Social Learning with Non-Bayesian Local Updates

Virginia Bordignon
EPFL
Lausanne, Switzerland
virginia.bordignon@epfl.ch

Mert Kayaalp
EPFL
Lausanne, Switzerland
mert.kayaalp@epfl.ch

Vincenzo Matta
University of Salerno
Fisciano, Italy
vmatta@unisa.it

Ali H. Sayed
EPFL
Lausanne, Switzerland
ali.sayed@epfl.ch

***Abstract*—In *non-Bayesian* social learning, the agents of a network form their belief about a hypothesis of interest by performing individual Bayesian updates, which are then shared with their neighbors and aggregated according to a suitable pooling rule. This social learning scheme is called non-Bayesian because the pooling rule cannot be Bayesian owing to the limitations arising from the distributed learning setting. However, traditional non-Bayesian learning relies on using a local Bayesian update rule. In this work, we move away from this assumption and consider instead non-Bayesian learning with *non-Bayesian updates*. Taking as a benchmark the optimal centralized posterior, we show that this modified strategy can outperform traditional social learning and that, intriguingly, it can attain the same error exponent as the optimal scheme under two opposite scenarios: when the data are independent across the agents and when there are agents with highly dependent data.**

***Index Terms*—Social learning, Bayesian update, Large deviations, Opinion formation, Distributed decision-making.**

I. Introduction and Related Work

The learning problem addressed in this work concerns the formation of opinions about some hypothesis of interest θ belonging to a discrete finite set Θ, based on the observation of a stream of data [1], [2]. Data are observed by a group of spatially dispersed agents that are allowed to exchange their beliefs according to a communication graph dictated by a certain network topology [3]. In the ideal case where $i)$ all data are available to all agents and $ii)$ the joint distribution of the data is available, the solution is well known. Specifically, the optimal belief is the posterior probability computed by means of Bayes' rule. Once this posterior is evaluated, one can then choose the hypothesis that results from the maximum-a-posteriori (MAP) rule, which is known to minimize the probability of erroneous decisions [4].

In practice, the aforementioned two conditions are seldom verified. Often, data are not only dispersed, but also statistically dependent across agents, and there is no knowledge about their joint distribution. Each agent has access to only the marginal distribution characterizing its own data. Furthermore, since communication is permitted only between neighbors, at each communication round the individual agents would have access to information arising from only a portion of the overall distribution, even under independence. Under these limitations, determining the optimal fully Bayesian solution is not feasible [5], thus motivating the emergence of the *non-Bayesian* social learning paradigm [6]–[10].

We will denote by $\boldsymbol{\xi}_{k,i} \in \mathbb{R}^{d_k}$ (we use bold font for random quantities) the streaming data arriving at agent $k = 1, 2, \dots, N$ at time $i \in \mathbb{N}$. Note that the data can have different dimensions d_k across the agents. The goal of each agent k is to construct a *belief vector* $\boldsymbol{\mu}_{k,i}$ at each time i, namely, a probability vector where $\boldsymbol{\mu}_{k,i}(\theta)$ is the probability assigned to hypothesis θ, and satisfying $\sum_{\theta\in\Theta} \boldsymbol{\mu}_{k,i}(\theta) = 1$. The data $\{\boldsymbol{\xi}_{k,i}\}_{k=1}^{N}$ are assumed independent and identically distributed over time. We assume each agent k has only access to a *local* likelihood $L_k(\xi|\theta)$, which is the *marginal* probability (density or mass) function of $\boldsymbol{\xi}_{k,i}$ given the hypothesis θ. Accordingly, if the true underlying hypothesis is denoted by θ_0, then the true marginal distribution governing the data $\boldsymbol{\xi}_{k,i}$ is $L_k(\xi|\theta_0)$.

