ON THE FUSION STRATEGIES FOR FEDERATED DECISION MAKING

Mert Kayaalp⋆, Yunus İnan⋆, Visa Koivunen†, Emre Telatar⋆, Ali H. Sayed⋆

⋆École Polytechnique Fédérale de Lausanne (EPFL)
†Aalto University

ABSTRACT

We consider the problem of information aggregation in federated decision making, where a group of agents collaborate to infer the underlying state of nature without sharing their private data with the central processor or each other. We analyze the non-Bayesian social learning strategy in which agents incorporate their individual observations into their opinions (i.e., soft-decisions) with Bayes rule, and the central processor aggregates these opinions by arithmetic or geometric averaging. Building on our previous work, we establish that both pooling strategies result in asymptotic normality characterization of the system, which, for instance, can be utilized to derive approximate expressions for the error probability. We verify the theoretical findings with simulations and compare both strategies.

1. INTRODUCTION AND RELATED WORK

The centralized strategy has been the traditional form for collaborative decision-making, where data is collected from all sources and processed at a central controller, fusion center or cloud. However, this approach raises concerns about data privacy and communication costs. In this work, we study the paradigm of federated decision-making, which enables cooperative decision-making without data collection [1–3]. For example, in healthcare applications, federated decision-making can be used to test possible hypotheses without gathering private data from siloed institutions [4]. Moreover, in classification tasks, agents (e.g., mobile device users) may have different views of a physical phenomenon such as scenery as in multi-view learning [5], and the server may want to classify the scene or infer the state of environment without collecting the local data of users.

The setting we consider in this work, which is described in Sec. 2, is an instance of locally Bayesian (a.k.a. non-Bayesian) social learning [6–10] that is based on general decentralized networks. Here, we only consider the special case of a star network topology where all agents are connected to a central node. In social learning, the goal is to infer the hypothesis that best describes the observations received from the environment. The true hypothesis is common for all agents and hence cooperation is of interest for all agents. To this end, at each time instant, (i) distributed agents process their private data locally and send their opinions (i.e., soft-decisions) to the server, (ii) the server pools the information received from agents and broadcasts the aggregated belief back to the agents, and (iii) the agents use the broadcasted belief as a prior belief for their subsequent opinions.

Note that this setting has close ties with the distributed detection literature [11–15], which typically involves sharing log-likelihood
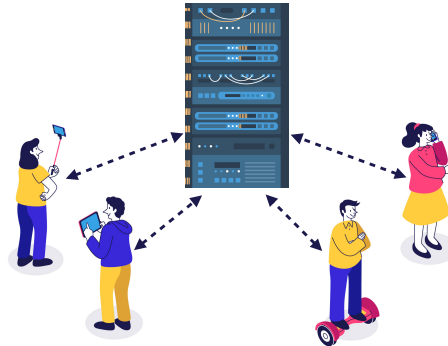
Fig. 1: Data types at the edge devices can be highly heterogeneous. Image from: freepik.com.

ratios for each data and assumes spatial independence and/or homogeneity across agents. In contrast, in the social learning framework, agents share beliefs (posteriors) formed over time, rather than a statistic about individual data. In this way, they can incorporate their individual priors as well. Furthermore, in social learning, data is not required to be independent between agents nor identically distributed. In fact, agents can possess different data types (e.g., text, video), as in multi-modal machine learning — see Fig. 1. The federated decision-making is also related to federated learning [16–19] as they both fall under the umbrella of federated analytics [1–3], which aims to enable cooperation among multiple users while preserving privacy of data. However, federated learning focuses on training machine learning models with labeled data samples, whereas the current setting involves making collaborative inferences without sharing raw observations.

In this paper, we address the following problem: How should the server aggregate information received from multiple distributed agents that process the data locally and share only their local soft decision statistics? Two of the most widely used methods for combining probability vectors are *arithmetic* averaging (AA), a.k.a. linear opinion pooling, and *geometric* averaging (GA), a.k.a. logarithmic opinion pooling. The distinct properties of these procedures are well-established in the literature — see, e.g., [20]. In our previous work [21], we investigated the performance of AA and GA strategies in the general social learning setting. For the special case of federated architectures, we derived an exact expression for the asymptotic convergence rates of beliefs. Different from prior works that only compare AA and GA with some assumptions, such as assuming only one step of fusion [22–24] or assuming specific/particular distributions like Gaussian or Poisson [22, 24, 25], our analysis considered repeated fusion of beliefs without any assumption on the distribution of data. In this work, we extend our previous findings and show that the asymptotic behavior of both AA and GA can be characterized

with normal distributions and the asymptotic error probabilities of both strategies can be established accordingly. The results provide further insight into the arithmetic and geometric fusion rules, and can be utilized for designing federated decision-making systems.

