Risk-Aware Decentralized Federated Learning with Intermittent Communications

Periklis Theodoropoulos, Christos Mavrokefalidis, Kostas Berberidis
Computer Engineering and Informatics Department, University of Patras, Greece
Emails: { p.theodorop}@ac.upatras.gr, { maurokef, berberid }@ceid.upatras.gr

***Abstract*—Decentralized Federated Learning (DFL) is a server-less federated learning framework that facilitates collaborative model training while ensuring data privacy, eliminating the need for orchestration from a centralized entity. In many real-world applications, unreliable communication networks result in uneven agent participation due to stochastic environmental conditions. In this study, we model communication links between nodes as erasure channels, where transmitted information may be received by neighboring nodes with a low probability. To address this challenge, we propose two risk-aware extensions of the traditional Combine-then-Adapt (CTA) technique, employing the Conditional Value-at-Risk (CVaR) over the distribution of nodes. Our experimental evaluations on MNIST and FashionMNIST datasets show that the proposed approaches achieve higher performance as compared with standard CTA, across various setups.**

***Index Terms*—Decentralized Federated Learning, Conditional Value-at-Risk, Risk-Aware Learning, Stochastic Optimization, Erasure Channels.**

I. Introduction

Decentralized Federated Learning (DFL) is a server-less distributed learning framework allowing multiple agents to collaboratively solve a global optimization problem using only their private local information [1]. Unlike classic federated learning frameworks, which rely on a central parameter server to coordinate the communication among the agents, by receiving their updates, aggregating parameters, and broadcasting the aggregated model back to them [2], DFL substitutes this centralized scheme with a peer-to-peer communication analog. This shift from the centralized structure of FL to a fully distributed architecture enhancing data privacy and reducing the risk of single points of failure [3].

In vanilla DFL, agents are connected via some graph topology that allows information exchange between neighboring agents [4]. The diffusion-based strategies (adapt-then-combine(ATC) and combine-then-adapt(CTA) protocols) allow the linked agents, through continuous interactions with their neighbors, to achieve comparable level of performance to that of a single centralized agent with access to all available data in the network [4]. Despite its advantages, DFL deals with uncertainty arising by intermittent communication between nodes, which may be caused by network outages, user inactivity or under random nodes communication [5]–[7]. The presence of non-independent and identically distributed (non-IID) data across nodes [8] further exacerbates the associated optimization problem.

Prior research has extensively studied the intermittent nature of node communication in DFL, proposing various methods to mitigate the related challenges. One line of this research focuses on the impact of imperfect communication channels on DFL, particularly in determining the optimal number of local aggregations per training round based on network topology and channel imperfections [9]. Additionally, the importance of robust communication in DFL frameworks to maintain model performance under volatile network conditions has been well studied. For instance, the so-called Soft-DSGD approach addresses unreliability caused by packet losses and transmission errors, by incorporating partially received messages into model updates and optimizing mixing weights based on the link reliability matrix [10].

This work is supported by the University of Patras and the EU project IoT-ECO (ID:101083018).

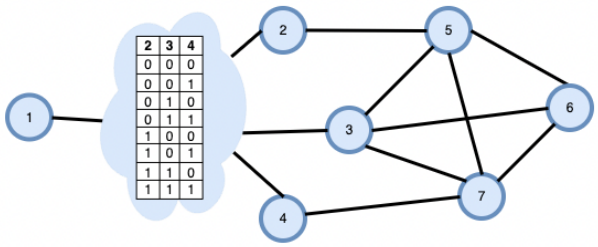


Fig. 1: Decentralized Federated Learning (DFL) under erasure communication channels. **0** indicates channel failure, while **1** indicates channel functionality.

Recent studies have also explored advancements in server-less FL under dynamic conditions. For example, an asynchronous network of agents where each agent independently decides on participation and neighbor selection is examined in [1]. Additionally, server-less FL schemes for multi-cell environments, under frequent test-time distribution shifts have been studied in [11], while test-time performance in server-less Vertical FL has been improved robustness through replication, gossiping, and selective feature omission [12]. Moreover, the impact of communication errors and constraints has been studied extensively for both centralized and decentralized FL. Specifically, the centralized FL scheme is studied under communication errors by modeling the communication channels as erasure channels [13], while a gradient recycling method has been proposed to enhance learning performance in resource-limited wireless networks [14]. Similarly, a federated policy evaluation problem where agents communicate with a central aggregator over finite-capacity uplink channels using a Bernoulli erasure model is investigated in [15]. Beyond FL, a risk-aware decentralized control framework has been explored to manage agent responsibility share, preventing collisions while maintaining efficient movement without direct communications, in [16]. Finally, a particularly promising approach is the decentralized over-the-air MIMO FL method, which enables consensus among an increasing number of nodes while dealing with bottlenecks associated with network expansion [17].

