymbol{E}\{\delta_{ni} \,|X_i = x, \| X - X_i \| \leq r_n\}$. Proceeding in the traditional manner, note that by the standard inequality

$$(a_1 + \cdots + a_k)^2 \leq k(a_1^2 + \cdots + a_k^2) \tag{18}$$

it follows that

$$\begin{aligned} &\boldsymbol{E}\{|\hat{\eta}_n(X) - \eta(X)|^2\} \\ &\quad \leq 2\boldsymbol{E}\left\{\left|2c_n\left(\frac{\sum_{i \in I_V} \delta_{ni}}{|I_V|} - \frac{1}{2}\right)\right.\right. \\ &\qquad \left.\left. - 2c_n\left(\frac{\sum_{i \in I_V} \bar{\eta}_n(X_i)}{|I_V|} - \frac{1}{2}\right)\right|^2\right\} \\ &\qquad + 2\boldsymbol{E}\left\{\left|2c_n\left(\frac{\sum_{i \in I_V} \bar{\eta}_n(X_i)}{|I_V|} - \frac{1}{2}\right) - \eta(X)\right|^2\right\} \\ &\quad \triangleq J_n + K_n. \end{aligned}$$

Starting with the first term

$$J_n = 8c_n^2 \boldsymbol{E}\left\{\left|\frac{\sum_{i\in I_V}(\delta_{ni}-\bar{\eta}_n(X_i))}{|I_V|}\right|^2\right\}$$
$$= 8c_n^2 \boldsymbol{E}\left\{\boldsymbol{E}\left\{\frac{\sum_{i\in I_V}(\delta_{ni}-\bar{\eta}_n(X_i))^2}{|I_V|^2}\,\middle|\, X, X_1, \ldots, X_n\right\}\right\}.$$

Here, the first equality follows from algebra; the second follows after noting that for all $i \in I_V$, $\boldsymbol{E}\{\delta_{ni}\,|X, X_1, \ldots, X_n|\} = \hat{\eta}_n(X_i)$, and canceling out cross-terms in the expansion of the squared sum in the numerator. Note that conditioned on X and X_i, δ_{ni} is Bernoulli with parameter $\bar{\eta}_n(X_i)$ for all $i \in I_V$. Thus, bounding the variance of a Bernoulli random variable, we continue the preceding expression as

$$\leq 2c_n^2 \boldsymbol{E}\left\{\frac{1}{|I_V|}1_{\{|I_V|>0\}}\right\}.$$

Here we have applied the convention $\frac{0}{0} = 0$. Conditioning on X and applying technical Lemma 4 (see the Appendix) to the binomial random variable $|I_V| = \sum_{i=1}^n 1_{\{X_i \in B_{r_n}(X)\}}$, it follows that

$$J_n \;\leq\; 2c_n^2 \boldsymbol{E}\left\{\frac{2}{n\boldsymbol{P}_{X_1}\{X_1 \in B_{r_n}(X)\}}\right\}. \tag{19}$$

Here, for convenience, we have exploited the fact that D_n is i.i.d. and reused the variable X_1. Since $\boldsymbol{P}_X$ is compactly supported, the expectation in (19) can be bounded by a term $O(\frac{1}{nr_n^d})$ using an argument typically used to demonstrate the consistency of kernel estimators [12]. For completeness, we include it here.

Since S, the support of $\boldsymbol{P}_X$, is compact, we can find $z_1, \ldots, z_{M_n} \in \mathbb{R}^d$ such that $S \subseteq \cup_i^{M_n} B_{r_n/2}(z_i)$ and $M_n \leq \frac{b_1}{r_n^d}$ for some constant b_1. Thus,

$$\begin{aligned}
&2c_n^2 \boldsymbol{E}\left\{\frac{2}{n\boldsymbol{P}_{X_1}\{X_1 \in B_{r_n}(X)\}}\right\} \\
&\quad\leq 4c_n^2 \sum_{i=1}^{M_n} \boldsymbol{E}\left\{\frac{1_{\{B_{r_n/2}(z_i)\}}(X)}{n\boldsymbol{P}_{X_1}\{X_1 \in B_{r_n}(X)\}}\right\} \\
&\quad\leq 4c_n^2 \sum_{i=1}^{M_n} \boldsymbol{E}\left\{\frac{1_{\{B_{r_n/2}(z_i)\}}(X)}{n\boldsymbol{P}_{X_1}\{X_1 \in B_{r_n/2}(z_i)\}}\right\} \\
&\quad= \frac{4c_n^2 M_n}{n} \\
&\quad\leq \frac{4b_1 c_n^2}{nr_n^d}.
\end{aligned}$$

Finally, by condition 3) of Theorem 3, it follows that $J_n \to 0$. Note that J_n is essentially the variance of the estimator. Much of the work thus far has been the same as showing that in traditional i.i.d. sampling process settings, the variance of the naive kernel is universally bounded by a term $O(\frac{1}{nr_n^d})$ when $\boldsymbol{P}_X$ is compactly supported and Y is bounded [12]. This observation is consistent with the preceding comments.