Notation. Random variables are written in boldface letters. We use "proportional to" symbol $\propto$ whenever the LHS of an equation is a proper normalization of the RHS. The KL-divergence between two distributions p and q is denoted by $D_{\text{KL}}(p||q)$.

2. FEDERATED DECISION MAKING

We consider a distributed hypothesis testing task in which a group $\mathcal{K}$ of K agents, aided by a central controller or server, performs inference on a phenomenon of interest. More formally, the agents seek to choose the true hypothesis θ° from a finite set of H hypotheses, $\Theta = \{1, 2, \ldots, H\}$. At each time instant i, agent k receives a personal observation $\boldsymbol{\xi}_{k,i}$ from the environment that is distributed according to the marginal likelihood function $L_k(\cdot|\theta^\circ)$ (since the true hypothesis is θ°). The likelihood functions are allowed to depend on the agent index k, so that the data distributions across the different agents can be heterogeneous. We assume that the environment is stationary, and that the data is independent and identically distributed (i.i.d.) over time at each agent. However, we do not make any assumption on spatial dependence, i.e., there is no assumption on the dependence among different agents. Each agent k only knows its own local likelihood models $L_k(\xi|\theta)$ for each $\theta \in \Theta$, which represent how likely an observation ξ is produced by θ. Furthermore, the server does not need to possess any information about the distribution of data over the agents.

The agents do not share their private raw data with the server: all data processing and ownership are local. Each agent k only shares its belief (i.e., opinion) $\boldsymbol{\psi}_{k,i}$ about the possible hypotheses, which is a probability mass function over Θ. In other words, $\boldsymbol{\psi}_{k,i}(\theta)$ represents the amount of confidence agent k has on the proposition "$\theta = \theta^\circ$", at time instant i. Once the server receives the intermediate beliefs from the distributed agents, it pools them and broadcasts the averaged belief back to the agents. In this work, we study the widely used linear (AA) and logarithmic (GA) pooling strategies, although there are other variations [20]. We now motivate the setting with two real-world applications.

Example 1 (**Crowd size estimation**). *Crowd counting aims to estimate the number of people in a given area using visual or other sensory data. The possible hypotheses are the size of the crowd, such as "the number of people in the event is between 500-1000". Agents, i.e., users cooperating with the server, obtain data in the form of photos and videos of their surroundings, or their mobile devices can measure the number of possible connections over Bluetooth. Due to privacy concerns, users may prefer not to share their raw data with the server. In addition, transmitting raw data can create communication bottlenecks at the server. Therefore, employing the social learning strategy, users would only share their beliefs over the set of hypotheses. They can form their beliefs by using local likelihood models, e.g., by using pre-trained neural network models on their devices to estimate the crowd size based on images [26].* ∎

Example 2 (**Flexible spectrum use**). *Spatially distributed sensors or mobile devices can try to estimate the number of targets or active wireless emitters within a particular area. For example, in the radio spectrum, the mobile devices can measure the local occupancy and field levels to distinguish between idle and congested spectrum, or*

map the radio environment. This practice can enhance the environmental awareness, and allocate the resources optimally in order to improve the wireless communication. Exchanging beliefs only rather than raw data can be advantageous in terms of privacy and communication costs. ∎

To form the beliefs, agents execute the non-Bayesian social learning strategy [6–9] under the special case of star topologies. At each time instant i, agents first process their observations in a *locally* Bayesian manner:

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

and then share these intermediate beliefs with the server. If the server employs AA for information fusion, the updated belief is a weighted arithmetic average of the intermediate beliefs from different agents:

$$\boldsymbol{\mu}_i(\theta) = \sum_{k \in \mathcal{K}} \pi_k \boldsymbol{\psi}_{k,i}(\theta) \qquad \text{(AA)}. \qquad (2)$$

Here, $\pi = [\pi_1, \ldots, \pi_K]^\mathsf{T}$ is a vector of the confidence weights π_k the server assigns to each agent k [12, 27]. We assume that these weights are positive, sum up to 1, and are constant over time. Alternatively, if the fusion rule is GA, the updated belief is a weighted geometric average:

$$\boldsymbol{\mu}_i(\theta) \propto \prod_{k \in \mathcal{K}} (\boldsymbol{\psi}_{k,i}(\theta))^{\pi_k} \qquad \text{(GA)}. \qquad (3)$$

The server then sends the updated belief to the agents, and they execute the same steps repeatedly over time by using the updated beliefs and locally observed data.