This work addresses the uncertainty in distributed learning due to unreliable networks, where this volatile nature of communication links can impair node cooperation in diffusion-based techniques. We propose two risk-aware DFL approaches leveraging the Conditional Value-at-Risk (CVaR) measure, applied over the node distribution [7], extending traditional CTA updates to enhance resilience against communication uncertainties by focusing on the worst-case node connections.

II. Problem Setup

We consider a server-less federated learning setting, with cooperating nodes $i \in [1, \dots, K]$ belonging to a network, described by a graph $\mathcal{G}(\mathcal{K}, \mathcal{E})$. With $\mathcal{K}$ we represent the set of the nodes, and with $\mathcal{E}$, we represent the communication links between the nodes. Furthermore, the nodes communicate through a diffusion-based protocol, where each node first combines the received information from its neighborhood $\mathcal{N}_i$, and then it adapts this combined information based on its private dataset D_i. For this work, we have focused on the multi-class classification problem, where each node $i \in [K]$ has a set of input vectors $x \in \mathcal{X} \subseteq R^d$ and the corresponding labels $y \in \mathcal{Y} = \{1, 2, \cdots, C\}$, where C is the total number of classes (patterns). As a result the private dataset $D_i(x, y) := D_i(\xi)$, with N_i number of data. In the context of fully distributed federated learning the nodes collaboratively minimize their local loss functions $l_i : R^C \times \mathcal{Y} \to R_+$, exchanging the parameters of their parameterized predictors only with their neighbors. Let us also define the family $\tilde{\Theta} := \{\phi : \mathcal{X} \times R^Q \to R^C\}$ of parameterized predictors with parameter $\theta \in R^Q$.

In the standard DFL the goal of the network would be to optimize the consensus problem of the participating nodes finding the optimal parameters θ solving the risk-neutral problem

$$\inf_{\theta \in R^Q} \mathbb{E}_{I \sim \mathcal{K}} \Big[\mathbb{E}_{\xi \sim D_I} \big[l_I(\phi(\xi; \theta)) \big] \Big], \tag{1}$$

whose empirical version reads as

$$\inf_{\theta \in R^Q} \left\{ \sum_{i=1}^{K} \rho_i \frac{1}{N_i} \sum_{j=1}^{N_i} l_i(\phi(\xi_j; \theta)) = \frac{1}{K} \sum_{i=1}^{K} f_i(\theta) \right\}, \tag{2}$$

where $\rho_i = \frac{1}{K}$, treating all nodes as equally important.

In this work, we want to capture a realistic network pathogenesis that corresponds to the intermittent nodes availability. Thus, we assume that the nodes communicate with each other through erasure channels. The communication links of the network has a probability to fail and as a result the transmitted information to be lost. To be more precise, let p_{li} be the erasure probability with which the packet fails to be received from node i, sending by node l, and $(1 - p_{li})$ is the probability the packet to be received from node i. The packet dropping processes are assumed to be independent of all other sources of randomness [15].

For the rest of the paper, we assume that a fixed topology governed by unreliable network connections (possible fading channels, MIMO channels). Due to this network unreliability (link failures), nodes have intermittent availability, which can lead to data scarcity across the entire network. This issue becomes particularly difficult, in scenarios characterized by strong data heterogeneity among nodes, where the data distribution of node i, denoted as D_i, is quite different from that of node l, D_l. In such cases, when network connectivity is poor, nodes that rarely participate in the network fail to exchange information with their neighbors. Consequently, the information of the data they hold cannot be diffused in the network.

In other words, from an optimization perspective these infrequently participating nodes may be treated as outliers in the risk-neutral objective (2). However, these nodes contain valuable information and should be considered as useful outliers. *Theoretically, this necessitates a risk measure over the node distribution that can capture information from the tail of the skewed distribution, ensuring that non-often available nodes contribute to the learning procedure.* Therefore, we assume that the network designer assigns to the nodes of the network, to optimize a risk-sensitive objective to ensure that the useful outliers will not be rejected. This technique is robust to the worst-case link failures.

To achieve this, we adopt the *Conditional Value-at-Risk* (CVaR) as our risk measure. As we will explain formally in (II-A) the intrinsic property of CVaR is to focus on the worst-case events (the tail of the skewed node distribution). This opposes the use of expectation in (1), which is biased toward the central region where the probability density is highest. By leveraging CVaR, we will have a risk-aware version of the diffusion-based strategy (6), ensuring that the decentralized federated learning framework remains sensitive to the impact of rarely available nodes.