Now, let us consider K_n. Fix $\epsilon > 0$. We will show that for all sufficiently large n, $K_n < \epsilon$. Let $\eta_\epsilon(x)$ be a bounded continuous function with bounded support such that $\boldsymbol{E}\{|\eta_\epsilon(X) - \eta(X)|^2\} \leq \frac{\epsilon}{12}$. Since $\boldsymbol{E}\{Y^2\} < \infty$ implies that $\eta(x) \in L^2(\boldsymbol{P}_X)$, such a function is assured to exist; the set of bounded continuous functions with bounded support is dense in $L^2(\mu)$ for all probability measures μ. By (18)

$$\begin{aligned}
K_n &\leq 4\boldsymbol{E}\left\{\left|2c_n\left(\frac{\sum_{i\in I_V}\bar{\eta}_n(X_i)}{|I_V|} - \frac{1}{2}\right) - \frac{\sum_{i\in I_V}\eta_\epsilon(X_i)}{|I_V|}\right|^2\right\} \\
&\quad + 4\boldsymbol{E}\left\{\left|\frac{\sum_{i\in I_V}\eta_\epsilon(X_i)}{|I_V|} - \frac{\sum_{i\in I_V}\eta_\epsilon(X)}{|I_V|}\right|^2\right\} \\
&\quad + 4\boldsymbol{E}\left\{\left|\frac{\sum_{i\in I_V}\eta_\epsilon(X)}{|I_V|} - \eta_\epsilon(X)\right|^2\right\} \\
&\quad + 4\boldsymbol{E}\{|\eta_\epsilon(X) - \eta(X)|^2\} \\
&\triangleq 4(K_{n1} + K_{n2} + K_{n3} + K_{n4}).
\end{aligned}$$

First, consider K_{n1}.

$$\begin{aligned}
K_{n1} &= \boldsymbol{E}\left\{\left|\frac{\sum_{i\in I_V}\left(2c_n\left(\bar{\eta}_n(X_i) - \frac{1}{2}\right) - \eta_\epsilon(X_i)\right)}{|I_V|}1_{\{|I_V|>0\}} - c_n 1_{\{|I_V|=0\}}\right|^2\right\} \\
&\leq 2\boldsymbol{E}\left\{\left|\frac{\sum_{i\in I_V}(2c_n(\bar{\eta}_n(X_i) - \frac{1}{2}) - \eta_\epsilon(X_i))}{|I_V|}1_{\{|I_V|>0\}}\right|^2\right\} \\
&\quad + 2\boldsymbol{E}\left\{c_n^2 1_{\{|I_V|=0\}}\right\},
\end{aligned}$$

with the equality following from algebra and the inequality from (18). Then, noting that when conditioned on X, $|I_V| = \sum_{i=1}^n 1_{\{X_i \in B_{r_n}(X)\}}$ is binomial with parameter $P_{X_1}\{X_1 \in B_{r_n}(X)|X\}$, we continue

$$\begin{aligned}
K_{n1} &\leq 2\boldsymbol{E}\left\{\left|\frac{\sum_{i\in I_V}\left(2c_n\left(\bar{\eta}_n(X_i) - \frac{1}{2}\right) - \eta_\epsilon(X_i)\right)}{|I_V|}\right|^2\right\} \\
&\quad + 2\boldsymbol{E}\left\{c_n^2\left(1 - \boldsymbol{P}_{X_1}\{X_1 \in B_{r_n}(X)\}\right)^n\right\} \\
&\leq 2c\boldsymbol{E}\left\{\left|2c_n\left(\bar{\eta}_n(X) - \frac{1}{2}\right) - \eta_\epsilon(X)\right|^2\right\} \\
&\quad + 2\boldsymbol{E}\left\{\frac{2c_n^2}{n\boldsymbol{P}_{X_1}\{X_1 \in B_{r_n}(X)\}}\right\}.
\end{aligned}$$

Here, the second inequality follows for some constant c, in part by applying technical Lemma 5 and in part by noting

$$(1-x)^n \leq \exp(-nx) \leq \frac{1}{nx}, \text{ for } 0 \leq x \leq 1 \text{ and } n = 1, 2, \ldots.$$

Continuing by applying (18), we have

$$\begin{aligned}
K_{n1} &\leq 2c\boldsymbol{E}\left\{\left|2c_n\left(\bar{\eta}_n(X) - \frac{1}{2}\right) - \eta(X)\right|^2\right\} \\
&\quad + \boldsymbol{E}\{|\eta_\epsilon(X) - \eta(X)|^2\} \\
&\quad + \boldsymbol{E}\left\{\frac{4c_n^2}{n\boldsymbol{P}_{X_1}\{X_1 \in B_{r_n}(X)\}}\right\}.
\end{aligned}$$

For our specific choice of agent decision rules, note that

$$\begin{aligned}
\bar{\eta}_n(x) &= \boldsymbol{E}\{T_n(Y)\,|\,X = x\} \\
&= \boldsymbol{E}\left\{\left(\frac{1}{2c_n}Y + \frac{1}{2}\right)1_{\{|Y|\leq c_n\}} + \frac{1}{2}1_{\{|Y|>c_n\}}\,|X = x\right\}.
\end{aligned}$$