Remark 1. *An important distinction between AA and GA lies in their handling of initial beliefs and their sensitivity to individual beliefs. In AA, only one agent having a positive initial belief on the true hypothesis is sufficient. In contrast, in GA, all agents must have positive initial beliefs on the truth in order not to discard it from the beginning. Furthermore, GA is more sensitive to small beliefs compared to AA. This is because, if some agents' beliefs are very small, they impact the product more significantly compared to the sum. In other words, GA grants the agents a veto power: if one agent transmits a belief entry as 0, the combined belief also becomes 0. In general, giving this level of autonomy to individual agents depends on the application at hand.* ∎

3. ASYMPTOTIC NORMALITY OF AA AND GA

In this section, we analyze the asymptotic behavior under both fusion strategies. In order to avoid pathological cases, we assume that each observation has finite information about the true hypothesis. That is to say, for each agent k and for each hypothesis $\theta \in \Theta$, we assume $D_{\text{KL}}(L_k(\cdot|\theta^\circ)||L_k(\cdot|\theta)) < \infty$. This implies that likelihood functions have the same support. Moreover, in order to uniquely distinguish the true hypothesis, we need the following condition.

Assumption 1 (**Global identifiability**). *For each wrong hypothesis $\theta \neq \theta^\circ$, there exists at least one clear-sighted agent k with $D_{\text{KL}}(L_k(\cdot|\theta^\circ)||L_k(\cdot|\theta)) > 0$.* ∎

Note that this assumption does not require *local* identifiability, which is an agent's ability of inferring θ° without any cooperation. As a result, all agents can benefit from cooperation. Under Assumption 1, it is already known that AA [6, 7, 28] and GA [8, 9] result

in consistent truth learning, that is, the beliefs on the true hypothesis converge to one almost surely. In this work, we further show that under both AA and GA fusion rules, the asymptotic error probabilities can be calculated with Gaussian cumulative distribution functions (CDFs). We start with the AA strategy.

Lemma 1 (**Asymptotic normality of AA**). *For each wrong hypothesis $\theta \neq \theta^\circ$, we define the mean and variance*

$$\rho_{\mathrm{A}} \triangleq -\mathbb{E}\left[\log\left(\sum_{k \in \mathcal{K}} \pi_k \boldsymbol{r}_{k,i} \right) \right], \ \sigma_{\mathrm{A}}^2 \triangleq \mathrm{Var}\left[\log\left(\sum_{k \in \mathcal{K}} \pi_k \boldsymbol{r}_{k,i} \right) \right],$$
(4)

in terms of the likelihood ratios

$$\boldsymbol{r}_{k,i} \triangleq \frac{L_k(\boldsymbol{\xi}_{k,i}|\theta)}{L_k(\boldsymbol{\xi}_{k,i}|\theta^\circ)}.$$
(5)

If $\sigma_{\mathrm{A}}^2 < \infty$, it holds under Assumption 1 that

$$\lim_{i \to \infty} \mathbb{P}\left(\frac{\log \boldsymbol{\mu}_i(\theta) + \rho_{\mathrm{A}} i}{\sigma_{\mathrm{A}} \sqrt{i}} \leq t \right) = \Phi(t),$$
(6)

where $\Phi(t)$ is the standard Gaussian cumulative distribution function.