Non-Bayesian social learning is composed of the following two steps. First, a *self-learning* step, where each agent constructs an *intermediate* belief vector $\boldsymbol{\psi}_{k,i}$ by updating the previous-lag belief vector $\boldsymbol{\mu}_{k,i-1}$ with the fresh information $\boldsymbol{\xi}_{k,i}$. In this first step, social learning is *locally Bayesian*, since it implements the *Bayesian update* rule:

$$\boldsymbol{\psi}_{k,i}(\theta) \propto \boldsymbol{\mu}_{k,i-1}(\theta) L_k(\boldsymbol{\xi}_{k,i}|\theta), \tag{1}$$

where the symbol $\propto$ denotes the proportionality constant that is necessary to make $\boldsymbol{\psi}_{k,i}$ a probability vector.

Subsequently, all agents share with their neighbors the intermediate beliefs, and there is a *cooperation* step, where each agent constructs the updated belief $\boldsymbol{\mu}_{k,i}$ by combining the beliefs received from the neighbors. Different pooling rules have been considered in the literature [6]–[11]. One strategy is *geometric averaging*, which is optimal under the minimization of a suitable weighted Kullback-Leibler divergence [12], and under appropriate behavioral constraints [13]. This pooling rule is given by:

$$\boldsymbol{\mu}_{k,i}(\theta) \propto \prod_{\ell=1}^{N} \boldsymbol{\psi}_{k,i}^{a_{\ell k}}(\theta), \tag{2}$$

where the combination weights $a_{\ell k}$ are convex, i.e., they are nonnegative and satisfy $\sum_{\ell=1}^{N} a_{\ell k} = 1$. Storing these weights into the combination matrix $A = [a_{\ell k}]$, we see that A is left stochastic and its support graph describes the effective communication topology between the agents.

It was established in previous works that the non-Bayesian social learning strategy learns consistently the truth (meaning that the belief mass placed on the true underlying hypothesis converges almost surely to 1 as $i \to \infty$) under mild regularity

This work was supported in part by SNSF grant 205121-184999.

assumptions on the likelihood functions and over strongly connected networks [9], [10], [14].

However, less attention has been paid in earlier works to the comparison of the opinions resulting from social learning with those arising from the ideal Bayesian posterior. The recent work [15] addresses the problem of tracking the Bayesian posterior under a hidden-Markov-model (HMM) where the underlying hypothesis varies at each social learning round. Here, in this manuscript, we address instead the comparison with the Bayesian posterior in terms of error probabilities in the opinion formation process, and under the standard setting adopted in social learning where the hypothesis is fixed over time. We provide the following main contributions.

First, we propose to replace the Bayesian update (1) with the following *non-Bayesian* update:

$$\boldsymbol{\psi}_{k,i}(\theta) \propto \boldsymbol{\mu}_{k,i-1}(\theta) L_k^{\gamma_k}(\boldsymbol{\xi}_{k,i}|\theta), \tag{3}$$

which raises the likelihood to a positive number γ_k. The resulting strategy will be referred to as non-Bayesian social learning with non-Bayesian update, and abbreviated as the NBNB (or NB^2) strategy. Note that the NB^2 strategy is equivalent to traditional non-Bayesian learning when we set $\gamma_k = 1$ for all $k = 1, 2, \dots, N$. In the context of distributed Bayesian filtering [15], [16] the idea of raising the likelihood to some agent-independent value $\gamma_k = \gamma$ was exploited to track the centralized Bayesian posterior. However, in tracking the Bayesian posterior an agent-independent value $\gamma_k = \gamma$ can be helpful only with doubly stochastic matrices, while in terms of error probabilities, we will see that an agent-independent value is not helpful. In particular, we will establish useful connections between agent-dependent values γ_k and attributes of the learning problem, such as the network topology and the dependence structure among the agents.

In [11] other update rules are proposed, whose critical feature is that the previous-lag belief $\boldsymbol{\mu}_{k,i-1}$ is raised to some positive number. Exponentiating the belief instead of the likelihood has a fundamentally different goal than the one considered in this work, namely, it is useful to infuse the social learning algorithm with *adaptation* [11].