Substituting this above and applying Jensen's inequality, we have

$$\begin{aligned} K_{n1} &\le 2c\boldsymbol{E}\left\{|\boldsymbol{E}\left\{Y1_{\{|Y|>c_n\}}\,|\,X\right\}|^2\right\} + \frac{\epsilon}{12} \\ &\quad + \boldsymbol{E}\left\{\frac{4c_n^2}{n\boldsymbol{P}_{X_1}\{X_1 \in B_{r_n}(X)\}}\right\} \\ &\le 2c\boldsymbol{E}\left\{\boldsymbol{E}\left\{Y^2 1_{\{|Y|>c_n\}}\,|\,X\right\}\right\} + \frac{\epsilon}{12} \\ &\quad + \boldsymbol{E}\left\{\frac{4c_n^2}{n\boldsymbol{P}_{X_1}\{X_1 \in B_{r_n}(X)\}}\right\} \\ &= 2c\boldsymbol{E}\left\{Y^2 1_{\{|Y|>c_n\}}\right\} + \frac{\epsilon}{12} \\ &\quad + \boldsymbol{E}\left\{\frac{4c_n^2}{n\boldsymbol{P}_{X_1}\{X_1 \in B_{r_n}(X)\}}\right\}. \end{aligned} \tag{20}$$

Since $f_n(y) = y^2 1_{\{|y|>c_n\}}$ is a monotonically decreasing sequence of functions and $f_n(y) \to 0$ everywhere, then by the Monotone Convergence Theorem, the first term in (20) converges to zero. The third term in (20) converges to zero by the same argument that was applied for J_n. Thus, $\limsup_{n\to\infty} K_{n_1} \le \frac{\epsilon}{12}$.

Observe that η_ϵ is uniformly continuous, since by construction, it is a bounded continuous function with bounded support. Let $\delta > 0$ be such that if $\| x - x' \| < \delta$, then $|\eta_\epsilon(x) - \eta_\epsilon(x')| \le \sqrt{\frac{\epsilon}{12}}$. Since $r_n \to 0$, for all sufficiently large n, $r_n < \delta$. Thus, for all sufficiently large n

$$\begin{aligned} K_{n2} &= \boldsymbol{E}\left\{\left|\frac{\sum_{i\in I_V}(\eta_\epsilon(X_i) - \eta_\epsilon(X))}{|I_V|}\right|^2\right\} \\ &\le \frac{\epsilon}{12}, \end{aligned}$$

since for all $i \in I_V$, $\| X_i - X \| \le r_n$. Next, consider K_{n3}. We have

$$\begin{aligned} K_{n3} &= \boldsymbol{E}\left\{\eta_\epsilon(X)^2 1_{\{|I_V|=0\}}\right\} \\ &\le \sup_x(\eta_\epsilon(x)^2)\boldsymbol{E}\left\{1_{\{|I_V|=0\}}\right\} \\ &\le \sup_x(\eta_\epsilon(x)^2)\boldsymbol{E}\left\{\frac{2c_n^2}{n\boldsymbol{P}_{X_1}\{X_1 \in B_{r_n}(X)\}}\right\} \end{aligned}$$

in the usual way, as we see that $K_{n3} \to 0$. Finally, $K_{n4} \le \frac{\epsilon}{12}$ by our choice of $\eta_\epsilon(x)$. Thus,

$$\begin{aligned} \limsup_{n\to\infty} K_n &\le 4\left(\frac{\epsilon}{12} + \frac{\epsilon}{12} + 0 + \frac{\epsilon}{12}\right) \\ &= \epsilon. \end{aligned}$$

Since ϵ was arbitrary, it is clear that K_n converges to zero. This completes the proof. □

VI. Distributed Regression Without Abstention

Finally, let us consider the model for distributed regression without abstention. Now, $\mathcal{Y} = \mathbb{R}$; agents will receive real-valued training data labels Y_i. However, when asked to respond with information, they will reply with either 0 or 1, as abstention is not an option.

In this section, we first establish natural regularity conditions for candidate fusion rules and specify a reasonable class of agent decision rules. As an important negative result, we then demonstrate that for any agent decision rule within this class, there does not exist a regular fusion rule that is L_2 consistent for every distribution $\boldsymbol{P}_{XY}$. This result establishes the impossibility of universal consistency in this model for distributed regression without abstention for a restricted, but reasonable class of decision rules.

To begin, consider the set of agent decision rules specified according to (1) for some $\bar{\delta}_n(\cdot)$. In this model without abstention, we require that the implicit responses satisfy $\boldsymbol{P}\{\delta_{ni} = \text{abstain}\} = 0$, but we impose no additional constraints on the agent decision rules. With the formalism introduced in Section II, this assumption is equivalent to assuming

$$\{\bar{\delta}_n(\cdot)\}_{n=1}^\infty \subset \mathcal{A} = \{\delta : \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \to [0,1]\}.$$

A fusion rule consists of a sequence of functions $\{\hat{\eta}_n\}_{n=1}^\infty$ mapping $\mathcal{X} \times \mathcal{S}^n$ to $\mathcal{Y} = \mathbb{R}$. Recall from Section II, that we can regard $\mathcal{S} = \{1, 0\}$ in this model without abstention. To proceed, we require some regularity on $\{\hat{\eta}_n(\cdot)\}_{n=1}^\infty$. Namely, let us consider only fusion rules that satisfy the following assumptions.

(A1) $\hat{\eta}_n(x, \cdot)$ is permutation invariant for all $x \in \mathcal{X}$. That is, for all $x \in \mathcal{X}$, any $b \in \{0,1\}^n$, and any permutation of $b, b' \in \{0,1\}^n, \hat{\eta}_n(x, b) = \hat{\eta}_n(x, b')$.

