Test-Time Collaborative Classification over Multi-Agent Networks

Ping Hu, *Member, IEEE,* Mert Kayaalp, *Member, IEEE* and Ali H. Sayed, *Fellow, IEEE*

***Abstract*—The increasing heterogeneity of multi-agent systems poses significant challenges for jointly training a global model across agents. At the same time, cooperative inference between agents has long been recognized as a powerful mechanism for distributed decision making over networks. Motivated by these observations, we propose a collaboration framework for distributed binary classification over multi-agent networks, where a set of independently trained agents, potentially differing in architecture, feature space, or modality, coordinate their actions during test time to form collective predictions. This coordination is achieved by exchanging local decision statistics through a distributed learning protocol. We develop a theoretical and experimental study of this independent training and cooperative inference paradigm, and examine its performance under different communication budgets and distributed learning rules. We establish classification error guarantees under sufficient, finite-round, and finite-precision communication, together with PAC-style generalization bounds. These results capture the influence of model heterogeneity, network topology, combination policy, and communication constraints on prediction accuracy. Taken together with the experimental results, they reveal both the price of independent training and the benefit of collective prediction for the proposed distributed decision making framework with models learned from data.**

***Index Terms*—Distributed learning and inference, collective intelligence, classification, probability of error, ensemble learning**

I. Introduction

DISTRIBUTED learning and inference is a key paradigm for large-scale and distributed systems, where data is generated and processed across multiple agents.[1] Within this context, two major lines of research have emerged: *distributed inference*, extensively studied in the statistics, signal processing, and control communities [1]–[3], and *distributed machine learning*, primarily advanced within the machine learning (ML) and data science communities [4], [5].

In distributed inference, a collection of agents is connected through a communication network. Each agent owns a family of statistical models that describes the distribution of its observations under a common, shared latent variable (e.g., a hidden state or hypothesis). The collective objective is to infer this latent variable through local computations and communication. To this end, agents iteratively update their beliefs or estimates by combining private observations with information received from their neighbors. Classical instances of this paradigm include distributed estimation [1]–[3], distributed detection [6]–[8], and social learning [9]. A key feature of these problems is that collaboration takes place during the *inference* phase. In contrast, distributed ML has mainly focused on collaboration during the *training* phase, as exemplified by federated learning and decentralized learning. In this setting, agents jointly train a global model on distributed data, which is later deployed for inference either centrally or across devices. The inference stage is thus governed by a single unified model, and the system operates as a *single* decision maker. This is fundamentally different from distributed inference, where each agent acts as an autonomous decision maker, and collaboration serves to improve the quality of individual or consensus predictions.

While these two research lines have evolved largely independently, they share some core motivation (e.g., scalability, decentralization, and privacy), and often employ similar tools, particularly those from distributed optimization [10], [11]. Recently, efforts have been made to bridge the two fields in order to leverage their complementary benefits. A promising line of work is inference with machine-learned models, as studied in [12]–[14]. In [12], this idea is examined in the context of hypothesis testing within a centralized setting involving a single agent. More relevantly, [13] and [14] study the social learning task where multiple agents train local classifiers on their own data to extract discriminative information for inference in the absence of explicit statistical models. In both cases, inference is performed using a stream of unlabeled observations. In this work, we consider instead a distributed binary classification setting in which each agent has access to only a *single* local testing sample associated with a common, unknown class label. The agents are collectively tasked with identifying this label from their own observations. This setting naturally arises in applications such as sensor fusion (e.g., vehicles capturing different views of the same object) and multi-view classification (e.g., image and text modalities of the same entity). The goal of this work is to develop a distributed classification framework for this setting.

A. Our work

To this end, we propose a collective prediction framework for distributed classification over multi-agent networks, which combines independent local training with collaborative inference. Specifically, during the *training* phase, each agent

Ping Hu and Ali H. Sayed are with the School of Engineering, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland (email: ping.hu@epfl.ch; ali.sayed@epfl.ch). Mert Kayaalp is now with the DTI, SUPSI, Dalle Molle Institute for Artificial Intelligence (IDSIA USI-SUPSI), 6962 Lugano, Switzerland, where he is affiliated with the UBS-IDSIA AI Lab (email: mert.kayaalp@idsia.ch). This work was done while he was a Ph.D. student and Post-doc at the Adaptive Systems Laboratory at EPFL. Mert Kayaalp's work was supported by UBS Switzerland AG and its affiliates through the UBS-IDSIA AI Lab. A preliminary version of this work was presented at the Algorithmic Collective Action Workshop, NeurIPS 2025.

[1]Throughout this paper, we use the term "distributed" to describe systems without a central coordinator. Closely related architectures are also termed "decentralized" in the machine learning literature.

trains a local classifier using its own dataset. This independent training process is motivated by several trends and challenges in modern ML systems. First, it has become increasingly common in large-scale and privacy-sensitive applications, such as edge computing [15] and foundation models [16]. For example, large language models or domain-specific experts may be trained independently across different entities (e.g., companies, institutions, clients, or devices) and later coordinated at inference time [17]. Second, feature and model heterogeneity often make joint training impractical. For example, agents may process different input modalities or rely on distinct model architectures tailored to their local data. During the *testing* phase, each agent receives a private testing sample and participates in a distributed inference protocol to infer the shared class label by exchanging predictive information with its neighbors. This collaborative framework introduces several challenges that are not present in the standard supervised classification setting:

(i) *How should agents aggregate predictions at test time?* Unlike standard ML pipelines that rely on a single trained model, each agent in our setting is equipped with an independently trained local model. This raises the question of how agents can effectively aggregate their individual predictions during inference.

(ii) *How does agent heterogeneity govern the generalization performance?* While classical supervised generalization theory typically analyzes a single learned predictor, our setting involves collective predictions based on multiple independently trained local classifiers. This raises the question of how agents' inherent heterogeneity and inter-agent interactions shape the generalization behavior of the entire distributed system.

(iii) *How does limited communication affect prediction?* Collective prediction typically relies on sufficient information exchange. When communication is limited in either the number of rounds or message precision, the resulting predictions may deviate from the sufficient-communication limit. This raises the question of how communication constraints affect the individual prediction accuracy.

To introduce our framework, we assume that agents produce soft predictions using their local classifiers and adopt the classical DeGroot model [18] as an interpretable and analytically tractable mechanism for information aggregation during inference. In this model, agents iteratively update their predictions by taking a weighted average of their neighbors' predictions according to a *combination policy* defined over a fixed communication topology [18]. The DeGroot model has been extensively studied in the distributed inference literature and is well known for achieving consensus in multi-agent systems [19], [20]. DeGroot updating has also been examined experimentally in human social networks [21]. Moreover, its linear update structure makes it particularly well-suited for distributed implementation with finite communication rounds.

The novelty of the present work is not in proposing a new information aggregation rule, but in formulating and analyzing this inference-time collaboration framework. Accordingly, this work develops a theoretical and experimental study with the following contributions:

(i) We establish theoretical guarantees for distributed classification over multi-agent networks under the independent training and collaborative inference paradigm. The derived bounds characterize how network structure, local model quality, and data heterogeneity influence predictive accuracy.

(ii) We analyze how communication constraints affect classification, deriving agent-dependent bounds for both finite-round and finite-precision communication. These results characterize how incomplete consensus, network topology, combination policy, and communication precision influence prediction accuracy.

(iii) We evaluate the proposed framework on two complementary benchmarks and empirically demonstrate the benefits of cooperative inference under different observation settings.

These contributions provide theoretical foundations for test-time collaboration in distributed classification and highlight connections between distributed inference and distributed ML.

B. Related Literature

1) Distributed machine learning: Most existing work in this domain has focused on collaboratively training a shared global model using data distributed across multiple devices or computational nodes. Two dominant approaches are federated learning, which relies on a central server to aggregate locally trained updates [22], [23], and decentralized learning, which dispenses with the server and synchronizes models through peer-to-peer communication over a fixed network graph [24], [25]. Variants of these methods address challenges such as non-i.i.d. data [26], [27], agent sampling [28], personalization [29], and communication efficiency during training [30], [31]. Despite their topological differences, in their canonical forms, both paradigms aim to produce a shared global model during training. At inference time, this model is either executed on a central server or deployed across devices, so the system effectively operates as a *centralized* decision maker, even when implementation is distributed. These joint-training frameworks are therefore complementary to the setting studied here: they aim to produce a shared global model during training, whereas our work focuses on inference-time collaboration among independently trained local models. A small body of work has explored collaboration during inference, where a set of independently trained models exchange information with *each other* to improve accuracy via trust-score design [32], [33]. Yet they all assume that all agents observe the *same* test input, which restricts agents' heterogeneity to model parameters or inductive biases. This setup closely parallels ensemble learning and multiple classifier combination discussed later.

In contrast, we study a setting in which each agent receives its own private testing sample, and these samples may be statistically dependent across agents. This formulation naturally accommodates heterogeneous feature spaces (e.g., multimodal sensors or domain-specific models) and enables distributed decision-making without feature alignment. Collaborative inference in this setting has received little attention, and our work provides a formal theoretical and experimental treatment.

2) Distributed inference: Distributed inference studies how a network of agents collaborates to estimate a latent variable based on private observations and local communication [1]–[3]. Applications include state estimation in smart power grids, cooperative perception in multi-robot systems, and opinion dynamics in social networks. These applications motivate different modeling frameworks depending on the nature of the latent variable. For discrete hypothesis spaces, a widely studied paradigm is social learning (SL), in which agents iteratively exchange local beliefs to identify the true hypothesis. SL naturally accommodates heterogeneous agents, since each agent may observe different types of signals modeled by distinct statistical distributions [9]. Classical SL methods assume that agents possess a family of likelihood models, which are used in updating beliefs via Bayes' rule (or suitable variants) [34]–[38]. While analytically elegant, this assumption is restrictive in real-world problems, where data distributions are complex and rarely admit closed-form models. To address this limitation, [13] proposed a social machine learning strategy, where each agent first trains a local classifier using its own labeled examples, and then performs the SL rule based on the trained classifier. The asymptotic and non-asymptotic performance of this strategy has been examined in [13], [14].

Our work follows this data-driven perspective but differs in the task formulation. In SL, agents typically process streams of unlabeled i.i.d. data to infer a shared latent hypothesis. In the supervised classification setting studied here, each agent receives only a *single* testing sample for collective prediction, and collaboration takes place through a finite-round exchange of learned local decision statistics at test time. This difference changes the operational role of communication and leads to analytical questions that are not directly addressed in existing SL studies or in classical single-agent supervised classification.

3) Multiple classifier combination: This is a classical line of research in ML that studies how to aggregate the predictions of multiple models to improve performance. While our framework is distributed and collaborative in nature, it bears a strong conceptual connection to this literature. Specifically, when all agents follow the DeGroot model for information aggregation at test time, their predictions converge (in the limit of infinite communication) to a consensus that is a *weighted average* of the local classifier outputs. This resembles the effect of a centralized fusion rule, a central topic in the literature on multiple classifier combination. Classical techniques in this area include voting schemes, bagging, boosting, and stacking [39], [40]. These approaches form the foundation of what is often referred to as *ensemble learning*, where multiple base classifiers are combined to improve generalization and robustness [41], [42]. More advanced strategies, such as mixture-of-experts models, combine classifiers through input-dependent gating [16]. These approaches, however, typically operate in a *centralized* setting during test time and assume a common feature representation across experts.

In contrast, our framework operates in a fully-distributed environment during both training and testing phases. When communication is limited to a finite number of rounds, consensus may not be reached, and predictions remain agent-dependent. This dynamic, iterative process differs fundamentally from the static, one-shot aggregation used in most ensemble methods.

Notation: We use boldface font to denote random variables and normal font for their realizations, e.g., $\boldsymbol{h}$ and h. $\mathbb{E}$ and $\mathbb{P}$ denote the expectation and probability operators, respectively. $1[\cdot]$ denotes the indicator function, and $\mathbb{1}$ denotes the all-ones vector.

II. Problem Formulation

We are given a network of K agents indexed by k and a binary classification task with class label γ. The sets of agents and labels are denoted by $\mathcal{K} \triangleq \{1, 2, \ldots, K\}$ and $\Gamma \triangleq \{+1, -1\}$, respectively. We assume that the semantic meaning of the labels is fixed and shared across the network during both training and inference. Each agent k holds a local training set $\mathcal{D}_k$ of N_k labeled examples $(\boldsymbol{h}_{k,n}, \boldsymbol{\gamma}_{k,n})$, where $\boldsymbol{h}_{k,n}$ is the n-th feature vector and $\boldsymbol{\gamma}_{k,n} \in \Gamma$ is its label. Let $\mathcal{H}_k$ denote the feature space of agent k. A network observation consists of one local view from each agent and therefore takes values in the product space

$$\mathcal{H} \triangleq \mathcal{H}_1 \times \cdots \times \mathcal{H}_K. \tag{1}$$

For each $\gamma \in \Gamma$, let P_γ denote the class-conditional distribution of a network observation $\boldsymbol{h} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K) \in \mathcal{H}$, and let $P_{\gamma,k}$ denote its k-th marginal. When this marginal admits a density or probability mass function, we denote it by $p_k(h|\gamma)$.

For the binary classification task, we assume a uniform class prior. Then, P_γ induces the joint distribution Q of $(\boldsymbol{h}, \boldsymbol{\gamma})$:

$$\mathbb{P}_Q(\boldsymbol{\gamma} = \gamma) = \frac{1}{2}, \quad \boldsymbol{h} \mid \{\boldsymbol{\gamma} = \gamma\} \sim P_\gamma, \quad \gamma \in \Gamma. \tag{2}$$

We denote by Q_k the corresponding marginal distribution of $(\boldsymbol{h}_k, \boldsymbol{\gamma})$ at agent k. Conditional on $\boldsymbol{\gamma}$, the local observations in $\boldsymbol{h}$ may be statistically dependent across agents. The agents are *heterogeneous* as they may have different feature spaces $\mathcal{H}_k$ (e.g., different views in multi-view learning) or distinct class-conditional models $p_k(h|\gamma)$ (e.g., different local distributions in personalized federated learning). The posterior log-ratio is sufficient for binary decision making. Under the uniform prior in (2), Bayes' rule gives

$$\Lambda_k(h) \triangleq \log \frac{p_k(+1|h)}{p_k(-1|h)} = \log \frac{p_k(h|+1)}{p_k(h|-1)}, \quad \forall h \in \mathcal{H}_k \tag{3}$$

where $p_k(\gamma|h)$ denotes the posterior probability. The function Λ_k is also known as the logit function in the literature.

A. The Training Phase

In the training phase, each agent k uses its own dataset $\mathcal{D}_k$ to train a local classifier. For each agent k, the samples in $\mathcal{D}_k$ are independently drawn according to Q_k. Moreover, samples with the same index n may be statistically dependent across agents, while the corresponding cross-agent sample collections are independent across indices. We denote by $f_k : \mathcal{H}_k \to \mathbb{R}$ a generic real-valued score function from an admissible function class $\mathcal{F}_k$. The function f_k serves as an approximation to the logit function Λ_k in (3). Formally,

$$f_k(h) = \log \frac{\widehat{p}_k(+1|h)}{\widehat{p}_k(-1|h)} \tag{4}$$

where $\widehat{p}_k(\gamma|h)$ denotes the estimated posterior probability. Given $\mathcal{D}_k$, the empirical risk associated with a candidate score function $f_k \in \mathcal{F}_k$ is

$$\boldsymbol{R}_{k,\mathsf{emp}}(f_k) \triangleq \frac{1}{N_k}\sum_{n=1}^{N_k}\Phi(\boldsymbol{\gamma}_{k,n} f_k(\boldsymbol{h}_{k,n})) \tag{5}$$

where Φ denotes the loss function used for training. Let $\mathcal{A}_k$ denote the local training algorithm at agent k and let $\mathcal{S}_k$ collect its internal randomness. The resulting trained score function is denoted by

$$\widehat{\boldsymbol{f}}_k \triangleq \mathcal{A}_k(\mathcal{D}_k, \mathcal{S}_k) \in \mathcal{F}_k. \tag{6}$$

We measure the optimization accuracy of $\widehat{\boldsymbol{f}}_k$ by its empirical optimality gap, namely, the difference between its empirical risk and the infimum empirical risk over $\mathcal{F}_k$:

$$\mathrm{Gap}_{k,\mathrm{opt}}(\widehat{\boldsymbol{f}}_k) \triangleq \boldsymbol{R}_{k,\mathsf{emp}}(\widehat{\boldsymbol{f}}_k) - \inf_{f_k\in\mathcal{F}_k}\boldsymbol{R}_{k,\mathsf{emp}}(f_k) \geq 0. \tag{7}$$

A zero gap corresponds to exact empirical risk minimization (ERM), while a positive gap quantifies the residual optimization error of the trained score function.

For the theoretical analysis, we use the following assumptions on the loss function Φ and the function class $\mathcal{F}_k$, which are commonly adopted in statistical learning theory [43], [44].

Assumption 1 (**Conditions on the loss function**). *The loss function $\Phi : \mathbb{R} \to \mathbb{R}_+$ is convex, non-increasing and differentiable at 0 with $\Phi'(0) < 0$. Also, it is L_Φ-Lipschitz.* □

Assumption 2 (**Boundedness of functions**). *There exists a constant $\beta > 0$ such that $|f_k(h)| \leq \beta$ for every $k \in \mathcal{K}$, $f_k \in \mathcal{F}_k$, and $h \in \mathcal{H}_k$.* □

The training phase is completed by empirically centering the learned score $\widehat{\boldsymbol{f}}_k$, following [13], [14]. Since independently trained local classifiers may exhibit different empirical offsets, we subtract each score's empirical training mean to better align the local outputs before collaboration. Specifically, we define the *empirical training mean as*

$$\boldsymbol{\mu}_{k,\mathsf{emp}}(f_k) \triangleq \frac{1}{N_k}\sum_{n=1}^{N_k} f_k(\boldsymbol{h}_{k,n}), \quad \forall f_k \in \mathcal{F}_k. \tag{8}$$

The resulting classifier generates a centered score:

$$\boldsymbol{c}_k(h) \triangleq \widehat{\boldsymbol{f}}_k(h) - \boldsymbol{\mu}_{k,\mathsf{emp}}(\widehat{\boldsymbol{f}}_k), \quad \forall h \in \mathcal{H}_k. \tag{9}$$

Therefore, after training, the sign of $\boldsymbol{c}_k(h)$ determines the local prediction at agent k for any unseen feature vector $h \in \mathcal{H}_k$. The role of empirical centering is examined in Section III-A.

B. The Prediction Phase

In the prediction phase, the network receives a fresh testing example $(\boldsymbol{h}^*, \boldsymbol{\gamma}^*)$ drawn from Q independently of the training data and local training-algorithm randomness, with each agent k observing only its local component $\boldsymbol{h}_k^* \in \mathcal{H}_k$. These local views may remain statistically dependent conditioned on $\boldsymbol{\gamma}^*$. The goal of the agents is to make a *collective* prediction about $\boldsymbol{\gamma}^*$ by collaborating with their neighbors. We use t to index the communication rounds and denote by $\boldsymbol{\lambda}_{k,t}$ the *decision statistic* of agent k at round t, whose sign determines its prediction. Following the DeGroot model, the agents iteratively combine their decision statistics over the network. Upon observing $\boldsymbol{h}_k^*$, agent k initializes its decision statistic using the local classifier $\boldsymbol{c}_k$:

$$\boldsymbol{\lambda}_{k,0} \triangleq \boldsymbol{c}_k(\boldsymbol{h}_k^*). \tag{10}$$

Then, each agent communicates with its neighbors and updates its decision statistic by aggregating the local decision statistics in the neighborhood [18]:

$$\boldsymbol{\lambda}_{k,t} = \sum_{\ell=1}^{K} a_{\ell k}\boldsymbol{\lambda}_{\ell,t-1} \tag{11}$$

where $a_{\ell k}$ denotes the combination weight agent k assigns to its neighbor ℓ, which satisfies

$$\sum_{\ell=1}^{K} a_{\ell k} = 1, \;\; a_{\ell k} > 0 \;\; \forall \ell \in \mathcal{N}_k, \;\text{ and }\; a_{\ell k} = 0 \;\; \forall \ell \notin \mathcal{N}_k \tag{12}$$

with $\mathcal{N}_k$ denoting the neighboring set of agent k. Communication proceeds in synchronous rounds over this fixed graph, and the exchanged messages are assumed to be delivered reliably. Accordingly, the communication model considered here does not include asynchronous updates, time-varying connectivity, or packet losses. Moreover, we impose the following assumption on the topology of the communication network to ensure information diffusion throughout the network.

Assumption 3 (**Strongly-connected graph**). *The graph of the communication network is strongly connected. That is, there exist paths with positive combination weights between any two distinct agents in both directions (the two paths need not be the same), and at least one agent has a self-loop, i.e., $a_{kk} > 0$ for some agent k [10].* □

Under this assumption and from the Perron-Frobenius theorem [45], the combination matrix $A = [a_{\ell k}]$ is primitive and admits a Perron vector π satisfying

$$A\pi = \pi, \;\; \sum_{k=1}^{K}\pi_k = 1, \;\text{ and }\; \pi_k > 0, \;\; \forall k \in \mathcal{K}. \tag{13}$$

The entries of π quantify the relative influence of the agents following the adopted combination policy. In network science terminology [46], [47], this Perron vector corresponds to the notion of eigenvector centrality associated with the matrix A. Moreover, the second-largest magnitude among all eigenvalues of A, denoted by σ_A, is strictly smaller than 1.

To evaluate the performance of this framework, we first introduce a feasibility condition associated with the classification task, defined in terms of the target risk for training. For each agent k, the target risk R_k^o is the infimum of the expected risk over the local function class $\mathcal{F}_k$:

$$\mathsf{R}_k^o \triangleq \inf_{f_k\in\mathcal{F}_k}\mathbb{E}_{(\boldsymbol{h}_k,\boldsymbol{\gamma})\sim Q_k}\Phi(\boldsymbol{\gamma} f_k(\boldsymbol{h}_k)). \tag{14}$$

The target risk for the *network* is defined as a weighted average of the individual target risks involving the Perron vector π:

$$\mathsf{R}^o \triangleq \sum_{k=1}^{K}\pi_k\mathsf{R}_k^o. \tag{15}$$

The reason for introducing π in defining R^o will be clear in our subsequent analysis for the classification error. The feasibility of the classification task requires the following assumption.

Assumption 4 (Feasibility). *The network target risk satisfies* $\mathsf{R}^o < \Phi(0)$. □

We note that $\Phi(0)$ represents the expected risk of the model $f_k = 0$, which corresponds to the case where agent k assigns labels $+1$ and -1 with equal probability to all feature vectors. Hence, $\mathsf{R}_k^o = \Phi(0)$ implies that the feature vectors of agent k provide no useful information for classification within the prescribed function class $\mathcal{F}_k$. Assumption 4 is a network-level feasibility condition: it does not require every agent to be individually informative, but only that the network target risk satisfies $\mathsf{R}^o < \Phi(0)$. Thus, the collective observations available in the network must contain enough class-discriminative information to outperform random guessing. If this condition fails, then the classification problem is effectively uninformative at the network level.

III. Performance Guarantees

This section develops four principal performance guarantees for the proposed framework: sufficient-communication classification, finite-round communication, finite-precision communication, and probably approximately correct (PAC)-style generalization. The analysis of the first three guarantees proceeds in two stages. We first characterize the training performance of the classifier network through an *expected network margin*. Two supporting propositions are derived for clarifying the role of empirical centering in (9) and establishing when approximate local training attains such a margin with high probability. We then use this common margin event to characterize the prediction error under the three communication settings. The final PAC-style theorem provides a complementary guarantee that relates population error directly to the performance observed on the training data. We conclude by clarifying the relation to conventional social learning recursions and summarizing the corresponding results without empirical centering.

A. Network Margin and Training Guarantee

To define the expected network margin used throughout the analysis, we first identify the aggregate classifier obtained after sufficient communication. Following the DeGroot averaging rule (11), agents agree on a *common* decision statistic denoted by $\boldsymbol{\lambda}_{\mathsf{ave}}$ after sufficient communication [18]. That is, it holds almost surely (a.s.) that

$$\boldsymbol{\lambda}_{\mathsf{ave}} \triangleq \lim_{t\to\infty} \boldsymbol{\lambda}_{k,t} \overset{\text{a.s.}}{=} \sum_{k=1}^{K} \pi_k \boldsymbol{c}_k(\boldsymbol{h}_k^*). \tag{16}$$

Accordingly, the *collective* prediction under sufficient communication is

$$\widehat{\gamma} \triangleq \operatorname{sgn}(\boldsymbol{\lambda}_{\mathsf{ave}}) \tag{17}$$

where any fixed tie rule may be used at zero. Equivalently, the distributed rule (11) functions as an aggregate classifier:

$$\boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*) \triangleq \sum_{k=1}^{K} \pi_k \boldsymbol{c}_k(\boldsymbol{h}_k^*). \tag{18}$$

This aggregate classifier resembles a π-weighted soft decision combination of K independently trained local classifiers. The analysis of $\boldsymbol{c}_{\mathsf{ave}}$ is nontrivial because the informativeness of the local scores may vary across agents and the local views may be statistically dependent. Existing theoretical results on soft decision combination often rely on specific distributional assumptions on the local scores [48]. We therefore develop an analysis that accommodates these heterogeneous and statistically dependent local observations.

Since the deployed local scores are empirically centered, we first examine how the empirical centering operation in (9) affects the aggregate score relative to the common zero decision threshold. For a generic collection of local score functions

$$f = (f_1, \ldots, f_K) \in \mathcal{F} \triangleq \mathcal{F}_1 \times \cdots \times \mathcal{F}_K, \tag{19}$$

we define the uncentered aggregate score by

$$f_{\mathsf{ave}}(\boldsymbol{h}) \triangleq \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{h}_k). \tag{20}$$

Let $\boldsymbol{\mu}_{\mathrm{raw}}^{\gamma}(f) \triangleq \mathbb{E}^{(\gamma)} f_{\mathsf{ave}}(\boldsymbol{h})$ denote its class-conditional mean, where $\mathbb{E}^{(\gamma)}$ denotes expectation under P_γ. We introduce the following network-level quantities associated with $\boldsymbol{\mu}_{\mathrm{raw}}^{\gamma}(f)$:

$$\mu_{\mathrm{mid}}(f) \triangleq \frac{\boldsymbol{\mu}_{\mathrm{raw}}^{+}(f) + \boldsymbol{\mu}_{\mathrm{raw}}^{-}(f)}{2}, \tag{21}$$

$$D(f) \triangleq \frac{\boldsymbol{\mu}_{\mathrm{raw}}^{+}(f) - \boldsymbol{\mu}_{\mathrm{raw}}^{-}(f)}{2}. \tag{22}$$

Moreover, we define the network empirical training mean by

$$\boldsymbol{\mu}_{\mathsf{emp}}(f) \triangleq \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_{k,\mathsf{emp}}(f_k). \tag{23}$$

Under the uniform class prior, for any fixed f independent of the training samples, the training setup in Section II-A implies $\mathbb{E}[\boldsymbol{\mu}_{\mathsf{emp}}(f)] = \mu_{\mathrm{mid}}(f)$. Hence, $\mu_{\mathrm{mid}}(f)$ is the population counterpart of the network empirical training mean. With the above definitions, we establish the following proposition.

Proposition 1 (Effect of additive centering). *For arbitrary additive centers* χ_k*, let* $\chi_\pi \triangleq \sum_k \pi_k \chi_k$ *and define*

$$s_{\chi,f}(\boldsymbol{h}) \triangleq \sum_{k=1}^{K} \pi_k \big[f_k(\boldsymbol{h}_k) - \chi_k \big]. \tag{24}$$

Its expected margin over both classes at threshold zero is

$$\begin{aligned} \Delta_\chi(f) &\triangleq \min\Big\{ \mathbb{E}^{(+1)} s_{\chi,f}(\boldsymbol{h}), -\mathbb{E}^{(-1)} s_{\chi,f}(\boldsymbol{h}) \Big\} \\ &= D(f) - |\chi_\pi - \mu_{\mathrm{mid}}(f)|. \end{aligned} \tag{25}$$

Hence the population midpoint $\chi_\pi = \mu_{\mathrm{mid}}(f)$ *maximizes this margin. In particular, the margins under raw and empirically centered scores satisfy*

$$\Delta_{\mathrm{raw}}(f) = D(f) - |\mu_{\mathrm{mid}}(f)|, \tag{26}$$

$$\Delta_{\mathrm{cen}}(f) = D(f) - |\boldsymbol{\mu}_{\mathsf{emp}}(f) - \mu_{\mathrm{mid}}(f)|. \tag{27}$$

Therefore, empirical centering (9) *improves the expected zero-threshold margin if and only if*

$$|\boldsymbol{\mu}_{\mathsf{emp}}(f) - \mu_{\mathrm{mid}}(f)| < |\mu_{\mathrm{mid}}(f)|. \tag{28}$$

Moreover, empirical centering is invariant to the agent-specific additive offsets: if $f'_k(h) = f_k(h) + o_k$, then

$$f'_k(h) - \boldsymbol{\mu}_{k,\mathsf{emp}}(f'_k) = f_k(h) - \boldsymbol{\mu}_{k,\mathsf{emp}}(f_k). \tag{29}$$

Consequently, the centered DeGroot recursion (11) *stays unchanged by these offsets at every communication round.*

Proof. See Appendix A.2. □

For the learned classifiers $\boldsymbol{c}_k$ from Section II-A, Proposition 1 applies with $f_k = \widehat{\boldsymbol{f}}_k$ and $\chi_k = \boldsymbol{\mu}_{k,\mathsf{emp}}(\widehat{\boldsymbol{f}}_k)$. It shows that empirical centering acts as a threshold alignment mechanism that changes the location of the aggregate score relative to the zero decision threshold while leaving the class-conditional mean separation $2D(\widehat{\boldsymbol{f}})$ unchanged. Its effect on the expected zero-threshold margin depends on how closely the realized empirical training mean $\boldsymbol{\mu}_{\mathsf{emp}}(\widehat{\boldsymbol{f}})$ aligns with the corresponding population midpoint $\mu_{\mathrm{mid}}(\widehat{\boldsymbol{f}})$, as characterized by (28). Moreover, empirical centering removes agent-specific additive offsets from the local scores and hence from the subsequent collaborative recursion.

We next characterize the expected network margin attained by the centered local classifiers. For any $f_k \in \mathcal{F}_k$, define its empirically centered score

$$c_{k,f_k}(h) \triangleq f_k(h) - \boldsymbol{\mu}_{k,\mathsf{emp}}(f_k), \tag{30}$$

and the corresponding class-conditional means

$$\boldsymbol{\mu}_k^+(f_k) \triangleq \mathbb{E}^{(+1)}_{\boldsymbol{h}_k} c_{k,f_k}(\boldsymbol{h}_k), \tag{31}$$

$$\boldsymbol{\mu}_k^-(f_k) \triangleq \mathbb{E}^{(-1)}_{\boldsymbol{h}_k} c_{k,f_k}(\boldsymbol{h}_k), \tag{32}$$

where $\mathbb{E}^{(\gamma)}_{\boldsymbol{h}_k}$ denotes expectation under $P_{\gamma,k}$. The corresponding network means are

$$\boldsymbol{\mu}^+(f) = \sum_{k=1}^K \pi_k \boldsymbol{\mu}_k^+(f_k), \quad \boldsymbol{\mu}^-(f) = \sum_{k=1}^K \pi_k \boldsymbol{\mu}_k^-(f_k). \tag{33}$$

Before conditioning on the training phase, these quantities are random because the deployed local functions and empirical training means $\boldsymbol{\mu}_{k,\mathsf{emp}}(f_k)$ depend on the realized training data $\mathcal{D}_k$ and, when applicable, local optimization randomness $\mathcal{S}_k$. The inclusion of the Perron vector π in (33) is consistent with the collective nature of the prediction phase, where the same weights determine the common decision statistic in (16).