We establish that the NB^2 strategy can outperform traditional non-Bayesian social learning. In particular, we first show that, when the data are independent across agents, the error probability decays exponentially fast to zero as $i \to \infty$, with an error exponent that is equal to that of the optimal Bayesian posterior for *doubly stochastic* combination matrices. In contrast, traditional social learning achieves a suboptimal error exponent when the combination matrix is *left stochastic* (and not doubly stochastic),[1] a setting that plays an important role in practice, especially over *directed* graphs, where it can be difficult to construct a doubly stochastic combination matrix. In this work we establish that the NB^2 strategy is able to attain the exponent of the optimal Bayesian posterior even with left stochastic matrices, when each agent knows its own Perron eigenvector entry.

Second, we examine the case where clusters of agents feature highly dependent data. For this case, we show that traditional non-Bayesian learning is suboptimal, whereas the NB^2 strategy can recover the optimal error exponent of the Bayesian posterior, provided that the clusters are known and that the likelihoods in each cluster are properly discounted so that the data in each cluster are counted only once in the learning process.

In this work we evaluate the performance of the NB^2 strategy by resorting to the theory of large deviations [18], [19]. A large-deviations analysis of distributed decision-making schemes was proposed in [10], [11], [20]–[22], under settings other than ours. In [20], the focus is on binary hypothesis testing, identical distribution across agents, and combination matrices that are doubly stochastic, symmetric, and random. The characterization in [10] is relative to traditional social learning. In [11], [21], [22], large-deviations theory was applied to adaptive learning schemes to characterize the exponential decay of the error probability as a function of the adaptation parameter [11], which is a problem different from the one addressed here.

II. Assumptions

Assumption 1 (Strongly Connected Network). *Given any pair of nodes (ℓ, k), paths with nonzero weights exist in both directions, i.e., from ℓ to k and vice versa (the two paths need not be the same), and at least one agent k in the entire network has a positive self-weight ($a_{kk} > 0$).* □

Strong connectivity implies that the left stochastic matrix A is primitive [3], [4]. Then, the Perron-Frobenius theorem implies the existence of a vector $v = [v_k]$, a.k.a. the Perron eigenvector, which satisfies the following conditions [3]:

$$Av = v, \qquad \sum_{k=1}^{N} v_k = 1, \qquad v_k > 0 \quad \forall k. \tag{4}$$

Assumption 2 (Positive Initial Beliefs). *For each $k = 1, 2, \dots, N$ and each $\theta \in \Theta$, $\mu_{k,0}(\theta) > 0$.* □

It can happen that hypotheses θ_0 and θ are indistinguishable to agent k, i.e., the distributions $L_k(\xi|\theta_0)$ and $L_k(\xi|\theta)$ are the same. Agents with indistinguishable hypotheses are unable to learn on their own. In social learning, agents can overcome *local* unidentifiability through collaboration. In particular, if agent k is able to distinguish one hypothesis θ from the true hypothesis θ_0, then we will see that this ability can diffuse across the network. For this to be possible, we resort to the standard *global* identifiability assumption.

Assumption 3 (Global Identifiability). *For every pair (θ_0, θ) there exists at least one agent k with distinct distributions $L_k(\xi|\theta_0)$ and $L_k(\xi|\theta)$.* □

Note that identifiability is formulated with reference to any true hypothesis θ_0. This is because we do not know which one

[1] The superiority of doubly stochastic matrices was shown in [17] in adaptive social learning, in terms of the error exponent that characterizes the decay of the error probability as a function of the adaptation parameter. Results for the standard non-adaptive setting that we address here are missing.

is the true hypothesis, and in a classification problem we want to be able to classify any hypothesis.