(A2) For every $x \in \mathcal{X}, \hat{\eta}_n(x, \cdot)$ is Lipschitz in the average Hamming distance. That is, there exists a constant C such that

$$|\hat{\eta}_n(x, b_1) - \hat{\eta}_n(x, b_2)| \le C\frac{1}{n}\sum_{i=1}^n |b_{1i} - b_{2i}| \tag{21}$$

for every $b_1, b_2 \in \{0,1\}^n$.

Once again, take $D_n = \{(X_i, Y_i, \delta_{ni})\}_{i=1}^n$ and define $L_n = \boldsymbol{E}\{|\hat{\eta}_n(X) - Y|^2\,|D_n\}$. That is, L_n is the conditional mean squared error, conditioned on the randomness in the training data and in the agent decision rules.

A. Main Result and Comments

The following provides a negative result.

Theorem 4: For every sequence of agent decision rules specified according to (1) with a point-wise convergent sequence of functions $\{\bar{\delta}_n(\cdot)\}_{n=1}^\infty \subset \mathcal{A}$, there is no fusion rule $\{\hat{\eta}_n(\cdot)\}_{n=1}^\infty$ satisfying assumptions (A1) and (A2) such that

$$\lim_{n\to\infty} \boldsymbol{E}\{L_n\} = L^\star \tag{22}$$

for every distribution $\boldsymbol{P}_{XY}$ satisfying $\boldsymbol{E}\{Y^2\} < \infty$.

Note that there is nothing particularly special about the one bit regime and regression. In fact, under the conditions of the theorem, universal consistency cannot be achieved in a multiclass classification problem with even three possible labels. However, we consider regression as it illustrates the point nicely.

The restriction to distributions satisfying $\boldsymbol{E}\{Y^2\} < \infty$ actually strengthens this negative result, for without such a condition, Theorem 4 is trivial. In the proof, a counterexample is derived where Y is binary valued, a much stronger case that also satisfies this condition.

Further, the requirement that $\{\bar{\delta}_n(\cdot)\}_{n=1}^\infty$ be pointwise convergent is mild and is only a technical point in the proof. Indeed, the result can be trivially extended to allow for weaker notions of convergence.

B. Proof of Theorem 4

The proof will proceed by specifying two random variables (X, Y) and (X', Y') with

$$\eta(x) = \boldsymbol{E}\{Y\,|X = x\} \ne \boldsymbol{E}\{Y'\,|X' = x\} = \eta'(x).$$

Asymptotically, however, the fusion center's estimate will be indifferent to whether the agents are trained with random data distributed according to $\boldsymbol{P}_{XY}$ or $\boldsymbol{P}_{X'Y'}$. This observation will contradict universal consistency and complete the proof.

Proof: To start, fix a pointwise convergent sequence of functions $\{\bar{\delta}_n(\cdot)\}_{n=1}^{\infty} \subseteq \mathcal{A}$, arbitrary $x_0, x_1 \in \mathcal{X}$, and distinct $y_0, y_1 \in \mathbb{R}$. Let us specify a distribution $\boldsymbol{P}_{XY}$. Let $\boldsymbol{P}_X\{x_0\} = q, \boldsymbol{P}_X\{x_1\} = 1-q$, and $\boldsymbol{P}_{Y|X}\{Y = y_i | X = x_i\} = 1$ for $i = 0, 1$. Clearly, for this distribution $\eta(x_i) = y_i$ for $i = 0, 1$.

Suppose that the ensemble is trained with random data distributed according to $\boldsymbol{P}_{XY}$ and that the fusion center wishes to classify $X = x_0$. According to the model, after broadcasting X to the agents, the fusion center will observe a random sequence of n bits $\{\delta_{ni}\}_{i=1}^n$. For all $i \in \{1, \ldots, n\}$ and all n

$$\begin{aligned}\boldsymbol{P}\{\delta_{ni} = 1 \,|X = x_0\} \\ = \bar{\delta}_n(x_0, x_0, y_0)q + \bar{\delta}_n(x_0, x_1, y_1)(1-q).\end{aligned} \tag{23}$$

Now, let us define a sequence of auxiliary random variables $\{(X'_n, Y')\}_{n=1}^{\infty}$ with distributions satisfying

$$\begin{aligned}&\boldsymbol{P}_{X'_n}\{x_1\} \\ &= \frac{\bar{\delta}_n(x_0, x_0, y_0)q + \bar{\delta}_n(x_0, x_1, y_1)(1-q) - \bar{\delta}_n(x_0, x_1, y_1)}{\bar{\delta}_n(x_0, x_0, y_1) - \bar{\delta}_n(x_0, x_1, y_0)} \\ &\qquad \boldsymbol{P}_{X'_n}\{x_0\} = 1 - \boldsymbol{P}_{X'_n}\{x_1\} \\ &\boldsymbol{P}_{Y'|X'_n}\{Y' = y_{1-i} \,|X'_n = x_i\} = 1, \quad i = 0, 1.\end{aligned} \tag{24}$$