For a collection of learned functions $\widehat{\boldsymbol{f}} = (\widehat{\boldsymbol{f}}_1, \ldots, \widehat{\boldsymbol{f}}_K)$, we say that the classifier network satisfies the δ-margin consistent training condition if

$$\boldsymbol{\mu}^+(\widehat{\boldsymbol{f}}) > \delta \quad \text{and} \quad \boldsymbol{\mu}^-(\widehat{\boldsymbol{f}}) < -\delta \tag{34}$$

where $\delta \geq 0$ is the prescribed *decision margin*. In view of (18) and (30)–(33), this condition requires that the class-conditional mean of the aggregate score $\boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*)$ lies on the correct side of the zero decision threshold for both classes, with a margin exceeding δ. The performance of the training phase is then characterized by the probability that condition (34) holds for the classifier network. Define the corresponding event

$$\mathcal{C}_\delta \triangleq \Big\{ \boldsymbol{\mu}^+(\widehat{\boldsymbol{f}}) > \delta,\, \boldsymbol{\mu}^-(\widehat{\boldsymbol{f}}) < -\delta \Big\}, \tag{35}$$

and let

$$P_{c,\delta} \triangleq \mathbb{P}(\mathcal{C}_\delta), \tag{36}$$

where the probability is taken over the randomness from training. Analyzing $P_{c,\delta}$ is non-trivial, as it depends on various factors, including the size of the training sets, the complexity of the function class, the loss function, and the combination policy. A lower bound on $P_{c,\delta}$ under exact ERM, i.e., zero optimization gap in (7), was established in [14] using tools from [13]. We next extend it to approximate local optimization. The auxiliary quantities $N_{\max}$, α, ρ, $\mathcal{E}_\Phi(r,\delta)$, and $\delta_{\max}(r)$ appearing below are defined in Appendix A.1.

Proposition 2 (δ**-margin consistent training under approximate optimization**). *Suppose that, for every k, there exists a deterministic tolerance $\varepsilon_{k,\mathsf{opt}} \geq 0$ such that*

$$\mathrm{Gap}_{k,\mathsf{opt}}(\widehat{\boldsymbol{f}}_k) \leq \varepsilon_{k,\mathsf{opt}} \tag{37}$$

almost surely, and define

$$\varepsilon_{\mathsf{opt}} \triangleq \sum_{k=1}^K \pi_k \varepsilon_{k,\mathsf{opt}}. \tag{38}$$

If $\mathsf{R}^o + \varepsilon_{\mathsf{opt}} < \Phi(0)$, $0 \leq \delta < \delta_{\max}(\mathsf{R}^o + \varepsilon_{\mathsf{opt}})$, and $\rho < \mathcal{E}_\Phi(\mathsf{R}^o + \varepsilon_{\mathsf{opt}}, \delta)$, then under Assumptions 1–4,

$$P_{c,\delta} \geq 1 - 2\exp\left\{ -\frac{8N_{\max}}{\alpha^2\beta^2} \Big(\mathcal{E}_\Phi(\mathsf{R}^o + \varepsilon_{\mathsf{opt}}, \delta) - \rho \Big)^2 \right\}. \tag{39}$$

Proof. See Appendix A.3. □

The key implication of Proposition 2 is that $P_{c,\delta}$ admits an exponential guarantee when the network Rademacher complexity ρ is smaller than $\mathcal{E}_\Phi(\mathsf{R}^o + \varepsilon_{\mathsf{opt}}, \delta)$. The relevant properties of these auxiliary quantities are collected in Appendix A.1. In particular, for a fixed feasible decision margin δ, $\mathcal{E}_\Phi(r,\delta)$ decreases as the risk argument r increases, while the feasible margin range characterized by $\delta_{\max}(r)$ also becomes smaller. Consequently, increasing $\varepsilon_{\mathsf{opt}}$ reduces $\mathcal{E}_\Phi(\mathsf{R}^o + \varepsilon_{\mathsf{opt}}, \delta)$, making the condition $\rho < \mathcal{E}_\Phi(\mathsf{R}^o + \varepsilon_{\mathsf{opt}}, \delta)$ more restrictive. Moreover, whenever this condition remains satisfied, the quantity $\mathcal{E}_\Phi(\mathsf{R}^o + \varepsilon_{\mathsf{opt}}, \delta) - \rho$ decreases, thereby reducing the exponential rate in (39). Thus, a larger optimization residual makes it more difficult to certify a prescribed decision margin and weakens the resulting training guarantee. Setting $\varepsilon_{k,\mathsf{opt}} = 0$ for all agents recovers the exact-ERM specialization in [14].

The bound also helps characterize the statistical cost of independent local training as reflected in the training guarantee, a defining feature of our framework when data pooling or joint training is unavailable. To quantify this cost, we compare independent training with a centralized benchmark under a fixed budget of available training samples. In the homogeneous setting considered in Appendix A.4, under matched optimization accuracy and when the relevant Rademacher complexity decreases with the number of training samples as described there, specializing Proposition 2 shows that the centralized benchmark yields a tighter lower bound on $P_{c,\delta}$ than an even split of the same samples among independently trained agents. The detailed comparison is provided in Appendix A.4.

B. Sufficient-Communication Classification Guarantee

We next turn to the classification performance of the trained network under sufficient communication. Section III-A characterizes the training performance through the probability that the learned network attains a prescribed decision margin δ. We now examine how this margin condition translates into the probability of classification error for a fresh testing example. To distinguish the randomness arising from training from that of prediction, let

$$\mathcal{T} \triangleq \sigma(\mathcal{D}_1, \ldots, \mathcal{D}_K, \mathcal{S}_1, \ldots, \mathcal{S}_K), \tag{40}$$

which collects the training data and the randomness of the local training algorithms. Conditional on $\mathcal{T}$, the learned classifiers $\boldsymbol{c}_k$ and their empirical training means are therefore fixed. Moreover, the testing sample $(\boldsymbol{h}^*, \boldsymbol{\gamma}^*)$ is independent of $\mathcal{T}$. Since the training event $\mathcal{C}_\delta$ associated with (34) is determined by the training phase, it is $\mathcal{T}$-measurable.

From (17), under sufficient communication, the collective prediction $\widehat{\boldsymbol{\gamma}}$ is determined by the sign of the aggregate score $\boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*)$. Let

$$\mathcal{M} \triangleq \{\boldsymbol{\gamma}^* \boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*) \le 0\} \tag{41}$$

denote the event that the signed aggregate score is nonpositive. Under any fixed tie-breaking rule at zero, the actual misclassification event is contained in $\mathcal{M}$. With $P_e \triangleq \mathbb{P}(\widehat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}^*)$, the tower property and the $\mathcal{T}$-measurability of $\mathcal{C}_\delta$ give

$$\begin{aligned} P_e &\le \mathbb{P}(\mathcal{M}) = \mathbb{E}[\mathbb{P}(\mathcal{M}|\mathcal{T})] \\ &= \mathbb{E}[1[\mathcal{C}_\delta]\mathbb{P}(\mathcal{M}|\mathcal{T})] + \mathbb{E}\big[1[\overline{\mathcal{C}_\delta}]\mathbb{P}(\mathcal{M}|\mathcal{T})\big] \\ &\le \mathbb{E}[1[\mathcal{C}_\delta]\mathbb{P}(\mathcal{M}|\mathcal{T})] + \mathbb{P}(\overline{\mathcal{C}_\delta}). \end{aligned} \tag{42}$$

This decomposition separates the conditional prediction error for training realizations in $\mathcal{C}_\delta$ from the probability that training fails to attain the prescribed δ-margin condition in (34). The latter is characterized by Proposition 2. Based on the event $\mathcal{C}_\delta$, the class-conditional mean of the aggregate score $\boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*)$ lies on the correct side of the zero decision threshold with a margin greater than δ. The remaining prediction error depends on the variability of the realized aggregate score $\boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*)$ around its corresponding class-conditional mean.

The local observations $\boldsymbol{h}_1^*, \ldots, \boldsymbol{h}_K^*$ are different views of the same testing example and may therefore remain statistically dependent even after conditioning on the class label. Rather than requiring conditional independence, we allow their joint class-conditional distribution P_γ to satisfy the following approximate tensorization property [49]. For a probability law P and a bounded nonnegative function G, define

$$\mathrm{Ent}_P(G) \triangleq \mathbb{E}_P[G \log G] - \mathbb{E}_P[G] \log \mathbb{E}_P[G]. \tag{43}$$

For each k, let $\boldsymbol{h}_{-k}$ collect all components of $\boldsymbol{h}$ except $\boldsymbol{h}_k$, and let $P_\gamma^{k|-k}(\cdot|\boldsymbol{h}_{-k})$ denote the conditional law of $\boldsymbol{h}_k$ given the remaining views under P_γ.

Assumption 5 (Class-conditional approximate tensorization of entropy (ATE)). *For every $\gamma \in \Gamma$, there exists $\tau_\gamma \ge 1$ such that*

$$\mathrm{Ent}_{P_\gamma}(G) \le \tau_\gamma \sum_{k=1}^{K} \mathbb{E}_{P_\gamma}\left[\mathrm{Ent}_{P_\gamma^{k|-k}(\cdot|\boldsymbol{h}_{-k})}\big(G(\cdot, \boldsymbol{h}_{-k})\big)\right] \tag{44}$$

for every bounded nonnegative measurable G. □

Assumption 5 allows statistically dependent local views across agents beyond the product case. The coefficient τ_γ serves as a dependence parameter for the ATE condition: product class-conditional distributions satisfy the condition with $\tau_\gamma = 1$, while larger values accommodate greater departures from this product structure. For convenience, define

$$\tau_{\max} \triangleq \max_{\gamma \in \Gamma} \tau_\gamma. \tag{45}$$

Together with the training guarantee in Proposition 2, the ATE condition (44) leads to the following sufficient-communication classification guarantee.

Theorem 1 (Sufficient-communication classification guarantee). *Suppose Assumptions 2, 3, and 5 hold, and let $\delta \ge 0$. If the agents label their testing samples $\boldsymbol{h}_k^*$ with $\widehat{\boldsymbol{\gamma}}$ according to* (17)*, then, almost surely on $\mathcal{C}_\delta$,*

$$\mathbb{P}(\mathcal{M}|\mathcal{T}) \le \exp\left\{-\frac{\delta^2}{2\tau_{\max}\beta^2 \sum_{k=1}^{K} \pi_k^2}\right\}. \tag{46}$$

If, in addition, the conditions of Proposition 2 hold, combining the two guarantees through (42) *yields an explicit upper bound on P_e.*

Proof. See Appendix B.2. □

Theorem 1 establishes the connection between the training performance and the reliability of subsequent collaborative prediction. On $\mathcal{C}_\delta$, the conditional prediction error is controlled by the decision margin δ, the dependence factor $\tau_{\max}$, and the Perron weights, while Proposition 2 characterizes the probability of attaining this favorable training event. Through (42), the two guarantees jointly characterize the overall classification performance. We discuss several implications of this result.

1) Relation to margin theory: The decision margin δ plays complementary roles in training and prediction. Proposition 2 shows that requiring a larger δ makes the training condition more difficult to attain, whereas Theorem 1 shows that, once it is attained, a larger δ yields a stronger conditional prediction guarantee. This tradeoff is reminiscent of the role of margins in classical margin-based generalization theory [44]. Given the classifier $\boldsymbol{c}_{\mathsf{ave}}$ from (18), the quantity $\boldsymbol{\gamma}^* \boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*)$ is known as the "margin" associated with the testing sample $\boldsymbol{h}^*$. The class-conditional expectations of the margin distribution associated with $\boldsymbol{c}_{\mathsf{ave}}$ are $\boldsymbol{\mu}^+(\widehat{\boldsymbol{f}})$ under class $+1$ and $-\boldsymbol{\mu}^-(\widehat{\boldsymbol{f}})$ under class -1. Therefore, the δ-margin condition in (34) provides a first-order statistical characterization of the margin distribution for classifier $\boldsymbol{c}_{\mathsf{ave}}$. According to Theorem 1, this characterization leads to a classification guarantee without specifying the exact distributions of the local decision statistics, in contrast to some distribution-specific soft combination analyses [48].

2) Role of the Perron weights: From (16), the Perron entries π_k determine the relative influence of the local classifiers in the sufficient-communication aggregate. Fusion weight design has been widely studied in ensemble learning, typically for classifiers sharing the same feature representation [40], [50], [51]. The extensions to heterogeneous representations are more problem-dependent [39], [40]. Here, rather than optimizing the

fusion weights under arbitrary heterogeneity, we characterize the effect of the Perron weights induced by the communication policy. Their role extends beyond the factor $\sum_k \pi_k^2$ in the conditional prediction bound (46): the training quantities ρ, R^o, α, and $\varepsilon_{\mathsf{opt}}$, as well as the achievable decision margin, also depend on π. Consequently, although the uniform Perron vector minimizes $\sum_k \pi_k^2$ when δ and $\tau_{\max}$ are fixed, it need not optimize the overall classification performance. Importantly, the simple averaging rule for constructing A, where each agent assigns equal combination weights to all neighbors, does not guarantee uniform π on arbitrary topologies. Several useful rules for constructing doubly-stochastic A are available in [10].

3) Benefits of cooperative prediction: To isolate the benefit of cooperation during prediction, suppose that the same target decision margin δ is attained collaboratively and by a non-cooperative agent. The conditional error bound in Theorem 1 is tighter under collaboration whenever

$$\tau_{\max} \sum_{k=1}^{K} \pi_k^2 < 1. \tag{47}$$

This condition reflects the balance between the gain from averaging the local scores and the penalty induced by their statistical dependence. In particular, for product class-conditional distributions, we may take $\tau_{\max} = 1$, and (47) reduces to $\sum_k \pi_k^2 < 1$, which holds for any nontrivial network satisfying Assumption 3. Cooperation also provides a distinct benefit through the network-level margin condition. The event $\mathcal{C}_\delta$ requires only the π-weighted class-conditional means to attain the prescribed two-sided margin; individual classifiers need not satisfy the same condition. Consequently, agents with weak local evidence can be compensated by sufficiently informative or complementary observations from other agents. This is consistent with Assumption 4, which likewise requires informativeness only at the network level.

C. Finite-Round Communication

Theorem 1 characterizes the sufficient-communication limit. When communication is restricted to a finite number of rounds, however, consensus may not yet be reached and the prediction remains agent-dependent. We next characterize how limited communication affects the classification performance at each agent.

According to (11), after t communication rounds, agent k uses the following decision statistic

$$\boldsymbol{\lambda}_{k,t} = \sum_{\ell=1}^{K} [A^t]_{\ell k} \boldsymbol{c}_\ell(\boldsymbol{h}_\ell^*), \tag{48}$$

and predicts the label

$$\widehat{\boldsymbol{\gamma}}_{k,t} \triangleq \operatorname{sgn}(\boldsymbol{\lambda}_{k,t}). \tag{49}$$

Define the corresponding event

$$\mathcal{M}_{k,t} \triangleq \{\boldsymbol{\gamma}^* \boldsymbol{\lambda}_{k,t} \le 0\}. \tag{50}$$

Under any fixed tie-breaking rule at zero, the misclassification event is contained in $\mathcal{M}_{k,t}$. Thus, with $P_{k,t} \triangleq \mathbb{P}(\widehat{\boldsymbol{\gamma}}_{k,t} \neq \boldsymbol{\gamma}^*)$, applying the same decomposition as in (42) gives

$$P_{k,t} \le \mathbb{P}(\mathcal{M}_{k,t}) \le \mathbb{E}[1[\mathcal{C}_\delta]\mathbb{P}(\mathcal{M}_{k,t}|\mathcal{T})] + \mathbb{P}(\overline{\mathcal{C}_\delta}). \tag{51}$$

On the event $\mathcal{C}_\delta$, a positive decision margin is guaranteed for the π-weighted aggregate under sufficient communication (18). At a finite communication round, from (48), agent k combines the local scores $\boldsymbol{c}_\ell(\boldsymbol{h}_\ell^*)$ using the weights $[A^t]_{\ell k}$. To characterize how much of this guaranteed margin is retained after t rounds, fix any $\sigma \in (\sigma_A, 1)$. Matrix theory ensures that there exists a constant $C(A,\sigma) > 0$ such that [52]

$$|[A^t]_{\ell k} - \pi_\ell| \le C(A,\sigma)\sigma^t \tag{52}$$

for all agents k, ℓ and communication rounds t. Accordingly, define the residual finite-round margin

$$r_t(\delta) \triangleq \delta - 2\beta K C(A,\sigma)\sigma^t. \tag{53}$$

The second term bounds the possible reduction of the guaranteed margin after only t communication rounds.

Theorem 2 (Finite-round classification guarantee). *Suppose Assumptions 2, 3 and 5 hold, and let $\delta > 0$. Fix $\sigma \in (\sigma_A, 1)$ and a corresponding $C(A,\sigma)$. Then, for every agent k and every integer $t \ge 0$ such that $r_t(\delta) > 0$, almost surely on $\mathcal{C}_\delta$,*

$$\mathbb{P}(\mathcal{M}_{k,t}|\mathcal{T}) \le \exp\left\{ -\frac{r_t^2(\delta)}{2\tau_{\max}\beta^2 \sum_{\ell=1}^{K} ([A^t]_{\ell k})^2} \right\}. \tag{54}$$

If, in addition, the conditions of Proposition 2 hold, combining the two guarantees through (51) *yields an explicit upper bound on $P_{k,t}$.*

Proof. See Appendix C. □

The condition $r_t(\delta) > 0$ provides a sufficient communication horizon for retaining a positive guaranteed margin. In particular, when $2\beta K C(A,\sigma) > \delta$, it is equivalent to

$$t > \frac{\log \frac{\delta}{2\beta K C(A,\sigma)}}{\log \sigma}. \tag{55}$$

Thus, faster decay of the transient term $C(A,\sigma)\sigma^t$ allows a positive margin to be guaranteed with fewer communication rounds. The bound in (54) is both round- and agent-dependent. Finite communication affects the prediction guarantee in two ways. First, the residual margin $r_t(\delta)$ reflects the reduction in the guaranteed decision margin caused by incomplete mixing. Second, the column $\{[A^t]_{\ell k}\}_{\ell=1}^K$ specifies the fusion profile at agent k and determines the corresponding concentration term. Therefore, finite-round performance depends on the full combination matrix A, through both the transient fusion weights and their convergence toward the Perron vector. The second-largest eigenvalue magnitude σ_A determines the admissible geometric rates in (52), while $C(A,\sigma)$ captures the associated matrix-power constant. This is consistent with the role of spectral properties in the convergence behavior of distributed averaging [53]. As $t \to \infty$, $A^t \to \pi \mathbb{1}^\top$ and $r_t(\delta) \to \delta$. Accordingly, the finite-round conditional bound converges to the sufficient-communication bound in Theorem 1.

D. Finite-Precision Communication

Theorem 2 considers finite-round communication with real-valued messages. In practical implementations, however, transmitted scalars are typically represented with finite precision. We therefore introduce a finite-bit version of the prediction-stage communication rule and characterize its effect on classification performance. The local training procedures, empirical centering, and the training event $\mathcal{C}_\delta$ remain unchanged.

Since both a bounded score and its empirical training mean lie in $[-\beta, \beta]$, every centered score satisfies

$$|c_{k,f_k}(h)| \le 2\beta. \tag{56}$$

Let $b \ge 1$ denote the available bits per transmitted scalar, and let $2 \le L_b \le 2^b$ reconstruction levels be used. Over the interval $[-2\beta, 2\beta]$, consider the uniform alphabet

$$\begin{aligned} \mathcal{V}_b &= \{v_j = -2\beta + j\Delta_b : \ j = 0, \dots, L_b - 1\}, \\ \Delta_b &= \frac{4\beta}{L_b - 1}. \end{aligned} \tag{57}$$

We consider here an adjacent-level stochastic rounding protocol [54]. For $u \in [v_j, v_{j+1}]$, $j = 0, \dots, L_b - 2$, the quantizer Q_b returns one of the two adjacent reconstruction levels v_j and v_{j+1} according to

$$\mathsf{Q}_b(u) = \begin{cases} v_j, & \text{with probability } \dfrac{v_{j+1} - u}{\Delta_b}, \\ v_{j+1}, & \text{with probability } \dfrac{u - v_j}{\Delta_b}. \end{cases} \tag{58}$$

An important feature of this protocol is that the quantizer Q_b is unbiased, that is,

$$\mathbb{E}[\mathsf{Q}_b(u)|u] = u. \tag{59}$$

Let $x_{k,t}^{(b)}$ denote the finite-precision decision statistic at agent k after t communication rounds. Starting from $x_{k,0}^{(b)} = \boldsymbol{c}_k(\boldsymbol{h}_k^*)$, each agent quantizes its current decision statistic before transmission, and the receiving agents combine the quantized values according to

$$q_{\ell,t}^{(b)} = \mathsf{Q}_b(x_{\ell,t}^{(b)}), \quad x_{k,t+1}^{(b)} = \sum_{\ell=1}^{K} a_{\ell k} q_{\ell,t}^{(b)}. \tag{60}$$

The random draws used for stochastic rounding are independent across agents and communication rounds, with each agent sending the same quantized value to all of its neighbors. After t rounds, agent k predicts according to

$$\widehat{\boldsymbol{\gamma}}_{k,t}^{(b)} \triangleq \operatorname{sgn}(x_{k,t}^{(b)}). \tag{61}$$

Define the event for the finite-precision decision statistic

$$\mathcal{M}_{k,t}^{(b)} \triangleq \{\boldsymbol{\gamma}^* x_{k,t}^{(b)} \le 0\}, \tag{62}$$

and let

$$P_{k,t}^{(b)} \triangleq \mathbb{P}(\widehat{\boldsymbol{\gamma}}_{k,t}^{(b)} \ne \boldsymbol{\gamma}^*) \tag{63}$$

denote the classification error probability. As in the real-valued finite-round case, the decomposition in (42), with $\mathcal{M}$ replaced by $\mathcal{M}_{k,t}^{(b)}$, separates the conditional prediction error $\mathbb{P}(\mathcal{M}_{k,t}^{(b)}|\mathcal{T})$ on $\mathcal{C}_\delta$ from the probability of the unfavorable training event $\overline{\mathcal{C}_\delta}$.

Since quantization is performed at every communication round, the perturbations introduced at earlier rounds are propagated through subsequent neighbor-aggregation steps. To state the resulting finite-precision guarantee compactly, define the accumulated quantization term

$$\begin{aligned} V_{k,t}^{\mathrm{q}} &\triangleq \frac{\Delta_b^2}{4} \sum_{r=1}^{t} \sum_{j=1}^{K} ([A^r]_{jk})^2 \\ &= \frac{4\beta^2}{(L_b - 1)^2} \sum_{r=1}^{t} \sum_{j=1}^{K} ([A^r]_{jk})^2. \end{aligned} \tag{64}$$

The quantity $V_{k,t}^{\mathrm{q}}$ depends jointly on the communication precision, the number of communication rounds, and the network combination policy.

Theorem 3 (Finite-precision classification guarantee). *Suppose Assumptions 2, 3, and 5 hold, and let $\delta > 0$. Suppose the agents follow the stochastic finite-precision recursion* (60)*. Fix $\sigma \in (\sigma_A, 1)$ and a corresponding $C(A, \sigma)$. Then, for every agent k and every integer $t \ge 0$ such that $r_t(\delta) > 0$, almost surely on $\mathcal{C}_\delta$,*

$$\mathbb{P}(\mathcal{M}_{k,t}^{(b)}|\mathcal{T}) \le \exp\left\{ -\frac{r_t^2(\delta)}{2\left[\tau_{\max}\beta^2 \sum_{\ell=1}^{K} ([A^t]_{\ell k})^2 + V_{k,t}^{\mathrm{q}}\right]} \right\}. \tag{65}$$

If, in addition, the conditions of Proposition 2 hold, combining the two bounds in (65) *and* (39) *yields an explicit upper bound on $P_{k,t}^{(b)}$.*

Proof. See Appendix D. □

Theorem 3 extends the real-valued finite-round guarantee (54) established in Theorem 2 by quantifying the additional effect of finite communication precision. The same residual finite-round margin term $r_t(\delta)$ from (53) appears in both guarantees, while finite precision introduces the additional term $V_{k,t}^{\mathrm{q}}$ in the denominator of the prediction exponent (65). This result also reveals a tradeoff pertaining to the number of communication rounds. Increasing t increases the guaranteed residual margin $r_t(\delta)$ by reducing the effect of incomplete mixing, but also allows additional quantization effects to accumulate through $V_{k,t}^{\mathrm{q}}$. Moreover, the combination policy A influences both effects through the finite-round fusion weights and the propagation of quantization noise across rounds. Consequently, at fixed precision, the bound need not improve monotonically with the number of communication rounds.

The communication precision also determines the communication cost. Let

$$E_{\mathrm{off}}(A) \triangleq \sum_{k=1}^{K} \sum_{\substack{\ell=1 \\ \ell \ne k}}^{K} \mathbb{1}[a_{\ell k} > 0] \tag{66}$$

denote the number of active directed off-diagonal links. If each quantizer index is represented by a fixed-length b-bit word, the per-round and total communication costs per testing example are

$$B_{\mathrm{round}} = bE_{\mathrm{off}}(A), \quad B_{\mathrm{total}}(t) = tbE_{\mathrm{off}}(A), \tag{67}$$

respectively, excluding packet headers and coding overhead. Self-loops correspond to the local operations and are therefore not counted. Moreover, for any fixed t, we have from (64) that

$$V_{k,t}^{\mathrm{q}} = O\big((L_b - 1)^{-2}\big). \tag{68}$$

Here, the $O(\cdot)$ notation describes the decay with respect to L_b for fixed t. Under the full-code convention $L_b = 2^b$, this becomes $V_{k,t}^{\mathrm{q}} = O(2^{-2b})$. Therefore, increasing the bit width increases the communication cost linearly, while the finite-precision term in the bound (65) decays as $O(2^{-2b})$. For fixed t, as $b \to \infty$, the finite-precision guarantee converges to the corresponding real-valued finite-round guarantee.

E. PAC-Style Generalization Guarantee

The preceding results characterize collaborative prediction through the lens of expected network margin. We now consider a complementary PAC-style perspective that relates the population classification error of the realized classifiers to their performance observed on the training data. This is the core idea of classical margin-based generalization bounds [44]. Specifically, to quantify this performance, for $\eta > 0$, we define the η-margin loss [44]

$$\Phi_\eta(x) = \min\bigg(1, \max\Big(0, 1 - \frac{x}{\eta}\Big)\bigg). \tag{69}$$

It satisfies

$$1[\mathrm{sgn}(f(\boldsymbol{h})) \neq \boldsymbol{\gamma}] \leq \Phi_\eta(\boldsymbol{\gamma} f(\boldsymbol{h})) \leq 1[\boldsymbol{\gamma} f(\boldsymbol{h}) \leq \eta]. \tag{70}$$

The margin parameter η used here is distinct from the decision-margin level δ in (34). For a training set $\{(h_n, \gamma_n)\}_{n=1}^m$, the empirical margin loss for an arbitrary function f is given by

$$P_{e,\mathsf{emp}}^{\eta}(f) \triangleq \frac{1}{m} \sum_{n=1}^{m} \Phi_\eta(\gamma_n f(h_n)). \tag{71}$$

By (70), $P_{e,\mathsf{emp}}^{\eta}(f)$ upper bounds the empirical classification error on the same sample. When the classifiers depend on the training data, their classification errors on a fresh example are understood conditionally on the realized training information. Specifically, for a local classifier $\boldsymbol{c}_k$ and the aggregate classifier $\boldsymbol{c}_{\mathsf{ave}}$, respectively, we define

$$P_e(\boldsymbol{c}_k) \triangleq \mathbb{P}\big(\mathrm{sgn}(\boldsymbol{c}_k(\boldsymbol{h}_k^*)) \neq \boldsymbol{\gamma}^* \,|\, \mathcal{T}\big), \tag{72}$$

$$P_e(\boldsymbol{c}_{\mathsf{ave}}) \triangleq \mathbb{P}\big(\mathrm{sgn}(\boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*)) \neq \boldsymbol{\gamma}^* \,|\, \mathcal{T}\big). \tag{73}$$

To evaluate the empirical margin loss of the aggregate classifier $\boldsymbol{c}_{\mathsf{ave}}$ with (71), the training samples must provide the aligned collection of local views on which $\boldsymbol{c}_{\mathsf{ave}}$ operates. Therefore, for the PAC analysis, we specialize the training setup to equal local sample sizes $N_k = N$ and aligned network examples. Specifically, let

$$\mathcal{D} \triangleq \big\{(\boldsymbol{h}_n, \boldsymbol{\gamma}_n)\big\}_{n=1}^N \tag{74}$$

consist of N i.i.d. draws from the joint distribution Q, where $\boldsymbol{h}_n = (\boldsymbol{h}_{1,n}, \ldots, \boldsymbol{h}_{K,n})$. The local training set at agent k is the corresponding projection

$$\mathcal{D}_k \triangleq \big\{(\boldsymbol{h}_{k,n}, \boldsymbol{\gamma}_n)\big\}_{n=1}^N. \tag{75}$$

This specialization leaves the local training procedures in Section II-A unchanged. The network examples are independent across n, whereas the K views within each example may remain statistically dependent.

Based on the network training set $\mathcal{D}$ in (74), we establish the following PAC-style generalization bounds for the classifiers $\boldsymbol{c}_k$ and $\boldsymbol{c}_{\mathsf{ave}}$ using their empirical margin losses. Here, ρ_k and ρ denote the individual and network Rademacher complexities defined in Appendix A.1, evaluated at the common sample size $N_k = N$.

Theorem 4 (PAC-style margin generalization guarantee). *Under Assumption 2 and the sampling setup above, fix any $\eta > 0$. Then, for any $\epsilon \in (0,1)$ and every fixed agent k,*

$$P_e(\boldsymbol{c}_k) \leq P_{e,\mathsf{emp}}^{\eta}(\boldsymbol{c}_k) + \frac{4}{\eta}\rho_k + \mathcal{R}(\epsilon, N) \tag{76}$$

with probability at least $1 - \epsilon$. Moreover,

$$P_e(\boldsymbol{c}_{\mathsf{ave}}) \leq P_{e,\mathsf{emp}}^{\eta}(\boldsymbol{c}_{\mathsf{ave}}) + \frac{4}{\eta}\rho + \mathcal{R}(\epsilon, N) \tag{77}$$

with probability at least $1 - \epsilon$, where

$$\mathcal{R}(\epsilon, N) \triangleq \frac{4\beta}{\eta\sqrt{N}} \left(1 + \sqrt{2 \log \frac{1}{\epsilon}}\right). \tag{78}$$

Proof. See Appendix E. □

Theorem 4 provides a complementary guarantee for the trained classifiers without investigating the prescribed network-margin event $\mathcal{C}_\delta$. It also does not require Assumption 5: dependence among the local views within each network example is incorporated through their joint distribution Q. Moreover, the guarantee does not require exact empirical risk minimization or a prescribed optimization gap. Therefore, it holds *uniformly* across classifiers returned by the local training procedures in Section II-A.

From (76) and (77), the local and collaborative bounds have the same form, but differ in their empirical margin losses and complexity terms. Structurally, $\boldsymbol{c}_k$ operates on a single local feature space $\mathcal{H}_k$, whereas $\boldsymbol{c}_{\mathsf{ave}}$ operates on the joint feature space $\mathcal{H}_1 \times \cdots \times \mathcal{H}_K$. The empirical margins of $\boldsymbol{c}_{\mathsf{ave}}$ can therefore reflect how information from different local views is combined. The theorem does not, however, impose a universal ordering between the local and collaborative bounds. Any collaborative advantage must be reflected in the aggregate margin loss $P_{e,\mathsf{emp}}^{\eta}(\boldsymbol{c}_{\mathsf{ave}})$ relative to the corresponding complexity terms ρ. This differs from the conventional ensemble setting, where multiple models typically operate on the same feature representation and the benefit of combining their outputs is often discussed in terms of variance reduction [44], [55].