III. Performance Analysis

As an error measure, we evaluate the probability that the maximum belief is *not* located at the true hypothesis:

$$p_{k,i,\theta_0} = \mathbb{P}[\boldsymbol{\mu}_{k,i}(\theta) \geq \boldsymbol{\mu}_{k,i}(\theta_0), \text{ for some } \theta \neq \theta_0]. \tag{5}$$

A detailed characterization of the error probability is a formidable task. Thus, we focus on an asymptotic analysis aimed at establishing that the error probability vanishes exponentially fast as $i \to \infty$. Once exponential decay is established, a compact descriptor of the learning performance is provided by the *error exponent*:

$$\mathscr{E}_{k,\theta_0} \triangleq \lim_{i\to\infty} -\frac{\log p_{k,i,\theta_0}}{i} \Leftrightarrow p_{k,i,\theta_0} = \exp\left\{-i\mathscr{E}_{k,\theta_0} + o(i)\right\}, \tag{6}$$

where $o(i)$ is a quantity such that $\lim_{i\to\infty} o(i)/i = 0$. Note that in general the exponent depends on the particular agent and true hypothesis.

Before addressing the exponential characterization of the error probability, we need to introduce some relevant quantities. First, we introduce the log-moment-generating-function (LMGF) of the log-likelihood ratio of agent k between hypotheses θ_0 and θ, with $\theta \neq \theta_0$:

$$\Lambda_{k,\theta_0\theta}(t) \triangleq \log \mathbb{E} \exp\left\{ t \log \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta_0)}{L_k(\boldsymbol{\xi}_{k,i}|\theta)} \right\}. \tag{7}$$

A random variable playing a fundamental role in our analysis is the following weighted average:

$$\boldsymbol{\lambda}_{i,\theta_0\theta} \triangleq \sum_{\ell=1}^{N} v_\ell \gamma_\ell \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)}. \tag{8}$$

The next theorem will be proved under the assumption that the data are independent across agents. Since the LMGF is additive for independent random variables, the LMGF of the average variable $\boldsymbol{\lambda}_{i,\theta_0\theta}$ is given by:

$$\Lambda_{\theta_0\theta}(t) \triangleq \log \mathbb{E} \exp\left\{ t \boldsymbol{\lambda}_{i,\theta_0\theta} \right\} = \sum_{\ell=1}^{N} \Lambda_{\ell,\theta_0\theta}(v_\ell \gamma_\ell t). \tag{9}$$

We also introduce the *Fenchel-Legendre transform* of $\Lambda_{\theta_0\theta}(t)$:

$$\Lambda^\star_{\theta_0\theta}(x) = \sup_{t\in\mathbb{R}} \Big[tx - \Lambda_{\theta_0\theta}(t) \Big], \tag{10}$$

which in the context of large deviations is referred to as the *rate function* [18], [19].

Theorem 1 (Error Exponents). *Consider Assumptions 1– 3. Assume that the data are independent across agents and that, for all θ_0, θ, with $\theta \neq \theta_0$:*

$$\Lambda_{\theta_0\theta}(t) < \infty, \quad \text{for all } t \in \mathbb{R}. \tag{11}$$

Then, for the NB^2 strategy, the error probabilities $p_{k,i}$ of all agents vanish as $i \to \infty$ with one and the same error exponent, which is given by:

$$\mathscr{E}_{\theta_0} \triangleq \min_{\theta\in\Theta} \Lambda^\star_{\theta_0\theta}(0) > 0. \tag{12}$$

Furthermore, this exponent coincides with the exponent of the centralized Bayesian posterior if the update parameters are chosen as:

$$\gamma_k \propto \frac{1}{v_k}. \tag{13}$$

Proof: Due to space limitations, we offer a sketch of the proof. By unfolding the recursion arising from repeated application of (3) and (2), we arrive at the equality:

$$\frac{1}{i} \log \frac{\boldsymbol{\mu}_{k,i}(\theta_0)}{\boldsymbol{\mu}_{k,i}(\theta)} = \sum_{\ell=1}^{N} \frac{1}{i} \sum_{n=1}^{i} [A^{n+1}]_{\ell k} \gamma_\ell \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,n}|\theta_0)}{L_\ell(\boldsymbol{\xi}_{\ell,n}|\theta)}, \tag{14}$$