A. Proposed problem

Following the reasoning of [7] to guarantee efficient classification and simultaneously mitigate the effect of data starvation, we use a risk-aware objective, the *Conditional Value-at-Risk* (CVaR) on the nodes distribution (3). The CVaR of a random variable Z at confidence level $\alpha \in (0, 1]$ is defined as [18]

$$\text{CVaR}^{\alpha}\big[Z\big] := \inf_{t \in R} \left\{ t + \frac{1}{\alpha} \mathbb{E}\big[(Z - t)_+\big] \right\}. \tag{3}$$

The hyper-parameter α governs the behavior of CVaR. Specifically, as α decreases, CVaR shifts further toward the tail of the distribution. In the limiting condition $\lim_{\alpha \to 0} \text{CVaR}^{\alpha}[Z]$ asymptotically approaches the essential supremum of Z (ess sup$Z$). Conversely, as α increases, CVaR moves closer to expectation (formally, $\text{CVaR}^{\alpha}\big[Z\big] = \mathbb{E}[Z]$, for $\alpha = 1$) [18]. The parameter t represents the threshold beyond which the random variable Z is considered extreme. Intuitively, the optimal t^* is essentially the Value-at-Risk (VaR) of Z, at confidence level α [18].

From the node viewpoint and by setting as the random variable Z the empirical training loss $f_i(\theta)$ of the node i the risk-aware problem can be written as

$$\begin{aligned} &\min_{\theta \in R^Q} \left\{ \text{CVaR}^{\alpha}[f_i(\theta)] = \inf_{t \in R} \left\{ t + \frac{1}{\alpha} \mathbb{E}[f_i(\theta) - t]_+ \right\} \right\} = \\ &\min_{\theta \in R^Q} \inf_{t \in R} \left\{ t + \frac{1}{\alpha} \sum_{i=1}^{K} \rho_i [f_i(\theta) - t]_+ \right\} = \\ &\min_{(\theta, t) \in R^Q \times R} \sum_{i=1}^{K} \rho_i \left\{ t + \frac{1}{\alpha} [f_i(\theta) - t]_+ \right\}. \end{aligned} \tag{4}$$

Since the positive part of (4) is by definition a non-differentiable maximization operator we may substitute it with a smooth approximation [19] and the smoothed objective becomes

$$\min_{(\theta, t) \in R^Q \times R} \sum_{i=1}^{K} \rho_i \left\{ \tilde{G}_i(\theta, t) := t + \frac{1}{\alpha} \Big(f_i(\theta) - t + \log\big(1 + e^{-(f_i(\theta) - t)}\big) \Big) \right\}. \tag{5}$$

Following the CTA diffusion technique [4], each node $i \in [K]$ employs steepest-descent iterations to minimize its local function $\tilde{G}_i(\theta, t)$. This process involves receiving parameters (θ_l, t_l) from its neighborhood $\mathcal{N}_i$, and then utilizing step-sizes $\eta_\theta^{'}, \eta_t^{'}$ for variables (θ, t), respectively, as follows

$$\underbrace{\begin{cases} \psi_i^n &= \sum_{l \in \mathcal{N}_i} a_{li} \theta_l^n \\ \lambda_i^n &= \sum_{l \in \mathcal{N}_i} a_{li} t_l^n \end{cases}}_{\textit{Cooperation}}, \underbrace{\begin{cases} \theta_i^{n+1} &= \psi_i^n - \eta_\theta^{'} \rho_i \nabla_\theta \tilde{G}_i(\psi_i^n, \lambda_i^n) \\ t_i^{n+1} &= \lambda_i^n - \eta_t^{'} \rho_i \nabla_t \tilde{G}_i(\psi_i^n, \lambda_i^n). \end{cases}}_{\textit{Adaptation}} \tag{6}$$

For the above updates we have to assume that the neighborhood $\mathcal{N}_i$ will include always the node i itself. This assumption take place due to the dynamic nature of the channel, which may result in the loss

of all neighboring nodes $l \in \mathcal{N}_i/\{i\}$ when transmitting to node i, as illustrated in Fig. (1). Furthermore, we assume that there exist a transition matrix A made by the coefficients $a_{li} \geq 0$ which represent the weight the node i assigns to the information receiving from the node l. This matrix A is considered left-stochastic ($\mathbf{1}^T A = \mathbf{1}^T$) [1].