Proof Sketch. Due to space limitations, we only give a sketch of the proof. The complete proof can be found in the extended version [29]. Consider an arbitrary $\theta \neq \theta^\circ$. For notational simplicity, define

$$\boldsymbol{x}_i \triangleq \log \boldsymbol{\mu}_i(\theta), \quad \boldsymbol{L}_i \triangleq \log\left(\sum_{k \in \mathcal{K}} \pi_k \boldsymbol{r}_{k,i} \right).$$
(7)

Since $\boldsymbol{L}_i$'s are assumed to be i.i.d. and of finite variance, the central limit theorem [30] yields:

$$\lim_{i \to \infty} \mathbb{P}\left(\frac{\sum_{j=1}^{i} \boldsymbol{L}_j + \rho_{\mathrm{A}} i}{\sigma_{\mathrm{A}} \sqrt{i}} \leq t \right) = \Phi(t).$$
(8)

Therefore, if we show that the distance between $\frac{1}{\sqrt{i}} \boldsymbol{x}_i$ and $\frac{1}{\sqrt{i}} \sum_{j=1}^{i} \boldsymbol{L}_j$ goes to zero with high probability, we are done. To that end, we utilize the result from [21] which states

$$\frac{1}{i} \boldsymbol{x}_i \to -\rho_{\mathrm{A}} \quad \text{as } i \to \infty, \text{ almost surely.}$$
(9)

In particular, we use the following bootstrapping argument on (9): There exists an i_0 after which (i.e., $\forall i \geq i_0$)

$$\boldsymbol{\mu}_i(\theta) \leq \exp\{-i(\rho_{\mathrm{A}} - \epsilon)\}, \ \forall \theta \neq \theta^\circ$$
(10)

with high probability. This result enables a finer study of the evolution of $\boldsymbol{x}_i$ as $i \to \infty$. ∎

Note that Lemma 1 can equivalently be given in terms of the log-belief ratios.

Corollary 1. *If we define the log-belief ratio $\boldsymbol{\lambda}_i(\theta) \triangleq \log \dfrac{\boldsymbol{\mu}_i(\theta^\circ)}{\boldsymbol{\mu}_i(\theta)}$,*

for each $\theta \neq \theta^\circ$, Lemma 1 implies that

$$\lim_{i \to \infty} \mathbb{P}\left(\frac{\boldsymbol{\lambda}_i(\theta) - \rho_A i}{\sigma_A \sqrt{i}} \leq t \right) = \Phi(t).$$
(11)

Proof Sketch. Due to space limitations, we only give a sketch of the proof. The complete proof can be found in the extended version [29]. Essentially, the proof follows from consistent truth learning, i.e., $\boldsymbol{\mu}_i(\theta^\circ) \to 1$ almost surely as $i \to \infty$. This implies that, as $i \to \infty$, $\log \boldsymbol{\mu}_i(\theta)$ and $\log \boldsymbol{\mu}_i(\theta) - \log \boldsymbol{\mu}_i(\theta^\circ)$ behave similarly. ∎

Next, we investigate the behavior of beliefs under the GA fusion rule.

Lemma 2 (**Asymptotic normality of GA**). *For each wrong hypothesis $\theta \neq \theta^\circ$, we define the mean*

$$\rho_{\mathrm{G}} \triangleq -\mathbb{E}\left[\sum_{k \in \mathcal{K}} \pi_k \log \boldsymbol{r}_{k,i} \right] = \sum_{k \in \mathcal{K}} \pi_k D_{\mathrm{KL}}(L_k(.|\theta^\circ)||L_k(.|\theta)) \quad (12)$$

and the variance

$$\sigma_{\mathrm{G}}^2 \triangleq \mathrm{Var}\left[\sum_{k \in \mathcal{K}} \pi_k \log \boldsymbol{r}_{k,i} \right].$$
(13)

If $\sigma_{\mathrm{G}}^2 < \infty$, then under Assumption 1 it holds that

$$\lim_{i \to \infty} \mathbb{P}\left(\frac{\boldsymbol{\lambda}_i(\theta) - \rho_{\mathrm{G}} i}{\sigma_{\mathrm{G}} \sqrt{i}} \leq t \right) = \Phi(t).$$
(14)

Proof. The proofs from [8, 9] for the convergence rate analysis can be easily extended to establish asymptotic normality. From (1) and (3), the log-belief ratio $\boldsymbol{\lambda}_i(\theta)$ evolves according to the recursion:

$$\boldsymbol{\lambda}_i(\theta) = \boldsymbol{\lambda}_{i-1}(\theta) - \sum_{k \in \mathcal{K}} \pi_k \log \boldsymbol{r}_{k,i}.$$
(15)