where, for simplicity, we assumed uniform initial beliefs for all agents (it can be seen that the additional transient term arising from a non-uniform assignment will not change the claim of the theorem). Since the data are independent over time and across agents, the LMGF of the random variable defined on the LHS of (14) is equal to:

$$\begin{aligned}\Lambda_{i,\theta_0\theta}(t) &\triangleq \log \mathbb{E} \exp\left\{ \frac{t}{i} \log \frac{\boldsymbol{\mu}_{k,i}(\theta_0)}{\boldsymbol{\mu}_{k,i}(\theta)} \right\} \\ &= \sum_{\ell=1}^{N} \sum_{n=1}^{i} \Lambda_{\ell,\theta_0\theta}([A^{n+1}]_{\ell k} \gamma_\ell t / i). \end{aligned} \tag{15}$$

If we show that

$$\lim_{i\to\infty} \frac{1}{i} \Lambda_{i,\theta_0\theta}(it) = \Lambda_{\theta_0\theta}(t), \tag{16}$$

then we can call upon the Gärtner-Ellis theorem, which implies the following Large Deviations Principle [18], [19]:

$$\lim_{i\to\infty} \log \mathbb{P}\left[\frac{1}{i} \log \frac{\boldsymbol{\mu}_{\ell,i}(\theta_0)}{\boldsymbol{\mu}_{\ell,i}(\theta)} \leq 0 \right] = -\inf_{x\leq 0} \Lambda^\star_{\theta_0\theta}(x) = -\Lambda^\star_{\theta_0\theta}(0), \tag{17}$$

where the last equality can be proved by exploiting the convexity properties of $\Lambda^\star_{\theta_0\theta}(x)$. Once we have the exponential characterization provided by (17), it can be shown by union-bound arguments that the exponent corresponding to the probability of error in (5) is given by the worst-case exponent $\mathscr{E}_{\theta_0}$ in (12). It remains to show (16). We can write:

$$\begin{aligned} &\frac{1}{i} \sum_{n=1}^{i} \Lambda_{\ell,\theta_0\theta}([A^{n+1}]_{\ell k} \gamma_\ell t) = \Lambda_{\ell,\theta_0\theta}(v_\ell \gamma_\ell t) \\ &+ \frac{1}{i} \sum_{n=1}^{i} \Big(\Lambda_{\ell,\theta_0\theta}([A^{n+1}]_{\ell k} \gamma_\ell t) - \Lambda_{\ell,\theta_0\theta}(v_\ell \gamma_\ell t) \Big). \end{aligned} \tag{18}$$

The first term on the RHS of (18) is the desired one, and hence it suffices to show that the second term vanishes as $i \to \infty$. This is true by the convergence of Cesáro means [23], since the individual summands in this second term vanish in view of the Perron-Frobenius theorem and the fact that the function $\Lambda_{\ell,\theta_0\theta}$ is continuous in $\mathbb{R}$ thanks to (11). The fact that (13) yields the exponent of the Bayesian posterior can be checked similarly. Finally, we observe that we have not proved that $\mathscr{E}_{\theta_0} > 0$. This property can be shown, with some additional steps, by starting from the observation that $\Lambda^\star_{\theta_0\theta}(x) = 0$ if, and only if, $x = \mathbb{E}\boldsymbol{\lambda}_{i,\theta_0\theta}$, and that this expectation is positive in view of global identifiability. ■

	Likelihood: $L_k(\xi\|\theta)$		
Agent k	$\theta = 1$	$\theta = 2$	$\theta = 3$
$1-3$	$g_1(\xi)$	$g_1(\xi)$	$g_3(\xi)$
$4-6$	$g_1(\xi)$	$g_3(\xi)$	$g_3(\xi)$
$7-10$	$g_1(\xi)$	$g_2(\xi)$	$g_1(\xi)$

TABLE I
IDENTIFIABILITY SETUP FOR THE PROBLEM IN THE LEFT PLOT OF FIG. 1.