F. Relation to Social Learning

The DeGroot rule (11) and conventional social learning (SL) strategies both aggregate information across a network, but differ in how evidence enters the inference process. In the present framework, agents repeatedly communicate decision statistics initialized from a fixed testing example, whereas conventional SL recursions continually incorporate newly arriving

observations. To clarify the relation between these two inference mechanisms, Appendix F considers a standard geometric-pooling SL recursion under a *static-observation* specialization, in which each agent reuses the same local testing observation at every communication round. This specialization places the SL rule in the same single-testing-example setting as the DeGroot rule and enables a direct comparison, while remaining distinct from the conventional streaming operation of SL.

Under this specialization, repeated SL updates accumulate the same local evidence rather than new temporal evidence. After normalization by the number of communication rounds, the SL decision statistic converges to the same π-weighted aggregate $\boldsymbol{c}_{\mathsf{ave}}$ as the DeGroot rule (11). Consequently, whenever $\boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*) \neq 0$, the two strategies yield the same prediction under sufficient communication. Their finite-round decision statistics are generally different, however, and therefore lead to different finite-round guarantees. The complete comparison and the corresponding finite-round SL guarantee are provided in Appendix F.

Raw-score counterparts. The preceding analysis considers the empirically centered local scores $\boldsymbol{c}_k(\boldsymbol{h}_k)$ defined in (9). As shown by Proposition 1, centering affects threshold alignment rather than being a structural requirement for collaboration. We refer to the learned scores $\widehat{\boldsymbol{f}}_k(\boldsymbol{h}_k)$ used without empirical centering as *raw scores*. In Appendix G, we provide an analogous analysis for raw scores. This analysis shows that the main collaborative guarantees persist without empirical centering, although the conditions for attaining a prescribed zero-threshold margin and the constants appearing in the finite-round bounds change. Neither representation universally dominates the other: centering can improve threshold alignment, whereas raw scores avoid estimating the empirical center and have a smaller worst-case score magnitude than the centered scores.

IV. Numerical Simulations

We evaluate the proposed test-time collaborative classification framework under two complementary observation models. The first is a controlled patch-partition benchmark based on CIFAR-10, where each agent observes a spatial patch extracted from a common image. The second is a multi-view benchmark constructed from rendered objects in ModelNet40, where each agent observes the same 3D object from a different viewpoint. The two benchmarks introduce different forms of local heterogeneity, namely spatial partitioning and viewpoint variation. We further examine dependence across agents, heterogeneous local models, communication topology, quantization, adaptive stopping, and corrupted reports. Unless otherwise stated, all curves are averaged over multiple Monte Carlo repetitions and the error bars indicate 95% confidence intervals. Additional implementation details and extended numerical results are provided in Appendix H.

A. CIFAR-10 patch-partition benchmark

We begin with a controlled patch-partition benchmark based on CIFAR-10 [56]. We consider the binary task of distinguishing *cats* from *dogs*. Each 32×32 RGB image is divided into a 3×3 grid of non-overlapping spatial patches, yielding $K = 9$ agents, and agent k receives only the patch associated with its grid location. In this way, the agents observe different local regions of the same image and collaborate only during inference. The communication network is taken to be a directed Erdös–Rényi graph [46], generated randomly at the beginning of the experiment. The observation map of the nine agents and the communication topology are illustrated in Fig. 1.

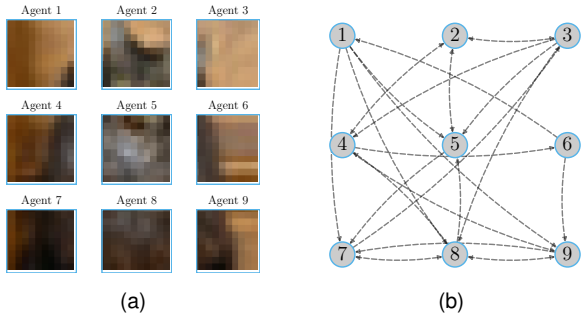


Fig. 1. CIFAR-10 patch-partition benchmark. Panel (a) shows the local patch observed by each agent for a representative cat image, and panel (b) shows the communication topology used for collaboration.

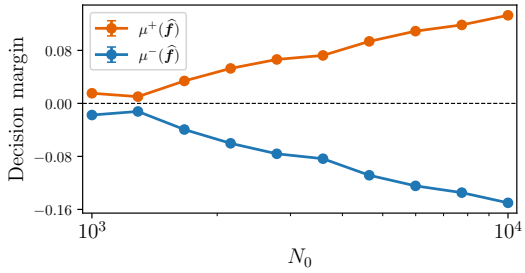
At each agent, a convolutional neural network (CNN) is trained independently on the corresponding patch dataset. For simplicity, we use the same local labeled dataset size N_0 for all agents. For each value of N_0, we construct balanced local datasets, train the classifiers independently, perform collaborative classification at test time, and average the results over multiple repetitions. The detailed training protocol is provided in Appendix H.1.

Figure 2 summarizes the main performance results for this CIFAR-10 benchmark. Panel (a) reports the empirical class-conditional decision statistics under different values of N_0. In the notation of Section III, these curves correspond to the empirical behavior of $\boldsymbol{\mu}^+(\widehat{\boldsymbol{f}})$ and $\boldsymbol{\mu}^-(\widehat{\boldsymbol{f}})$ in (33). As N_0 increases, $\boldsymbol{\mu}^+(\widehat{\boldsymbol{f}})$ moves farther above zero and $\boldsymbol{\mu}^-(\widehat{\boldsymbol{f}})$ moves farther below zero, so that the separation between the two curves grows overall, consistent with a larger achievable decision margin in the sense of condition (34). This suggests that the network of trained classifiers becomes more informative as more training samples are available, which is beneficial for classification at test time.

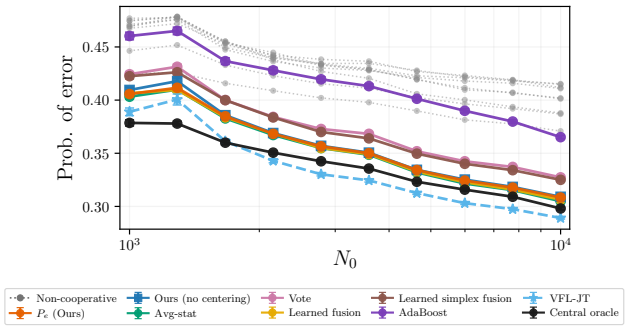
Panel (b) shows the corresponding classification performance. We compare the proposed method with several rules from ensemble learning and decision fusion, together with the non-cooperative baseline, an explicit centering ablation *Ours (no centering)*, the VFL-JT baseline, and a central oracle trained on the full image as a full-information benchmark. The VFL-JT baseline uses a score-level joint-training scheme inspired by [57], where the local models and a server-side fusion head are optimized jointly using aligned labeled samples. Table I summarizes the training and inference coordination of the methods considered in the main baseline comparison. Since several of these reference rules depend on centralized aggregation of local outputs, the comparison is carried out in the regime of sufficient communication, where the distributed recursion (11) approaches its π-weighted consensus limit given

TABLE I
TRAINING AND INFERENCE COORDINATION OF THE METHODS CONSIDERED IN THE MAIN BASELINE COMPARISONS.

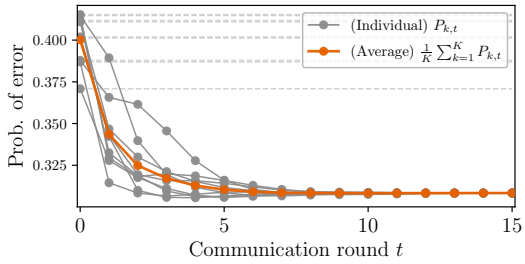
Method	Training coordination	Additional fitting	Inference coordination
Non-cooperative	None	None	None
Avg-stat / Vote	None	None	One-shot centralized averaging / voting
Learned (simplex) fusion	None	Fusion rule fitted on aligned validation outputs	One-shot centralized fusion
AdaBoost	Centralized	Boosting-based fitting of local predictors	One-shot weighted fusion
Ours / Ours (no centering)	None	None	Iterative peer-to-peer scalar exchange
VFL-JT	Joint training on aligned labeled samples	Server-side fusion head learned jointly	One-shot server fusion
Central oracle	Centralized full-information training	None	Centralized full-information prediction



(a) Decision margin



(b) Probability of error



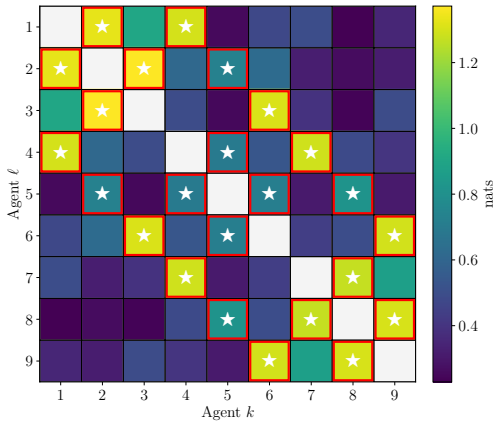
(c) Instantaneous probability of error

Fig. 2. Main CIFAR-10 patch-partition results. Panel (a) reports the empirical class-conditional decision statistics, panel (b) compares the probability of error under different values of N_0, and panel (c) shows the error evolution over the communication rounds for $N_0 = 10000$.

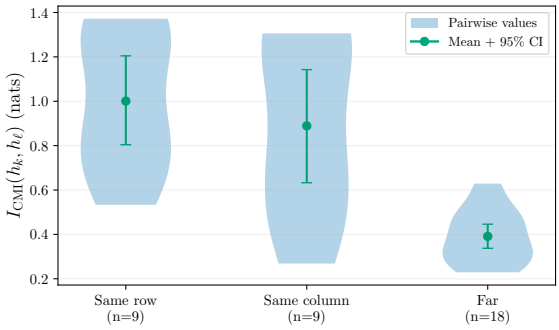
in (16). From Fig. 2b, the probability of error decreases as N_0 increases, consistent with the increased decision margins observed in Fig. 2a. At the same time, the proposed method consistently improves upon non-cooperative prediction and remains competitive with the stronger fusion baselines. With joint training, VFL-JT achieves lower error than the proposed method over the tested range. We also note that under sufficient communication, *Avg-stat* coincides with the proposed method when A is doubly-stochastic, while *Ours* and *Ours (no centering)* exhibit similar performance over the tested range.

Panel (c) illustrates the evolution of $P_{k,t}$ during the collaboration process. As the communication round t grows, the decision statistic $\boldsymbol{\lambda}_{k,t}$ converges to $\boldsymbol{\lambda}_{\mathsf{ave}}$, and $P_{k,t}$ approaches the error of the limiting aggregate classifier. It is worth noting that $P_{k,0}$ at $t = 0$ corresponds to the error incurred by agent k in the non-cooperative scenario. The light grey dashed horizontal lines in Fig. 2c indicate these baseline error levels for reference. The average of $P_{k,t}$ across all K agents is also shown in Fig. 2c, which decreases over the communication rounds before stabilizing. This highlights the improvement in the network-wide performance under collective prediction.

We next present two additional studies for the CIFAR-10 patch-partition benchmark.



(a) Conditional mutual information matrix



(b) Grouped conditional mutual information

Fig. 3. Conditional mutual information between CIFAR-10 patches. Panel (a) shows the estimated pairwise conditional mutual information matrix across the nine patch agents. White stars mark patch pairs that are adjacent on the 3×3 grid shown in Fig. 1a, while red boxes indicate, in each row, the largest off-diagonal entries whose number matches the number of grid neighbors of the corresponding patch. Panel (b) shows the grouped summary over same-row, same-column, and far patch pairs.

1) Conditional mutual information between patches: Since all patches are extracted from the same natural image, nearby patches may remain statistically dependent even after conditioning on the class label. To quantify this effect, we estimate the class-conditional mutual information for each label value and then average according to the class probabilities:

$$I_{\mathrm{CMI}}(h_k, h_\ell) \triangleq \sum_{\gamma \in \Gamma} \mathbb{P}(\boldsymbol{\gamma} = \gamma)\, I(h_k; h_\ell | \boldsymbol{\gamma} = \gamma) \tag{79}$$

where h_k denotes the raw pixel vector of the patch observed by agent k, $\boldsymbol{\gamma}$ denotes the class label, and $I(h_k; h_\ell|\boldsymbol{\gamma} = \gamma)$ is the mutual information under the conditional distribution given $\boldsymbol{\gamma} = \gamma$. The reported estimates are computed by first reducing the patch dimension through principal component analysis (PCA) [45] and then applying a k-nearest-neighbor (kNN) mutual information estimator [58]. More details about this estimation process are provided in Appendix H.1.3.

Figure 3 shows that the estimated conditional mutual information is clearly nonzero for many patch pairs and tends to be strongest among spatially related patches. Thus, the CIFAR-10 patch-partition benchmark induces a structured observation model in which nearby local views exhibit stronger statistical dependence than distant ones. This observation motivates the dependence-aware prediction analysis in Section III. While the CMI diagnostic is not an estimate of the ATE coefficient in Assumption 5, it does provide empirical evidence that the local views are not independent in this benchmark.

2) Heterogeneous local model families: We next relax the assumption that all agents use the same local architecture. Different agents are then assigned different local model families, while the communication graph and the collaboration rule remain unchanged. The local models include the patch-level convolutional network used in the main CIFAR-10 benchmark, a logistic-regression classifier on raw patch pixels, and a small residual convolutional network [59]. The detailed protocol and the precise model specifications are given in Appendix H.1.4.

Figure 4 reports a balanced assignment study in which the model families are rotated across the 3×3 grid so that each family appears equally often at every spatial location. The results indicate that collaboration remains beneficial across all three model families and across all patch locations, although the magnitude of the improvement varies with both the family and the position. Furthermore, under one representative heterogeneous assignment, the main performance results summarized in Fig. 5 show that the same qualitative behavior observed in the homogeneous benchmark is preserved here: the empirical decision margins become more separated as N_0 increases, the collaborative classifier improves over non-cooperative inference, and most of the performance gain is realized within the first few communication rounds.

B. ModelNet40 multi-view benchmark

In this section, we consider a multi-view benchmark derived from ModelNet40 [60]. Unlike the CIFAR-10 patch-partition setting, where the agents observe different *spatial* regions of the same image, here each agent receives one rendered *view* of the same 3D object. The heterogeneity across agents therefore arises from viewpoint rather than patch location.

For each object, we generate $K = 12$ RGB renderings from fixed camera positions whose azimuth angles are uniformly spaced around the object, with a common elevation angle of 30°. One rendered view is then assigned to each agent. In the experiments, we consider the binary task pair of *dresser* and *nightstand*. The multi-view construction and a randomly-generated communication topology used in our experiments are illustrated in Fig. 6. Each local classifier is a lightweight convolutional network, referred to as *MicroCNN*; further experimental details are given in Appendix H.2.

Figure 7 summarizes the main performance results for this ModelNet40 benchmark. In panel (a), the empirical statistics $\boldsymbol{\mu}^+(\widehat{\boldsymbol{f}})$ and $\boldsymbol{\mu}^-(\widehat{\boldsymbol{f}})$ remain separated across N_0, indicating that the classifier network remains informative under the viewpoint heterogeneity. Compared with the CIFAR-10 benchmark in Fig. 2, their variation with N_0 is milder and exhibits more visible fluctuations, which is consistent with the smaller amount of training data available in this benchmark. The corresponding probability of error curves in Fig. 7b also decrease overall as N_0 increases, in agreement with the larger decision margins attained at larger values of N_0. Meanwhile, the improvement of the proposed method is more limited at smaller N_0, where some local training sets may be too small for the agents to learn informative classifiers from their views. In this regime, some non-cooperative agents can achieve lower error when acting alone than under collaboration. The relative ordering of the score-fusion baselines is also more sensitive than in the CIFAR-10 benchmark, which may reflect the fact that classifiers trained on different viewpoints need not produce scores on comparable scales. In particular, since *Avg-stat* corresponds to a special case of the proposed method under sufficient communication when A is doubly-stochastic, the gap between *Avg-stat* and *Ours* indicates that the choice of the combination policy A can have a visible effect on collaborative performance. The instantaneous error curves in Fig. 7c further show that most of the performance gain is achieved within the first few communication rounds.

1) Temperature scaling and score calibration: The sensitivity of the score fusion baselines in Fig. 7 suggests that score calibration across viewpoints deserves closer examination. To this end, we consider temperature scaling as a post-hoc calibration step in our experiments [61]. For each agent, a positive temperature is fitted on the local validation set by minimizing the cross-entropy loss of the rescaled logits, as detailed in Appendix H.2.1. The local decision statistic is then formed from the difference between the two calibrated logits, followed by the same centering step (9) as in the uncalibrated setting. Since temperature scaling does not change the predicted class of an individual local classifier, its role here is not to alter the local decisions themselves, but to improve the comparability of the real-valued decision statistics exchanged during collaboration.

Figure 8 shows that, relative to Fig. 7, the class-conditional decision statistics become more clearly separated after calibration, and the probability of error curve of the proposed method is slightly improved over most of the tested range of N_0. This suggests that, under viewpoint heterogeneity, score calibration can improve the alignment of the exchanged

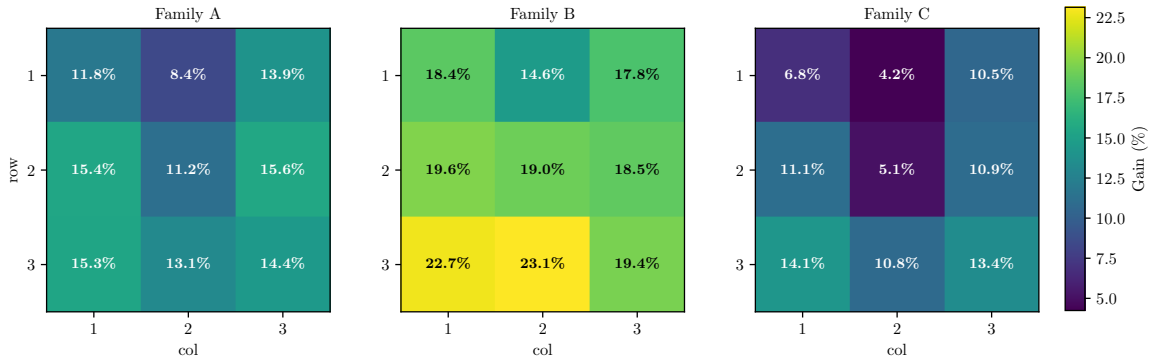
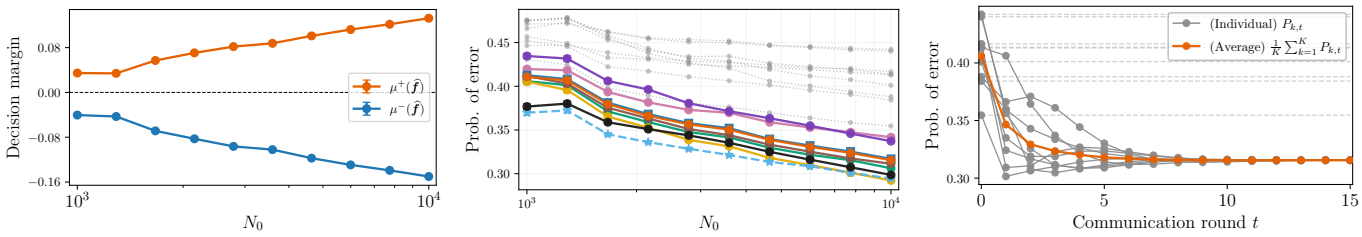
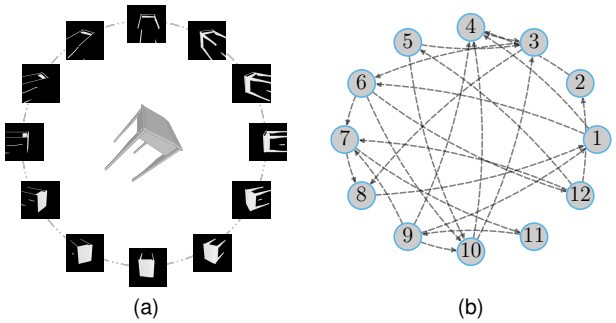


Fig. 4. Collaboration gain under heterogeneous local model families on CIFAR-10. Each cell shows the relative gain when the corresponding model family is assigned to that patch location under the balanced rotation protocol.



(a) Decision margin (b) Probability of error (c) Instantaneous probability of error

Fig. 5. Performance of heterogeneous local model families on CIFAR-10. The curve labels in panel (b) and the value of N_0 in panel (c) are the same as those in Fig. 2.



(a) (b)

Fig. 6. ModelNet40 multi-view benchmark. Panel (a) shows the 12 rendered views associated with a nightstand object, and panel (b) shows the communication topology used for collaboration.

decision statistics across agents and thereby improve collaborative prediction. VFL-JT is not included in this study because temperature scaling of its local scores is coupled with the jointly optimized fusion head. Additional figures are provided in Appendix H.2.1.

2) Controlled correlation stress test: In the multi-view setting, different agents observe the same underlying object and can therefore produce statistically dependent local decision statistics. To examine how the collaborative classifier behaves as this dependence becomes stronger, we perform a controlled stress test directly on the local decision statistics, with more details provided in Appendix H.2.2. Let $x_{i,k}$ denote the clean local decision statistic produced by agent k for the i-th testing sample. To control the correlation between $x_{i,k}$ across agents, we consider a perturbed local decision statistic at each agent, denoted by $x'_{i,k}$:

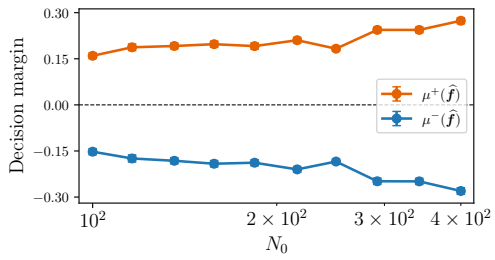
$$x'_{i,k} = x_{i,k} + \sigma_{\text{pert}}(\omega_c(r)\, u_i + \omega_i(r)\, e_{i,k}) \tag{80}$$

where $u_i \sim \mathcal{N}(0,1)$ is a shared standard Gaussian perturbation for sample i, $e_{i,k} \sim \mathcal{N}(0,1)$ are independent standard Gaussian perturbations across samples and agents, and

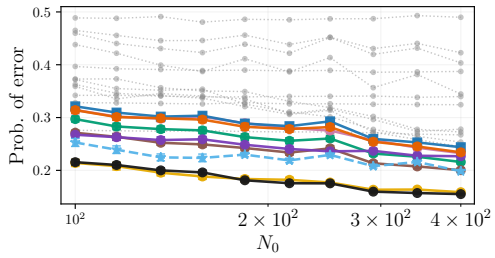
$$\omega_c(r) = \frac{r}{\sqrt{1+r^2}}, \quad \omega_i(r) = \frac{1}{\sqrt{1+r^2}}. \tag{81}$$

Here, $r \geq 0$ controls the relative strength of the common-mode component, while $\sigma_{\text{pert}} > 0$ controls the overall perturbation scale. Since $\omega_c(r)^2 + \omega_i(r)^2 = 1$, varying r changes the balance between common and idiosyncratic perturbations without substantially changing the total injected variance. This construction follows a standard common-factor decomposition with shared and idiosyncratic components. At test time, the agents employ the same rule (11), but with the perturbed local decision statistics $x'_{i,k}$ in place of the clean statistics $x_{i,k}$.

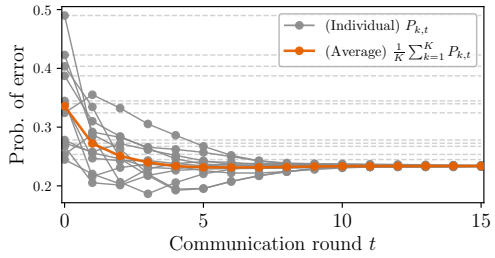
Figure 9 shows that the average pairwise correlations across agents, both unconditional and label-conditional, increase with r. The corresponding collaboration gain, which is measured by the error reduction relative to non-cooperative inference after $t = 20$ communication rounds, decreases with r for each fixed perturbation amplitude. This is consistent with the fact that stronger dependence makes the local decision statistics more redundant and therefore reduces the benefit of information sharing across the network. At the same time, the degradation



(a) Decision margin



(b) Probability of error



(c) Instantaneous probability of error

Fig. 7. Main ModelNet40 multi-view results. Panel (a) reports the empirical class-conditional decision statistics, panel (b) compares the probability of error under different values of N_0, and panel (c) shows the error evolution over the communication rounds for $N_0 = 400$.

is gradual, indicating that the collaborative classifier continues to provide measurable gains over a broad range of dependence levels.

C. Communication design, adaptive stopping, and robustness

In this section, we examine several aspects of the communication protocol using the ModelNet40 benchmark. In particular, we study how collaborative performance is influenced by the graph topology, the combination policy, the communication precision, adaptive stopping rules, and corrupted messages.

1) Topology, combination policy, and communication precision: We begin by examining the impact of the communication design across three graph topologies: ring, grid, and directed Erdös–Rényi. For each topology, the combination policy A is constructed using two different schemes: the uniform averaging rule and the Metropolis rule. For this quantization study only, we represent the already trained probabilistic classifier by the bounded posterior-probability difference

$$g_k(h) \triangleq \widehat{p}_k(+1|h) - \widehat{p}_k(-1|h) = \tanh\left(\frac{\widehat{\boldsymbol{f}}_k(h)}{2}\right) \in [-1, 1]. \tag{82}$$

No retraining process is required for this representation. Under this transformation, the resulting centered local score

$$c_k^g(h) \triangleq g_k(h) - \frac{1}{N_k}\sum_{n=1}^{N_k} g_k(h_{k,n}) \tag{83}$$

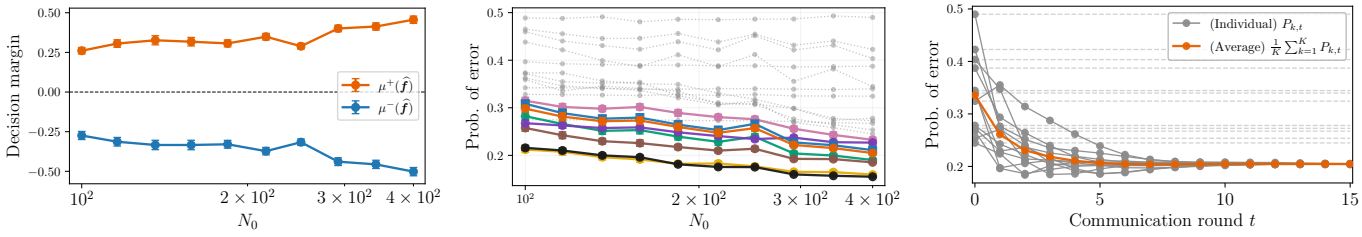
satisfies $|c_k^g(h)| \leq 2$. We compare the resulting matched full-precision recursion with 4-, 6-, and 8-bit stochastic communication. Further details of the quantization protocol are provided in Appendix H.3.1.

Figure 10 reports the evolution of the average probability of error under these communication designs. All settings benefit from communication in the early rounds, but both the finite-round improvement and the error level eventually reached depend visibly on the topology, the construction rule of A, and the communication precision. Among the graph families considered here, the grid generally attains the lowest error, while the ring also performs competitively. By contrast, the directed Erdös–Rényi topology is more sensitive to the choice of combination policy and communication precision. The precision hierarchy is systematic: 8-bit stochastic communication is practically indistinguishable from the matched bounded full-precision reference, 6-bit communication produces a modest intermediate degradation, and 4-bit communication produces the largest loss. More detailed grouped views of these results are provided in Appendix H.3.1.

2) Adaptive stopping rules: In the previous experiments, the collaboration framework is evaluated under a fixed communication budget t chosen in advance. In practice, however, selecting this budget is not straightforward: a larger t increases communication cost, while a smaller t may be insufficient for collaboration. This motivates the study of adaptive stopping rules, in which the communication pattern is allowed to depend on the evolution of local decision statistics. We consider two local event-triggered families: a Δ-*trigger* rule based on the change in the communicated statistic, and a *label-stability with confidence* rule based on the local prediction and a confidence threshold (see Appendix H.3.2). Since transmissions may stop at different times for different agents and different samples, we measure the resulting communication cost by the total number of transmissions per testing sample.

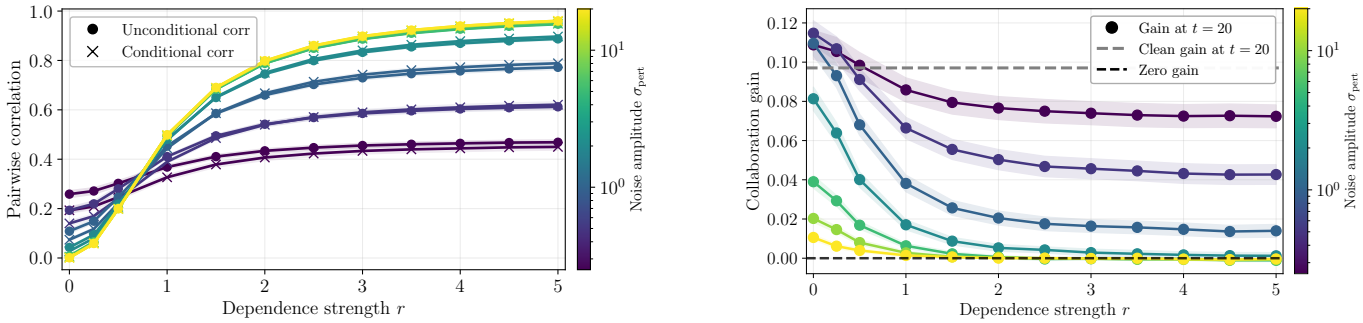
Figure 11 illustrates the tradeoff between the probability of error and the total communication cost for four values of the maximum communication budget $t_{\max}$. In each panel, only the Pareto-efficient operating points are shown, namely those for which no alternative configuration attains both lower communication cost and lower probability of error. Both stopping rule families exhibit a clear communication–accuracy tradeoff, but the label-stability rule generally yields a better Pareto frontier than the Δ-trigger rule, especially in the low communication regime. The star markers denote the best points selected from the two families of stopping rules. Their positions indicate that communication can often be terminated well before reaching the maximum communication budget.

3) Robustness to noisy, faulty, and adversarial agents: We next examine how the collaborative inference mechanism behaves when a subset of agents transmits corrupted messages at test time. The perturbation is introduced only at the level of



(a) Decision margin (b) Probability of error (c) Instantaneous probability of error

Fig. 8. Performance of temperature scaling on ModelNet40. The curve labels in panel (b) and the value of N_0 in panel (c) are the same as those in Fig. 7.



(a) Average pairwise correlation (b) Collaboration gain

Fig. 9. Controlled correlation stress test on ModelNet40. Panel (a) shows the average pairwise Pearson correlation across agent decision statistics, both unconditional and label-conditional, and panel (b) shows the collaboration gain after $t = 20$ communication rounds.

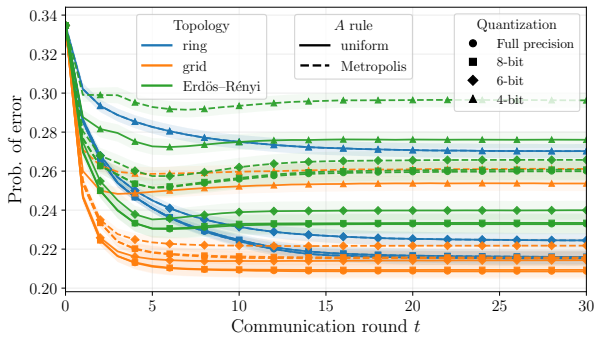
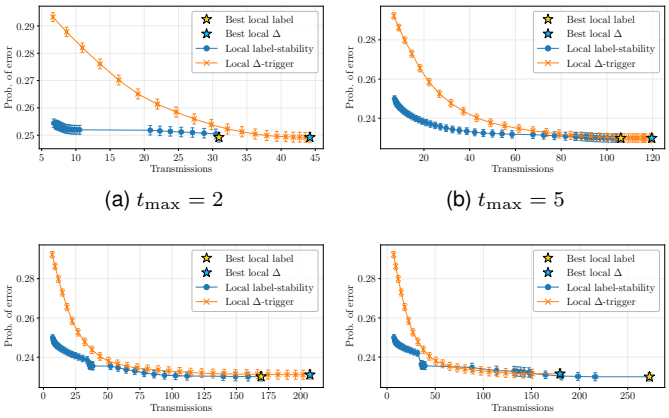


Fig. 10. Effect of topology, the construction rule of A, and quantization noise.