Here, $\eta'(x_i) = \boldsymbol{E}\{Y' \,|X'_n = x_i\} = y_{1-i}$. Suppose that the ensemble was trained with random data distributed according to $\boldsymbol{P}_{X'_n Y'}$ and let $\{\delta_{ni}^{(n)}\}_{i=1}^n$ denote the random response variables of the agents. Then, we have

$$\begin{aligned}&\boldsymbol{P}\left\{\delta_{ni}^{(n)} = 1 \middle| X'_n = x_0\right\} \\ &= \frac{\bar{\delta}_n(x_0, x_0, y_1)}{\bar{\delta}_n(x_0, x_0, y_1) - \bar{\delta}_n(x_0, x_1, y_0)}\Big(\bar{\delta}_n(x_0, x_0, y_0)q \\ &\qquad + \bar{\delta}_n(x_0, x_1, y_1)(1-q) - \bar{\delta}_n(x_0, x_1, y_1)\Big) \\ &\quad + \frac{\bar{\delta}_n(x_0, x_1, y_0)}{\bar{\delta}_n(x_0, x_0, y_1) - \bar{\delta}_n(x_0, x_1, y_0)}\Big(1 - \bar{\delta}_n(x_0, x_0, y_0)q \\ &\qquad + \bar{\delta}_n(x_0, x_1, y_1)(1-q) - \bar{\delta}_n(x_0, x_1, y_1)\Big) \\ &= \boldsymbol{P}\{\delta_{ni} = 1 \,|X = x_0\}\end{aligned} \tag{25}$$

for all n. Thus, conditioned on the observation to be labeled by the ensemble X (or X'_n), the fusion center will observe an identical stochastic process regardless of whether the ensemble was trained with data distributed according to $\boldsymbol{P}_{XY}$ or $\boldsymbol{P}_{X'_n Y'}$ for any fixed n. Note, this observation is true despite the fact that $\eta(x) \neq \eta'(x)$.

Finally, let (X', Y') be such that

$$\begin{aligned}\boldsymbol{P}_{X'}\{x_1\} &= \lim_{n\to\infty} \boldsymbol{P}_{X'_n}\{x_1\} \\ \boldsymbol{P}_{X'}\{x_0\} &= 1 - \boldsymbol{P}_{X'}\{x_1\} \\ \boldsymbol{P}_{Y'|X'}\{Y' = y_{1-i} \,|X' = x_i\} &= 1, \quad i = 0, 1.\end{aligned} \tag{26}$$

Again, $\eta'(x_i) = \boldsymbol{E}\{Y' \,|X' = x_i\} = y_{1-i}$. These limits are assured to exist by the assumption that $\{\bar{\delta}_n(\cdot)\}_{n=1}^{\infty}$ is a pointwise converging sequence of functions. Finally, let $\{\delta'_{ni}\}_{i=1}^n$ denote the random response random variables for the ensemble agents trained with data distributed according to $\boldsymbol{P}_{X'Y'}$.

By standard orthogonality arguments [12], for the ensemble to be universally consistent, we must have both

$$\boldsymbol{E}\left\{|\hat{\eta}_n(X, \{\delta_{ni}\}_{i=1}^n) - \eta(X)|^2\right\} \to 0 \tag{27}$$

and

$$\boldsymbol{E}\left\{|\hat{\eta}_n(X', \{\delta'_{ni}\}_{i=1}^n) - \eta'(X')|^2\right\} \to 0. \tag{28}$$

Let us assume that (27) holds; we now demonstrate that necessarily

$$\boldsymbol{E}\left\{|\hat{\eta}_n(X', \{\delta'_{ni}\}_{i=1}^n) - \eta(X')|^2\right\} \to 0. \tag{29}$$

Since $\eta(x) \neq \eta'(x)$, (29) contradicts (28) and the proposition of universal consistency. To show (29), it suffices to focus on the L^2 risk conditioned on X', due to the convenient point-mass structure of $\boldsymbol{P}_{X'}$. To proceed, note that by (18), for any $b \in \{0,1\}^n$

$$\begin{aligned}&\boldsymbol{E}\left\{|\hat{\eta}_n(X', \{\delta'_{ni}\}_{i=1}^n) - \eta(X')|^2 \,|X' = x_0\right\} \\ &\quad \leq 2\boldsymbol{E}\left\{|\hat{\eta}_n(X', b) - \eta(X')|^2 \,|X' = x_0\right\} \\ &\qquad + 2\boldsymbol{E}\left\{|\hat{\eta}_n(X', \{\delta'_{ni}\}_{i=1}^n) - \hat{\eta}_n(X', b)|^2 \,|X' = x_0\right\} \\ &\quad \triangleq 2T_1(b) + 2T_2(b).\end{aligned}$$