1) Average-CVaR Approach: In our initial approach, we assume that each node i receives a subset $S_i \subseteq \mathcal{N}_i$ of its neighborhood $\mathcal{N}_i$, where $|S_i|$ denotes the cardinality of S_i. Consequently, the cooperation term in (6) averages the received parameters by assigning weights $a_{li} = \frac{1}{|S_i|}$ and the adaptation term retains its original form using $\eta_\theta = \eta_\theta' \rho_i$, $\eta_t = \eta_t' \rho_i$ as follows,

$$\begin{cases} \psi_i^n &= \sum_{l \in S_i} \frac{1}{|S_i|} \theta_l^n \\ \lambda_i^n &= \sum_{l \in S_i} \frac{1}{|S_i|} t_l^n \end{cases}, \begin{cases} \theta_i^{n+1} &= \psi_i^n - \eta_\theta \nabla_\theta \tilde{G}_i(\psi_i^n, \lambda_i^n) \\ t_i^{n+1} &= \lambda_i^n - \eta_t \nabla_t \tilde{G}_i(\psi_i^n, \lambda_i^n). \end{cases} \tag{7}$$

2) CVaR-CVaR Approach: Our second approach is motivated by the equivalent definition of Conditional Value-at-Risk $\text{CVaR}^\alpha[Z]$ of a random variable Z, related with the Value-at-Risk $\text{VaR}^\alpha[Z]$ [18], we have that

$$\text{CVaR}^\alpha[Z] = \mathbb{E}\big[Z | Z \geq \text{VaR}^\alpha[Z]\big]. \tag{8}$$

Intuitively, we can consider a more risk-sensitive cooperation technique in the expression (7), once each node $i \in K$ has received a set of parameters $\{\theta_l, t_l\}_{l \in S_i}$ and the corresponding empirical training losses $\{f_l(\theta_l)\}_{l \in S_i}$, then each node $i \in K$ can aggregate only the parameters $\{\theta_l, t_l\}_{l \in S_i}$ which correspond to the worst training losses $\{f_l(\theta_l)\}_{l \in S_i}$. This cooperation technique can be justified from the nature of the Conditional Value-at-Risk which is a risk measure that reject the samples that do not exceed the Value-at-Risk of the random variable. Mathematically, we have a set $W_i \subseteq S_i \subseteq \mathcal{N}_i$ such that

$$W_i = \Big\{ l : f_l(\theta) \geq \text{VaR}^\alpha[f_l(\theta)], \textit{ for every } l \in S_i \Big\}, \tag{9}$$

which corresponds to the worst training losses $\{f_l(\theta_l)\}_{l \in S_i}$. So, the new risk-aware update becomes,

$$\begin{cases} \psi_i^n &= \sum_{l \in W_i} \frac{1}{|W_i|} \theta_l^n \\ \lambda_i^n &= \sum_{l \in W_i} \frac{1}{|W_i|} t_l^n \end{cases}, \begin{cases} \theta_i^{n+1} &= \psi_i^n - \eta_\theta \nabla_\theta \tilde{G}_i(\psi_i^n, \lambda_i^n) \\ t_i^{n+1} &= \lambda_i^n - \eta_t \nabla_t \tilde{G}_i(\psi_i^n, \lambda_i^n). \end{cases} \tag{10}$$

The risk-sensitive diffusion-based updates in (7),(10) serve as the foundation for Algorithm of Table (I). The first branch of Algorithm of Table (I) introduces a risk-sensitive modification to the traditional Combine-then-Adapt diffusion technique. In each communication round, the node i receives the parameters $(\theta_l, t_l), l \in S_i \subseteq \mathcal{N}_i$. Then, it combines these parameters assigning weight $\frac{1}{|S_i|}$ to each received parameter. This term is subsequently used for adaptation, where mini-batch-SGD updates are applied over multiple mini-batches of size b, randomly selected from the private dataset D_i to optimize the risk-aware objective in (5). Furthermore, building upon the expression (10) we implement the second branch of Algorithm of Table (I). In this approach, we require each node broadcast not only its learnable parameters but also the value of its training loss. This modification is computationally inexpensive, as it involves transmitting only an additional scalar value. Therefore, we can safely assume that the additional network overload is minimal. Upon receiving the parameters $(\theta_l, t_l, f_l(\theta_l)), l \in S_i \subseteq \mathcal{N}_i$, node i sorts the empirical losses $f_l(\theta_l), l \in S_i$, using the order statistics technique, and keeps only the parameters (θ_l, t_l) corresponding to empirical losses $f_l(\theta_l)$ that exceed the Value-at-Risk at level α (defined by the set W_i in (9)). Then it aggregates assigning weight $\frac{1}{|W_i|}$ to each parameter. Finally, node i uses the aggregated parameters to optimize the risk-sensitive CTA update in (10).