(a) $t_{\max} = 2$ (b) $t_{\max} = 5$ (c) $t_{\max} = 10$ (d) $t_{\max} = 20$

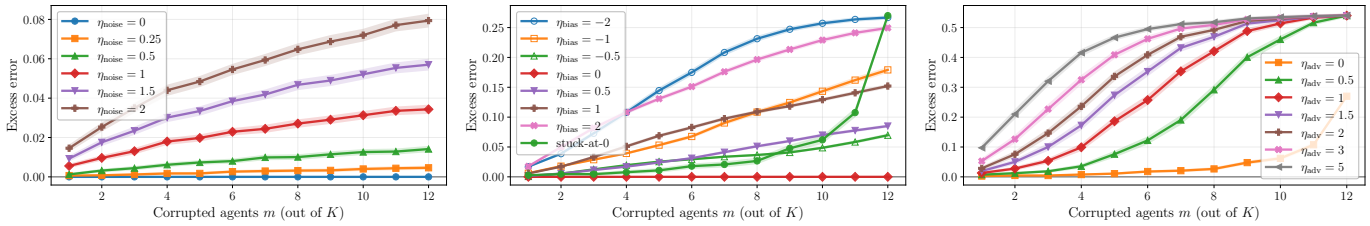
Fig. 11. Tradeoff between the probability of error and the total number of transmissions (per testing sample) under local adaptive stopping rules.

the local decision statistics before communication begins. For each corruption scenario, we compare the resulting probability of error with that of the corresponding clean run obtained from the same trained models, the same testing set, and the same combination policy. The degradation is measured by the *paired excess error*, namely the increase in probability of error relative to the clean run within the same Monte Carlo repetition. In each repetition, we corrupt exactly $m \in \{1, \ldots, K\}$ agents and consider three families of perturbations: benign Gaussian noise, faulty agents (including stuck-at-zero and constant-bias failures), and adversarial reports modeled by scaled sign flips. Additional details on the perturbation models are provided in Appendix H.3.3.

Figure 12 summarizes the robustness results after 20 rounds of communication. In all three perturbation settings, the excess error increases with both the number of corrupted agents and the perturbation severity. Among the three cases, adversarial scaled sign flips are the most damaging, indicating that persistently misleading decision statistics can have a particularly strong effect on collaborative prediction. By contrast, benign Gaussian noise and faulty agents yield a more gradual degradation, especially when only a few agents are affected.

V. Concluding Remarks

In this work, we proposed a distributed test-time collaborative classification framework for multi-agent networks. We considered a group of independently trained agents, poten-



(a) Benign Gaussian noise (b) Faulty agents (c) Adversarial reports

Fig. 12. Robustness of collaborative classification to corrupted agents, measured by the paired excess error at the communication round $t = 20$. Panel (a) varies the Gaussian noise multiplier η_{noise}, panel (b) varies the fault type and the normalized bias level η_{bias}, and panel (c) varies the adversarial scale η_{adv}.

tially with heterogeneous feature spaces and model architectures, that collaborates at inference time by exchanging local beliefs with graph neighbors to produce collective predictions. Our analysis established classification error bounds that characterize the impact of the agent heterogeneity, network topology, combination policies, and communication budgets on collective prediction performance. These bounds provide theoretical insight into the interplay between local training and distributed inference in multi-agent systems, in addition to highlighting the benefits of collaborative classification relative to the non-cooperative scenario. Importantly, this study shows that cooperation can be valuable at test time even when the local models are trained independently, which is becoming an increasingly important paradigm in real-world implementations of distributed machine learning. Future directions include extensions to multi-class settings, theoretical analysis of adaptive and communication-efficient protocols, privacy-preserving implementations, and more general communication settings, including asynchronous updates, time-varying topologies, and lossy communication.

REFERENCES

[1] A. H. Sayed, S.-Y. Tu, J. Chen, X. Zhao, and Z. J. Towfic, "Diffusion strategies for adaptation and learning over networks: an examination of distributed strategies and network behavior," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 155–171, 2013.

[2] S. Kar and J. M. F. Moura, "Consensus + innovations distributed inference over networks: cooperation and sensing in networked systems," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 99–109, 2013.

[3] M. I. Jordan, J. D. Lee, and Y. Yang, "Communication-efficient distributed statistical inference," *Journal of the American Statistical Association*, vol. 114, no. 526, pp. 668–681, 2019.

[4] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer, "A survey on distributed machine learning," *ACM Computing Surveys*, vol. 53, no. 2, pp. 1–33, 2020.

[5] D. Jin, N. Kannengießer, S. Rank, and A. Sunyaev, "Collaborative distributed machine learning," *ACM Computing Surveys*, vol. 57, no. 4, pp. 95:1–95:36, 2024.

[6] V. Matta, P. Braca, S. Marano, and A. H. Sayed, "Diffusion-based adaptive distributed detection: Steady-state performance in the slow adaptation regime," *IEEE Transactions on Information Theory*, vol. 62, no. 8, pp. 4710–4732, 2016.

[7] D. Bajović, J. M. F. Moura, J. Xavier, and B. Sinopoli, "Distributed inference over directed networks: Performance limits and optimal design," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3308–3323, 2016.

[8] 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.

[9] V. Matta, V. Bordignon, and A. H. Sayed, *Social Learning: Opinion Formation and Decision-Making over Graphs*, ser. NOW Book Series on Information and Learning Sciences. Now Publishers, 2025.

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

[11] A. Nedic, "Distributed gradient methods for convex machine learning problems in networks: Distributed optimization," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 92–101, 2020.

[12] P. Braca, L. M. Millefiori, A. Aubry, S. Marano, A. De Maio, and P. Willett, "Statistical hypothesis testing based on machine learning: Large deviations analysis," *IEEE Open Journal of Signal Processing*, vol. 3, pp. 464–495, 2022.

[13] V. Bordignon, S. Vlaski, V. Matta, and A. H. Sayed, "Learning from heterogeneous data based on social interactions over graphs," *IEEE Transactions on Information Theory*, vol. 69, no. 5, pp. 3347–3371, 2023.

[14] P. Hu, V. Bordignon, M. Kayaalp, and A. H. Sayed, "Non-asymptotic performance of social machine learning under limited data," *Signal Processing*, vol. 230, p. 109849, 2025.

[15] S. F. Yilmaz, B. Hasircioğlu, L. Qiao, and D. Gündüz, "Private collaborative edge inference via over-the-air computation," *IEEE Transactions on Machine Learning in Communications and Networking*, vol. 3, pp. 215–231, 2025.

[16] P. Yadav, C. Raffel, M. Muqeeth, L. Caccia, H. Liu, T. Chen, M. Bansal, L. Choshen, and A. Sordoni, "A survey on model MoErging: Recycling and routing among specialized experts for collaborative learning," *Transactions on Machine Learning Research*, 2025.

[17] C. Mavromatis, P. Karypis, and G. Karypis, "Pack of LLMs: Model fusion at test-time via perplexity optimization," in *Conference on Language Modeling*, 2024.

[18] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[19] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[20] D. Acemoglu and A. Ozdaglar, "Opinion dynamics and learning in social networks," *Dynamic Games and Applications*, vol. 1, no. 1, pp. 3–49, 2011.

[21] A. G. Chandrasekhar, H. Larreguy, and J. P. Xandri, "Testing models of social learning on networks: Evidence from two experiments," *Econometrica*, vol. 88, no. 1, pp. 1–32, 2020.

[22] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[23] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv:1610.05492*, 2016.

[24] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[25] Z. Jiang, A. Balu, C. Hegde, and S. Sarkar, "Collaborative deep learning in fixed topology networks," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[26] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv:1806.00582*, 2018.

[27] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 2351–2363.

[28] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[29] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 3557–3568.

[30] D. Wen, K.-J. Jeon, and K. Huang, "Federated dropout–A simple approach for enabling federated learning on resource constrained devices," *IEEE wireless communications letters*, vol. 11, no. 5, pp. 923–927, 2022.

[31] S. Xie, D. Wen, X. Liu, C. You, T. Ratnarajah, and K. Huang, "Convergence analysis for federated dropout," in *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, 2024, pp. 5313–5318.

[32] C. Mendler-Dünner, W. Guo, S. Bates, and M. I. Jordan, "Test-time collective prediction," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 13 719–13 731.

[33] O. A. Alzubi, J. A. A. Alzubi, S. Tedmori, H. Rashaideh, and O. Almomani, "Consensus-based combining method for classifier ensembles," *The International Arab Journal of Information Technology*, vol. 15, no. 1, pp. 76–86, 2018.

[34] 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.

[35] 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.

[36] 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.

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

[38] M. Kayaalp, Y. İnan, E. Telatar, and A. H. Sayed, "On the arithmetic and geometric fusion of beliefs for distributed inference," *IEEE Transactions on Automatic Control*, vol. 69, no. 4, pp. 2265–2280, 2024.

[39] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[40] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, 2014.

[41] P. L. Bartlett, Y. Freund, W. S. Lee, and R. E. Schapire, "Boosting the margin: A new explanation for the effectiveness of voting methods," *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[42] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, 1st ed. Chapman and Hall/CRC, 2012.

[43] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[44] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. MIT press, 2018.

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

[46] M. Newman, *Networks*. Oxford University Press, 2018.

[47] T. G. Lewis, *Network Science: Theory and Applications*. John Wiley & Sons, 2011.

[48] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 281–286, 2002.

[49] P. Caputo, G. Menz, and P. Tetali, "Approximate tensorization of entropy at high temperature," *Annales de la Faculté des Sciences de Toulouse: Mathématiques*, vol. 24, no. 4, pp. 691–716, 2015.

[50] G. Fumera and F. Roli, "A theoretical and experimental analysis of linear combiners for multiple classifier systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 942–956, 2005.

[51] T. Heskes, "Selecting weighting factors in logarithmic opinion pools," in *Advances in Neural Information Processing Systems*, vol. 10, 1997.

[52] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge University Press, 2012.

[53] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.

[54] T. C. Aysal, M. J. Coates, and M. G. Rabbat, "Distributed average consensus using probabilistic quantization," in *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, Madison, WI, USA, Aug. 2007, pp. 640–644.

[55] G. Fumera, F. Roli, and A. Serrau, "A theoretical analysis of bagging as a linear combination of classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1293–1299, 2008.

[56] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[57] Y. Hu, D. Niu, J. Yang, and S. Zhou, "FDML: A collaborative machine learning framework for distributed features," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2019, pp. 2232–2240.

[58] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[60] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1912–1920.

[61] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, pp. 1321–1330.

Ping Hu received her Ph.D. in electrical engineering from EPFL, Switzerland in 2026. She is currently a postdoctoral researcher with the Adaptive Systems Laboratory, EPFL. Her research interests include distributed learning, inference and decision-making over networked systems.

Mert Kayaalp received his Ph.D. in computer and communication sciences from EPFL, Switzerland in 2025. He is currently a research scientist at Dalle Molle Institute for Artificial Intelligence (IDSIA USI-SUPSI), Switzerland, where he is affiliated with the UBS-IDSIA AI Lab. His research focuses on machine learning and signal analysis in networked and dynamical systems.

Ali H. Sayed is professor and former dean of engineering at EPFL, Switzerland, where he also directs the Adaptive Systems Laboratory. He served before as a distinguished professor and chair of electrical engineering with UCLA. He is a member of the US National Academy of Engineering (NAE) and The World Academy of Sciences (TWAS). He served as a president of the IEEE Signal Processing Society in 2018 and 2019. An author of over 650 scholarly publications and 10 books, his research involves several areas including adaptation and learning theories, statistical inference, and multi-agent systems. His work has been recognized with several awards including the 2022 IEEE Fourier Technical Field Award and the 2020 IEEE Wiener Society Award. He is also a fellow of EURASIP and the American Association for the Advancement of Science (AAAS).

Supplementary Material for "Test-Time Collaborative Classification over Multi-Agent Networks"

Ping Hu, Mert Kayaalp, and Ali H. Sayed

Organization of the Supplementary Material

This supplementary material provides detailed proofs, auxiliary results, and additional experimental information supporting the main paper. The theoretical appendices follow the order of the main development. Appendix A presents the network margin and training analysis and collects the auxiliary quantities used in the Theorems and subsequent proofs. Appendices B–D establish the sufficient-communication, finite-round, and finite-precision classification guarantees, respectively. Appendix E provides the Probably Approximately Correct (PAC)-style generalization analysis, while Appendix F develops the relation to social learning. Appendix G collects the corresponding raw score results when the empirical centering step (9) is omitted. Finally, Appendix H presents additional experimental details. Cross-references to theorem, equation, section, and figure numbers in the main paper are retained throughout.

Appendix A
Network Margin and Training Guarantee

This appendix provides the technical details supporting the network margin and training guarantees in Section III-A of the main paper. We first collect the complexity and margin quantities used in the finite-sample analysis. We then prove the effect of additive centering and the δ-margin training guarantee under approximate optimization. The final subsection compares the resulting bound with homogeneous centralized training.

1. Definitions of Auxiliary Quantities

This subsection collects the definitions and relevant properties of several auxiliary quantities used throughout the theoretical analysis.

a) Rademacher complexity: We begin with introducing the Rademacher complexity of a general function class $\mathcal{G}$. Let $\mathcal{Z} \triangleq \{z_1, z_2, \ldots, z_m\}$ be a fixed sample set of size m. The empirical Rademacher complexity of $\mathcal{G}$ with respect to $\mathcal{Z}$ is defined as

$$\widehat{\mathfrak{R}}_{\mathcal{Z}}(\mathcal{G}) \triangleq \mathbb{E}_{\boldsymbol{r}} \sup_{g\in\mathcal{G}} \left| \frac{1}{m} \sum_{n=1}^{m} \boldsymbol{r}_n g(z_n) \right| \tag{A.1}$$

where $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m$ are independent Rademacher random variables. For an i.i.d. random sample set $\mathcal{Z}'$ of size m, the Rademacher complexity of $\mathcal{G}$ is

$$\mathfrak{R}_m(\mathcal{G}) \triangleq \mathbb{E}_{\mathcal{Z}'} \widehat{\mathfrak{R}}_{\mathcal{Z}'}(\mathcal{G}). \tag{A.2}$$

For the local function class $\mathcal{F}_k$ and training set size N_k, we define the individual Rademacher complexity by

$$\rho_k \triangleq \mathfrak{R}_{N_k}(\mathcal{F}_k), \tag{A.3}$$

and the corresponding network Rademacher complexity by the weighted average involving the Perron vector π:

$$\rho \triangleq \sum_{k=1}^{K} \pi_k \rho_k. \tag{A.4}$$

Here, $\mathfrak{R}_{N_k}(\mathcal{F}_k)$ is evaluated for N_k i.i.d. local feature samples drawn from the feature marginal induced by Q_k.

b) Sample-size factor: We define two quantities related to the size of local training sets:

$$N_{\max} \triangleq \max_k N_k, \quad \alpha \triangleq \sum_{k=1}^{K} \pi_k \frac{N_{\max}}{N_k}. \tag{A.5}$$

The factor α captures the effect of unequal local training set sizes. Since $N_{\max}/N_k \geq 1$ and $\sum_k \pi_k = 1$, we have $\alpha \geq 1$. In particular, $\alpha = 1$ when all local training sets have the same size.

c) Function $\mathcal{E}_\Phi(r,\delta)$: Following Eqs. (A.16), (A.18), and (A.20) of [1], we use the same construction with a generic risk level $r < \Phi(0)$. For any prescribed margin $\delta \geq 0$, let $d_\delta^\star(r)$ denote the unique nonnegative solution of

$$d - \delta - \frac{\Phi(d) - r}{2L_\Phi} = 0. \tag{A.6}$$

The corresponding auxiliary function is

$$\mathcal{E}_\Phi(r,\delta) \triangleq \frac{d_\delta^\star(r) - \delta}{4} = \frac{\Phi(d_\delta^\star(r)) - r}{8L_\Phi}. \tag{A.7}$$

For $r = \mathsf{R}^o$, this reduces to the quantity $\mathcal{E}_\Phi(\mathsf{R}^o,\delta)$ introduced in [1]. An important property of the quantity $d_\delta^\star(r)$ is that, for fixed r, it increases with the prescribed margin δ. Since Φ is non-increasing under Assumption 1, $\mathcal{E}_\Phi(r,\delta)$ decreases as δ increases. For a fixed feasible δ, the same construction shows that $\mathcal{E}_\Phi(r,\delta)$ decreases as the risk level r increases.

d) Maximum feasible margin $\delta_{\max}(r)$: According to the definition in Eq. (A.21) of [1], we have

$$d_r \triangleq \inf\{x \geq 0 : \Phi(x) \leq r\}, \tag{A.8}$$

with $d_r = +\infty$ if the set is empty. The quantity $\delta_{\max}(r)$ is defined as the largest prescribed margin for which the solution $d_\delta^\star(r)$ remains below d_r, namely,

$$\delta_{\max}(r) \triangleq \sup\{\delta \geq 0 : d_\delta^\star(r) < d_r\}. \tag{A.9}$$

From (A.7), $d_\delta^\star(r) < d_r$ is precisely the condition that ensures $\mathcal{E}_\Phi(r,\delta) > 0$. Therefore, $0 \leq \delta < \delta_{\max}(r)$ specifies the feasible range of prescribed margins. Since d_r is non-increasing in r under Assumption 1, this feasible range becomes more restrictive as the risk level r increases. In Proposition 2, the generic risk level is $r = \mathsf{R}^o + \varepsilon_{\mathsf{opt}}$.

2. Proof of Proposition 1

Proof. Since $\boldsymbol{\mu}^{+}_{\mathrm{raw}} = \mu_{\mathrm{mid}} + D$ and $\boldsymbol{\mu}^{-}_{\mathrm{raw}} = \mu_{\mathrm{mid}} - D$, the two shifted expected margins in (25) are

$$D + (\mu_{\mathrm{mid}} - \chi_\pi) \quad \text{and} \quad D - (\mu_{\mathrm{mid}} - \chi_\pi). \tag{A.10}$$

Their minimum is $D - |\mu_{\mathrm{mid}} - \chi_\pi|$, which proves (25). Since the absolute value term is minimized at zero, $\Delta_\chi(f)$ is maximized when $\chi_\pi = \mu_{\mathrm{mid}}$. Setting respectively $\chi_\pi = 0$ and $\chi_\pi = \boldsymbol{\mu}_{\mathsf{emp}}(f)$ gives the two identities in (26), and comparing them yields the condition (28). Finally, $\boldsymbol{\mu}_{k,\mathsf{emp}}(f'_k) = \boldsymbol{\mu}_{k,\mathsf{emp}}(f_k) + o_k$, and hence (29) follows directly. Therefore, the centered local scores are identical before and after the additive offsets, and substituting them into (11) gives the same DeGroot iterates at every communication round. □

Remarks on centering. The achieved two-sided expected margin $\Delta_\chi(f)$ should be distinguished from the prescribed target $\delta \geq 0$ used in the subsequent guarantees. In particular, $\Delta_\chi(f)$ may be negative. An additive shift changes only the location of the two class means relative to threshold zero and leaves their separation $2D(f)$ unchanged. Hence, if $D(f) \leq 0$, no additive center can produce a positive two-sided expected margin. The proposition concerns this expected margin criterion and does not imply that centering universally improves classification accuracy. For example, if the score is already the exact equal-prior log-likelihood ratio, zero is the Bayes threshold and a nonzero shift can worsen the resulting decision rule.

3. Proof of Proposition 2

First, we recall the δ-margin consistent training event associated with condition (34):

$$\mathcal{C}_\delta \triangleq \left\{ \boldsymbol{\mu}^{+}(\widehat{\boldsymbol{f}}) > \delta,\, \boldsymbol{\mu}^{-}(\widehat{\boldsymbol{f}}) < -\delta \right\}, \tag{A.11}$$

and $P_{c,\delta} \triangleq \mathbb{P}(\mathcal{C}_\delta)$. The proof extends the exact-ERM analysis of [1], which in turn uses uniform deviation bounds established in [2]. We retain the uniform concentration and margin-to-risk arguments from that analysis, while accounting explicitly for the nonzero empirical optimization gaps of the classifiers returned by the local training algorithms from (6). The key new step is to show that these gaps replace the target risk R^o in the exact-ERM argument by the effective risk level $\mathsf{R}^o + \varepsilon_{\mathsf{opt}}$.

Proof. Uniform deviations. We introduce the empirical and population surrogate risks for the network as

$$\boldsymbol{R}_{\mathsf{emp}}(f) \triangleq \sum_{k=1}^{K} \pi_k \boldsymbol{R}_{k,\mathsf{emp}}(f_k), \tag{A.12}$$

$$\boldsymbol{R}(f) \triangleq \sum_{k=1}^{K} \pi_k \mathbb{E}\Phi(\boldsymbol{\gamma} f_k(\boldsymbol{h}_k)). \tag{A.13}$$

Since $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_K$ and the population risk separates across agents,

$$\inf_{f\in\mathcal{F}} \boldsymbol{R}(f) = \sum_{k=1}^{K} \pi_k \inf_{f_k \in \mathcal{F}_k} \mathbb{E}_{(\boldsymbol{h}_k,\boldsymbol{\gamma})\sim Q_k} \Phi(\boldsymbol{\gamma} f_k(\boldsymbol{h}_k)) = \mathsf{R}^o \tag{A.14}$$

where R^o is defined in (15). Let

$$U \triangleq \sup_{f\in\mathcal{F}} |\boldsymbol{R}_{\mathsf{emp}}(f) - \boldsymbol{R}(f)|, \tag{A.15}$$

$$V_\mu \triangleq \sup_{f\in\mathcal{F}} |\boldsymbol{\mu}_{\mathsf{emp}}(f) - \mu_{\mathrm{mid}}(f)|. \tag{A.16}$$

The following two uniform deviation bounds are the concentration results used in the exact-ERM analysis of [1], recalled there from Theorem 3 of [2]. For every $x > 4\rho$,

$$\mathbb{P}(V_\mu \geq x) \leq \exp\left\{ -\frac{N_{\max}(x - 4\rho)^2}{2\alpha^2\beta^2} \right\}, \tag{A.17}$$

and, for every $x > 4L_\Phi\rho$,

$$\mathbb{P}(U \geq x) \leq \exp\left\{ -\frac{N_{\max}(x - 4L_\Phi\rho)^2}{2\alpha^2\beta^2 L_\Phi^2} \right\}. \tag{A.18}$$

We use these established bounds directly.

Effect of approximate empirical optimization. From the definition in (7), condition (37) is equivalent to

$$\boldsymbol{R}_{k,\mathsf{emp}}(\widehat{\boldsymbol{f}}_k) \leq \inf_{f_k\in\mathcal{F}_k} \boldsymbol{R}_{k,\mathsf{emp}}(f_k) + \varepsilon_{k,\mathsf{opt}} \tag{A.19}$$

for every agent k. Multiplying by π_k and summing gives

$$\boldsymbol{R}_{\mathsf{emp}}(\widehat{\boldsymbol{f}}) \leq \sum_{k=1}^{K} \pi_k \inf_{f_k\in\mathcal{F}_k} \boldsymbol{R}_{k,\mathsf{emp}}(f_k) + \varepsilon_{\mathsf{opt}}. \tag{A.20}$$

Because the coordinates f_k can be selected independently over the Cartesian product class $\mathcal{F}$, the weighted sum of the local infima is the infimum of the network empirical risk. Hence,

$$\boldsymbol{R}_{\mathsf{emp}}(\widehat{\boldsymbol{f}}) \leq \inf_{f\in\mathcal{F}} \boldsymbol{R}_{\mathsf{emp}}(f) + \varepsilon_{\mathsf{opt}}. \tag{A.21}$$

The infimum defining R^o need not be attained. By the definition of U, for every $f \in \mathcal{F}$, $\boldsymbol{R}_{\mathsf{emp}}(f) \leq \boldsymbol{R}(f) + U$. Hence, using (A.21),

$$\begin{aligned} \boldsymbol{R}(\widehat{\boldsymbol{f}}) &\leq \boldsymbol{R}_{\mathsf{emp}}(\widehat{\boldsymbol{f}}) + U \\ &\leq \inf_{f\in\mathcal{F}} \boldsymbol{R}_{\mathsf{emp}}(f) + \varepsilon_{\mathsf{opt}} + U \\ &\leq \inf_{f\in\mathcal{F}} \{\boldsymbol{R}(f) + U\} + \varepsilon_{\mathsf{opt}} + U \\ &= \mathsf{R}^o + \varepsilon_{\mathsf{opt}} + 2U \end{aligned} \tag{A.22}$$

where the last equality follows because U does not depend on f and $\mathsf{R}^o = \inf_{f\in\mathcal{F}} \boldsymbol{R}(f)$. Therefore,

$$\boldsymbol{R}(\widehat{\boldsymbol{f}}) \leq \mathsf{R}^o + \varepsilon_{\mathsf{opt}} + 2U. \tag{A.23}$$

For convenience, define the effective risk level

$$r \triangleq \mathsf{R}^o + \varepsilon_{\mathsf{opt}}. \tag{A.24}$$

Thus,

$$\boldsymbol{R}(\widehat{\boldsymbol{f}}) \leq r + 2U. \tag{A.25}$$

This is the step that differs from the exact-ERM analysis, for which $r = \mathsf{R}^o$.

Reduction of margin failure to the deviation events. By Proposition 1,

$$\Delta_{\mathrm{cen}}(\widehat{\boldsymbol{f}}) = D(\widehat{\boldsymbol{f}}) - \left| \boldsymbol{\mu}_{\mathsf{emp}}(\widehat{\boldsymbol{f}}) - \mu_{\mathrm{mid}}(\widehat{\boldsymbol{f}}) \right|. \tag{A.26}$$

Fix any constant $d > \delta$. If $D(\widehat{\boldsymbol{f}}) > d$ and $V_\mu < d - \delta$ hold simultaneously, then

$$\Delta_{\mathrm{cen}}(\widehat{\boldsymbol{f}}) > d - (d - \delta) = \delta. \tag{A.27}$$

Consequently,

$$\overline{\mathcal{C}_\delta} \subseteq \big\{V_\mu \geq d - \delta\big\} \cup \big\{D(\widehat{\boldsymbol{f}}) \leq d\big\}. \tag{A.28}$$

We next use the same surrogate risk reduction as in the proof of the exact-ERM result in [1]. Under the uniform class prior, convexity of Φ and Jensen's inequality give

$$\begin{aligned}
\boldsymbol{R}(f) &= \frac{1}{2}\sum_{k=1}^{K} \pi_k \Big[\mathbb{E}^{(+1)}\Phi(f_k(\boldsymbol{h}_k)) + \mathbb{E}^{(-1)}\Phi(-f_k(\boldsymbol{h}_k))\Big] \\
&\geq \frac{1}{2}\Big[\Phi\big(\boldsymbol{\mu}^+_{\mathrm{raw}}(f)\big) + \Phi\big(-\boldsymbol{\mu}^-_{\mathrm{raw}}(f)\big)\Big] \\
&\geq \Phi\left(\frac{\boldsymbol{\mu}^+_{\mathrm{raw}}(f) - \boldsymbol{\mu}^-_{\mathrm{raw}}(f)}{2}\right) \\
&= \Phi(D(f)).
\end{aligned} \tag{A.29}$$

Thus, since Φ is non-increasing, $D(\widehat{\boldsymbol{f}}) \leq d$ implies $\boldsymbol{R}(\widehat{\boldsymbol{f}}) \geq \Phi(d)$. Combining this implication with (A.23) yields

$$\big\{D(\widehat{\boldsymbol{f}}) \leq d\big\} \subseteq \left\{U \geq \frac{\Phi(d) - r}{2}\right\}. \tag{A.30}$$

Using (A.28) and (A.30), followed by the established deviation bounds in (A.17)–(A.18), gives

$$\begin{aligned}
\mathbb{P}(\overline{\mathcal{C}_\delta}) \leq \exp\left\{-\frac{N_{\max}(d - \delta - 4\rho)^2}{2\alpha^2\beta^2}\right\} \\
+ \exp\left\{-\frac{N_{\max}}{2\alpha^2\beta^2 L_\Phi^2}\left(\frac{\Phi(d) - r}{2} - 4L_\Phi\rho\right)^2\right\},
\end{aligned} \tag{A.31}$$

for every $d > \delta$ satisfying $d - \delta > 4\rho$ and $(\Phi(d) - r)/2 > 4L_\Phi\rho$.

Balancing the two deviations. We now use the same balancing construction as in [1], with the exact target risk R^o replaced by the effective risk r defined in (A.24). Specifically, choose $d = d^\star_\delta(r)$ from Appendix A.1, which satisfies

$$d - \delta = \frac{\Phi(d) - r}{2L_\Phi}. \tag{A.32}$$

Under the conditions of Proposition 2, this choice is admissible. By (A.7), the two terms in (A.31) are then equal, and

$$\mathbb{P}(\overline{\mathcal{C}_\delta}) \leq 2\exp\left\{-\frac{8N_{\max}}{\alpha^2\beta^2}\Big(\mathcal{E}_\Phi(r,\delta) - \rho\Big)^2\right\}. \tag{A.33}$$

Since $P_{c,\delta} = 1 - \mathbb{P}(\overline{\mathcal{C}_\delta})$ and $r = \mathsf{R}^o + \varepsilon_{\mathsf{opt}}$, we obtain (39). When $\varepsilon_{\mathsf{opt}} = 0$, the effective risk reduces to $r = \mathsf{R}^o$, which recovers the exact-ERM result of [1]. □

Interpretation of the optimization-gap condition. Proposition 2 is conditional on the empirical optimality gap bounds in (37) for the score functions $\widehat{\boldsymbol{f}}_k$ returned by local training. It does not provide an algorithm-specific guarantee that a prescribed $\varepsilon_{k,\mathsf{opt}}$ is attained. In particular, when $\mathcal{F}_k$ is parameterized by a neural network, convexity of the scalar surrogate loss Φ does not imply convexity of the empirical training objective with respect to the network parameters. Any optimization guarantee for a particular training algorithm is therefore separate from the statistical statement of Proposition 2. Once such tolerances are available, their effect enters the bound through the quantity $\varepsilon_{\mathsf{opt}}$.

4. Homogeneous Centralized-Training Comparison

While the bound in Proposition 2 is derived for independent local training, where each agent trains its own classifier without exchanging model updates, it is instructive to compare this setup against a centralized benchmark. For clarity, we focus on the homogeneous case where all agents share the same feature space, model class, and local population distribution. In the heterogeneous case, different agents may observe different feature distributions or spaces and use distinct architectures, so a single centralized model is not directly comparable without introducing additional modeling choices.

For this comparison, suppose a total of N_{tot} i.i.d. training samples from the common local population is available. A centralized learner uses all N_{tot} samples to train one classifier, whereas independent training partitions the same sample budget across K agents. For a symmetric comparison, we consider an even split,

$$N_k = \frac{N_{\mathsf{tot}}}{K}, \quad k = 1, \ldots, K. \tag{A.34}$$

We next examine the quantities that enter the bound in (39).