In particular, let us select $b \in \{0,1\}^n$ randomly such that the components are i.i.d. with $b_i \sim \boldsymbol{P}\{\delta_{ni} \,|X = x_0\}$ for all $i = 1, \ldots, n$. Note that if we can show that $\boldsymbol{E}_b\{T_1(b) + T_2(b)\} \to 0$, then the result holds by the probabilistic method. First consider $T_1(b)$. Note that we have

$$\begin{aligned}\boldsymbol{E}_b\{T_1(b)\} &= \boldsymbol{E}\left\{|\hat{\eta}_n(X', b) - \eta(X')|^2 \,|X' = x_0\right\} \\ &= \boldsymbol{E}\left\{\left|\hat{\eta}_n(X, \{\delta_{ni}\}_{i=1}^n) - \eta(X)\right|^2 \middle| X = x_0\right\},\end{aligned}$$

by our selection of b. Thus, $\boldsymbol{E}_b\{T_1(b)\}$ must converge to zero by the assumption that (27) holds true. Considering $T_2(b)$, note that

$$\begin{aligned}&\boldsymbol{E}_b\{T_2(b)\} \\ &= \boldsymbol{E}\left\{|\hat{\eta}_n(X', b) - \hat{\eta}_n(X', \{\delta'_{ni}\}_{i=1}^n)|^2 \,|X' = x_0\right\} \\ &\leq C^2\boldsymbol{E}\left\{\left|\frac{1}{n}\sum_{i=1}^n b_i - \frac{1}{n}\sum_{i=1}^n \delta'_{ni}\right|^2 \middle| X' = x_0\right\} \\ &\leq 3C^2\boldsymbol{E}\left\{\left|\frac{1}{n}\sum_{i=1}^n b_i - \boldsymbol{P}\{\delta_{ni} = 1 \,|X = x_0\}\right|^2\right\} \qquad (30) \\ &\quad + 3C^2\boldsymbol{E}\left\{\left|\frac{1}{n}\sum_{i=1}^n \delta'_{ni} - \boldsymbol{P}\{\delta'_{ni} = 1 \,|X' = x_0\}\right|^2 \middle| X' = x_0\right\} \qquad (31) \\ &\quad + 3C^2|\boldsymbol{P}\{\delta_{ni} = 1 \,|X = x_0\} - \boldsymbol{P}\{\delta'_{ni} = 1 \,|X' = x_0\}|^2. \qquad (32)\end{aligned}$$

Here, the first inequality follows from assumptions (A1) and (A2) and the second inequality follows by (18). Note that since $\{b_i\}_{i=1}^n$ is i.i.d. with $b_i \sim \boldsymbol{P}\{\delta_{ni} = 1 \,|X = x_0\}$

$$3C^2 \boldsymbol{E}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} b_i - \boldsymbol{P}\{\delta_{ni} = 1\,|X = x_0\}\right|^2\right\} \leq \frac{3C^2}{4n}$$

after bounding the variance of a binomial random variable; therefore, (30) must converge to zero. A similar argument can be applied to (31). Next, from (25)

$$\begin{aligned}&|\boldsymbol{P}\{\delta_{ni} = 1\,|X = x_0\} - \boldsymbol{P}\{\delta'_{ni} = 1\,|X' = x_0\}|^2 \\ &\quad = \left|\boldsymbol{P}\left\{\delta_{ni}^{(n)} = 1\,|X'_n = x_0\right\} - \boldsymbol{P}\{\delta'_{ni} = 1\,|X' = x_0\}\right|^2.\end{aligned}$$

Thus, (32) must converge to zero by our design of (X', Y') in (26). Finally, we have demonstrated that (29) holds true; by the discussion above, this completes the proof. □

VII. Conclusion and Future Work

Motivated by sensor networks and other distributed settings, this paper has presented several models for distributed learning. The models differ from classical works in statistical pattern recognition by allocating observations of an i.i.d. sampling process to individual learning agents. By limiting the ability of the agents to communicate, we constrain the amount of information available to the ensemble and to the fusion center for use in classification or regression. This setting models a distributed environment and presents new questions to consider with regard to universal consistency.

Insofar as these models present a useful picture of distributed scenarios, this paper has answered several questions about whether or not the guarantees provided by Stone's theorem in centralized environments hold in distributed settings. The models have demonstrated that when agents are allowed to communicate $\log_2(3)$ bits per decision, the ensemble can achieve universal consistency in both binary classification and regression frameworks in the limit as the number of agents increases without bound. In the binary classification case, we have demonstrated this property as a special case of naive kernel classifiers. In the regression case, we have shown this to hold true with randomized agent decision rules. When investigating the necessity of these $\log_2(3)$ bits, we have found that in the binary classification framework only one bit per agent per classification was necessary for universal consistency, and the analysis provided an interesting comparison for naive kernel methods in the traditional framework. For regression, we have established the impossibility of universal consistency in the one bit regime for a natural, but restricted class of candidate rules.

With regard to future research in distributed learning, there are numerous directions of interest. As these results are useful only if they accurately depict some aspect of distributed environments, other perhaps more reflective models are important to consider. In particular, the current models assume that a reliable physical layer exists where bits transmitted from the agents are guaranteed to arrive unperturbed at the fusion center. Future research may consider a richer model for this communication, perhaps within an information-theoretic (i.e., Shannon-theoretic) formalism. Further, the current models consider simplified network models where the fusion center communicates with agents via a broadcast medium and each agent has a direct, albeit limited, channel to the fusion center. Future research may focus on network models that allow for interagent communication. Consistent with the spirit of sensor networks, we might allow agents to communicate locally among themselves (or perhaps, hierarchically) before coordinating a response to the fusion center. In general, models of this form would weaken (A) in the discussion in Section II by allowing for correlated agent responses. A related assumption in this work is that the underlying data is i.i.d. Extending the results to another sampling process is important since in many distributed applications, the data observed by the agents may be correlated. In this vein, connections to results in statistical pattern recognition results under non-i.i.d. sampling processes would be interesting and important to resolve.