To understand the proportionality sign in (13), observe that scaling γ_k by a constant amounts to scaling $\boldsymbol{\lambda}_{i,\theta_0\theta}$ by a constant, which is immaterial in terms of error exponents. For doubly stochastic matrices ($v_k = 1/N$), condition (13) becomes $\gamma_k \propto N$. This means that, for doubly stochastic matrices, both the NB^2 strategy and traditional social learning ($\gamma_k = 1$) achieve the exponent of the Bayesian posterior.

The situation changes for left stochastic matrices, for which the Bayesian error exponent is not achieved in general by traditional social learning, and is achieved by the NB^2 strategy with the choice in (13). The Perron eigenvector, which is necessary to apply (13), can be estimated by means of the following distributed consensus protocol [24]. Assume each agent k is initialized with a certain value $w_{k,0}$. Let us stack these values into the $N \times 1$ vector $w_0 = [w_{1,0}, w_{2,0}, \dots, w_{N,0}]^\top$ and apply recursively the updates $w_{k,i} = \sum_{\ell=1}^{N} a_{\ell k} w_{\ell,i-1}$ for all agents $k = 1, 2, \dots, N$. In terms of the vectors $w_i = [w_{1,i}, w_{2,i}, \dots, w_{N,i}]^\top$ we obtain:

$$w_i = A^\top w_{i-1} \Leftrightarrow w_i = (A^\top)^i w_0 \xrightarrow{i\to\infty} \sum_{\ell=1}^{N} v_\ell w_{\ell,0}, \tag{19}$$

with the convergence following from the Perron-Frobenius theorem. If agent k is initialized with value 1 and all other agents with 0, all agents converge exponentially fast to v_k, thus learning in a few rounds the kth Perron eigenvector entry.

IV. ILLUSTRATIVE EXAMPLES

In the left plot of Fig. 1 we consider the following problem. We have a family of Laplace probability density functions with unit scale parameter and three different means, namely,

$$g_m(\xi) = \frac{1}{2} e^{-|\xi - 0.1\,m|}, \qquad \text{for } m \in \{1, 2, 3\}. \tag{20}$$

The distributions of the agents are chosen from among these Laplace densities, in the specific way reported in Table I, which corresponds to a globally (but not locally) identifiable problem. We consider the network topology displayed in Fig. 1, equipped with a left stochastic combination matrix obtained through the uniform averaging rule [3].

To show a unique error measure in a single plot, we consider the error probability averaged over all agents and hypotheses:

$$p_i \triangleq \frac{1}{N|\Theta|} \sum_{k=1}^{N} \sum_{\theta_0 \in \Theta} p_{k,i,\theta_0}, \tag{21}$$

whose exponent can be proved to be the worst-case (i.e., the minimum) $\mathscr{E}_{\theta_0}$ across all θ_0. The left plot in Fig. 1 shows the average error probability as a function of time for: *i*) the NB^2 strategy with optimized choice (13); *ii*) traditional social learning (SL); and *iii*) the centralized Bayesian posterior. We see that the NB^2 strategy outperforms traditional social learning, and attains the same error exponent as the optimal Bayesian posterior. We have verified (not shown for space limitations) that discrepancies across the error probabilities of different agents can arise, as it must be since it was shown that in distributed learning we can attain the centralized error exponent, but we cannot eliminate differences in the error probabilities of the agents, which are embodied in higher-order sub-exponential corrections [11].