(i) *Network complexity* ρ: For many standard base models, such as feedforward neural networks with norm-constrained weights and kernel methods over bounded balls in reproducing kernel Hilbert spaces, the Rademacher complexity admits the standard $O(1/\sqrt{m})$ dependence on the training set size m [2]–[4]. For the homogeneous setting with an even split, all local complexities are identical and therefore

$$\rho^{\mathsf{ind}} = \mathfrak{R}_{N_{\mathsf{tot}}/K}(\mathcal{F}), \quad \rho^{\mathsf{cen}} = \mathfrak{R}_{N_{\mathsf{tot}}}(\mathcal{F}) \tag{A.35}$$

where $\mathcal{F}$ denotes the common function class. Hence, for model classes whose Rademacher complexity decreases with sample size as above,

$$\rho^{\mathsf{cen}} \leq \rho^{\mathsf{ind}}. \tag{A.36}$$

(ii) *Function* $\mathcal{E}_\Phi(\mathsf{R}^o + \varepsilon_{\mathsf{opt}}, \delta)$: Because the local population distribution and model class are common, the individual target risks are identical, and hence the target risk R^o is the same for centralized and independent training. The loss function Φ is also unchanged. Let $\varepsilon^{\mathsf{cen}}_{\mathsf{opt}}$ denote the empirical optimization tolerance of the centralized learner and let

$$\varepsilon^{\mathsf{ind}}_{\mathsf{opt}} \triangleq \sum_{k=1}^{K} \pi_k \varepsilon_{k,\mathsf{opt}} \tag{A.37}$$

denote the corresponding π-weighted tolerance for independent training. If the two procedures are compared at matched optimization accuracy,

$$\varepsilon^{\mathsf{cen}}_{\mathsf{opt}} = \varepsilon^{\mathsf{ind}}_{\mathsf{opt}} \triangleq \varepsilon, \tag{A.38}$$

then $\mathcal{E}_\Phi(\mathsf{R}^o + \varepsilon, \delta)$ is identical in the two bounds. If their optimization tolerances differ, the corresponding effective risk

levels must instead be retained separately. The theory does not presume that either training procedure solves its empirical objective more accurately.

(iii) *Admissible margin range*: At matched optimization accuracy, the effective risk $\mathsf{R}^o + \varepsilon$ and the loss Φ are the same under the two setups. Consequently, $\delta_{\max}(\mathsf{R}^o + \varepsilon)$, and hence the admissible range of δ, is also identical.

The common model class also yields the same score bound β. Suppose now that δ lies in the common admissible range and that

$$\rho^{\mathsf{ind}} < \mathcal{E}_\Phi(\mathsf{R}^o + \varepsilon, \delta). \tag{A.39}$$

Under the dependence of training set size described above, $\rho^{\mathsf{cen}} \le \rho^{\mathsf{ind}}$, so Proposition 2 applies to both training setups. For centralized training, taking $K = 1$, $N_{\max} = N_{\mathsf{tot}}$, and $\alpha = 1$ gives

$$P^{\mathsf{cen}}_{c,\delta} \ge 1 - 2\exp\left\{-\frac{8N_{\mathsf{tot}}}{\beta^2}\Big(\mathcal{E}_\Phi(\mathsf{R}^o+\varepsilon,\delta) - \rho^{\mathsf{cen}}\Big)^2\right\}. \tag{A.40}$$

For independent training with the even split, $N_{\max} = N_{\mathsf{tot}}/K$ and $\alpha = 1$, which gives

$$P^{\mathsf{ind}}_{c,\delta} \ge 1 - 2\exp\left\{-\frac{8N_{\mathsf{tot}}}{K\beta^2}\Big(\mathcal{E}_\Phi(\mathsf{R}^o+\varepsilon,\delta) - \rho^{\mathsf{ind}}\Big)^2\right\}. \tag{A.41}$$

Therefore, under this homogeneous comparison with matched optimization accuracy, the centralized benchmark benefits from both a larger effective size of training set in the exponent and, for the model classes considered above, a Rademacher complexity that is no larger. Consequently, it yields a tighter lower bound on $P_{c,\delta}$ than that obtained by evenly splitting the same sample budget among independently trained models. This comparison makes explicit a statistical cost associated with independent training.

APPENDIX B
SUFFICIENT COMMUNICATION

This appendix provides the prediction argument underlying the sufficient-communication guarantee in Section III-B of the main paper. We first record the bounded-difference concentration inequality obtained by combining the class-conditional approximate tensorization of entropy (ATE) assumption with the entropy method. We then apply this inequality under each class-conditional testing law P_γ and combine the resulting prediction bound with the training guarantee from Proposition 2. The same ATE concentration tool will also be used in the subsequent prediction analyses of other appendices.

We recall the σ-field generated by the training data and local training randomness in (40):

$$\mathcal{T} \triangleq \sigma(\mathcal{D}_1, \dots, \mathcal{D}_K, \mathcal{S}_1, \dots, \mathcal{S}_K). \tag{B.1}$$

1. Bounded-Difference Inequality under ATE

We first record the bounded-difference concentration consequence of ATE used throughout the prediction analysis. It follows by applying the standard entropy-method bounded-difference argument of [5] with the approximate tensorization inequality of [6] in place of exact tensorization. We use this established consequence without reproving it.

Lemma B.1 (Bounded differences under ATE). *Let $X = (X_1, \dots, X_K)$ have law P satisfying ATE with coefficient τ, in the sense of* (44)*. Suppose a bounded measurable function F satisfies*

$$|F(x) - F(x')| \le d_k \tag{B.2}$$

whenever x and x' differ only in coordinate k. Then, for every $\theta \in \mathbb{R}$,

$$\log \mathbb{E}\exp\{\theta(F(X) - \mathbb{E}F(X))\} \le \frac{\tau\theta^2}{8}\sum_{k=1}^K d_k^2. \tag{B.3}$$

Consequently, if $\sum_{k=1}^K d_k^2 > 0$, then, for every $u \ge 0$,

$$\mathbb{P}\big(F(X) - \mathbb{E}F(X) \le -u\big) \le \exp\left\{-\frac{2u^2}{\tau\sum_{k=1}^K d_k^2}\right\}. \tag{B.4}$$

The same bound holds for the upper tail. If $\sum_{k=1}^K d_k^2 = 0$, then F is constant and the concentration statement is trivial.

2. Proof of Theorem 1

Fix a realization of the training phase that belongs to $\mathcal{C}_\delta$. Conditional on $\mathcal{T}$, the learned score functions and empirical training means are fixed, whereas the fresh testing pair remains independent of the training phase. Conditional on $\boldsymbol{\gamma}^* = \gamma$, the joint testing observation $\boldsymbol{h}^*$ therefore has law P_γ. We bound the prediction error separately under the two class-conditional laws and then average over the uniform class prior. To lighten notation, throughout this proof we write $\widehat{f}_k$, c_k, and c_{ave} for the corresponding realized training-dependent quantities; the same convention will be used in the later prediction proofs.

Proof. Suppose first that $\boldsymbol{\gamma}^* = +1$. From the definition of the aggregate classifier c_{ave} in (18),

$$\begin{aligned}\mathbb{E}^{(+1)}_{\boldsymbol{h}^*} c_{\mathsf{ave}}(\boldsymbol{h}^*) &= \sum_{k=1}^K \pi_k \mathbb{E}^{(+1)}_{\boldsymbol{h}^*_k} c_k(\boldsymbol{h}^*_k) \\ &= \mu^+(\widehat{f}) > \delta\end{aligned} \tag{B.5}$$

where the last inequality follows because the fixed training realization belongs to $\mathcal{C}_\delta$.

Now consider another observation collection $\widehat{\boldsymbol{h}}$ that differs from $\boldsymbol{h}^*$ only in coordinate k. From (9) and Assumption 2,

$$\begin{aligned}\big|c_{\mathsf{ave}}(\boldsymbol{h}^*) - c_{\mathsf{ave}}(\widehat{\boldsymbol{h}})\big| &= \pi_k\big|c_k(\boldsymbol{h}^*_k) - c_k(\widehat{\boldsymbol{h}}_k)\big| \\ &= \pi_k\big|\widehat{f}_k(\boldsymbol{h}^*_k) - \widehat{f}_k(\widehat{\boldsymbol{h}}_k)\big| \\ &\le 2\beta\pi_k.\end{aligned} \tag{B.6}$$

This reveals that the coordinate oscillations of c_{ave} are $d_k = 2\beta\pi_k$. Conditional on $\boldsymbol{\gamma}^* = +1$ and $\mathcal{T}$, the testing observation has law P_{+1}, which satisfies ATE with coefficient τ_{+1}. Hence, following (B.5), Lemma B.1 gives

$$\begin{aligned}&\mathbb{P}\Big(c_{\mathsf{ave}}(\boldsymbol{h}^*) \le 0 \,\Big|\, \boldsymbol{\gamma}^* = +1, \mathcal{T}\Big) \\ &= \mathbb{P}\Big(c_{\mathsf{ave}}(\boldsymbol{h}^*) - \mathbb{E}^{(+1)}_{\boldsymbol{h}^*} c_{\mathsf{ave}}(\boldsymbol{h}^*) \le -\mu^+(\widehat{f}) \,\Big|\, \boldsymbol{\gamma}^* = +1, \mathcal{T}\Big)\end{aligned}$$

$$\leq \exp\left\{-\frac{\delta^2}{2\tau_{+1}\beta^2\sum_{k=1}^K \pi_k^2}\right\}. \tag{B.7}$$

The final inequality follows from $\mu^+(\widehat{f}) > \delta$ on $\mathcal{C}_\delta$. For $\boldsymbol{\gamma}^* = -1$, we similarly have

$$\mathbb{E}^{(-1)}_{\boldsymbol{h}^*} c_{\mathsf{ave}}(\boldsymbol{h}^*) = \mu^-(\widehat{f}) < -\delta \tag{B.8}$$

on $\mathcal{C}_\delta$, and the same coordinate oscillations $d_k = 2\beta\pi_k$ apply. Accordingly, we have

$$\mathbb{P}\Big(c_{\mathsf{ave}}(\boldsymbol{h}^*) \geq 0 \,\Big|\, \boldsymbol{\gamma}^* = -1, \mathcal{T}\Big) \leq \exp\left\{-\frac{\delta^2}{2\tau_{-1}\beta^2\sum_{k=1}^K \pi_k^2}\right\}. \tag{B.9}$$

Since $\boldsymbol{\gamma}^*$ is independent of $\mathcal{T}$ and has the uniform prior, the following holds

$$\begin{aligned}
\mathbb{P}(\mathcal{M}|\mathcal{T}) &= \frac{1}{2}\mathbb{P}\Big(c_{\mathsf{ave}}(\boldsymbol{h}^*) \leq 0 \,\Big|\, \boldsymbol{\gamma}^* = +1, \mathcal{T}\Big) \\
&\quad + \frac{1}{2}\mathbb{P}\Big(c_{\mathsf{ave}}(\boldsymbol{h}^*) \geq 0 \,\Big|\, \boldsymbol{\gamma}^* = -1, \mathcal{T}\Big) \\
&\leq \frac{1}{2}\sum_{\gamma\in\Gamma} \exp\left\{-\frac{\delta^2}{2\tau_\gamma\beta^2\sum_{k=1}^K \pi_k^2}\right\} \\
&\leq \exp\left\{-\frac{\delta^2}{2\tau_{\max}\beta^2\sum_{k=1}^K \pi_k^2}\right\}
\end{aligned} \tag{B.10}$$

almost surely on $\mathcal{C}_\delta$. This proves the conditional bound (46) in Theorem 1. Since $\mathcal{C}_\delta$ is $\mathcal{T}$-measurable and the preceding conditional bound holds almost surely on $\mathcal{C}_\delta$, we have

$$\begin{aligned}
\mathbb{E}[\mathbb{1}[\mathcal{C}_\delta]\mathbb{P}(\mathcal{M}|\mathcal{T})] &\leq \exp\left\{-\frac{\delta^2}{2\tau_{\max}\beta^2\sum_{k=1}^K \pi_k^2}\right\}\mathbb{P}(\mathcal{C}_\delta) \\
&\leq \exp\left\{-\frac{\delta^2}{2\tau_{\max}\beta^2\sum_{k=1}^K \pi_k^2}\right\}.
\end{aligned} \tag{B.11}$$

Combining (B.11) with the decomposition in (42) gives

$$P_e \leq \exp\left\{-\frac{\delta^2}{2\tau_{\max}\beta^2\sum_{k=1}^K \pi_k^2}\right\} + \mathbb{P}(\overline{\mathcal{C}_\delta}). \tag{B.12}$$

If the assumptions of Proposition 2 also hold, substituting the bound on $\mathbb{P}(\overline{\mathcal{C}_\delta})$ from (39) into (B.12) gives

$$\begin{aligned}
P_e &\leq \exp\left\{-\frac{\delta^2}{2\tau_{\max}\beta^2\sum_{k=1}^K \pi_k^2}\right\} \\
&\quad + 2\exp\left\{-\frac{8N_{\max}}{\alpha^2\beta^2}\Big(\mathcal{E}_\Phi(\mathsf{R}^o + \varepsilon_{\mathsf{opt}}, \delta) - \rho\Big)^2\right\}.
\end{aligned} \tag{B.13}$$

This is the explicit classification error bound under sufficient communication. □

APPENDIX C
FINITE-ROUND COMMUNICATION

Appendix B establishes the classification guarantee for the limiting case of sufficient communication. At a finite communication round, a similar concentration argument applies. The proof below reuses the conditioning convention and ATE bounded-difference inequality from Appendix B

Proof. Fix a realization of the training phase that belongs to $\mathcal{C}_\delta$. As in Appendix B.2, the learned functions, their empirical training means, and therefore the centered score functions are fixed conditional on $\mathcal{T}$. We proceed first with the case $\boldsymbol{\gamma}^* = +1$. From (50), we have

$$\mathbb{P}\Big(\mathcal{M}_{k,t} \,\Big|\, \boldsymbol{\gamma}^* = +1, \mathcal{T}\Big) = \mathbb{P}\Big(\boldsymbol{\lambda}_{k,t}(\boldsymbol{h}^*) \leq 0 \,\Big|\, \boldsymbol{\gamma}^* = +1, \mathcal{T}\Big) \tag{C.1}$$

where the argument of $\boldsymbol{h}^*$ in the notation $\boldsymbol{\lambda}_{k,t}(\boldsymbol{h}^*)$ emphasizes the dependence on the testing observation collection. Consider another observation collection $\widehat{\boldsymbol{h}}$ that differs from $\boldsymbol{h}^*$ only in coordinate ℓ, (48) gives

$$\begin{aligned}
\big|\boldsymbol{\lambda}_{k,t}(\boldsymbol{h}^*) - \boldsymbol{\lambda}_{k,t}(\widehat{\boldsymbol{h}})\big| &= \Big|[A^t]_{\ell k}\big(c_\ell(\boldsymbol{h}^*_\ell) - c_\ell(\widehat{\boldsymbol{h}}_\ell)\big)\Big| \\
&\leq 2\beta[A^t]_{\ell k}.
\end{aligned} \tag{C.2}$$

Thus, the coordinate oscillation associated with the ℓth testing view is $d_\ell = 2\beta[A^t]_{\ell k}$.

We next control the deviation in the class-conditional mean caused by incomplete mixing. Since $|\widehat{f}_\ell| \leq \beta$ and its empirical training mean lies in $[-\beta, \beta]$, the realized centered class mean satisfies $|\mu^+_\ell(\widehat{f}_\ell)| \leq 2\beta$. Hence, using (52),

$$\begin{aligned}
&\Big|\mathbb{E}^{(+1)}_{\boldsymbol{h}^*}\boldsymbol{\lambda}_{k,t}(\boldsymbol{h}^*) - \mu^+(\widehat{f})\Big| \\
&= \left|\sum_{\ell=1}^K \big([A^t]_{\ell k} - \pi_\ell\big)\mu^+_\ell(\widehat{f}_\ell)\right| \\
&\leq 2\beta\sum_{\ell=1}^K \Big|[A^t]_{\ell k} - \pi_\ell\Big| \\
&\leq 2\beta KC(A,\sigma)\sigma^t.
\end{aligned} \tag{C.3}$$

Therefore, we have $\mu^+(\widehat{f}) > \delta$ when $\mathcal{C}_\delta$ is true. Consequently, for every t satisfying $r_t(\delta) > 0$, the following holds:

$$\begin{aligned}
\mathbb{E}^{(+1)}_{\boldsymbol{h}^*}\boldsymbol{\lambda}_{k,t}(\boldsymbol{h}^*) &> \delta - 2\beta KC(A,\sigma)\sigma^t \\
&= r_t(\delta) > 0.
\end{aligned} \tag{C.4}$$

Conditional on $\boldsymbol{\gamma}^* = +1$ and $\mathcal{T}$, the testing observation has law P_{+1}. Therefore, using the similar arguments used in (B.7) yields

$$\mathbb{P}\Big(\mathcal{M}_{k,t} \,\Big|\, \boldsymbol{\gamma}^* = +1, \mathcal{T}\Big) \leq \exp\left\{-\frac{r_t^2(\delta)}{2\tau_{+1}\beta^2\sum_{\ell=1}^K \big([A^t]_{\ell k}\big)^2}\right\}. \tag{C.5}$$

Repeating the steps in (C.3)–(C.5) for the case $\boldsymbol{\gamma}^* = -1$, we can show that almost surely on $\mathcal{C}_\delta$,

$$\mathbb{P}(\mathcal{M}_{k,t}|\mathcal{T}) \leq \exp\left\{-\frac{r_t^2(\delta)}{2\tau_{\max}\beta^2\sum_{\ell=1}^K \big([A^t]_{\ell k}\big)^2}\right\}, \tag{C.6}$$

which proves the bound in (54). If the conditions of Proposition 2 also hold, substituting the bound on $\mathbb{P}(\overline{\mathcal{C}_\delta})$ from (39) gives the following explicit upper bound for classification after t communication rounds:

$$P_{k,t} \leq \exp\left\{-\frac{r_t^2(\delta)}{2\tau_{\max}\beta^2\sum_{\ell=1}^K \big([A^t]_{\ell k}\big)^2}\right\}$$

$$+ 2\exp\left\{-\frac{8N_{\max}}{\alpha^2\beta^2}\Big(\mathcal{E}_\Phi(\mathsf{R}^o+\varepsilon_{\mathsf{opt}},\delta)-\rho\Big)^2\right\}. \quad \text{(C.7)}$$

This completes the proof of Theorem 2. □

APPENDIX D
FINITE-PRECISION COMMUNICATION

Appendix C establishes the finite-round classification guarantee under exact real-valued message exchange. Theorem 3 additionally accounts for stochastic finite-precision communication. Relative to the finite-round analysis, two new issues must be controlled: i) quantization errors accumulate as they propagate through the network, and ii) their distributions are input-dependent, so that the resulting errors need not be independent across rounds or from the fresh testing views. We first derive a conditional sub-Gaussian bound for the accumulated quantization perturbation by unrolling the recursion and conditioning on the quantization history. We then combine this bound with the finite-round concentration bound established in Appendix C.

Proof. For $t = 0$, no quantization has yet occurred: $x_{k,0}^{(b)} = c_k(\boldsymbol{h}_k^*) = \boldsymbol{\lambda}_{k,0}$ and $V_{k,0}^{\mathrm{q}} = 0$. The claimed conditional bound therefore reduces to the corresponding real-valued finite-round bound. We henceforth consider $t \geq 1$.

To isolate the effect of stochastic quantization, we condition on the training realization and the fresh testing example, so that only the rounding randomness remains. Let

$$\boldsymbol{x}_t^{(b)} \triangleq (x_{1,t}^{(b)},\dots,x_{K,t}^{(b)})^\top, \quad \boldsymbol{x}_t^{\mathrm{fp}} \triangleq (x_{1,t}^{\mathrm{fp}},\dots,x_{K,t}^{\mathrm{fp}})^\top \quad \text{(D.1)}$$

where $\boldsymbol{x}_t^{\mathrm{fp}}$ corresponds to the matched full-precision recursion obtained from the same initialization by replacing Q_b with the identity map. In particular,

$$x_{k,t}^{\mathrm{fp}} = \boldsymbol{\lambda}_{k,t} \quad \text{(D.2)}$$

where $\boldsymbol{\lambda}_{k,t}$ is defined in (48). Before analyzing the rounding errors, note that the prescribed quantization range is preserved by the recursion. From (56), $x_{k,0}^{(b)} = c_k(\boldsymbol{h}_k^*) \in [-2\beta, 2\beta]$. The outputs of the quantizer Q_b remain in the same interval, and (60) forms each subsequent decision statistic as a convex combination of the transmitted values. Hence, by induction,

$$x_{k,s}^{(b)} \in [-2\beta, 2\beta] \quad \text{for every agent } k \text{ and round } s. \quad \text{(D.3)}$$

This indicates that the input to the stochastic quantizer remains within its prescribed range at every round.

Define the sender-level rounding errors

$$\varepsilon_{\ell,s}^{\mathrm{q}} \triangleq q_{\ell,s}^{(b)} - x_{\ell,s}^{(b)}, \quad \boldsymbol{\varepsilon}_s^{\mathrm{q}} \triangleq (\varepsilon_{1,s}^{\mathrm{q}},\dots,\varepsilon_{K,s}^{\mathrm{q}})^\top. \quad \text{(D.4)}$$

Let $W = A^\top$. The finite-precision and matched full-precision recursions satisfy

$$\boldsymbol{x}_{s+1}^{(b)} = W\boldsymbol{x}_s^{(b)} + W\boldsymbol{\varepsilon}_s^{\mathrm{q}}, \quad \boldsymbol{x}_{s+1}^{\mathrm{fp}} = W\boldsymbol{x}_s^{\mathrm{fp}}, \quad \text{(D.5)}$$

with the same initialization. Expanding the recursion therefore gives

$$\boldsymbol{x}_t^{(b)} - \boldsymbol{x}_t^{\mathrm{fp}} = \sum_{s=0}^{t-1} W^{t-s}\boldsymbol{\varepsilon}_s^{\mathrm{q}}. \quad \text{(D.6)}$$

Therefore, a rounding error generated at round s is propagated through W^{t-s} before contributing to the decision statistic at round t.

Let $\mathcal{Q}_s$ denote the σ-field generated by $\mathcal{T}$, the fresh testing pair $(\boldsymbol{h}^*, \boldsymbol{\gamma}^*)$, and all quantizer randomness through round $s-1$. Conditional on $\mathcal{Q}_s$, the current decision statistics of each agent and therefore the active quantization bins are fixed. Following the stochastic rounding protocol (58), the sender-level rounding errors at round s are conditionally independent across senders, satisfy

$$\mathbb{E}\Big[\varepsilon_{\ell,s}^{\mathrm{q}} \,\Big|\, \mathcal{Q}_s\Big] = 0, \quad \text{(D.7)}$$

and each takes values in an interval of length at most Δ_b. Hence, for any $\mathcal{Q}_s$-measurable weights $w_1,\dots,w_K$, Hoeffding's lemma [7] gives, for every $\theta \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\theta\sum_{\ell=1}^K w_\ell\varepsilon_{\ell,s}^{\mathrm{q}}\right)\middle|\mathcal{Q}_s\right] \leq \exp\left\{\frac{\theta^2\Delta_b^2}{8}\sum_{\ell=1}^K w_\ell^2\right\}. \quad \text{(D.8)}$$

Given a fixed terminal pair (k,t), the kth coordinate of (D.6) can be written as

$$x_{k,t}^{(b)} - x_{k,t}^{\mathrm{fp}} = \sum_{s=0}^{t-1} U_s, \quad U_s \triangleq \sum_{j=1}^K [A^{t-s}]_{jk}\varepsilon_{j,s}^{\mathrm{q}}. \quad \text{(D.9)}$$

Applying (D.8) with $w_j = [A^{t-s}]_{jk}$ gives

$$\mathbb{E}\big[e^{\theta U_s} \,\big|\, \mathcal{Q}_s\big] \leq \exp\left\{\frac{\theta^2\Delta_b^2}{8}\sum_{j=1}^K([A^{t-s}]_{jk})^2\right\}. \quad \text{(D.10)}$$

It is worth noting that the errors at different rounds need not be independent because the quantization bin at a later round depends on the preceding randomized updates. For this reason, in the following, we accumulate the bounds through repeated conditioning rather than by multiplying unconditional moment generating functions (MGFs). According to the law of total expectation, conditioning first on $\mathcal{Q}_{t-1}$ gives

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\theta\sum_{s=0}^{t-1}U_s\right)\middle|\mathcal{T},\boldsymbol{h}^*,\boldsymbol{\gamma}^*\right] \\
&= \mathbb{E}\left[\exp\left(\theta\sum_{s=0}^{t-2}U_s\right)\mathbb{E}\big[e^{\theta U_{t-1}}\,\big|\,\mathcal{Q}_{t-1}\big]\middle|\mathcal{T},\boldsymbol{h}^*,\boldsymbol{\gamma}^*\right] \\
&\leq \exp\left\{\frac{\theta^2\Delta_b^2}{8}\sum_{j=1}^K([A]_{jk})^2\right\}\mathbb{E}\left[\exp\left(\theta\sum_{s=0}^{t-2}U_s\right)\middle|\mathcal{T},\boldsymbol{h}^*,\boldsymbol{\gamma}^*\right].
\end{aligned} \quad \text{(D.11)}$$

Repeating the same conditioning step for $U_{t-2},\dots,U_0$ yields

$$\begin{aligned}
&\mathbb{E}\Big[\exp\Big(\theta[x_{k,t}^{(b)} - x_{k,t}^{\mathrm{fp}}]\Big)\Big|\mathcal{T},\boldsymbol{h}^*,\boldsymbol{\gamma}^*\Big] \\
&\leq \exp\left\{\frac{\theta^2\Delta_b^2}{8}\sum_{r=1}^t\sum_{j=1}^K([A^r]_{jk})^2\right\} \\
&= \exp\left\{\frac{\theta^2 V_{k,t}^{\mathrm{q}}}{2}\right\}
\end{aligned} \quad \text{(D.12)}$$

where $V_{k,t}^{\mathrm{q}}$ is defined in (64). This derivation requires conditional independence only among the sender-level rounding

draws within each round; temporal independence of the rounding errors is not needed.

We now combine the accumulated quantization perturbation with the fluctuation induced by the fresh testing views. Fix $\gamma \in \Gamma$ and a training realization in $\mathcal{C}_\delta$, and define

$$\mu_{k,t}^{(\gamma)} \triangleq \mathbb{E}\Big[\gamma x_{k,t}^{\mathrm{fp}} \,\big|\, \mathcal{T}, \boldsymbol{\gamma}^* = \gamma\Big], \quad Y_{k,t}^{(\gamma)} \triangleq \gamma x_{k,t}^{\mathrm{fp}} - \mu_{k,t}^{(\gamma)}. \tag{D.13}$$

Since $x_{k,t}^{\mathrm{fp}} = \boldsymbol{\lambda}_{k,t}$, the finite-round mean argument leading to (C.4) in Appendix C, applied to either class, gives

$$\mu_{k,t}^{(\gamma)} \geq r_t(\delta). \tag{D.14}$$

For fixed $\mathcal{T}$ and $\boldsymbol{\gamma}^* = \gamma$, changing only the ℓth testing view changes $\gamma x_{k,t}^{\mathrm{fp}}$ by at most $2\beta[A^t]_{\ell k}$. Applying the MGF form of Lemma B.1, namely (B.3), under P_γ therefore gives

$$\mathbb{E}\Big[e^{\theta Y_{k,t}^{(\gamma)}} \,\big|\, \mathcal{T}, \boldsymbol{\gamma}^* = \gamma\Big] \leq \exp\left\{\frac{\theta^2}{2}\tau_\gamma \beta^2 \sum_{\ell=1}^{K}([A^t]_{\ell k})^2\right\} \tag{D.15}$$

for any $\theta \in \mathbb{R}$. Define the signed quantization perturbation:

$$Z_{k,t}^{\mathrm{q},\gamma} \triangleq \gamma\big(x_{k,t}^{(b)} - x_{k,t}^{\mathrm{fp}}\big). \tag{D.16}$$

Although the quantizer uses fresh auxiliary randomness, the distribution of $Z_{k,t}^{\mathrm{q},\gamma}$ depends on the realized testing example through the input-dependent quantization bins. Thus, $Z_{k,t}^{\mathrm{q},\gamma}$ and $Y_{k,t}^{(\gamma)}$ are not assumed to be independent. Instead, we condition first on the complete fresh testing example. Since $Y_{k,t}^{(\gamma)}$ is then fixed, the tower property and (D.12) give

$$\begin{aligned}
&\mathbb{E}\Big[e^{\theta(Y_{k,t}^{(\gamma)} + Z_{k,t}^{\mathrm{q},\gamma})} \,\big|\, \mathcal{T}, \boldsymbol{\gamma}^* = \gamma\Big] \\
&= \mathbb{E}\Big[e^{\theta Y_{k,t}^{(\gamma)}} \mathbb{E}\Big[e^{\theta Z_{k,t}^{\mathrm{q},\gamma}} \,\big|\, \mathcal{T}, \boldsymbol{h}^*, \boldsymbol{\gamma}^* = \gamma\Big]\Big|\mathcal{T}, \boldsymbol{\gamma}^* = \gamma\Big] \\
&\leq \exp\left\{\frac{\theta^2 V_{k,t}^{\mathrm{q}}}{2}\right\} \mathbb{E}\Big[e^{\theta Y_{k,t}^{(\gamma)}} \,\big|\, \mathcal{T}, \boldsymbol{\gamma}^* = \gamma\Big] \\
&\leq \exp\left\{\frac{\theta^2}{2}\left[\tau_\gamma \beta^2 \sum_{\ell=1}^{K}([A^t]_{\ell k})^2 + V_{k,t}^{\mathrm{q}}\right]\right\}.
\end{aligned} \tag{D.17}$$

Therefore, the sub-Gaussian variance proxies associated with the testing views and quantization errors add in the conditional MGF bound. In particular, no unconditional independence between the two fluctuations is required. Suppose $r_t(\delta) > 0$. Since

$$\gamma x_{k,t}^{(b)} = \mu_{k,t}^{(\gamma)} + Y_{k,t}^{(\gamma)} + Z_{k,t}^{\mathrm{q},\gamma}, \tag{D.18}$$

the event $\mathcal{M}_{k,t}^{(b)} = \{\boldsymbol{\gamma}^* x_{k,t}^{(b)} \leq 0\}$ defined in (62), conditional on $\boldsymbol{\gamma}^* = \gamma$, together with (D.14), implies

$$Y_{k,t}^{(\gamma)} + Z_{k,t}^{\mathrm{q},\gamma} \leq -r_t(\delta). \tag{D.19}$$

Let

$$V_{k,t}^{(\gamma)} \triangleq \tau_\gamma \beta^2 \sum_{\ell=1}^{K}([A^t]_{\ell k})^2 + V_{k,t}^{\mathrm{q}}. \tag{D.20}$$

For any $\lambda > 0$, Chernoff's method and (D.17) give

$$\begin{aligned}
&\mathbb{P}\Big(Y_{k,t}^{(\gamma)} + Z_{k,t}^{\mathrm{q},\gamma} \leq -r_t(\delta) \,\big|\, \mathcal{T}, \boldsymbol{\gamma}^* = \gamma\Big) \\
&\leq \exp\Big\{-\lambda r_t(\delta) + \frac{\lambda^2}{2} V_{k,t}^{(\gamma)}\Big\}.
\end{aligned} \tag{D.21}$$

Optimizing at $\lambda = r_t(\delta)/V_{k,t}^{(\gamma)}$ yields

$$\begin{aligned}
&\mathbb{P}(\mathcal{M}_{k,t}^{(b)} \,|\, \mathcal{T}, \boldsymbol{\gamma}^* = \gamma) \\
&\leq \exp\left\{-\frac{r_t^2(\delta)}{2\left[\tau_\gamma \beta^2 \sum_{\ell=1}^{K}([A^t]_{\ell k})^2 + V_{k,t}^{\mathrm{q}}\right]}\right\}.
\end{aligned} \tag{D.22}$$

Since the fresh label is independent of $\mathcal{T}$ and has the uniform prior, averaging the two class-conditional bounds gives, almost surely on $\mathcal{C}_\delta$,

$$\begin{aligned}
\mathbb{P}(\mathcal{M}_{k,t}^{(b)} \,|\, \mathcal{T}) &\leq \frac{1}{2}\sum_{\gamma \in \Gamma} \exp\left\{-\frac{r_t^2(\delta)}{2\left[\tau_\gamma \beta^2 \sum_{\ell=1}^{K}([A^t]_{\ell k})^2 + V_{k,t}^{\mathrm{q}}\right]}\right\} \\
&\leq \exp\left\{-\frac{r_t^2(\delta)}{2\left[\tau_{\max} \beta^2 \sum_{\ell=1}^{K}([A^t]_{\ell k})^2 + V_{k,t}^{\mathrm{q}}\right]}\right\},
\end{aligned} \tag{D.23}$$

which proves (65). Using the decomposition in (42) with $\mathcal{M}$ replaced by $\mathcal{M}_{k,t}^{(b)}$, if the conditions of Proposition 2 also hold, substituting its bound on $\mathbb{P}(\overline{\mathcal{C}_\delta})$ gives

$$\begin{aligned}
P_{k,t}^{(b)} \leq\ & \exp\left\{-\frac{r_t^2(\delta)}{2\left[\tau_{\max} \beta^2 \sum_{\ell=1}^{K}([A^t]_{\ell k})^2 + V_{k,t}^{\mathrm{q}}\right]}\right\} \\
&+ 2\exp\left\{-\frac{8N_{\max}}{\alpha^2\beta^2}\Big(\mathcal{E}_\Phi(\mathsf{R}^o + \varepsilon_{\mathsf{opt}}, \delta) - \rho\Big)^2\right\}.
\end{aligned} \tag{D.24}$$

This is the explicit unconditional finite-precision classification bound referred to in Theorem 3 and completes the proof. □

Appendix E
PAC-Style Generalization Guarantee

This appendix proves Theorem 4 of the main paper. Unlike the prediction guarantees in Appendices B–D, the PAC-style result does not rely on the expected margin event $\mathcal{C}_\delta$ or the class-conditional ATE assumption. Its sampling requirement is instead that the network examples $(\boldsymbol{h}_n, \boldsymbol{\gamma}_n)$ are i.i.d. across n, while the K local views within one example may remain statistically dependent.