Expanding the recursion over time, we obtain

$$\boldsymbol{\lambda}_i(\theta) = -\sum_{j=1}^{i} \left(\sum_{k \in \mathcal{K}} \pi_k \log \boldsymbol{r}_{k,j} \right) + \boldsymbol{\lambda}_0(\theta).$$
(16)

Since (i) $\boldsymbol{\lambda}_0(\theta) < \infty$ due to initial beliefs being positive, (ii) $\boldsymbol{r}_{k,i}$'s are i.i.d., and (iii) $\sigma_{\mathrm{G}}^2 < \infty$, an application of the central limit theorem [30] concludes the proof. ∎

In Lemmas 1 and 2, we have established the asymptotic normality for each wrong hypothesis, which allows approximating the error probabilities when agents or the central processor employs a maximum a-posteriori (MAP) rule to estimate θ°. It is important to note that under spatial independence, GA matches the optimal Bayes posterior, which attains minimum error probability, when the confidence weight π_k on each agent k is uniform. Additionally, the convergence rates ρ_{G} and ρ_{A} were already established in previous works [8, 9] and [21] respectively. Notice that $\rho_{\mathrm{G}} \geq \rho_{\mathrm{A}}$ — applying Jensen's inequality to the equation on the left in (4), we obtain (12). However, this does not directly imply that GA has a better performance in probability of error criterion since it depends on the variances as well. Our simulations indicate, however, that GA has lower error probability in general. We leave it to future work to determine whether GA is superior to AA in terms of asymptotic error probability under any data distribution or confidence weight configuration — such as in the convergence rate case.

4. NUMERICAL RESULTS

We consider a binary hypothesis testing problem with $K = 10$ distributed agents connected to a fusion center. In the first experiment, agents receive i.i.d. Gaussian observations with mean 0 under the true hypothesis θ°, and with mean 1 under $\theta \neq \theta^\circ$. For both hypotheses, the correlation matrix is the identity matrix I_K, i.e., data is independent across agents. A possible application of this experiment is in flexible spectrum use, for example in cognitive radios. The possible hypotheses are whether the spectrum is idle or not. If there is noise only, the signal distributions would be central (i.e., 0 mean).
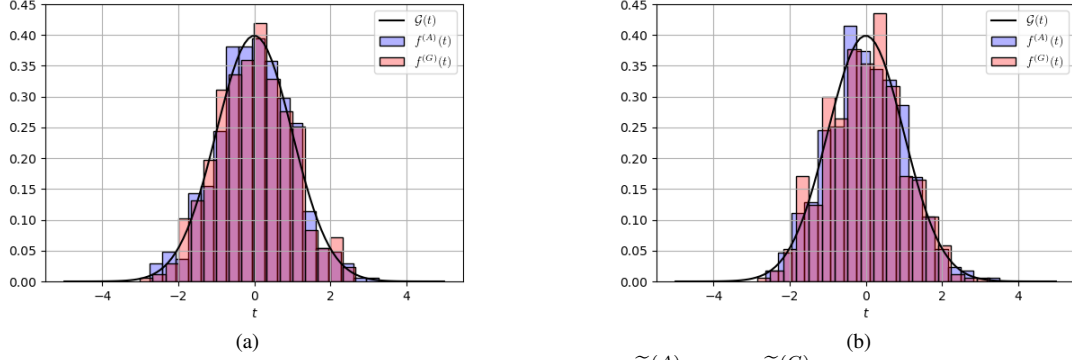
Fig. 2: (a) For the first experiment with Gaussian data, the histograms of $\widetilde{\boldsymbol{\lambda}}_i^{(A)}(\theta)$ and $\widetilde{\boldsymbol{\lambda}}_i^{(G)}(\theta)$ are drawn as shaded bar plots with colors blue and red respectively. The standard normal density $\mathcal{G}(t)$ is shown as solid black line. (b) For the second experiment with exponentially distributed data, the histograms of $\widetilde{\boldsymbol{\lambda}}_i^{(A)}(\theta)$ and $\widetilde{\boldsymbol{\lambda}}_i^{(G)}(\theta)$ are drawn as shaded bar plots with colors blue and red respectively. The standard normal density $\mathcal{G}(t)$ is shown as solid black line.
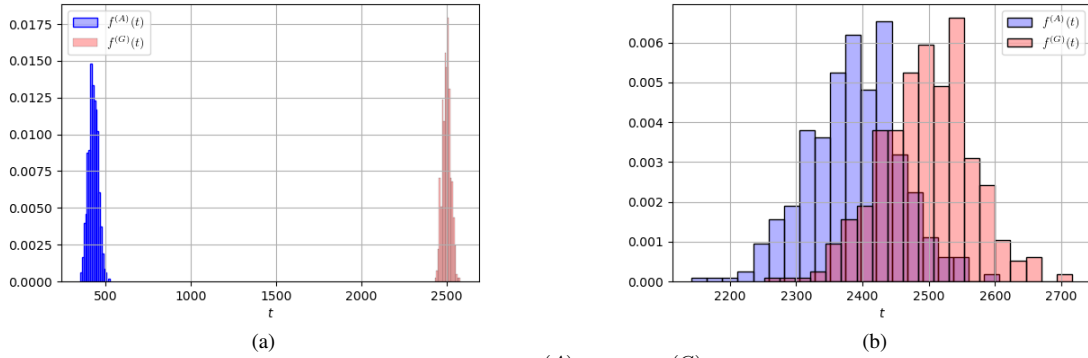