TABLE I: Algorithm **CVaR-CTA**

Initialize $\theta_i^1 = \theta$ and $t_i^1 = t$, for all nodes i. **Set** K, E, T, H, a, b.

for each *global round* $e \in [1, \ldots, E]$ **do**
  for each *com. round* $n \in [1, \ldots, T]$ **do**
    for node $i \in [K]$ in parallel **do**

if *Average-CVaR* is applied **then**
  $S_i \leftarrow$ received subset of $\mathcal{N}_i$
  Combine Step: using (7)
  Adapt Step: using (7)
end if

if *CVaR-CVaR* is applied **then**
  $S_i \leftarrow$ received subset of $\mathcal{N}_i$
  $W_i \leftarrow$ Based on (9)
  Combine Step: using (10)
  Adapt Step: using (10)
end if

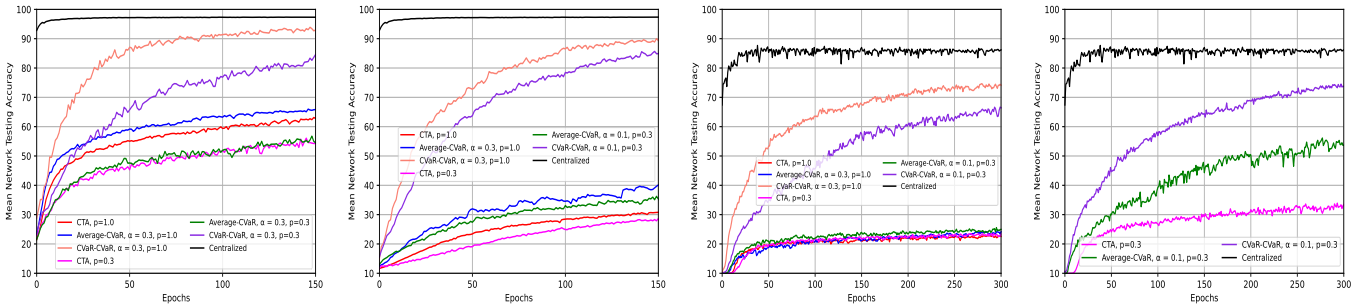
    end for
  end for
end for
procedure ADAPT($\theta_i^{e,n}$,$t_i^{e,n}$)
  $\mathcal{B} \leftarrow$ *split each* D_i *into batches of size* b
  $\begin{bmatrix} \theta_i^{e,n} \\ t_i^{e,n} \end{bmatrix} = \begin{bmatrix} \theta_i^{e,H} \\ t_i^{e,H} \end{bmatrix} \leftarrow$ *Update based on* (7) *or* (10)
end procedure

At this stage, it is important to highlight some critical aspects for the two branches of Algorithm of Table (I). Notably, both versions are reduced to the traditional CTA strategy when the parameter $\alpha = 1$. This follows from the fact that, as previously discussed, the objective (4) simplifies to (2) when $\alpha = 1$. Secondly, as demonstrated in the experimental results, the primary contribution of the risk-sensitive objective lies in the cooperation mechanism. As a result, the first version of Algorithm CVaR-CTA (I) which employs an average aggregation cooperation technique is less effective than the second version, which uses a *CVaR*-based selection scheme for the cooperation step. Finally, it is worth noting that the second approach introduces some computational challenges. Specifically, it relies on computation of Value-at-Risk (*VaR*) (8), which correspond to the upper α-quantile of the sorted list of the training losses. Exact quantile computation may be impractical due to high computational cost, particularly in networks with a large number of neighbors. Lastly, this second approach may also face interpolation issues when the desired quantile falls between discrete data points.

It is noted that, in the DFL literature there is widely employed also the ATC diffusion-based technique. However, due to space limitation, this work focuses exclusively on the CTA diffusion strategy and its risk-sensitive extensions. Future extensions of this research will explore the ATC protocol and its risk-aware variants.

B. Motivation Example

We now illustrate the fundamental difference between risk-neutral and risk-aware DFL objectives. In the DFL framework, each node $i \in [K]$ has access to information of its neighborhood $\mathcal{N}_i$ [4]. Consequently, in the risk-neutral case (traditional CTA method), the optimization problem (1) is formulated as the minimization of the cost function $\sum_{i=1}^K f_i(\theta)$, where $f_i(\theta)$ represents the local cost function at node i [20]. Since each node i communicates exclusively with its neighbors, we introduce the nonnegative coefficients c_{li} that define the relationship between node i and its neighbors $l \in \mathcal{N}_i$, satisfying that $c_{li} \geq 0$, $\sum_{l \in \mathcal{N}_i} c_{li} = 1$, *and* $c_{li} = 0$, *if* $l \notin \mathcal{N}_i$. Using the



(a) **MNIST** (Line) (b) **MNIST** (Ring) (c) **FashionMNIST** (Ring) (d) **FashionMNIST** (Random)

Fig. 2: Mean Network Testing Accuracy for **MNIST** and **FashionMNIST**.