The main technical issue is the empirical centering operation in (9), which leads the resulting classifier class to be sample-dependent. The proof handles this issue in three steps. We first bound the sensitivity of the uniform population–empirical risk gap to replacing one network example. We then embed every realized centered aggregate in a fixed offset-augmented function class and control its expected uniform deviation by Rademacher complexity. Finally, McDiarmid's inequality yields a high-probability bound, which is specialized to the η-margin loss Φ_η in Section III-E.

Proof. We prove the collaborative bound in (77), with the local bound following from the same argument by taking a single agent. Recall that

$$\mathcal{D} = \{(\boldsymbol{h}_n, \boldsymbol{\gamma}_n)\}_{n=1}^{N} \tag{E.1}$$

contains N i.i.d. complete network examples drawn from Q, and that $\mathcal{D}_k = \{(\boldsymbol{h}_{k,n}, \boldsymbol{\gamma}_n)\}_{n=1}^N$ is its projection at agent k. The proof treats each complete network example as one independent sample coordinate and therefore never requires independence among the local views within an example.

For $f_k \in \mathcal{F}_k$, we define its empirical training mean on $\mathcal{D}_k$ and centered score by

$$\bar{f}_k(\mathcal{D}_k) \triangleq \frac{1}{N}\sum_{n=1}^{N} f_k(\boldsymbol{h}_{k,n}), \tag{E.2}$$

$$\boldsymbol{c}_{f_k,\mathcal{D}_k}(h_k) \triangleq f_k(h_k) - \bar{f}_k(\mathcal{D}_k). \tag{E.3}$$

For $f = (f_1, \ldots, f_K) \in \mathcal{F}$, let

$$\boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}(h) \triangleq \sum_{k=1}^{K} \pi_k \boldsymbol{c}_{f_k,\mathcal{D}_k}(h_k), \tag{E.4}$$

and, for a generic L_Φ-Lipschitz loss Φ, define

$$\boldsymbol{R}^{\mathcal{D}}_{\mathsf{emp}}(f) \triangleq \frac{1}{N}\sum_{n=1}^{N} \Phi\big(\boldsymbol{\gamma}_n \boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}(\boldsymbol{h}_n)\big), \tag{E.5}$$

$$\boldsymbol{R}^{\mathcal{D}}(f) \triangleq \mathbb{E}_{(\boldsymbol{h},\boldsymbol{\gamma})\sim Q} \Phi\big(\boldsymbol{\gamma} \boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}(\boldsymbol{h})\big). \tag{E.6}$$

The superscript $\mathcal{D}$ emphasizes that, even for fixed f, the deployed centered aggregate score associated with $\boldsymbol{c}_{\mathsf{ave}}$ depends on the training sample. We therefore study

$$\Omega(\mathcal{D}) \triangleq \sup_{f\in\mathcal{F}} \big|\boldsymbol{R}^{\mathcal{D}}(f) - \boldsymbol{R}^{\mathcal{D}}_{\mathsf{emp}}(f)\big|. \tag{E.7}$$

a) Bounded difference of $\Omega(\mathcal{D})$: Let $\mathcal{D}'$ be obtained from $\mathcal{D}$ by replacing only the j-th complete network example $(\boldsymbol{h}_j, \boldsymbol{\gamma}_j)$ by $(\boldsymbol{h}'_j, \boldsymbol{\gamma}'_j)$. Using

$$\Big|\sup_f |a_f| - \sup_f |b_f|\Big| \le \sup_f |a_f - b_f| \tag{E.8}$$

and the triangle inequality, we obtain

$$\begin{aligned}\big|\Omega(\mathcal{D}) - \Omega(\mathcal{D}')\big| \le{}& \sup_{f\in\mathcal{F}} \big|\boldsymbol{R}^{\mathcal{D}}(f) - \boldsymbol{R}^{\mathcal{D}'}(f)\big| \\ &+ \sup_{f\in\mathcal{F}} \big|\boldsymbol{R}^{\mathcal{D}}_{\mathsf{emp}}(f) - \boldsymbol{R}^{\mathcal{D}'}_{\mathsf{emp}}(f)\big|.\end{aligned} \tag{E.9}$$

Since only one observation changes in each projection $\mathcal{D}_k$ and $|f_k| \le \beta$ by Assumption 2, one gets from (E.2) that

$$\big|\bar{f}_k(\mathcal{D}_k) - \bar{f}_k(\mathcal{D}'_k)\big| \le \frac{2\beta}{N}. \tag{E.10}$$

For any $f \in \mathcal{F}$ and any h, it follows that

$$\begin{aligned}\big|\boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}(h) - \boldsymbol{c}^{\mathcal{D}'}_{\mathsf{ave},f}(h)\big| &= \left|\sum_{k=1}^{K} \pi_k \big(\bar{f}_k(\mathcal{D}'_k) - \bar{f}_k(\mathcal{D}_k)\big)\right| \\ &\le \sum_{k=1}^{K} \pi_k \big|\bar{f}_k(\mathcal{D}'_k) - \bar{f}_k(\mathcal{D}_k)\big| \\ &\le \frac{2\beta}{N}\end{aligned} \tag{E.11}$$

where we used $\sum_{k=1}^K \pi_k = 1$. By the triangle inequality and the Lipschitz property of Φ,

$$\sup_{f\in\mathcal{F}} \big|\boldsymbol{R}^{\mathcal{D}}(f) - \boldsymbol{R}^{\mathcal{D}'}(f)\big| \le \frac{2\beta L_\Phi}{N}. \tag{E.12}$$

For the empirical risk, we distinguish the unchanged samples from the replaced one. For an unchanged coordinate $n \ne j$, we have $(\boldsymbol{h}_n, \boldsymbol{\gamma}_n) = (\boldsymbol{h}'_n, \boldsymbol{\gamma}'_n)$, and therefore

$$\begin{aligned}&\Big|\boldsymbol{\gamma}_n \boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}(\boldsymbol{h}_n) - \boldsymbol{\gamma}'_n \boldsymbol{c}^{\mathcal{D}'}_{\mathsf{ave},f}(\boldsymbol{h}'_n)\Big| \\ &= \Big|\boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}(\boldsymbol{h}_n) - \boldsymbol{c}^{\mathcal{D}'}_{\mathsf{ave},f}(\boldsymbol{h}_n)\Big| \\ &\le \frac{2\beta}{N}.\end{aligned} \tag{E.13}$$

At the replaced coordinate, $|f_k| \le \beta$ and $|\bar{f}_k(\mathcal{D}_k)| \le \beta$ imply

$$|\boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}(h)| \le 2\beta. \tag{E.14}$$

Hence, we have

$$\begin{aligned}&\Big|\boldsymbol{\gamma}_j \boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}(\boldsymbol{h}_j) - \boldsymbol{\gamma}'_j \boldsymbol{c}^{\mathcal{D}'}_{\mathsf{ave},f}(\boldsymbol{h}'_j)\Big| \\ &\le \big|\boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}(\boldsymbol{h}_j)\big| + \Big|\boldsymbol{c}^{\mathcal{D}'}_{\mathsf{ave},f}(\boldsymbol{h}'_j)\Big| \\ &\le 4\beta.\end{aligned} \tag{E.15}$$

Applying the Lipschitz property of Φ to these two cases yields

$$\begin{aligned}\big|\boldsymbol{R}^{\mathcal{D}}_{\mathsf{emp}}(f) - \boldsymbol{R}^{\mathcal{D}'}_{\mathsf{emp}}(f)\big| &\le \frac{L_\Phi}{N}\left[(N-1)\frac{2\beta}{N} + 4\beta\right] \\ &\le \frac{6\beta L_\Phi}{N},\end{aligned} \tag{E.16}$$

for every $f \in \mathcal{F}$. Combining the two terms in (E.12) and (E.16) gives

$$\big|\Omega(\mathcal{D}) - \Omega(\mathcal{D}')\big| \le \frac{8\beta L_\Phi}{N}. \tag{E.17}$$

This replacement argument uses only bounded local scores and independence across complete network examples. No factorization of the joint law Q across agents is required, which implies that the conditional independence assumption of the local views across agents is not required.

b) Expectation of $\Omega(\mathcal{D})$: A direct symmetrization of the function class consisting of $\boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}$, $\forall f \in \mathcal{F}$ would incorrectly treat a sample-dependent class as fixed. To address this issue, we instead embed every realized centered classifier $\boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}$ in a *deterministic* augmented class. Define

$$a_f(\mathcal{D}) \triangleq \sum_{k=1}^{K} \pi_k \bar{f}_k(\mathcal{D}_k). \tag{E.18}$$

Since $|a_f(\mathcal{D})| \le \beta$ and

$$\boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}(h) = f_{\mathsf{ave}}(h) - a_f(\mathcal{D}), \quad f_{\mathsf{ave}}(h) \triangleq \sum_{k=1}^{K} \pi_k f_k(h_k), \tag{E.19}$$

every realized centered classifier belongs to the fixed class

$$\mathcal{G}_{\mathrm{off}} \triangleq \Big\{h \mapsto f_{\mathsf{ave}}(h) - a : f \in \mathcal{F},\ a \in [-\beta, \beta]\Big\}. \tag{E.20}$$

Define the corresponding fixed-class uniform deviation

$$\Delta_{\mathrm{off}}(\mathcal{D}) \triangleq \sup_{g\in\mathcal{G}_{\mathrm{off}}} \left|\mathbb{E}\Phi(\boldsymbol{\gamma} g(\boldsymbol{h})) - \frac{1}{N}\sum_{n=1}^{N} \Phi(\boldsymbol{\gamma}_n g(\boldsymbol{h}_n))\right|. \tag{E.21}$$

Then

$$\Omega(\mathcal{D}) \le \Delta_{\mathrm{off}}(\mathcal{D}). \tag{E.22}$$

To apply the contraction inequality in its standard form, define

$$\widetilde{\Phi}(u) \triangleq \Phi(u) - \Phi(0). \tag{E.23}$$

Then $\widetilde{\Phi}(0) = 0$, $\widetilde{\Phi}$ remains L_Φ-Lipschitz, and replacing Φ by $\widetilde{\Phi}$ does not change any population–empirical risk difference. Let

$$\mathcal{D}^{\mathrm{g}} \triangleq \left\{(\boldsymbol{h}_n^{\mathrm{g}}, \boldsymbol{\gamma}_n^{\mathrm{g}})\right\}_{n=1}^N \tag{E.24}$$

be an independent ghost sample drawn from the same joint distribution Q. Since $\mathcal{G}_{\mathrm{off}}$ is independent of both $\mathcal{D}$ and $\mathcal{D}^{\mathrm{g}}$, Jensen's inequality gives

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}} \Delta_{\mathrm{off}}(\mathcal{D}) \\
&= \mathbb{E}_{\mathcal{D}} \sup_{g \in \mathcal{G}_{\mathrm{off}}} \left| \mathbb{E}_{\mathcal{D}^{\mathrm{g}}} \frac{1}{N} \sum_{n=1}^N \left[\widetilde{\Phi}(\boldsymbol{\gamma}_n^{\mathrm{g}} g(\boldsymbol{h}_n^{\mathrm{g}})) - \widetilde{\Phi}(\boldsymbol{\gamma}_n g(\boldsymbol{h}_n)) \right] \right| \\
&\le \mathbb{E}_{\mathcal{D}, \mathcal{D}^{\mathrm{g}}} \sup_{g \in \mathcal{G}_{\mathrm{off}}} \left| \frac{1}{N} \sum_{n=1}^N \left[\widetilde{\Phi}(\boldsymbol{\gamma}_n^{\mathrm{g}} g(\boldsymbol{h}_n^{\mathrm{g}})) - \widetilde{\Phi}(\boldsymbol{\gamma}_n g(\boldsymbol{h}_n)) \right] \right|.
\end{aligned} \tag{E.25}$$

Introducing N independent Rademacher variables $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N$ and using the usual symmetrization argument, we obtain

$$\begin{aligned}
\mathbb{E}\Omega(\mathcal{D}) &\le \mathbb{E}\Delta_{\mathrm{off}}(\mathcal{D}) \\
&\le 2\mathbb{E}_{\mathcal{D}, \boldsymbol{r}} \sup_{g \in \mathcal{G}_{\mathrm{off}}} \left| \frac{1}{N} \sum_{n=1}^N \boldsymbol{r}_n \widetilde{\Phi}(\boldsymbol{\gamma}_n g(\boldsymbol{h}_n)) \right| \\
&\le 4L_\Phi \mathbb{E}_{\mathcal{D}, \boldsymbol{r}} \sup_{g \in \mathcal{G}_{\mathrm{off}}} \left| \frac{1}{N} \sum_{n=1}^N \boldsymbol{r}_n \boldsymbol{\gamma}_n g(\boldsymbol{h}_n) \right|
\end{aligned} \tag{E.26}$$

where the last step follows from the contraction inequality for Rademacher averages [3], [7]. For $g(h) = f_{\mathsf{ave}}(h) - a \in \mathcal{G}_{\mathrm{off}}$, using the definition of f_{ave} in (19), we further have

$$\begin{aligned}
&\sup_{g \in \mathcal{G}_{\mathrm{off}}} \left| \frac{1}{N} \sum_{n=1}^N \boldsymbol{r}_n \boldsymbol{\gamma}_n g(\boldsymbol{h}_n) \right| \\
&\le \sum_{k=1}^K \pi_k \sup_{f_k \in \mathcal{F}_k} \left| \frac{1}{N} \sum_{n=1}^N \boldsymbol{r}_n \boldsymbol{\gamma}_n f_k(\boldsymbol{h}_{k,n}) \right| \\
&\quad + \beta \left| \frac{1}{N} \sum_{n=1}^N \boldsymbol{r}_n \boldsymbol{\gamma}_n \right|.
\end{aligned} \tag{E.27}$$

Conditional on the sample set $\mathcal{D}$, the variables $\boldsymbol{r}_n \boldsymbol{\gamma}_n$ are again independent Rademacher signs. Therefore, using the definition $\rho = \sum_k \pi_k \rho_k$ from Appendix A.1,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}, \boldsymbol{r}} \sup_{g \in \mathcal{G}_{\mathrm{off}}} \left| \frac{1}{N} \sum_{n=1}^N \boldsymbol{r}_n \boldsymbol{\gamma}_n g(\boldsymbol{h}_n) \right| \\
&\le \rho + \beta \mathbb{E}_{\boldsymbol{r}} \left| \frac{1}{N} \sum_{n=1}^N \boldsymbol{r}_n \right| \\
&\le \rho + \frac{\beta}{\sqrt{N}}
\end{aligned} \tag{E.28}$$

Thus,

$$\mathbb{E}\Omega(\mathcal{D}) \le 4L_\Phi \rho + \frac{4\beta L_\Phi}{\sqrt{N}}. \tag{E.29}$$

c) High-probability bound: Applying McDiarmid's inequality [8] to (E.17) gives, for every $x > 0$,

$$\mathbb{P}\big(\Omega(\mathcal{D}) - \mathbb{E}\Omega(\mathcal{D}) \ge x\big) \le \exp\left\{ -\frac{Nx^2}{32\beta^2 L_\Phi^2} \right\}. \tag{E.30}$$

Taking

$$x = 4\beta L_\Phi \sqrt{\frac{2\log(1/\epsilon)}{N}} \tag{E.31}$$

and combining with (E.29), we obtain, with probability at least $1 - \epsilon$, uniformly for all $f \in \mathcal{F}$,

$$\begin{aligned}
\left| \boldsymbol{R}^{\mathcal{D}}(f) - \boldsymbol{R}^{\mathcal{D}}_{\mathsf{emp}}(f) \right| &\le 4L_\Phi \rho + \frac{4\beta L_\Phi}{\sqrt{N}} \\
&\quad + 4\beta L_\Phi \sqrt{\frac{2\log(1/\epsilon)}{N}}.
\end{aligned} \tag{E.32}$$

We now take $\Phi = \Phi_\eta$, the η-margin loss in (69), for which $L_\Phi = 1/\eta$. Using the definition of $\mathcal{R}(\epsilon, N)$ in (78), the preceding display becomes

$$\left| \boldsymbol{R}^{\mathcal{D}}(f) - \boldsymbol{R}^{\mathcal{D}}_{\mathsf{emp}}(f) \right| \le \frac{4\rho}{\eta} + \mathcal{R}(\epsilon, N). \tag{E.33}$$

For a given training set $\mathcal{D}$, $\boldsymbol{R}^{\mathcal{D}}(f)$ is the expected η-margin loss of the deployed classifier $\boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}$, while $\boldsymbol{R}^{\mathcal{D}}_{\mathsf{emp}}(f)$ denotes its empirical value. Hence (70) gives

$$P_e(\boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}) \le P^{\eta}_{e,\mathsf{emp}}(\boldsymbol{c}^{\mathcal{D}}_{\mathsf{ave},f}) + \frac{4\rho}{\eta} + \mathcal{R}(\epsilon, N). \tag{E.34}$$

The high-probability event is uniform over $f \in \mathcal{F}$ and depends only on the aligned network training set $\mathcal{D}$. Therefore, given a fixed $\mathcal{D}$, (E.34) may be evaluated at any data-dependent outputs $\widehat{\boldsymbol{f}}_k = \mathcal{A}_k(\mathcal{D}_k, \mathcal{S}_k) \in \mathcal{F}_k$, for every realization of the local training randomness, without any empirical optimality condition. This establishes (77). Applying the same argument to the projected sample $\mathcal{D}_k$ for a fixed agent k gives (76) and completes the proof. □

Remark on a refinement. The preceding proof is formulated for a generic L_Φ-Lipschitz loss and controls the effect of replacing one complete network example through the bounded score range. This produces the factor βL_Φ in the confidence term. For the η-margin loss Φ_η, a sharper concentration argument is available because $0 \le \Phi_\eta \le 1$.

Specifically, after embedding the empirically centered classifiers in the fixed offset-augmented class $\mathcal{G}_{\mathrm{off}}$, define the one-sided population–empirical deviation

$$\Delta_{\mathrm{off},+}(\mathcal{D}) \triangleq \sup_{g \in \mathcal{G}_{\mathrm{off}}} \left\{ \mathbb{E}\Phi_\eta(\boldsymbol{\gamma} g(\boldsymbol{h})) - \frac{1}{N} \sum_{n=1}^N \Phi_\eta(\boldsymbol{\gamma}_n g(\boldsymbol{h}_n)) \right\}. \tag{E.35}$$

Standard one-sided symmetrization and the non-absolute contraction inequality in [7] give

$$\mathbb{E}\Delta_{\mathrm{off},+}(\mathcal{D}) \le \frac{2}{\eta} \left(\rho + \frac{\beta}{\sqrt{N}} \right). \tag{E.36}$$

Moreover, replacing one complete network example changes this fixed-class uniform deviation by at most $1/N$, since the η-margin loss takes values in $[0, 1]$. Thus, the bounded-difference constant for $\Delta_{\mathrm{off},+}(\mathcal{D})$ is $1/N$, and McDiarmid's inequality

yields the alternative confidence term

$$\sqrt{\frac{\log(1/\epsilon)}{2N}}. \tag{E.37}$$

Consequently, the collaborative bound associated with the η-margin loss can be refined to

$$P_e(\boldsymbol{c}_{\mathsf{ave}}) \le P^{\eta}_{e,\mathsf{emp}}(\boldsymbol{c}_{\mathsf{ave}}) + \frac{2\rho}{\eta} + \frac{2\beta}{\eta\sqrt{N}} + \sqrt{\frac{\log(1/\epsilon)}{2N}}. \tag{E.38}$$

The analogous refinement for a local classifier is

$$P_e(\boldsymbol{c}_k) \le P^{\eta}_{e,\mathsf{emp}}(\boldsymbol{c}_k) + \frac{2\rho_k}{\eta} + \frac{2\beta}{\eta\sqrt{N}} + \sqrt{\frac{\log(1/\epsilon)}{2N}}. \tag{E.39}$$

These refinements use the boundedness of Φ_η and a one-sided fixed-class argument. Theorem 4 retains the bound obtained from the generic Lipschitz-loss argument above.

Appendix F
Relation to Social Learning

This appendix returns to the prediction dynamics and relates the DeGroot collaboration rule (11) used in the main paper to social learning (SL) based on geometric pooling. We will first formulate the corresponding static-observation specialization. Then, we establish its finite-round classification guarantee, and compare the two update mechanisms.

1. Static Observation Specialization

A defining feature of the DeGroot model (11) is that agents update their opinions $\boldsymbol{\lambda}_{k,t}$ solely by averaging those of their neighbors, without acquiring new observations over time. In contrast, SL rules are designed for inference in *streaming* environments, where agents continuously receive new observations and combine them with information aggregated from their neighbors. To connect these two inference mechanisms, we consider the geometric-pooling SL mechanism [9], specialized to the learned local decision statistics used in social machine learning [1], [2]:

$$\boldsymbol{\lambda}^{\mathsf{SL}}_{k,t} = \sum_{\ell=1}^{K} a_{\ell k}\big(\boldsymbol{\lambda}^{\mathsf{SL}}_{\ell,t-1} + \boldsymbol{c}_\ell(\boldsymbol{h}_{\ell,t})\big) \tag{F.1}$$

where $\boldsymbol{h}_{\ell,t}$ denotes the private observation received by agent ℓ at round t. The linear recursion in (F.1) is the log-domain form of the corresponding geometric SL update. Unrolling it gives

$$\boldsymbol{\lambda}^{\mathsf{SL}}_{k,t} = \sum_{\ell=1}^{K}[A^t]_{\ell k}\boldsymbol{\lambda}^{\mathsf{SL}}_{\ell,0} + \sum_{i=1}^{t}\sum_{\ell=1}^{K}[A^{t+1-i}]_{\ell k}\boldsymbol{c}_\ell(\boldsymbol{h}_{\ell,i}). \tag{F.2}$$

We assume no prior information about the class label, so that $\boldsymbol{\lambda}^{\mathsf{SL}}_{k,0} = 0$. In the static observation specialization, $\boldsymbol{h}_{\ell,t} = \boldsymbol{h}^*_\ell$ for every t, and therefore

$$\boldsymbol{\lambda}^{\mathsf{SL}}_{k,t} = \sum_{\ell=1}^{K}\sum_{i=1}^{t}[A^i]_{\ell k}\boldsymbol{c}_\ell(\boldsymbol{h}^*_\ell). \tag{F.3}$$

As the same testing example is reused at every round, the repeated terms in (F.3) do not represent new temporal evidence. Instead, they repeatedly accumulate the statistics associated with the same K local views.

Since $[A^i]_{\ell k} \to \pi_\ell$, its Cesàro averages satisfy

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{t}\boldsymbol{\lambda}^{\mathsf{SL}}_{k,t} &= \sum_{\ell=1}^{K}\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t}[A^i]_{\ell k}\boldsymbol{c}_\ell(\boldsymbol{h}^*_\ell)\\ &= \sum_{\ell=1}^{K}\pi_\ell\boldsymbol{c}_\ell(\boldsymbol{h}^*_\ell) = \boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*).\end{aligned} \tag{F.4}$$

Hence, whenever $\boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*) \neq 0$, the prediction of every agent eventually agrees with the collective prediction based on the aggregate classifier $\boldsymbol{c}_{\mathsf{ave}}$. The sufficient-communication guarantee of Theorem 1 therefore applies to the limiting SL prediction as well. For the finite-round analysis, we define

$$\widehat{\boldsymbol{\gamma}}^{\mathsf{SL}}_{k,t} = \operatorname{sgn}(\boldsymbol{\lambda}^{\mathsf{SL}}_{k,t}), \quad \mathcal{M}^{\mathsf{SL}}_{k,t} = \{\boldsymbol{\gamma}^*\boldsymbol{\lambda}^{\mathsf{SL}}_{k,t} \le 0\}. \tag{F.5}$$

Similar to Theorems 1–3, we analyze the conditional probability $\mathbb{P}[\mathcal{M}^{\mathsf{SL}}_{k,t}\,|\,\mathcal{T}]$. To quantify the cumulative deviation of the finite-round weights from the Perron vector, define

$$\mathsf{B}_A \triangleq \max_{k\in\mathcal{K}}\sum_{i=1}^{\infty}\sum_{\ell=1}^{K}|[A^i]_{\ell k} - \pi_\ell|, \tag{F.6}$$

which is finite under Assumption 3. For every $\sigma \in (\sigma_A, 1)$ and corresponding $C(A,\sigma)$ in (52), it holds that

$$\mathsf{B}_A \le \frac{KC(A,\sigma)\sigma}{1-\sigma}. \tag{F.7}$$

With this definition, we establish the classification guarantee for the SL rule.

Proposition F.1 (**Static-observation SL classification guarantee**). *Suppose Assumptions 2, 3, and 5 hold, and let $\delta > 0$. Define*

$$\kappa \triangleq 2\beta\mathsf{B}_A. \tag{F.8}$$

Almost surely on $\mathcal{C}_\delta$, for every agent $k \in \mathcal{K}$ and every integer $t > \kappa/\delta$,

$$\mathbb{P}(\mathcal{M}^{\mathsf{SL}}_{k,t}\,|\,\mathcal{T}) \le \exp\left\{-\frac{(t\delta-\kappa)^2}{2\tau_{\max}\beta^2\sum_{\ell=1}^{K}\big(\sum_{i=1}^{t}[A^i]_{\ell k}\big)^2}\right\}. \tag{F.9}$$

This bound converges to the sufficient-communication bound for $\boldsymbol{c}_{\mathsf{ave}}$ as $t \to \infty$.

2. Proof of Proposition F.1

Proof. The proof reuses the conditioning convention and ATE bounded-difference inequality from Appendix B. Relative to the finite-round DeGroot analysis presented in Appendix C, the only change is that the SL statistic uses the cumulative weights $\sum_{i=1}^{t}[A^i]_{\ell k}$ rather than the single-round weights $[A^t]_{\ell k}$.

Fix a realization of the training phase in $\mathcal{C}_\delta$. Conditional on $\mathcal{T}$, the trained classifiers are fixed, while the fresh testing vector follows P_γ under $\boldsymbol{\gamma}^* = \gamma$. Since the same testing view is reused at every round, its repeated appearances are not treated as independent coordinates. Instead, changing view ℓ changes all of its accumulated contributions simultaneously.

From (F.3), changing only coordinate ℓ of the testing vector changes the statistic by at most

$$\begin{aligned}\left|\boldsymbol{\lambda}_{k,t}^{\mathsf{SL}}(\boldsymbol{h}^*) - \boldsymbol{\lambda}_{k,t}^{\mathsf{SL}}(\widehat{\boldsymbol{h}})\right| &\le \sum_{i=1}^{t}[A^i]_{\ell k}\left|c_\ell(\boldsymbol{h}_\ell^*) - c_\ell(\widehat{\boldsymbol{h}}_\ell)\right| \\ &\le 2\beta\sum_{i=1}^{t}[A^i]_{\ell k}. \end{aligned} \tag{F.10}$$

Thus Lemma B.1 applies with coordinate oscillations

$$d_{\ell,k,t}^{\mathsf{SL}} = 2\beta\sum_{i=1}^{t}[A^i]_{\ell k}. \tag{F.11}$$

We next control the class-conditional mean. For either $\gamma \in \Gamma$,

$$\begin{aligned}&\left|\mathbb{E}_{\boldsymbol{h}^*}^{(\gamma)}\boldsymbol{\lambda}_{k,t}^{\mathsf{SL}}(\boldsymbol{h}^*) - t\mu^\gamma(\widehat{f})\right| \\ &= \left|\sum_{i=1}^{t}\sum_{\ell=1}^{K}\big([A^i]_{\ell k} - \pi_\ell\big)\mu_\ell^\gamma(\widehat{f}_\ell)\right| \\ &\le 2\beta\sum_{i=1}^{t}\sum_{\ell=1}^{K}\big|[A^i]_{\ell k} - \pi_\ell\big| \\ &\le 2\beta\mathsf{B}_A = \kappa \end{aligned} \tag{F.12}$$

where $|\mu_\ell^\gamma(\widehat{f}_\ell)| \le 2\beta$ follows from $|c_\ell| \le 2\beta$. For $\boldsymbol{\gamma}^* = +1$, the event $\mathcal{C}_\delta$ implies $\mu^+(\widehat{f}) > \delta$, so for $t > \kappa/\delta$,

$$\mathbb{E}_{\boldsymbol{h}^*}^{(+1)}\boldsymbol{\lambda}_{k,t}^{\mathsf{SL}}(\boldsymbol{h}^*) > t\delta - \kappa > 0. \tag{F.13}$$

The lower-tail form of Lemma B.1 therefore gives

$$\begin{aligned}&\mathbb{P}\big(\mathcal{M}_{k,t}^{\mathsf{SL}}\,\big|\,\boldsymbol{\gamma}^* = +1, \mathcal{T}\big) \\ &\le \exp\left\{-\frac{(t\delta-\kappa)^2}{2\tau_{+1}\beta^2\sum_{\ell=1}^{K}\big(\sum_{i=1}^{t}[A^i]_{\ell k}\big)^2}\right\}. \end{aligned} \tag{F.14}$$

A similar bound is obtained for the case $\boldsymbol{\gamma}^* = -1$ by replacing τ_{+1} with τ_{-1}. The inequality in (F.9) is then established using the definition of $\tau_{\max}$ in (45). Finally,

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t}[A^i]_{\ell k} = \pi_\ell, \tag{F.15}$$

which implies

$$\lim_{t\to\infty}\frac{1}{t^2}\sum_{\ell=1}^{K}\left(\sum_{i=1}^{t}[A^i]_{\ell k}\right)^2 = \sum_{\ell=1}^{K}\pi_\ell^2. \tag{F.16}$$

Since $(t\delta-\kappa)^2/t^2 \to \delta^2$, we obtain

$$\lim_{t\to\infty}\frac{(t\delta-\kappa)^2}{2\tau_{\max}\beta^2\sum_{\ell=1}^{K}\big(\sum_{i=1}^{t}[A^i]_{\ell k}\big)^2} = \frac{\delta^2}{2\tau_{\max}\beta^2\sum_{\ell=1}^{K}\pi_\ell^2}, \tag{F.17}$$

which is the exponent that appears in the bound (46) for the aggregate classifier $\boldsymbol{c}_{\mathsf{ave}}$ under sufficient communication. This completes the proof. □

3. Comparison with the DeGroot Rule

By Jensen's inequality, the following holds:

$$\begin{aligned}\sum_{\ell=1}^{K}\left(\sum_{i=1}^{t}[A^i]_{\ell k}\right)^2 &\le t\sum_{\ell=1}^{K}\sum_{i=1}^{t}\big([A^i]_{\ell k}\big)^2 \\ &\le t\sum_{\ell=1}^{K}\sum_{i=1}^{t}[A^i]_{\ell k} = t^2 \end{aligned} \tag{F.18}$$

where the last equality uses left-stochasticity of A. Hence, for every agent,

$$\mathbb{P}\big(\mathcal{M}_{k,t}^{\mathsf{SL}}\,|\,\mathcal{T}\big) \le \exp\left\{-\frac{(t\delta-\kappa)^2}{2\tau_{\max}\beta^2 t^2}\right\} \tag{F.19}$$

almost surely on $\mathcal{C}_\delta$ whenever $t > \kappa/\delta$.