Fig. 3: (a) For the uncorrelated Gaussian case, histograms of $\boldsymbol{\lambda}_i^{(A)}(\theta)$ and $\boldsymbol{\lambda}_i^{(G)}(\theta)$ are drawn as shaded bar plots with colors blue and red respectively. (b) For the correlated Gaussian case, histograms of $\boldsymbol{\lambda}_i^{(A)}(\theta)$ and $\boldsymbol{\lambda}_i^{(G)}(\theta)$ are drawn as shaded bar plots with colors blue and red respectively.

Otherwise, if there is a signal present, it would not be 0 mean. We choose the combination vector as

$$\pi = [0.13, 0.2, 0.09, 0.15, 0.08, 0.05, 0.1, 0.05, 0.1, 0.05]. \quad (17)$$

If we take the time horizon $i = 5000$ and repeat the experiment for 500 realizations, we obtain the histograms given in Figure 2a. Specifically, we obtain the normalized log-belief ratios:

$$\widetilde{\boldsymbol{\lambda}}_i^{(A)}(\theta) \triangleq \frac{\boldsymbol{\lambda}_i^{(A)}(\theta) - \rho_A i}{\sqrt{i}\sigma_A}, \quad \widetilde{\boldsymbol{\lambda}}_i^{(G)}(\theta) \triangleq \frac{\boldsymbol{\lambda}_i^{(G)}(\theta) - \rho_G i}{\sqrt{i}\sigma_G} \quad (18)$$

and plot their histograms in density forms (denoted by $f^{(A)}(t)$ and $f^{(G)}(t)$, respectively) to compare them with the standard normal density $\mathcal{G}(t) \triangleq \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$. In addition to the visual matching supporting our claims, we performed a Shapiro-Wilk test for testing the normality of $\widetilde{\boldsymbol{\lambda}}_i^{(A)}(\theta)$ and $\widetilde{\boldsymbol{\lambda}}_i^{(G)}(\theta)$. The resulting p-values turn out to be 0.66 and 0.54, which strongly support the null hypothesis of normal distribution.

For the second experiment, we assume that agents receive i.i.d. exponential observations with mean 1 under θ° and with mean 0.5 under $\theta \neq \theta^\circ$. The other problem parameters are the same as the first experiment. The resulting histograms are given in Figure 2b. The Shapiro-Wilk test yields p-values 0.93 and 0.53, respectively, again, strongly supporting the null hypothesis of normal distribution.

We also conducted a third experiment to verify that correlations between the agents affect ρ_A but do not affect ρ_G (see (12)). To compare correlated versus uncorrelated data across agents, we simulate the setting from the first experiment, except that this time we set the correlation matrix as $\Sigma = 0.95\mathbb{1}_K \mathbb{1}_K^\top + 0.05I_K$. Note that this choice makes the data highly correlated. Then, we plot the histograms of $\boldsymbol{\lambda}_i^{(A)}(\theta)$ and $\boldsymbol{\lambda}_i^{(G)}(\theta)$ (without normalization) for both uncorrelated and correlated cases in Figure 3a and 3b. We can observe that $\rho_G = 0.5$ is unchanged by the correlations — the observation that $f^{(G)}$'s are concentrated around $i \times \rho_G = 2500$ supports this fact. In contrast, ρ_A increases drastically under the presence of positive correlations. Nevertheless, in both cases considered, GA outperforms AA in terms of the error probability.