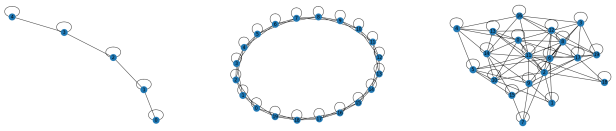


Fig. 3: Line, Ring and Random Graph Topologies

coefficients c_{li}, we define a local cost function for each node $l \in \mathcal{N}_i$, as a weighted combination of the individual costs of its neighboring nodes [20]. Specifically, $f_l^{loc}(\theta) = \sum_{i=1}^{K} c_{li} f_i(\theta)$. Furthermore, the cumulative sum of these local cost functions satisfies

$$\begin{aligned}\sum_{i=1}^{K} f_i(\theta) &= \sum_{i=1}^{K}\left(\sum_{l=1}^{K} c_{li}\right) f_i(\theta) = \sum_{l=1}^{K}\sum_{i=1}^{K} c_{li} f_i(\theta) = \sum_{l=1}^{K} f_l^{loc}(\theta) \\ &= f_i^{loc}(\theta) + \sum_{l=1, l\neq i}^{K} f_l^{loc}(\theta).\end{aligned} \tag{11}$$

In the risk-aware setting (4), the optimization problem is formulated as the minimization of $\sum_{i=1}^{K} \tilde{G}_i(\theta, t)$. Consequently, the local cost becomes as $\tilde{G}_l^{loc}(\theta, t) = \sum_{i=1}^{K} c_{li} \tilde{G}_i(\theta, t)$. Moreover, the cumulative sum of the risk-sensitive local cost functions satisfies the relationship

$$\sum_{i=1}^{K} \tilde{G}_i(\theta, t) = \tilde{G}_i^{loc}(\theta, t) + \sum_{l=1, l\neq i}^{K} \tilde{G}_l^{loc}(\theta, t). \tag{12}$$

With an appropriate choice of the hyper-parameter α (in (5)), it can be shown that the positive part of the risk-aware objective in (5) is activated only for the upper α-quantile of the empirical losses $f_i(\theta)$ (for each fixed θ) [7]. Therefore, the relation in (12) simplifies to the expression (11) when $\alpha = 1$. However, for strictly positive and small values of α, and for fixed values of θ and t, (12) serves as a robust generalization of (11) [21], [18].

III. Convergence Analysis

We now describe a short sketch of the convergence of our proposed Algorithm of Table (I). In the DFL literature, convergence analysis involves two key steps [22]. The first step establishes that the pair parameter of each node (θ_i^n, t_i^n) remains close enough to the centroid pair model $(\bar{\theta}^n, \bar{t}^n)$. The second step demonstrates that the distance error between the centroid pair model $(\bar{\theta}^n, \bar{t}^n)$ and the optimal pair model (θ^*, t^*) is upper bounded. Adopting a network-based viewpoint, we stack the parameters (θ_i^n, t_i^n) into matrices $\Theta = [\theta_1^n, \theta_2^n, \ldots, \theta_K^n]^T$ and $T = [t_1^n, t_2^n, \ldots, t_K^n]^T$. Furthermore, we define the corresponding average matrices $\bar{\Theta} = [\bar{\theta}^n, \bar{\theta}^n, \ldots, \bar{\theta}^n]^T$, and $\bar{T} = [\bar{t}^n, \bar{t}^n, \ldots, \bar{t}^n]^T$. Using this matrix notation [4], the CTA update for the parameter θ can be expressed as $\Theta^{n+1} = \mathcal{A}^T\Theta^n - \eta_\theta \nabla_\theta \tilde{G}\big(\mathcal{A}^T\Theta^n, \mathcal{A}^T T^n\big)$, and for the parameter t can be expressed as $T^{n+1} = \mathcal{A}^T T^n - \eta_t \nabla_t \tilde{G}\big(\mathcal{A}^T\Theta^n, \mathcal{A}^T T^n\big)$. Moreover, we define the consensus matrix $J := \frac{\mathbf{1}\mathbf{1}^T}{K}$, which enables us to express the centroid update as $\bar{\Theta}^n = J\Theta^n$,$\bar{T}^n = JT^n$. Additionally, we introduce the error terms $D_\theta^n := \Theta^n - \bar{\Theta}^n$ and $D_t^n = T^n - \bar{T}^n$. Furthermore, we assume that there exist a mixing parameter $0 \leq r \leq 1$ such that for a left stochastic matrix A (and its Kronecker product $\mathcal{A} = A \otimes I$) we have $\|(\mathcal{A}^T - J\mathcal{A}^T)D^n\| \leq r\|D^n\|$. Finally, assuming bounded gradients, $\|\nabla_\theta G(\mathcal{A}^T\Theta^n, \mathcal{A}^T T^n)\| \leq C_\theta, \|\nabla_t G(\mathcal{A}^T\Theta^n, \mathcal{A}^T T^n)\| \leq C_t$ and further mathematical manipulations, we obtain

$$\left\|D_\theta^{n+1}\right\|^2 + \left\|D_t^{n+1}\right\|^2 \leq 2r^2\left(\left\|D_\theta^n\right\|^2 + \left\|D_t^n\right\|^2\right) + 2\left(\eta_\theta^2 C_\theta^2 + \eta_t^2 C_t^2\right) \tag{13}$$