The finite-round DeGroot and static observation SL statistics are generally different: the former uses the weights $[A^t]_{\ell k}$ in (48), whereas the latter accumulates $\sum_{i=1}^{t}[A^i]_{\ell k}$ in (F.3). Accordingly, neither finite-round analytical bound is asserted to vary monotonically with t, because its denominator depends on the corresponding finite-round mixing weights. Nevertheless, (16) and (F.4) show that, whenever $\boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*) \neq 0$, both mechanisms eventually yield the same collective prediction $\widehat{\boldsymbol{\gamma}}$ in (17). Their error bounds likewise converge to the sufficient-communication bound.

A particularly transparent connection arises for the fully connected consensus matrix $A = \pi\mathbb{1}^\top$. In this case, $A^m = A$ for every integer $m \ge 1$, and for every $t \ge 1$,

$$\boldsymbol{\lambda}_{k,t} = \sum_{\ell=1}^{K}\pi_\ell\boldsymbol{c}_\ell(\boldsymbol{h}_\ell^*) = \boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*), \tag{F.20}$$

whereas

$$\boldsymbol{\lambda}_{k,t}^{\mathsf{SL}} = t\sum_{\ell=1}^{K}\pi_\ell\boldsymbol{c}_\ell(\boldsymbol{h}_\ell^*) = t\boldsymbol{c}_{\mathsf{ave}}(\boldsymbol{h}^*). \tag{F.21}$$

Therefore, the two statistics have the same sign and produce the same prediction after one communication round. Moreover, exact mixing is achieved after one communication round, since

$$[A^t]_{\ell k} = \pi_\ell, \quad t \ge 1. \tag{F.22}$$

Hence, for $t \ge 1$, using the exact identity $[A^t]_{\ell k} = \pi_\ell$ directly in the finite-round DeGroot argument makes the transient mean error identically zero, and the resulting specialized concentration bound coincides with the sufficient-communication bound. For the SL rule, $\mathsf{B}_A = 0$ and hence $\kappa = 0$, while

$$\sum_{i=1}^{t}[A^i]_{\ell k} = t\pi_\ell. \tag{F.23}$$

Its finite-round bound therefore reduces to the same sufficient-communication bound for every $t \ge 1$. Finally, the two recursions use different initial conditions because they represent different inference semantics. For SL, $\boldsymbol{\lambda}_{k,0}^{\mathsf{SL}} = 0$ is the state before the first observation is incorporated. For the DeGroot rule, $\boldsymbol{\lambda}_{k,0} = \boldsymbol{c}_k(\boldsymbol{h}_k^*)$ is the local decision statistic available before communication. This discrepancy at $t = 0$ does not affect the preceding finite-round or limiting comparison.

APPENDIX G
RAW-SCORE COUNTERPARTS

The main analysis uses the empirically centered local scores $\boldsymbol{c}_k$ in (9). As discussed in Proposition 1, empirical centering acts as a threshold alignment mechanism rather than a structural requirement for collaboration. This appendix provides

the corresponding results when the learned scores $\widehat{\boldsymbol{f}}_k$ are used directly, without empirical centering.

The raw aggregate score f_{ave} and its class-conditional means $\boldsymbol{\mu}^{\gamma}_{\mathrm{raw}}$ are already defined in (20). Moreover, the raw expected zero-threshold margin is $\Delta_{\mathrm{raw}}(f)$ in (26). Since the arguments closely parallel those developed for the centered scores c_{ave} in Appendices A–F, we state only the changes needed for the raw-score counterparts.

1. Raw-Score Training Guarantee

For $0 \le \delta < \beta$ and a generic risk level r, we define

$$\mathcal{E}^{\mathrm{raw}}_{\Phi}(r,\delta) \triangleq \frac{\Phi(\delta)+\Phi(\beta)-2r}{16L_{\Phi}}. \tag{G.1}$$

For the learned functions $\widehat{\boldsymbol{f}} = \{\widehat{\boldsymbol{f}}_1,\dots,\widehat{\boldsymbol{f}}_K\}$, we denote the raw-score δ-margin training event by

$$\begin{aligned}\mathcal{C}^{\mathrm{raw}}_{\delta} &\triangleq \big\{\Delta_{\mathrm{raw}}(\widehat{\boldsymbol{f}}) > \delta\big\}\\ &= \big\{\boldsymbol{\mu}^{+}_{\mathrm{raw}}(\widehat{\boldsymbol{f}}) > \delta,\, \boldsymbol{\mu}^{-}_{\mathrm{raw}}(\widehat{\boldsymbol{f}}) < -\delta\big\}\end{aligned} \tag{G.2}$$

where the equality follows from (21), (22) and (26). This is the raw-score counterpart of the centered condition (34). Similar to Proposition 2, we can establish a bound on the achievability of (G.2) as follows.

Corollary G.1 (Raw-score δ-margin consistent training). *Suppose Assumptions 1–3 hold and the approximate optimization condition* (37) *is satisfied. Let*

$$r \triangleq \mathsf{R}^{o} + \varepsilon_{\mathsf{opt}} \tag{G.3}$$

where $\varepsilon_{\mathsf{opt}}$ is defined in (38)*. If*

$$0 \le \delta < \beta, \quad r < \frac{\Phi(\delta)+\Phi(\beta)}{2}, \quad \rho < \mathcal{E}^{\mathrm{raw}}_{\Phi}(r,\delta) \tag{G.4}$$

then

$$\mathbb{P}(\mathcal{C}^{\mathrm{raw}}_{\delta}) \ge 1 - \exp\left\{-\frac{8N_{\max}}{\alpha^2\beta^2}\Big(\mathcal{E}^{\mathrm{raw}}_{\Phi}(r,\delta)-\rho\Big)^2\right\}. \tag{G.5}$$

Proof. The analysis on the approximate optimization remains unchanged from the proof of Proposition 2. With the argument in Appendix A.3, we have

$$\boldsymbol{R}(\widehat{\boldsymbol{f}}) \le r + 2U, \quad r = \mathsf{R}^{o} + \varepsilon_{\mathsf{opt}} \tag{G.6}$$

where U is the uniform risk deviation in (A.15). It remains to relate failure of the δ-margin condition (G.2) to the population surrogate risk. We introduce the following two class-conditional risks:

$$\boldsymbol{R}^{+}(f) \triangleq \sum_{k=1}^{K} \pi_k \mathbb{E}^{(+1)}\Phi(f_k(\boldsymbol{h}_k)), \tag{G.7}$$

$$\boldsymbol{R}^{-}(f) \triangleq \sum_{k=1}^{K} \pi_k \mathbb{E}^{(-1)}\Phi(-f_k(\boldsymbol{h}_k)). \tag{G.8}$$

Under the uniform class prior, $\boldsymbol{R}(f) = [\boldsymbol{R}^{+}(f) + \boldsymbol{R}^{-}(f)]/2$. By the same Jensen argument used in (A.29),

$$\boldsymbol{R}^{+}(f) \ge \Phi\big(\boldsymbol{\mu}^{+}_{\mathrm{raw}}(f)\big), \quad \boldsymbol{R}^{-}(f) \ge \Phi\big(-\boldsymbol{\mu}^{-}_{\mathrm{raw}}(f)\big). \tag{G.9}$$

Moreover, Assumption 2 and the non-increasing property of Φ imply

$$\boldsymbol{R}^{+}(f) \ge \Phi(\beta), \quad \boldsymbol{R}^{-}(f) \ge \Phi(\beta). \tag{G.10}$$

If $\mathcal{C}^{\mathrm{raw}}_{\delta}$ fails, one of the two class-conditional risks is at least $\Phi(\delta)$, while the other is at least $\Phi(\beta)$. Therefore,

$$\overline{\mathcal{C}^{\mathrm{raw}}_{\delta}} \subseteq \left\{\boldsymbol{R}(\widehat{\boldsymbol{f}}) \ge \frac{\Phi(\delta)+\Phi(\beta)}{2}\right\}. \tag{G.11}$$

Combining (G.11) with (G.6) gives, on $\overline{\mathcal{C}^{\mathrm{raw}}_{\delta}}$,

$$U \ge \frac{\Phi(\delta)+\Phi(\beta)-2r}{4} = 4L_{\Phi}\mathcal{E}^{\mathrm{raw}}_{\Phi}(r,\delta). \tag{G.12}$$

Applying the uniform deviation bound (A.18) at this threshold proves (G.5). □

Compared with Proposition 2, the change in the bound (G.5) occurs in the reduction from margin event failure to population risk. The centered analysis additionally controls the empirical training mean, whereas the raw margin in (26) is determined directly by the two raw class-conditional means.

2. Sufficient-Communication Counterpart

Set $f = \widehat{\boldsymbol{f}}$ in the raw aggregate score (20), and define

$$\mathcal{M}^{\mathrm{raw}} \triangleq \{\boldsymbol{\gamma}^{*} f_{\mathsf{ave}}(\boldsymbol{h}^{*}) \le 0\}. \tag{G.13}$$

This is the raw-score counterpart of the event $\mathcal{M}$ in (42). When $\mathcal{C}^{\mathrm{raw}}_{\delta}$ is true, the class-conditional mean of $f_{\mathsf{ave}}(\boldsymbol{h}^{*})$ lies on the correct side of zero by more than δ. Moreover, changing only testing view k changes $f_{\mathsf{ave}}(\boldsymbol{h}^{*})$ by at most $2\beta\pi_k$, which is the same coordinate oscillation used for the centered aggregate in Appendix B.2. Applying Lemma B.1 under the two class-conditional laws P_{γ} and using the definition of $\tau_{\max}$ gives

$$\mathbb{P}(\mathcal{M}^{\mathrm{raw}}\,|\,\mathcal{T}) \le \exp\left\{-\frac{\delta^2}{2\tau_{\max}\beta^2\sum_{k=1}^{K}\pi_k^2}\right\} \tag{G.14}$$

almost surely on $\mathcal{C}^{\mathrm{raw}}_{\delta}$. Thus, for the same prescribed margin δ, the prediction-side bound (G.14) has exactly the same form as its centered counterpart in (46). Applying the same tower-property decomposition as in (42), with $\mathcal{M}$ and $\mathcal{C}_{\delta}$ replaced by $\mathcal{M}^{\mathrm{raw}}$ and $\mathcal{C}^{\mathrm{raw}}_{\delta}$, respectively, gives

$$\begin{aligned}\mathbb{P}(\mathcal{M}^{\mathrm{raw}}) &= \mathbb{E}[\mathbb{P}(\mathcal{M}^{\mathrm{raw}}\,|\,\mathcal{T})]\\ &\le \mathbb{E}[1[\mathcal{C}^{\mathrm{raw}}_{\delta}]\mathbb{P}(\mathcal{M}^{\mathrm{raw}}\,|\,\mathcal{T})] + \mathbb{P}(\overline{\mathcal{C}^{\mathrm{raw}}_{\delta}}).\end{aligned} \tag{G.15}$$

Combining (G.15) with (G.14) and the raw-score training guarantee (G.5) gives the corresponding unconditional guarantee.

3. Finite-Round Counterpart

Without empirical centering, the finite-round statistic corresponding to (48) becomes

$$\boldsymbol{\lambda}^{\mathrm{raw}}_{k,t} \triangleq \sum_{\ell=1}^{K}[A^{t}]_{\ell k}\widehat{\boldsymbol{f}}_{\ell}(\boldsymbol{h}^{*}_{\ell}), \tag{G.16}$$

with

$$\mathcal{M}^{\mathrm{raw}}_{k,t} \triangleq \{\boldsymbol{\gamma}^{*}\boldsymbol{\lambda}^{\mathrm{raw}}_{k,t} \le 0\}. \tag{G.17}$$

The finite-round mixing argument pertaining to $\boldsymbol{\lambda}_{k,t}^{\mathrm{raw}}$ follows Appendix C. The coordinate oscillation remains $2\beta[A^t]_{\ell k}$. The only change in the expectation term follows from

$$\left|\mathbb{E}^{(\gamma)}\widehat{\boldsymbol{f}}_\ell(\boldsymbol{h}_\ell)\right| \le \beta, \tag{G.18}$$

instead of the bound 2β for empirically centered scores $\boldsymbol{c}_\ell(\boldsymbol{h}_\ell)$. Using the matrix-power estimate in (52),

$$\left|\mathbb{E}^{(\gamma)}\boldsymbol{\lambda}_{k,t}^{\mathrm{raw}} - \boldsymbol{\mu}_{\mathrm{raw}}^{\gamma}(\widehat{\boldsymbol{f}})\right| \le \beta KC(A,\sigma)\sigma^t. \tag{G.19}$$

Accordingly, define the raw-score counterpart of the residual margin in (53) by

$$r_t^{\mathrm{raw}}(\delta) \triangleq \delta - \beta KC(A,\sigma)\sigma^t. \tag{G.20}$$

Repeating the argument used to establish Theorem 2 gives, for every agent k and every $t \ge 0$ such that $r_t^{\mathrm{raw}}(\delta) > 0$,

$$\mathbb{P}(\mathcal{M}_{k,t}^{\mathrm{raw}} \,|\, \mathcal{T}) \le \exp\left\{-\frac{[r_t^{\mathrm{raw}}(\delta)]^2}{2\tau_{\max}\beta^2\sum_{\ell=1}^K([A^t]_{\ell k})^2}\right\} \tag{G.21}$$

almost surely on $\mathcal{C}_\delta^{\mathrm{raw}}$. Therefore, relative to the bound in (54), only the residual margin changes:

$$r_t(\delta) = \delta - 2\beta KC(A,\sigma)\sigma^t \longrightarrow r_t^{\mathrm{raw}}(\delta) = \delta - \beta KC(A,\sigma)\sigma^t. \tag{G.22}$$

The corresponding unconditional guarantee follows by combining (G.21) with (G.5), in the same manner as the centered finite-round decomposition (51).

4. Finite-Precision Counterpart

We next consider the raw-score counterpart of the finite-precision recursion (60). Since $|\widehat{\boldsymbol{f}}_k(h)| \le \beta$ by Assumption 2, the raw inputs can be quantized over $[-\beta,\beta]$ rather than the centered range $[-2\beta,2\beta]$ in (56). Accordingly, we employ a narrower raw-score alphabet

$$\mathcal{V}_b^{\mathrm{raw}} \triangleq \left\{v_j^{\mathrm{raw}} = -\beta + j\Delta_b^{\mathrm{raw}} : j = 0,\dots,L_b-1\right\} \tag{G.23}$$

where

$$\Delta_b^{\mathrm{raw}} \triangleq \frac{2\beta}{L_b-1} = \frac{\Delta_b}{2}. \tag{G.24}$$

We apply the same adjacent-level stochastic rounding protocol as in (58) using this raw-score alphabet. The initialization and all quantizer outputs lie in $[-\beta,\beta]$, and the update in (60) is a convex combination. Therefore, this interval is preserved throughout the recursion. The accumulated perturbation term corresponding to (64) is

$$\begin{aligned} V_{k,t}^{\mathrm{q,raw}} &\triangleq \frac{(\Delta_b^{\mathrm{raw}})^2}{4}\sum_{r=1}^t\sum_{j=1}^K([A^r]_{jk})^2 \\ &= \frac{\beta^2}{(L_b-1)^2}\sum_{r=1}^t\sum_{j=1}^K([A^r]_{jk})^2 \\ &= \frac{1}{4}V_{k,t}^{\mathrm{q}}. \end{aligned} \tag{G.25}$$

Let $x_{k,t}^{(b),\mathrm{raw}}$ denote the recursion (60) initialized by

$$x_{k,0}^{(b),\mathrm{raw}} = \widehat{\boldsymbol{f}}_k(\boldsymbol{h}_k^*) \tag{G.26}$$

and using the raw quantization range above. Define

$$\mathcal{M}_{k,t}^{(b),\mathrm{raw}} \triangleq \{\boldsymbol{\gamma}^* x_{k,t}^{(b),\mathrm{raw}} \le 0\}. \tag{G.27}$$

The conditional-MGF argument in Appendix D is unchanged. Replacing $r_t(\delta)$ and $V_{k,t}^{\mathrm{q}}$ in (65) respectively by $r_t^{\mathrm{raw}}(\delta)$ and $V_{k,t}^{\mathrm{q,raw}}$ therefore gives

$$\begin{aligned} &\mathbb{P}(\mathcal{M}_{k,t}^{(b),\mathrm{raw}} \,|\, \mathcal{T}) \\ &\le \exp\left\{-\frac{[r_t^{\mathrm{raw}}(\delta)]^2}{2\left[\tau_{\max}\beta^2\sum_{\ell=1}^K([A^t]_{\ell k})^2 + V_{k,t}^{\mathrm{q,raw}}\right]}\right\} \end{aligned} \tag{G.28}$$

almost surely on $\mathcal{C}_\delta^{\mathrm{raw}}$ whenever $r_t^{\mathrm{raw}}(\delta) > 0$. Thus, relative to the centered result (65), the raw-score counterpart changes only the residual finite-round margin $r_t^{\mathrm{raw}}(\delta)$ and the quantization proxy $V_{k,t}^{\mathrm{q,raw}}$. The unconditional guarantee follows by combining (G.28) with (G.5), as in Theorem 3.

5. PAC-Style Generalization Counterpart

We next consider the PAC-style result under the aligned network training set (74) and its local projections (75). The raw aggregate f_{ave} in (20) does not contain the sample-dependent empirical training mean that appears in the centered aggregate. Therefore, the fixed-class refinement technique described at the end of Appendix E applies directly without the additional offset variable.

For the raw local and aggregate scores, respectively, define

$$P_e(f_k) \triangleq \mathbb{P}\big(\mathrm{sgn}(f_k(\boldsymbol{h}_k^*)) \ne \boldsymbol{\gamma}^* \,|\, \mathcal{T}\big), \tag{G.29}$$

$$P_e(f_{\mathsf{ave}}) \triangleq \mathbb{P}\big(\mathrm{sgn}(f_{\mathsf{ave}}(\boldsymbol{h}^*)) \ne \boldsymbol{\gamma}^* \,|\, \mathcal{T}\big). \tag{G.30}$$

We establish the analogous PAC-style bounds for f_{ave} and f_k as follows.

Corollary G.2 (Raw-score margin generalization bound). *Under the sampling setup of Theorem 4, fix $\eta > 0$ and $\epsilon \in (0,1)$. Then, with probability at least $1-\epsilon$, uniformly for all $f \in \mathcal{F}$,*

$$P_e(f_{\mathsf{ave}}) \le P_{e,\mathsf{emp}}^{\eta}(f_{\mathsf{ave}}) + \frac{2\rho}{\eta} + \sqrt{\frac{\log(1/\epsilon)}{2N}}. \tag{G.31}$$

Moreover, for every fixed agent k, with probability at least $1-\epsilon$, uniformly for all $f_k \in \mathcal{F}_k$,

$$P_e(f_k) \le P_{e,\mathsf{emp}}^{\eta}(f_k) + \frac{2\rho_k}{\eta} + \sqrt{\frac{\log(1/\epsilon)}{2N}}. \tag{G.32}$$

Proof. Following (E.35), we define the one-sided deviation

$$\Delta_+^{\mathrm{raw}}(\mathcal{D}) \triangleq \sup_{f\in\mathcal{F}}\Big\{\mathbb{E}\Phi_\eta(\boldsymbol{\gamma} f_{\mathsf{ave}}(\boldsymbol{h})) - P_{e,\mathsf{emp}}^{\eta}(f_{\mathsf{ave}})\Big\} \tag{G.33}$$

where Φ_η and $P_{e,\mathsf{emp}}^{\eta}$ are defined in (69) and (71), respectively. Since the function class consisting of f_{ave}, $\forall f \in \mathcal{F}$ is fixed, the one-sided symmetrization and contraction argument used in the refinement of Appendix E gives

$$\mathbb{E}\Delta_+^{\mathrm{raw}}(\mathcal{D}) \le \frac{2}{\eta}\sum_{k=1}^K \pi_k\rho_k = \frac{2\rho}{\eta}. \tag{G.34}$$

Moreover, $0 \le \Phi_\eta \le 1$, so replacing one complete network example changes $\Delta_+^{\mathrm{raw}}(\mathcal{D})$ by at most $1/N$. McDiarmid's inequality therefore gives, with probability at least $1-\epsilon$,

$$\Delta_+^{\mathrm{raw}}(\mathcal{D}) \le \frac{2\rho}{\eta} + \sqrt{\frac{\log(1/\epsilon)}{2N}}. \tag{G.35}$$

Using (70) proves (G.31). Since this high-probability event is uniform over $f \in \mathcal{F}$, it may be evaluated at $f = \widehat{\boldsymbol{f}}$ for every realization of the local training randomness. Applying the same argument to the projected sample $\mathcal{D}_k$ gives (G.32). □

Relative to the centered refinement in Appendix E, the offset-complexity term is absent in Corollary G.2 because the raw-score classifier does not contain an empirical training mean. As already characterized by (26) and (27), this difference does not imply a universal ordering of the resulting classifiers, since empirical centering also changes their margins relative to the zero decision threshold.

6. Static-Observation Social Learning Counterpart

Replacing the centered decision statistic $\boldsymbol{\lambda}_{k,t}^{\mathsf{SL}}$ in (F.3) by the learned raw scores gives

$$\boldsymbol{\lambda}_{k,t}^{\mathsf{SL,raw}} \triangleq \sum_{\ell=1}^{K}\sum_{i=1}^{t}[A^i]_{\ell k}\widehat{\boldsymbol{f}}_\ell(\boldsymbol{h}_\ell^*). \tag{G.36}$$

We define

$$\mathcal{M}_{k,t}^{\mathsf{SL,raw}} \triangleq \{\boldsymbol{\gamma}^*\boldsymbol{\lambda}_{k,t}^{\mathsf{SL,raw}} \le 0\}. \tag{G.37}$$

Using the cumulative mixing constant B_A in (F.6), we denote

$$\kappa_{\mathrm{raw}} \triangleq \beta\mathsf{B}_A = \frac{\kappa}{2} \tag{G.38}$$

where $\kappa = 2\beta\mathsf{B}_A$ is defined in (F.8). The proof of Proposition F.1 carries over directly. The cumulative coordinate oscillations are unchanged, while $|\mathbb{E}^{(\gamma)}\widehat{\boldsymbol{f}}_\ell(\boldsymbol{h}_\ell)| \le \beta$ reduces the cumulative mean-transient term in (F.12) from κ to κ_{raw}. Consequently, almost surely on $\mathcal{C}_\delta^{\mathrm{raw}}$, for every agent k and every integer $t > \kappa_{\mathrm{raw}}/\delta$,

$$\mathbb{P}(\mathcal{M}_{k,t}^{\mathsf{SL,raw}}\,|\,\mathcal{T}) \le \exp\left\{-\frac{(t\delta-\kappa_{\mathrm{raw}})^2}{2\tau_{\max}\beta^2\sum_{\ell=1}^{K}\left(\sum_{i=1}^{t}[A^i]_{\ell k}\right)^2}\right\}. \tag{G.39}$$

This is the direct raw-score counterpart of (F.9). Since

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t}[A^i]_{\ell k} = \pi_\ell, \tag{G.40}$$

the bound (G.39) converges to the bound associated with raw scores under sufficient communication (G.14).

Appendix H
Additional Experimental Details

This appendix provides additional implementation details and supplementary figures for the experiments conducted in Section IV. The presentation is grouped into the CIFAR-10 benchmark, the ModelNet40 benchmark, and the communication and robustness studies.

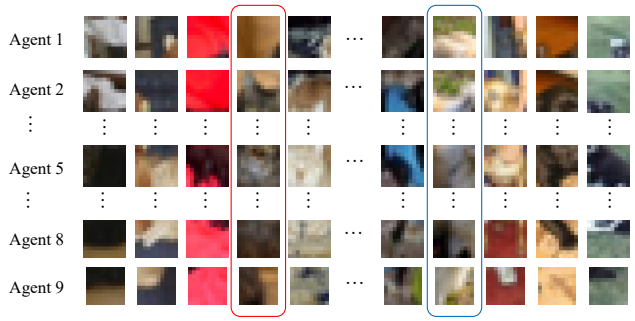


Fig. H.1. Illustration of the local labeled datasets in the CIFAR-10 patch-partition benchmark. The highlighted columns illustrate two representative images from the two classes contributing different local patches to all agents.

1. CIFAR-10 Benchmark

1) Experimental Protocol: This subsection provides additional implementation details for the CIFAR-10 patch-partition benchmark introduced in Section IV-A. Each agent trains a convolutional neural network (CNN) independently on the patch associated with its grid location. The network consists of two convolutional layers followed by fully connected layers and produces a two-dimensional output $z_k(\boldsymbol{h}_k) \in \mathbb{R}^2$ for each input patch $\boldsymbol{h}_k$ [1]. Its two entries, denoted by $z_{k,+1}(\boldsymbol{h}_k)$ and $z_{k,-1}(\boldsymbol{h}_k)$, are the logits associated with the two classes, which define the class-posterior probabilities through the soft-max rule:

$$\widehat{p}_k(\gamma|\boldsymbol{h}_k) = \frac{\exp(z_{k,\gamma}(\boldsymbol{h}_k))}{\exp(z_{k,+1}(\boldsymbol{h}_k)) + \exp(z_{k,-1}(\boldsymbol{h}_k))} \tag{H.1}$$

for $\gamma \in \{+1,-1\}$. Here, these logits refer to the raw outputs of the neural network, and should not be confused with the theoretical logit function in (3). In the binary case, their difference satisfies

$$z_{k,+1}(\boldsymbol{h}_k) - z_{k,-1}(\boldsymbol{h}_k) = \log\frac{\widehat{p}_k(+1|\boldsymbol{h}_k)}{\widehat{p}_k(-1|\boldsymbol{h}_k)}. \tag{H.2}$$

Therefore, in this implementation, the scalar quantity used for collaboration is first formed by taking the difference between the two logits. This logit difference is then centered according to (9) to obtain the local decision statistic $\boldsymbol{c}_k(\boldsymbol{h}_k)$.

For each value of N_0, we first sample a balanced set of CIFAR-10 images containing equal numbers of cats and dogs. The same selected images are then used across all agents, with each agent retaining only the patch corresponding to its own grid location. In this way, N_0 denotes the size of the local labeled dataset at each agent before the train/validation split. We reserve 20% of each local dataset for validation and use the remaining 80% for training. Figure H.1 illustrates this construction: each row corresponds to the local dataset of one agent, and each column shows the local patches extracted from a common image across agents.

Each local CNN classifier is trained independently using the Adam optimizer with learning rate 10^{-4}, mini-batch size 256, and cross-entropy loss for at most 100 epochs. Early stopping is applied in all CIFAR-10 experiments: training is monitored on the validation set and terminated once the validation loss ceases to improve over several consecutive validation checks,

and the model parameters corresponding to the best validation loss are retained.

For each value of N_0, all reported quantities in Fig. 2 are averaged over 200 Monte Carlo repetitions. In each repetition, the balanced image set is resampled, the train/validation split is regenerated, and all local classifiers are retrained from scratch. In the main baseline comparison, the communication graph is generated once and then kept fixed across repetitions. Each agent is also assumed to have a self-loop in the communication graph, which is omitted from Fig. 1b for clarity. The corresponding combination policy A is constructed using the uniform averaging rule:

$$a_{\ell k} = \frac{1}{|\mathcal{N}_k|}, \quad \forall \ell \in \mathcal{N}_k \tag{H.3}$$

where $|\mathcal{N}_k|$ denotes the cardinality of $\mathcal{N}_k$.

2) Baselines: Figure 2b compares the following methods.

a) Non-cooperative: Each agent predicts independently from its own local decision statistic, without any communication or aggregation.

b) Avg-stat: The local decision statistics are averaged uniformly across agents, and the final prediction is obtained by thresholding the resulting aggregate. This provides a simple averaging reference, and coincides with the proposed method when the combination policy A is doubly-stochastic and the communication is sufficient.

c) Vote: Each agent first generates a hard class prediction based on its local decision statistic. The final output is then determined by majority vote across the agents.

d) Learned fusion: A centralized affine fusion rule is fitted on a held-out validation set. If $\boldsymbol{c}(h) \in \mathbb{R}^K$ denotes the vector of local decision statistics for sample h, the baseline predicts according to the sign of

$$w^\top \boldsymbol{c}(h) + b_{\mathrm{fus}} \tag{H.4}$$

where both the fusion weights $w \in \mathbb{R}^K$ and the bias $b_{\mathrm{fus}} \in \mathbb{R}$ are learned from validation data.

e) Learned simplex fusion: This baseline also learns a centralized linear fusion rule from validation data, but under the structural constraints

$$w_k \geq 0, \quad \sum_{k=1}^{K} w_k = 1, \tag{H.5}$$

with no additive bias term. It may therefore be interpreted as a learned convex combination of the local decision statistics.

f) AdaBoost: A centralized boosting model is trained using the local predictors as constituent learners. Unlike the proposed method, this baseline introduces cooperation during training rather than only at test time.

g) VFL-JT: To provide a task-matched joint training comparator for the partitioned CIFAR-10 data, we implement a score-level vertical federated joint-training scheme inspired by the feature-distributed learning framework in [10]. For an aligned example, branch k receives the same local patch as the corresponding independently trained agent and outputs a scalar binary score s_k. The coordinating server forms

$$u_{\mathrm{VFL}} = b_{\mathrm{VFL}} + \sum_{k=1}^{K} w_k s_k, \tag{H.6}$$

and the branch parameters together with the fusion head are optimized jointly using the global binary logistic loss. During training, the local scores are sent to the server and the corresponding loss derivatives with respect to the local scores are returned to the branches. Raw patches and local parameter tensors are not exchanged, and no parameter averaging is performed.

TABLE H.1
COMMUNICATION COST OF VFL-JT AND THE PROPOSED SCHEME.

Method	Training	Inference per sample
VFL-JT	$KE_{\mathrm{VFL}}(2N_{\mathrm{tr}} + N_{\mathrm{val}})$	K
Proposed	0	$tE_{\mathrm{off}}(A)$

Because VFL-JT and the proposed method communicate at different stages, we report their training- and inference-time communication separately. Let E_{VFL} denote the realized number of VFL-JT training epochs, and N_{tr} and N_{val} the numbers of aligned training and validation samples, respectively. Recall that $E_{\mathrm{off}}(A)$ defined in (66) denotes the number of directed non-self communication links. The resulting communication costs, measured by the number of transmissions, are summarized in Table H.1.

For VFL-JT, each training sample involves the transmission of K local scores to the server and the return of K corresponding loss derivatives, while each validation sample requires the K forward score transmissions. For the proposed method, communication occurs only during test-time collaboration over the directed non-self links of the network. If each communicated scalar is represented using b bits, the corresponding bit costs are obtained by multiplying the entries in Table H.1 by b.

h) Central oracle: A centralized classifier from the same CNN family is trained on the full 32×32 CIFAR-10 image. In this implementation, the first convolutional layer is adapted to the full-image input, and the first fully connected layer is resized accordingly to match the resulting feature dimension. This baseline is included only as a full-information reference and is not constrained by the distributed observation model.

i) Ours (no centering): This ablation implements the same collaboration rule as the proposed method, but removes the centering step in (9). In other words, each agent uses the trained score $\widehat{\boldsymbol{f}}_k$ rather than the centered decision statistic $\boldsymbol{c}_k$.