Finally, from a learning perspective, the questions we have considered in this paper have been focused on the statistical issue of universal consistency. Though such a consideration seems to be one natural first step, other comparisons between centralized and distributed learning are essential, perhaps with respect to convergence rate and the finite data reality that exists in any practical system. Such questions open the door for agents to receive multiple training examples and may demand more complicated local decision algorithms; in particular, it may be interesting to study local regularization strategies for agents in an ensemble. Future work may explore these and other questions frequently explored in traditional, centralized learning systems, with the hope of further understanding the nature of distributed learning under communication constraints.

Appendix

This appendix includes important facts that are commonly used in the study of nonparametric statistics and are similarly applied in the proofs above. Lemma 1 is a basic result from probability theory and is included for clarity. Lemma 2 follows from Theorem 23.2 and Lemma 23.6 in [12] applied to the naive kernel. The proof of Theorem 6.2 in [7] contains the fundamental steps needed to prove Lemma 3. Lemma 4 can be found as Lemma 4.1 in [12]. Lemma 5 follows from arguments used in proving Theorem 5.1 in [12] applied to the naive kernel.

Lemma 1: Suppose $\{X_n\}_{n=1}^{\infty}$ is a sequence of random variables such that $X_n \to X$ in probability. Then, for any sequence of events $\{A_n\}_{n=1}^{\infty}$ with $\liminf \boldsymbol{P}\{A_n\} > 0$

$$\boldsymbol{P}\{|X_n - X| > \epsilon\,|A_n\} \to 0$$

for all $\epsilon > 0$.

Proof: After noting that

$$\begin{aligned}\boldsymbol{P}\{|X_n - X| > \epsilon\} &= \boldsymbol{P}\{|X_n - X| > \epsilon\,|A_n\}\boldsymbol{P}\{A_n\} \\ &\quad + \boldsymbol{P}\{|X_n - X| > \epsilon\,|\bar{A}_n\}\boldsymbol{P}\{\bar{A}_n\} \\ &\geq \boldsymbol{P}\{|X_n - X| > \epsilon\,|A_n\}\boldsymbol{P}\{A_n\},\end{aligned}$$

the lemma follows trivially from the fact that $\liminf \boldsymbol{P}\{A_n\} > 0$ and $X_n \to X$ in probability. The proof follows similarly if $X_n \to \infty$ in probability. □

Lemma 2: Let $X \sim \boldsymbol{P}_X$ be an $\mathbb{R}^d$-valued random variable and fix any function $f \in L(\boldsymbol{P}_X)$. For an arbitrary sequence of real numbers $\{r_n\}_{n=1}^{\infty}$, define a sequence of functions $f_n(x) = \boldsymbol{E}\{f(X)\,|X \in B_{r_n}(x)\}$. If $r_n \to 0$, then $f_n(X) \to f(X)$ in probability.

Lemma 3: Let $X \sim \boldsymbol{P}_X$ be an $\mathbb{R}^d$-valued random variable and define $\{r_n\}_{n=1}^{\infty}$ and $\{a_n\}_{n=1}^{\infty}$ as arbitrary sequences of real numbers such that $r_n \to 0$ and $a_n \to \infty$. If $(r_n)^d a_n \to \infty$, then

$$a_n \int 1_{B_{r_n}(X)}(y) P_X(dy) \to \infty \text{ i.p.}$$

Lemma 4: Suppose $B(n,p)$ is a binomially distributed random variable with parameters n and p. Then

$$\boldsymbol{E}\left\{\frac{1}{B(n,p)1_{\{B(n,p)>0\}}}\right\} \quad \leq \quad \frac{2}{(n+1)p}.$$

Lemma 5: There is a constant c such that for any measurable function f, any $\mathbb{R}^d$-valued random variable X, and any sequence $\{r_n\}_{n=1}^{\infty}$

$$\boldsymbol{E}\left\{\frac{\sum_{i=1}^{n} 1_{\{X_i \in B_{r_n}(X)\}} f(X_i)}{\sum_{i=1}^{n} 1_{\{X_i \in B_{r_n}(X)\}}}\right\} \quad \leq \quad c\boldsymbol{E}\{f(X)\}$$

for all n.

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Commun. Mag.*, vol. 40, no. 8, pp. 102–114, Aug. 2002.

[2] M. Barkat and P. K. Varshney, "Decentralized CFAR signal detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 25, pp. 141–149, Mar. 1989.

[3] R. Blum, S. Kassam, and H. V. Poor, "Distributed detection with multiple sensors II: Advanced topics," *Proc. IEEE*, vol. 85, no. 1, pp. 64–79, Jan. 1997.

[4] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 26, no. 2, pp. 123–140, 1996.

[5] T. M. Cover, "Rates of convergence for nearest neighbor procedures," in *Proc. 1st Annu. Hawaii Conf. Systems Theory*, Honolulu, HI, Jan. 1968, pp. 413–415.

[6] A. D'Costa and A. M. Sayeed, "Collaborative signal processing for distributed classification in sensor networks," in *Proc. IPSN'03 (Lecture Notes in Computer Science)*, F. Zhao and L. Guibas, Eds. Berlin, Germany: Springer-Verlag, 2003.

[7] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York, NY: Springer, 1996.