5. CONCLUSION

In this work, we considered two pooling algorithms for federated decision making, namely AA and GA. The choice between AA and GA may depend on the specific task at hand. We study their performance in terms of error probability, and establish asymptotic normality which is a valuable tool that can help avoid the need to simulate the algorithm for large time horizons. Note that the parameters in the normal approximations can be obtained either analytically or through Monte Carlo simulations, which is more efficient compared to the exact simulation of the entire system. We leave the study for asymptotic normality of general decentralized networks to a future work.

6. REFERENCES

[1] A. R. Elkordy, Y. H. Ezzeldin, S. Han, S. Sharma, C. He, S. Mehrotra, and S. Avestimehr, "Federated analytics: A survey," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 1, pp. 1–33, 2023.

[2] D. Wang, S. Shi, Y. Zhu, and Z. Han, "Federated analytics: Opportunities and challenges," *IEEE Network*, vol. 36, no. 1, pp. 151–158, 2021.

[3] D. Ramage and S. Mazzocchi, "Federated analytics: Collaborative data science without data collection." [Online]. Available: https://ai.googleblog.com/2020/05/federated-analytics-collaborative-data.html

[4] H. Hallock, S. E. Marshall, P. A. C. 't Hoen, J. F. Nygård, B. Hoorne, C. Fox, and S. Alagaratnam, "Federated networks for distributed analysis of health data," *Frontiers in Public Health*, vol. 9, 2021.

[5] 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*, pp. 1–25, Dec. 2022.

[6] 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.

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

[8] A. Nedić, 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.

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

[11] J. N. Tsitsiklis, "Decentralized detection," in *In Advances in Statistical Signal Processing*. JAI Press, 1993, pp. 297–344.

[12] P. K. Varshney, *Distributed Detection and Data Fusion*. Springer, 2012.

[13] Q. Zou, S. Zheng, and A. H. Sayed, "Cooperative sensing via sequential detection," *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 6266–6283, 2010.

[14] D. Bajovic, D. Jakovetic, J. M. Moura, J. Xavier, and B. Sinopoli, "Large deviations performance of consensus+ innovations distributed detection with non-Gaussian observations," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5987–6002, 2012.

[15] Y. İnan, M. Kayaalp, A. H. Sayed, and E. Telatar, "A fundamental limit of distributed hypothesis testing under memoryless quantization," in *Proc. IEEE International Conference on Communications*, 2022, pp. 4824–4829.

[16] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. International Conference on Artificial Intelligence and Statistics*, vol. 54, April 2017, pp. 1273–1282.

[17] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[18] P. Kairouz, H. B. McMahan *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[19] E. Rizk, S. Vlaski, and A. H. Sayed, "Federated learning under importance sampling," *IEEE Transactions on Signal Processing*, vol. 70, pp. 5381–5396, 2022.

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

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

[22] T. Li, H. Fan, J. García, and J. M. Corchado, "Second-order statistics analysis and comparison between arithmetic and geometric average fusion: Application to multi-sensor target tracking," *Information Fusion*, vol. 51, pp. 233–243, 2019.

[23] T. Li, Y. Xin, Y. Song, E. Song, and H. Fan, "Some statistic and information-theoretic results on arithmetic average density fusion," *arXiv:2110.01440*, 2021.

[24] L. Gao, G. Battistelli, and L. Chisci, "Multiobject fusion with minimum information loss," *IEEE Signal Processing Letters*, vol. 27, pp. 201–205, Jan. 2020.

[25] K. C. Chang, C.-Y. Chong, and S. Mori, "Analytical and computational evaluation of scalable distributed fusion algorithms," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 46, no. 4, pp. 2022–2034, 2010.

[26] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5099–5108.

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

[28] A. Jadbabaie, P. Molavi, and A. Tahbaz-Salehi, "Information heterogeneity and the speed of learning in social networks," *Columbia Business School Research Paper*, no. 13-28, pp. 1–38, May 2013.

[29] M. Kayaalp, Y. Inan, V. Koivunen, E. Telatar, and A. H. Sayed, "On the fusion strategies for federated decision making," *arXiv:2303.06109*, 2023.

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