3) Conditional Mutual Information: In Section IV-A1, the estimation of the mutual information $I(h_k; h_\ell | \gamma = y)$ appearing in (79) is carried out in two stages. For the reported results in Figs. 3a and 3b, the conditional mutual information is estimated on the training split using all 10,000 selected training samples. We also repeated the same analysis on the testing split and obtained the same qualitative conclusions. First, for each agent k, the raw patch vectors are stacked into a data matrix whose rows correspond to samples and whose columns correspond to pixel values, and principal component analysis (PCA)

is then applied to this matrix [11]. A reduced representation of dimension 20 is retained. The PCA transformation is fitted once per agent on the chosen split, and the resulting reduced coordinates are then partitioned according to the class label. Second, for each class $y \in \{-1, +1\}$, the mutual information between the two reduced patch representations is estimated by a k-nearest-neighbor (kNN) estimator with neighborhood parameter $k_{\mathrm{NN}} = 5$ [12]. The two class-conditional estimates are then combined according to (79). All reported values are given in nats. We also evaluated PCA dimensions 10, 20, and 30, together with neighborhood parameters $k_{\mathrm{NN}} = 3$, 5, and 10. Across these choices, the same qualitative conclusions are found.

4) Heterogeneous Models: This subsection provides additional details for the heterogeneous version of the CIFAR-10 patch-partition benchmark in Section IV-A2. We consider three local model families:

1) **Family** A: the patch-level CNN adopted in the main CIFAR-10 experiments;
2) **Family** B: logistic regression on the flattened raw patch pixels;
3) **Family** C: a small residual convolutional network with adaptive pooling [13].

All other aspects of the experiment remain unchanged from the homogeneous CIFAR-10 benchmark, including the local datasets, the train/validation splitting protocol, the communication graph, the combination policy, and the training hyperparameters.

Because different patch locations are not equally informative, the placement of a model family on the 3×3 grid can affect its apparent performance. To reduce this spatial bias, we use a balanced assignment protocol based on the base layout:

$$\mathbf{F}^{(1)} = \begin{array}{|c|c|c|}\hline \mathsf{A} & \mathsf{B} & \mathsf{C} \\ \hline \mathsf{B} & \mathsf{C} & \mathsf{A} \\ \hline \mathsf{C} & \mathsf{A} & \mathsf{B} \\ \hline \end{array} \tag{H.7}$$

where A, B, and C denote the three model families. A cyclic shift of this pattern yields two additional layouts:

$$\mathbf{F}^{(2)} = \begin{array}{|c|c|c|}\hline \mathsf{B} & \mathsf{C} & \mathsf{A} \\ \hline \mathsf{C} & \mathsf{A} & \mathsf{B} \\ \hline \mathsf{A} & \mathsf{B} & \mathsf{C} \\ \hline \end{array}, \quad \mathbf{F}^{(3)} = \begin{array}{|c|c|c|}\hline \mathsf{C} & \mathsf{A} & \mathsf{B} \\ \hline \mathsf{A} & \mathsf{B} & \mathsf{C} \\ \hline \mathsf{B} & \mathsf{C} & \mathsf{A} \\ \hline \end{array}. \tag{H.8}$$

In the study of collaboration gain at a fixed N_0 illustrated by Fig. 4, the values are averaged over these three layouts so that each family appears equally often at each spatial location. In the N_0-sweep of Fig. 5, the base layout $\mathbf{F}^{(1)}$ is kept fixed.

For each agent k, we quantify the performance gain from collaboration at round t by

$$\Delta_k(t) \triangleq P_{k,0} - P_{k,t} \tag{H.9}$$

where we recall that $P_{k,t}$ denotes the probability of error of agent k at communication round t. For the heatmap shown in Fig. 4, we report the corresponding *relative* error reduction at the representative communication round $t = 15$:

$$\Delta_k^{(\%)}(15) \triangleq 100 \cdot \frac{\Delta_k(15)}{\max\{P_{k,0}, \epsilon_{\mathrm{safe}}\}} \tag{H.10}$$

where $\epsilon_{\mathrm{safe}} = 10^{-6}$ is a small numerical safeguard.

2. ModelNet40 Benchmark

The ModelNet40 experiments follow the same general pipeline as the CIFAR-10 benchmark in Appendix H.1, except for the differences described below.

The dataset is constructed from CAD models by rendering multiple views of each object using Blender. For each object, we generate $K = 12$ RGB views from fixed camera positions whose azimuth angles are uniformly spaced around the object, with a common elevation angle of 30°. All rendered images are generated at resolution 224×224 with transparent background.

The local classifier is a lightweight convolutional network, referred to as *MicroCNN*. It consists of a single convolutional layer with kernel size 7×7 and stride 4, followed by a ReLU nonlinearity, global average pooling, an optional dropout layer, and a final linear layer producing two logits. As in the CIFAR-10 experiments, the local score is formed from the difference between two logits and then centered according to (9).

For each value of N_0, we sample a balanced set of objects containing $N_0/2$ examples from each class and split it into training and validation subsets using the same 20% validation fraction as in the CIFAR-10 study. The local models are trained independently using cross-entropy loss and validation-based early stopping, but with SGD optimizer and learning rate 0.05. For the VFL-JT comparison, we use the same score-level joint-training construction described in Appendix H.1.2, with $K = 12$ view-specific MicroCNN branches.

a) Centralized full-observation reference: In the ModelNet40 benchmark, the central oracle has access to all 12 views of the same object. The model consists of K view-specific MicroCNN backbones, one for each rendered view, whose feature vectors are concatenated and passed to a centralized multilayer head for final prediction. The oracle is initialized from the independently trained local models and then trained in two stages. In the first stage, all backbone parameters are frozen and only the fusion head is optimized. In the second stage, all parameters are unfrozen and the full model is fine-tuned jointly using a smaller learning rate. Both training stages use cross-entropy loss, the Adam optimizer, and validation-based early stopping.

1) Temperature Scaling: For each agent k, we denote by $z_k(\boldsymbol{h}_k) \in \mathbb{R}^2$ the two-dimensional output of its local classifier for feature vector $\boldsymbol{h}_k$, whose entries are the logits associated with the two classes. A positive scalar temperature $T_k > 0$ is fitted on the validation set by minimizing the cross-entropy loss of the rescaled logits:

$$\min_{T_k > 0} \frac{1}{N_{\mathrm{val}}} \sum_{n=1}^{N_{\mathrm{val}}} \ell_{\mathrm{CE}}\left(\frac{z_k(\boldsymbol{h}_{k,n})}{T_k}, \gamma_n\right) \tag{H.11}$$

where ℓ_{CE} denotes the cross-entropy loss, and $(\boldsymbol{h}_{k,n}, \gamma_n)$ denotes the n-th labeled example in the validation set. In the implementation, we optimize over $\log T_k$ in order to enforce

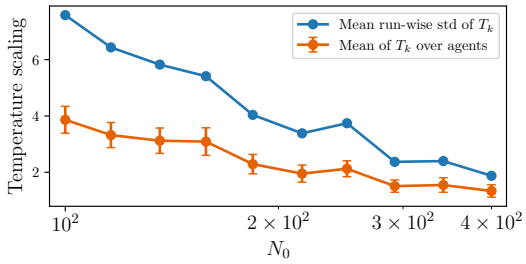


Fig. H.2. Temperature scaling diagnostics for the ModelNet40 benchmark. The orange curve shows the average of T_k over agents, with error bars indicating 95% confidence intervals across Monte Carlo repetitions. The blue curve shows the average run-wise standard deviation of $\{T_k\}_{k=1}^K$, which quantifies the variability of the fitted temperatures across agents.

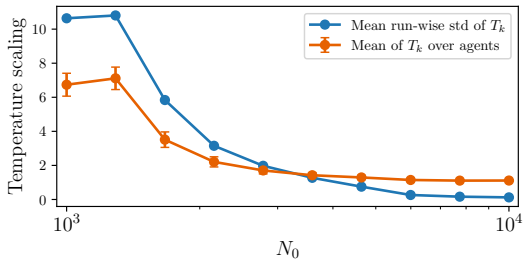
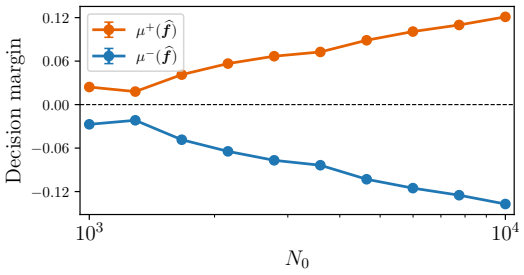
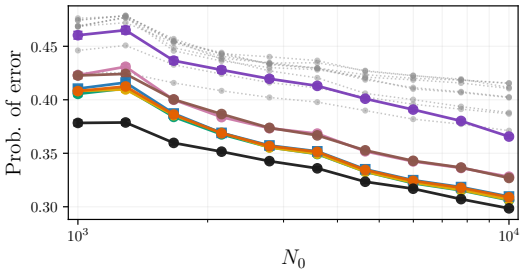


Fig. H.3. Temperature scaling diagnostics for the CIFAR-10 benchmark.



(a) Decision margin



(b) Probability of error

Fig. H.4. Performance of temperature scaling on CIFAR-10.

positivity. After fitting T_k, the classifier output is replaced by $z_k(\boldsymbol{h}_k)/T_k$. The local score used for collaboration is then formed by taking the difference between the two calibrated logits and is subsequently centered according to (9).

a) Additional temperature scaling diagnostics: Figure H.2 summarizes the fitted temperature parameters on the ModelNet40 benchmark across training set sizes N_0. For each value of N_0, we report the average of T_k over the agents together with a 95% confidence interval across 200 Monte Carlo repetitions, as well as the average run-wise standard deviation of $\{T_k\}_{k=1}^K$. A learned temperature $T_k = 1$ leaves the logits unchanged, $T_k > 1$ softens them by reducing their magnitude, and $T_k < 1$ sharpens them. The figure shows that the fitted temperatures are generally greater than one, and that both their average level and their variability across agents decrease as N_0 increases. This trend supports the conclusion that calibration is most critical in data-constrained regimes and becomes more stable as local classifier's reliability improves with additional training data.

Figure H.3 reports the corresponding temperature diagnostics for the CIFAR-10 benchmark considered in Fig. 2. Similar to ModelNet40, the fitted temperatures are larger and more variable at smaller training set sizes. As N_0 increases, however, they move closer to one, indicating that the logits require less rescaling in the higher-data regime. The corresponding results of decision margin and probability of error with temperature scaling are shown in Figs. H.4a and H.4b, respectively. Compared with Figs. 2a and 2b, temperature scaling tends to increase the achieved decision margin and slightly lowers the probability of error, with the effect being most visible at smaller N_0.

2) Correlation Stress Test: This subsection provides additional details for the controlled correlation stress test studied in Section IV-B2. The perturbed decision statistics $\{x'_{i,k}\}$ are generated from the clean centered statistics $\{x_{i,k}\}$ using the perturbation model in (80) and (81). After perturbation, the decision statistics over the test set are collected into a matrix

$$X \in \mathbb{R}^{N_{\text{test}} \times K}, \tag{H.12}$$

whose (i,k)-th entry is the perturbed decision statistic of agent k for the i-th testing sample, where N_{test} denotes the number of testing samples. To quantify the induced dependence, we compute the unconditional average pairwise Pearson correlation across agents:

$$\text{corr}_{\text{uncond}}(X) \triangleq \frac{2}{K(K-1)} \sum_{1 \le k < \ell \le K} \widehat{r}_{k\ell}(X) \tag{H.13}$$

where $\widehat{r}_{k\ell}(X)$ denotes the sample Pearson correlation between columns k and ℓ of X. The label-conditional counterpart is obtained by restricting X to each class separately and then averaging the corresponding pairwise correlations over the two classes.

In addition to these dependence measures, Fig. 9 reports the collaboration gain after t communication rounds, measured by

$$P_0 - P_t = \frac{1}{K} \sum_{k=1}^{K} (P_{k,0} - P_{k,t}) \tag{H.14}$$

where P_0 denotes the *average* probability of error across agents in the non-cooperative setting and P_t is the corresponding average probability of error after t rounds of collaboration at the same perturbation level.

For completeness, Fig. H.5 reports the corresponding non-cooperative and collaborative probabilities of error as functions of the dependence strength r. For a fixed perturbation ampli-

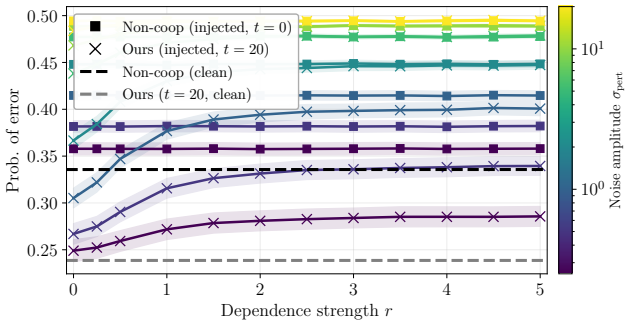
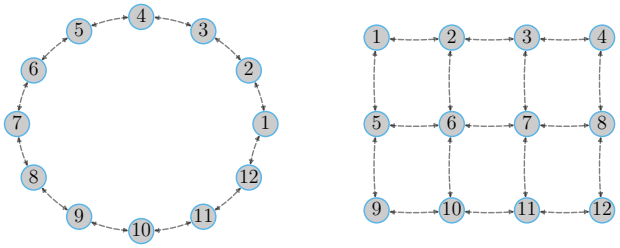


Fig. H.5. Probability of error under the controlled correlation stress test. For each perturbation amplitude σ_{pert}, the figure compares the non-cooperative error at $t = 0$ and the collaborative error after $t = 20$ communication rounds as functions of the dependence strength r. The clean non-cooperative and collaborative error levels are also shown for reference.



(a) Ring (b) Grid

Fig. H.6. Ring and grid communication topologies used in Fig. 10. Self-loops are omitted from the plots for visual clarity.

tude σ_{pert}, varying r changes the balance between the shared and idiosyncratic perturbation components while preserving the marginal perturbation distribution at each agent. This explains why the non-cooperative error remains approximately constant as r varies. In contrast, the collaborative error generally increases with r, since a stronger shared component makes the perturbations across agents more correlated and reduces the benefit of combining their information. Consequently, the gap between the non-cooperative and collaborative curves narrows as r increases, consistent with the reduction in collaboration gain observed in Fig. 9. Increasing σ_{pert}, on the other hand, increases the perturbation magnitude and raises the overall error levels. The clean reference curves show the corresponding performance without the injected perturbations.

3. *Communication and Robustness*

This subsection collects the implementation details for the communication design, adaptive stopping, and robustness studies in Section IV-C of the main paper.

1) Topology and Quantization: This subsection provides additional implementation details for the communication design study in Section IV-C. Figure H.6 illustrates the ring and grid communication topologies used in the experiments.

For each topology-rule pair, the graph and the corresponding combination policy A are generated once and then kept fixed throughout the Monte Carlo study. In this study, we choose $N_0 = 400$, which corresponds to all available training samples for the selected class pair. The training and testing sets are therefore kept fixed across repetitions, while the train/validation split and the training randomness are varied. For each finite-bit condition, the results are additionally averaged over 30 independent quantizer realizations within each of the 200 training repetitions. In this way, the comparison isolates the effect of the communication design from variability due to changes in the underlying graph or data.

For the Metropolis rule used in the simulations, the communication weights are defined by

$$a_{\ell k} = \begin{cases} \frac{1}{\max\{|\mathcal{N}_k|,|\mathcal{N}_\ell|\}}, & \text{if } \ell \neq k, \ell \in \mathcal{N}_k, \\ 1 - \sum_{m \in \mathcal{N}_k \setminus \{k\}} a_{mk}, & \text{if } \ell = k, \\ 0, & \text{otherwise.} \end{cases} \tag{H.15}$$

For the undirected ring and grid topologies, this is the standard Metropolis rule in the literature [14], which ensures that the resulting combination policy A is doubly-stochastic. In the case of the directed Erdös–Rényi topology, the inclusion of self-loops at all agents guarantees that A is well defined and left-stochastic by construction (H.15). However, because the graph is directed, the policy does not generally remain doubly-stochastic.

For the finite-precision study, we use the bounded posterior-probability difference $g_k(h)$ defined in (82). This statistic is obtained directly from the same trained classifier, and no retraining is performed. Its empirical training mean $\mu_{k,\mathsf{emp}}(g_k)$ is recomputed from the training statistics in this representation, so that the centered statistic lies in $[-2, 2]$. For $b \in \{4, 6, 8\}$, we use $L_b = 2^b$ uniformly spaced reconstruction levels on $[-2, 2]$, with step size

$$\Delta_b = \frac{4}{2^b - 1}. \tag{H.16}$$

The adjacent-level stochastic rounding protocol Q_b and the quantize-before-average recursion (60) are those defined in Section III-D. Quantization is applied at each sender before averaging, including the first exchange. For each testing sample and communication round, every sender draws one fresh quantization outcome independently across senders, rounds, and samples, and broadcasts the same quantized value to all of its outgoing neighbors. Since both stochastic quantization and left-stochastic averaging preserve the interval $[-2, 2]$, no overload clipping is required.

The matched full-precision reference uses the same trained classifiers, bounded scores, empirical centers, graph, and combination policy, but replaces the stochastic quantizer by the identity map. Thus, the finite-bit branches differ from the full-precision reference only through stochastic quantization.

For the finite-bit branches, the per-round communication cost is measured in transmitted bits per testing sample:

$$B_{\text{round}} = bE_{\text{off}}(A) \tag{H.17}$$

where $E_{\text{off}}(A)$ is defined in (66) and excludes self-loops from the communication count.

a) Additional discussion of Fig. 10: Figures H.7–H.9 provide alternative views of the results in Fig. 10, organized according to graph family, the construction rule of A, and

communication precision. These views help separate the role of each factor in collaborative classification. Across the graph families considered here, the uniform averaging rule and the Metropolis rule generally produce different combination matrices A. One exception is the ring graph, for which the two rules generate the same matrix A. Accordingly, Fig. H.7a presents coincident trajectories at all tested precisions. For the grid graph, Fig. H.7b shows that the uniform averaging rule performs better than the Metropolis rule, although the gap becomes small in the heavily quantized (4-bit) regime. For the tested directed Erdös–Rényi graph, Fig. H.7c shows that uniform averaging performs better than the Metropolis rule under the matched full-precision reference and all three finite-bit settings.

Figure H.8 relates these performance differences to the spectral properties of the corresponding combination matrices. In our experiments, across both construction rules, the second-largest eigenvalue magnitude σ_A is largest for the ring graph, intermediate for the grid, and smallest for the directed Erdös–Rényi graph. Since σ_A governs the speed of linear mixing associated with the local decision statistics in (11), this ordering helps explain the convergence behaviors observed in Figs. H.8a and H.8b: the ring trajectories approach their limiting regime more slowly than those of the grid or Erdös–Rényi graphs. This spectral comparison concerns the transient mixing behavior. The classification performance also depends on the Perron vector of A and the distribution of heterogeneous local scores.

The precision-specific views in Fig. H.9 further clarify the role of quantization. Since the ring and grid graphs are undirected, the Metropolis rule yields doubly-stochastic matrices for both graph families, and hence the same uniform Perron vector. Therefore, in the matched bounded full-precision setting, their decision statistics converge to the same limit. This is reflected in Fig. H.9a, where the corresponding curves approach nearly the same long-run error level. However, the ring graph exhibits a slower transient due to its larger σ_A. As the communication precision decreases, this coincidence is no longer exact. The finite-bit results exhibit a gradual precision hierarchy: the 8-bit trajectories remain close to full precision, 6-bit communication produces an intermediate degradation, and 4-bit communication produces the largest loss. Finally, although the ring graph converges more slowly in the early rounds, it attains a lower long-run error than the directed Erdös–Rényi graph in these experiments. In particular, the two graphs have the same number of communication links, and hence the same per-round communication cost at a fixed bit-width following (H.17). This comparison further illustrates the importance of graph topology in collaborative inference.

2) Adaptive Stopping: This subsection provides additional details for the adaptive stopping study in Section IV-C2. Our implementation follows a local event-triggered communication protocol. For the i-th testing sample and each agent k, let $x_{i,k}^{\mathrm{sent}}(t)$ denote the last value that agent k has transmitted up to round t. At the next round, agent k computes a candidate statistic given by

$$x_{i,k}^{\mathrm{new}}(t) = \sum_{\ell \in \mathcal{N}_k} a_{\ell k}\, x_{i,\ell}^{\mathrm{sent}}(t-1). \tag{H.18}$$

At initialization, we set $x_{i,k}^{\mathrm{sent}}(0) = x_{i,k}(0)$, where $x_{i,k}(0)$ is the local decision statistic before communication. The stopping decision is then made from $x_{i,k}^{\mathrm{new}}(t)$ together with the agent's local memory. If the trigger fires, the stored communicated value is updated; otherwise the previous transmitted value is kept. Therefore, an agent that remains silent at one round still computes a fresh candidate statistic and may become active again later if the trigger condition is violated.

We investigate two local stopping-rule families. The first is a Δ-*trigger* rule, which is based directly on the change in the communicated statistic. Under this rule, agent k transmits the candidate statistic $x_{i,k}^{\mathrm{new}}(t)$ to its neighbors at round t only if

$$\left|x_{i,k}^{\mathrm{new}}(t) - x_{i,k}^{\mathrm{sent}}(t-1)\right| > \varepsilon_{\mathrm{send}} \tag{H.19}$$

where $\varepsilon_{\mathrm{send}} > 0$ denotes the threshold parameter. The second is a *label-stability with confidence* rule. Let

$$y_{i,k}^{\mathrm{new}}(t) = \mathrm{sign}\big(x_{i,k}^{\mathrm{new}}(t)\big) \tag{H.20}$$

be the label induced by the candidate statistic, and let $s_{i,k}(t)$ denote the number of consecutive rounds up to round t over which this induced label has remained unchanged. In the implementation, the counter is updated by comparing the induced labels at two consecutive rounds. Thus, $L = 1$ corresponds to one unchanged transition, $L = 2$ to two consecutive unchanged transitions, and so on. Agent k does not transmit sample i at round t when

$$s_{i,k}(t) \geq L, \quad |x_{i,k}^{\mathrm{new}}(t)| \geq \tau_{\mathrm{conf}} \tag{H.21}$$

where L is a patience parameter and τ_{conf} is a confidence threshold.

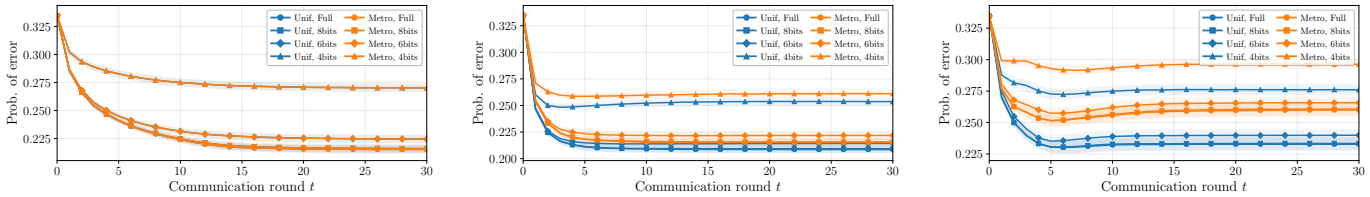
To evaluate the performance of the two stopping rules, we sweep $\varepsilon_{\mathrm{send}}$ over a dense grid for the Δ-trigger rule, and sweep $(L, \tau_{\mathrm{conf}})$ over dense grids for the label-stability rule. Since transmissions may stop at different times for different agents and different samples, the communication cost for each rule is measured by the total number of transmissions per testing sample, obtained by counting the realized transmissions over directed non-self links across the communication rounds. The Pareto-efficient operating points shown in Fig. 11 are then extracted in the plane of probability of error versus the total number of transmissions.

3) Robustness: This subsection provides additional details for the robustness study in Section IV-C3. For the i-th testing sample, let $x_{i,k}(0)$ denote the local decision statistic of agent k before communication. The collaborative recursion then evolves according to

$$x_{i,k}(t) = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} x_{i,\ell}(t-1). \tag{H.22}$$

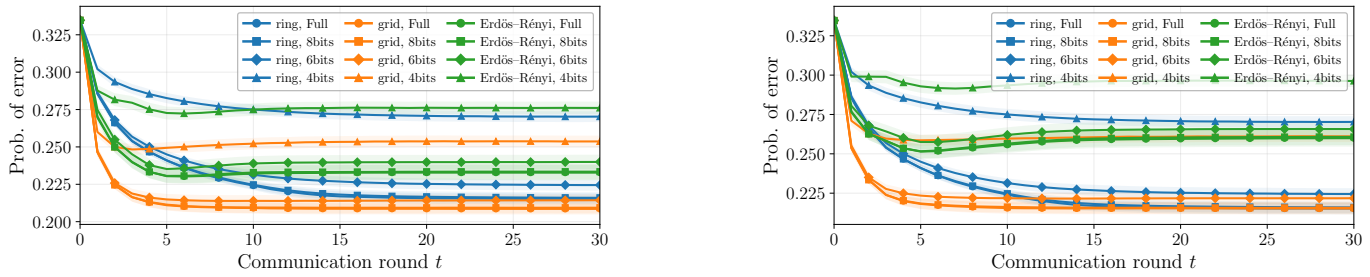
For each perturbation scenario, we replace the initial statistics $\{x_{i,k}(0)\}$ by corrupted values $\{x_{i,k}^{\mathrm{corr}}(0)\}$ and compare the resulting performance with the corresponding clean trajectory.

In each repetition, $m \in \{1, \ldots, K\}$ agents are corrupted,



(a) Ring topology (b) Grid topology (c) Directed Erdös–Rényi topology

Fig. H.7. Effect of the construction rule of A and quantization on classification performance under different graph topologies.



(a) Uniform averaging rule (b) Metropolis rule

Fig. H.8. Effect of graph topology and quantization on classification performance under different construction rules of A.

chosen uniformly without replacement, and kept fixed throughout that repetition. We consider three perturbation families.

a) Benign noisy agents: For each corrupted agent k, independent Gaussian perturbations are injected into the initial statistics:

$$x_{i,k}^{\mathrm{corr}}(0) = x_{i,k}(0) + \alpha_{\mathrm{noise}}\zeta_{i,k}, \quad \zeta_{i,k} \sim \mathcal{N}(0,1). \tag{H.23}$$

To normalize the perturbation level α_{noise}, we first compute the sample standard deviation of $\{x_{i,k}(0)\}_{i=1}^{N_{\mathrm{test}}}$ for each agent k, and then take the median of these values across agents. Denoting the resulting scale estimate by $\widehat{s}$, we set

$$\alpha_{\mathrm{noise}} = \eta_{\mathrm{noise}}\widehat{s}, \tag{H.24}$$

with a dimensionless multiplier η_{noise}.

b) Faulty agents: We consider both stuck-at-zero failures and constant-bias failures. In the stuck-at-zero model,

$$x_{i,k}^{\mathrm{corr}}(0) = 0, \tag{H.25}$$

whereas in the constant-bias model,

$$x_{i,k}^{\mathrm{corr}}(0) = x_{i,k}(0) + b_{\mathrm{fault}}, \quad b_{\mathrm{fault}} = \eta_{\mathrm{bias}}\widehat{s}. \tag{H.26}$$

Here, η_{bias} is the dimensionless bias multiplier varied in the experiments.

c) Adversarial reports: For each corrupted agent k, we apply a scaled sign flip,

$$x_{i,k}^{\mathrm{corr}}(0) = -\eta_{\mathrm{adv}}x_{i,k}(0), \quad \eta_{\mathrm{adv}} \geq 0. \tag{H.27}$$

This model includes several special cases: $\eta_{\mathrm{adv}} = 0$ corresponds to silencing, $\eta_{\mathrm{adv}} = 1$ to a pure sign flip, and $\eta_{\mathrm{adv}} > 1$ to an amplified wrong-sign report.

REFERENCES

[1] P. Hu, V. Bordignon, M. Kayaalp, and A. H. Sayed, "Non-asymptotic performance of social machine learning under limited data," *Signal Processing*, vol. 230, p. 109849, 2025.

[2] V. Bordignon, S. Vlaski, V. Matta, and A. H. Sayed, "Learning from heterogeneous data based on social interactions over graphs," *IEEE Transactions on Information Theory*, vol. 69, no. 5, pp. 3347–3371, 2023.

[3] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 463–482, 2002.

[4] B. Neyshabur, R. Tomioka, and N. Srebro, "Norm-based capacity control in neural networks," in *Conference on Learning Theory*. PMLR, 2015, pp. 1376–1401.

[5] S. Boucheron, G. Lugosi, and P. Massart, "Concentration inequalities using the entropy method," *The Annals of Probability*, vol. 31, no. 3, pp. 1583–1614, 2003.

[6] P. Caputo, G. Menz, and P. Tetali, "Approximate tensorization of entropy at high temperature," *Annales de la Faculté des Sciences de Toulouse: Mathématiques*, vol. 24, no. 4, pp. 691–716, 2015.

[7] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. MIT press, 2018.

[8] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics, 1989*, ser. London Mathematical Society Lecture Note Series, J. Siemons, Ed. Cambridge University Press, 1989, vol. 141, pp. 148–188.

[9] 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.

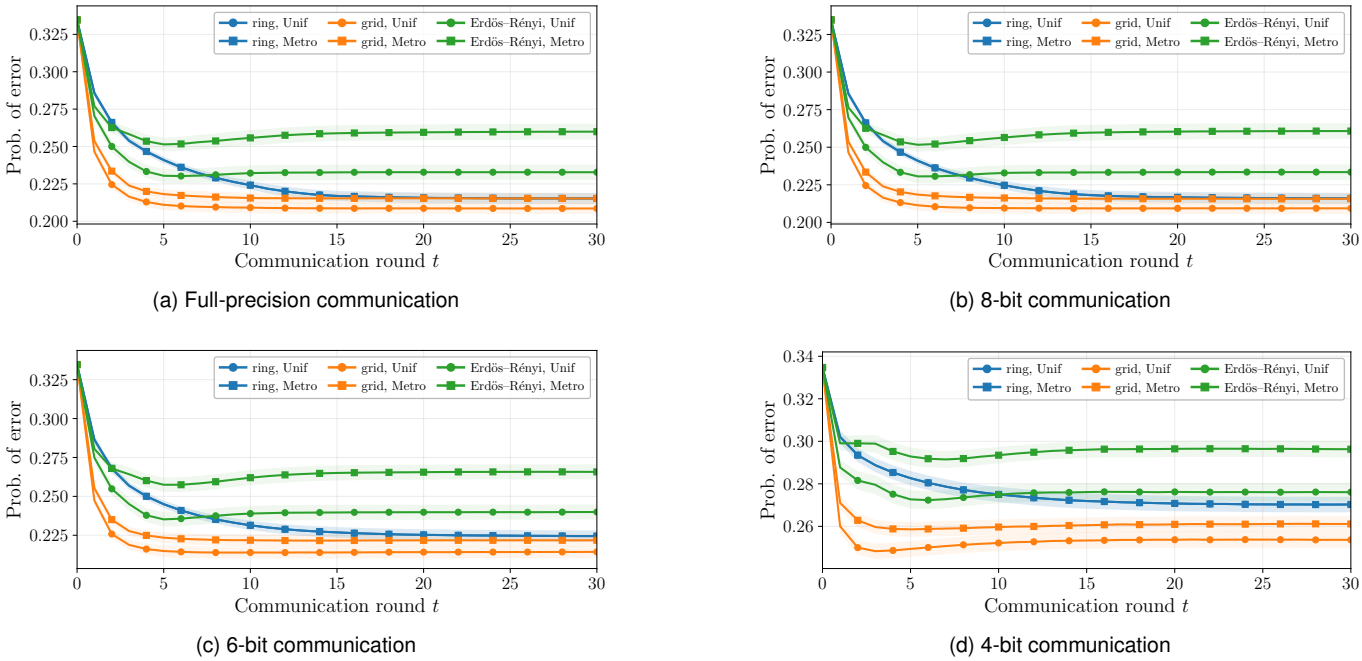
[10] Y. Hu, D. Niu, J. Yang, and S. Zhou, "FDML: A collaborative machine learning framework for distributed features," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2019, pp. 2232–2240.

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

[12] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[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.



(a) Full-precision communication

(b) 8-bit communication

(c) 6-bit communication

(d) 4-bit communication

Fig. H.9. Effect of graph topology and the construction rule of A under different communication precisions.