[8] M. Dong, L. Tong, and B. M. Sadler, "Information retrieval and processing in sensor networks: Deterministic scheduling vs. random access," *IEEE Trans. Inf. Theory*, submitted for publication.

[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Comp. Syst. Sci.*, vol. 55, pp. 119–139, 1997.

[10] Y. Freund, R. E. Schapire, Y. Singer, and M. K. Warmuth, "Using and combining predictors that specialize," in *Proc. 29th Annual ACM Symp. Theory of Computing*, El Paso, TX, May 1997, pp. 334–343.

[11] W. Greblicki and M. Pawlak, "Necessary and sufficient conditions for bayes risk consistency of recursive kernel classification rule," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 3, pp. 408–412, May 1987.

[12] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer-Verlag, 2002.

[13] R. Jacobs, M. I. Jordan, S. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 125–130, 1991.

[14] M. Kearns and H. S. Seung, "Learning from a population of hypotheses," *Mach. Learn.*, vol. 18, pp. 255–276, 1995.

[15] J. Kittler, M. Hatef, P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[16] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*. New York: Dover, 1975.

[17] J. H. Kotecha, V. Ramachandran, and A. Sayeed, "Distributed multi-target classification in wireless sensor networks," *IEEE J. Sel. Areas in Commun. (Special Issue on Self-Organizing Distributed Collaborative Sensor Networks)*, vol. 23, no. 4, pp. 703–713, Apr. 2005.

[18] A. Krzyżak, "The rates of convergence of kernel regression estimates and classification rules," *IEEE Trans. Inf. Theory*, vol. IT-32, no. 5, pp. 668–679, Sep. 1986.

[19] S. R. Kulkarni and S. E. Posner, "Rates of convergence of nearest neighbor estimation under arbitrary sampling," *IEEE Trans. Inf. Theory*, vol. 41, no. 4, pp. 1028–1039, Jul. 1995.

[20] S. R. Kulkarni, S. E. Posner, and S. Sandilya, "Data-dependent k_n-nn and kernel estimators consistent for arbitrary processes," *IEEE. Trans. Inf. Theory*, vol. 48, no. 10, pp. 2785–2788, Oct. 2002.

[21] S. Kumar, F. Zhao, and D. Shephard, "Collaborative signal and information processing in microsensor networks," *IEEE Signal Process. Mag.*, vol. 19, no. 2, pp. 13–14, 2002.

[22] A. Lazarevic and Z. Obradovic, "The distributed boosting algorithm," in *KDD'01: Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*. San Francisco, CA: ACM Press, 2001, pp. 311–316.

[23] D. Li, K. Wong, Y. H. Hu, and A. Sayeed, "Detection, classification, and tracking of targets," *IEEE Signal Process. Mag.*, vol. 19, no. 2, pp. 17–29, 2002.

[24] G. Lugosi, "Learning with an unreliable teacher," *Pattern Recogn.*, vol. 25, pp. 79–87, 1992.

[25] G. Morvai, S. R. Kulkarni, and A. B. Nobel, "Regression estimation from an individual stable sequence," *Statistics*, vol. 33, pp. 99–119, 1999.

[26] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Decentralized detection and classification using kernel methods," in *Proc. 21st Int. Conf. Machine Learning*, Banff, AB, Canada, 2004.

[27] A. B. Nobel, "Limits to classification and regression estimation from ergodic processes," *Ann. Statist.*, vol. 27, pp. 262–273, 1999.

[28] A. B. Nobel and T. M. Adams, "On regression estimation from ergodic samples with additive noise," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2895–2902, Nov. 2001.

[29] A. B. Nobel, G. Morvai, and S. Kulkarni, "Density estimation from an individual sequence," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 537–541, Mar. 1998.

[30] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "Consistency in a model for distributed learning with specialists," in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, Jun./Jul. 2004, p. 465.

[31] ——, "Consistency in models for communication constrained distributed learning," in *Learning Theory, Proc. 17th Annu. Conference on Learning Theory, COLT 2004*, J. Shawe-Taylor and Y. Singer, Eds. New York: Springer-Verlag, 2004.

[32] ——, "Distributed learning in wireless sensor networks," in *Proc. 42nd Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Sep. 2004.

[33] G. Roussas, "Nonparametric estimation in markov processes," *Ann. Inst. Statist. Math.*, vol. 21, pp. 73–87, 1967.

[34] S. N. Simic, "A learning theory approach to sensor networks," *IEEE Pervasive Computing*, vol. 2, no. 4, pp. 44–49, 2003.

[35] C. J. Stone, "Consistent nonparametric regression," *Ann. Statist.*, vol. 5, pp. 595–645, 1977.

[36] J. N. Tsitsiklis, "Decentralized detection," in *Advances in Statistical Signal Processing*. Greenwich, CT: JAI Press, 1993, pp. 297–344.

[37] P. K. Varshney, *Distributed Detection and Data Fusion*. New York: Springer-Verlag, 1996.

[38] V. V. Veeravalli, "Decentralized quickest change detection," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1657–1656, May 2001.

[39] S. Yakowitz, "Nonparametric density and regression estimation from Markov sequences without mixing assumptions," *J. Multivar. Anal.*, vol. 30, pp. 124–136, 1989.

[40] ——, "Nearest neighbor regression estimation for null-recurrent Markov time series," *Stoch. Processes Appl.*, vol. 48, pp. 311–318, 1993.