A. Highly Dependent Data

Theorem 1 assumes that the data are independent across agents. In order to see the potential benefits of the NB^2 strategy also with dependent data, consider the case where there exist clusters of agents that observe highly correlated data, while the data are independent across distinct clusters. Assume there are M clusters, and denote them by $\mathcal{C}(1), \mathcal{C}(2), \dots, \mathcal{C}(M)$. The cluster agent k belongs to will be denoted by $\mathcal{C}_k$. For the limiting case where the observations of different agents in the same cluster have unit correlation, it is obvious that the optimal Bayesian rule should consider a single likelihood per each cluster. In this case, the log-likelihood ratio pertaining to the ith data samples and to all agents can be written as:

$$\sum_{\ell=1}^{N} \frac{1}{|\mathcal{C}_\ell|} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)}. \tag{22}$$

To see why relation (22) holds, assume for instance that cluster $\mathcal{C}(1)$ is formed by agents j and ℓ. Then, we have that:

$$\begin{aligned}
&\frac{1}{|\mathcal{C}_j|} \log \frac{L_j(\boldsymbol{\xi}_{j,i}|\theta_0)}{L_j(\boldsymbol{\xi}_{j,i}|\theta)} + \frac{1}{|\mathcal{C}_\ell|} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)} \\
&= \frac{1}{2} \log \frac{L_j(\boldsymbol{\xi}_{j,i}|\theta_0)}{L_j(\boldsymbol{\xi}_{j,i}|\theta)} + \frac{1}{2} \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)} = \log \frac{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta_0)}{L_\ell(\boldsymbol{\xi}_{\ell,i}|\theta)},
\end{aligned} \tag{23}$$

where in the last step we used the fact that the data of the two agents coincide, since they have unit correlation. It can be shown (the proof is based on arguments similar to Theorem 1, but is omitted for space limitations) that the error exponent of the optimal Bayesian posterior for this limiting case can be achieved by the NB^2 strategy with the choice:

$$\gamma_k \propto \frac{1}{v_k |\mathcal{C}_k|}. \tag{24}$$

In contrast, traditional non-Bayesian learning neglects the dependence and simply treats the data in the cluster as if they were independent. This way, we are giving to the data in the cluster more relevance than what they would deserve according to the optimal Bayesian posterior.

The right plot in Fig. 1 shows an example with the following setting: the data samples of agent 1 come from a unit-scale Laplace distribution with mean equal to 0.1; the data samples of all other agents from a Laplace distribution with mean equal to 0.05, and these agents (from 2 to 10) form a cluster

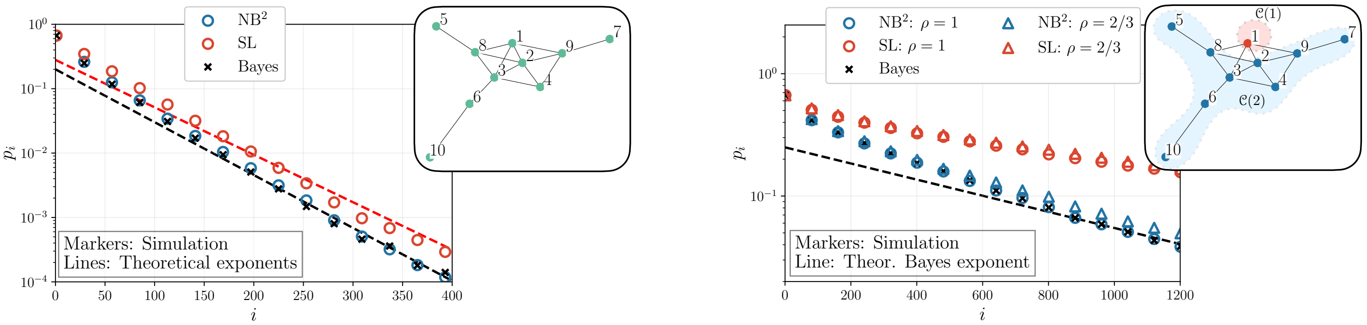

Fig. 1. *Left*. Average error probability (Eq. (21)) as a function of time, for the independent data case and a left stochastic matrix. The network is shown in the smaller panel. *Right*. Average error probability as a function of time, for the highly dependent data case and a doubly stochastic matrix. The shaded areas in the smaller panel represent the clusters of agents. In both plots, we compare the NB^2 strategy, traditional social learning, and the Bayesian posterior.