For the second branch of Algorithm (I), the bound follows the same structural form. However, the key difference lies in the parameter r, where we now adopt a more robust choice, denoted as r_{robust}, along with the corresponding constants $C_{robust,\theta}, C_{robust,t}$.

For the second step, we assume $\tilde{G}_i(\theta, t)$ is convexity and L-smooth, with L_θ, L_t for θ and t, respectively. We further assume bounded gradient variances $\bar{\sigma}_\theta, \bar{\sigma}_t$, for both parameters. Building on the results in [22], we obtain that

$$\begin{aligned}&\mathbb{E}\left[\left\|\bar{\theta}^{n+1} - \theta^*\right\|^2 + \left(\bar{t}^{n+1} - t^*\right)^2\right] \leq \cdots \leq \left\|\bar{\theta}^n - \theta^*\right\|^2 + \left|\bar{t}^n - t^*\right|^2 \\ &+ \frac{1}{K}(\eta_\theta \bar{\sigma}_\theta^2 + \eta_t \bar{\sigma}_t^2) - 2\eta_\theta\left(\tilde{G}(\bar{\theta}^n, \cdot) - \tilde{G}^*\right) - 2\eta_t\left(\tilde{G}(\cdot, \bar{t}^n) - \tilde{G}^*\right) \\ &+ \frac{1}{K}\sum_{i=1}^{K}\left(L_\theta\eta_\theta\left\|\bar{\theta}^n - \sum_{l\in\mathcal{N}_i} a_{li}\theta_l^n\right\|^2 + L_t\eta_t\left|\bar{t}^n - \sum_{l\in\mathcal{N}_i} a_{li} t_l^n\right|^2\right) \\ &+ \frac{2}{K}\sum_{i=1}^{K}\left(\eta_\theta L_\theta^2\left\|\sum_{l\in\mathcal{N}_i} a_{li}\theta_l^n - \bar{\theta}^n\right\|^2 + \eta_t L_t^2\left|\sum_{l\in\mathcal{N}_i} a_{li} t_l^n - \bar{t}^n\right|^2\right) \\ &+ \frac{4}{K}(\eta_\theta L_\theta + \eta_t L_t)\left(\tilde{G}(\bar{\theta}^n, \bar{t}^n) - \tilde{G}^*\right).\end{aligned} \tag{14}$$

Our convergence analysis demonstrates that network stability ensures linear decay of consensus error across nodes. The optimality gap is governed by the gradient noise and the step-sizes selection, revealing a trade-off between the convergence speed and steady-state accuracy.

IV. Experimental Results

Our experiments conducted on three distinct network topologies, a line (5 nodes), a ring (20 nodes) and a random Erdos-Renyi (20 nodes) graph (3). We assess the proposed Algorithm of Table (I) using the MNIST and Fashion-MNIST datasets, where each consists of 60.000 training samples across 10 distinct classes. The experiments conducted with data heterogeneity 0 (strong non-i.i.d.) for both datasets, in accordance with [23]. This data partitioning strategy intentionally simulates network conditions where intermittent connectivity contributes to localized data scarcity. The network connectivity is modeled through two probability regimes, the fully reliable ($p_{li} = p = 1.0$) with guaranteed neighbor transmission

reception, and the unreliable ($p_{li} = p = 0.3$) where nodes retain only 30% probability of receiving transmissions from neighbors, excluding self-connections which remain consistently available. Each experiment was simulated 10 times with mean testing accuracy over the network of nodes measured after each epoch. The code for all experiments is available at [24].

For MNIST results in Fig. [(2a), (2b)], global epochs are set as $E = 150$ (line and ring topology), with $T = 6$ inter-epoch communication rounds for neighbor information exchange. A neural network with two fully connected hidden layers with neurons $(200, 200)$ [25] is implemented, using constant step-sizes $\eta_\theta = \eta_t = 10^{-3}$ for *Average-CVaR* version of Algorithm of Table (I) and $\eta_\theta = \eta_t = 10^{-4}$ for *CVaR-CVaR* branch, with $H = 10$ local iterations of batch size $b = 100$ per user. As demonstrated in Fig. [(2a), (2b)] the performance of traditional CTA techniques in a reliable network setting ($p = 1.0$) exceeds the CTA techniques in an unreliable network ($p = 0.3$) across both graph topologies. Specifically, for both reliable ($p = 1.0$) and unreliable ($p = 0.3$) networks, the *CVaR-CVaR* approach outperforms the *Average-CVaR* approach, which is better than traditional CTA. Ultimately, the superior performance of the proposed method is achieved for the *CVaR-CVaR* approach, highlighting the effectiveness of the risk-aware node selection version.