with dependent data. The figure shows two cases: the limiting case where all data in the cluster are the same (correlation coefficient $\rho = 1$), and the more practical case where some zero-mean Gaussian noise with unit variance is added to each data sample in order to make them highly correlated but not equal ($\rho = 2/3$). We also consider a doubly stochastic matrix (namely, a Metropolis matrix [3]) to emphasize the role of dependence rather than of the combination policy. We see from the right plot in Fig. 1 that the NB^2 strategy significantly outperforms traditional social learning. Notably, this happens even in the non-limiting scenario. Moreover, in the non-limiting scenario the NB^2 strategy attains the same error exponent of the Bayesian posterior that assumes perfect correlation among the data within the same cluster.

V. Conclusion

In traditional social learning, before sharing their opinions, the agents act in an *individually-optimal* manner by performing local Bayesian updates. We proposed a new scheme that employs *non-Bayesian* updates and that, in some useful cases, attains the same performance as the optimal Bayesian posterior, while the traditional scheme cannot. This improvement is achieved by adapting the local updates to relevant attributes of the distributed learning environment, particularly the Perron eigenvector and the joint statistical dependence structure. Possible extensions include the characterization of alternative performance measures, and a behavioral interpretation to explain in which contexts the agents understand that is advantageous to depart from what is optimal individually, and adapt their behavior to the *social* setting.

References

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

[2] C. Chamley, A. Scaglione, and L. Li, "Models for the diffusion of beliefs in social networks: An overview," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 16–29, 2013.

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

[4] A. H. Sayed, *Inference and Learning from Data*, 3 vols., Cambridge University Press, 2022.

[5] J. Hązła, A. Jadbabaie, E. Mossel, and M. A. Rahimian, "Bayesian decision making in groups is hard," *Operations Research*, vol. 69, no. 2, pp. 632–654, 2021.

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

[7] 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.

[8] H. Salami, B. Ying, and A. H. Sayed, "Social learning over weakly connected graphs," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 2, pp. 222–238, 2017.

[9] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-Bayesian learning," *IEEE Trans. Autom. Control*, vol. 62, no. 11, pp. 5538–5553, 2017.

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

[11] V. Bordignon, V. Matta, and A. H. Sayed, "Adaptive social learning," *IEEE Trans. Inf. Theory*, vol. 67, no. 9, pp. 6053–6081, 2021.

[12] G. Koliander, Y. El-Laham, P. M. Djurić, and F. Hlawatsch, "Fusion of probability density functions," *Proceedings of the IEEE*, vol. 110, no. 4, pp. 404–453, 2022.

[13] P. Molavi, A. Tahbaz-Salehi, and A. Jadbabaie, "A theory of non-Bayesian social learning," *Econometrica*, vol. 86, no. 2, pp. 445–490, 2018.

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

[15] M. Kayaalp, V. Bordignon, S. Vlaski, and A. H. Sayed, "Hidden Markov modeling over graphs," in *Proc. IEEE DSLW*, Singapore, 2022, pp. 1– 6.

[16] O. Hlinka, O. Slučiak, F. Hlawatsch, P. M. Djurić, and M. Rupp, "Likelihood consensus and its application to distributed particle filtering," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4334-4349, 2012.

[17] P. Hu, V. Bordignon, S. Vlaski, and A. H. Sayed, "Optimal combination policies for adaptive social learning," in *Proc. IEEE ICASSP*, Singapore, 2022, pp. 5842-5846.

[18] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Springer, 1998.

[19] F. den Hollander, *Large Deviations*, American Mathematical Society, 2008.

[20] D. Bajović, D. Jakovetić, J. M. F. Moura, J. Xavier, and B. Sinopoli, "Large deviations performance of consensus+innovations distributed detection with non-Gaussian observations," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5987–6002, 2012.

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

[22] V. Matta, P. Braca, S. Marano, and A. H. Sayed, "Distributed detection over adaptive networks: Refined asymptotics and the role of connectivity," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 4, pp. 442–460, 2016.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 1991.

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