For FashionMNIST experiments in Fig. [(2c),(2d)] global epochs are set as $E = 300$ for the ring and random graph topologies, with $T = 10$ inter-epoch communication rounds for neighbor parameter information exchange. A CNN architecture follows [25], using two 5×5 convolutional layers (with 6 and 16 channels, respectively, each followed with 2×2 max pooling) and two fully connected layers with 120 and 84 neurons. The local iterations, the batch size and the step-sizes are fixed at $H = 10, b = 100, \eta_\theta = \eta_t = 10^{-3}$. As illustrated, in this scenario in Fig. (2c), both the traditional CTA technique and the *Average-CVaR* approach fail to effectively solve the problem in both reliable ($p = 1.0$) and unreliable ($p = 0.3$) networks setting. However, the *CVaR-CVaR* version of Algorithm (I) demonstrates improved performance, with the reliable network case surpassing the performance of the unreliable network, as expected. Finally, as shown in Fig. (2d), the *CVaR-CVaR* approach consistently outperforms the *Average-CVaR* method, which in turn performs better than the classic CTA procedure, in the random Erdos-Renyi graph topology (2d), under the unreliable ($p = 0.3$) network setting.

V. Conclusion

In this study, we investigate Decentralized Fedeted Learning (DFL) under limited node connectivity. We consider communication links as erasure channels where each node broadcasts its model parameters, however neighboring nodes receive these parameters with a low probability. Under this setting, we proposed two risk-aware extensions of the Combine-then-Adapt (CTA) technique. The first approach aggregates the received parameters and optimizes the *CVaR* objective, while the second approach selectively rejects a part of the received parameters based on their *CVaR* behavior before adapting the risk-sensitive function. Through experimental evaluation conducted on Mnist and FashionMnist datasets, we demonstrate that the risk-sensitive CTA approaches outperform the classic CTA technique, particularly when employing risk-sensitive parameter selection.

References

[1] E. Rizk, K. Yuan, and A. H. Sayed, "Asynchronous diffusion learning with agent subsampling and local updates," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 9246–9250.

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

[3] E. Georgatos, C. Mavrokefalidis, and K. Berberidis, "Neighbor selecting cooperation strategies for distributed learning," in *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2024, pp. 2242–2246.

[4] A. H. Sayed, *Inference and Learning from Data: Foundations*. Cambridge University Press, 2022, vol. 1.

[5] J. Wang et al., "A field guide to federated optimization," *arXiv preprint arXiv:2107.06917*, 2021.

[6] Y. Ruan, X. Zhang, S.-C. Liang, and C. Joe-Wong, "Towards flexible device participation in federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 3403–3411.

[7] P. Theodoropoulos, K. E. Nikolakakis, and D. Kalogerias, "Federated learning under restricted user availability," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 7055–7059.

[8] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[9] W. Li, T. Lv, W. Ni, J. Zhao, E. Hossain, and H. V. Poor, "Decentralized federated learning over imperfect communication channels," *IEEE Transactions on Communications*, 2024.

[10] H. Ye, L. Liang, and G. Y. Li, "Decentralized Federated Learning with Unreliable Communications," *IEEE journal of selected topics in signal processing*, vol. 16, no. 3, pp. 487–500, 2022.

[11] K. Cho, J. Park, A. Rizwan, and M. Choi, "Serverless federated learning in multi-cell scenarios robust to test-time distribution shifts," in *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2024, pp. 241–245.

[12] P. Valdeira, S. Wang, and Y. Chi, "Vertical federated learning with missing features during training and inference," *arXiv preprint arXiv:2410.22564*, 2024.

[13] M. Shirvanimoghaddam, A. Salari, Y. Gao, and A. Guha, "Federated learning with erroneous communication links," *IEEE communications letters*, vol. 26, no. 6, pp. 1293–1297, 2022.

[14] Z. Chen, W. Yi, Y. Liu, and A. Nallanathan, "Robust federated learning for unreliable and resource-limited wireless networks," *IEEE Transactions on Wireless Communications*, 2024.

[15] N. Dal Fabbro, A. Mitra, and G. J. Pappas, "Federated TD learning over finite-rate erasure channels: Linear speedup under markovian sampling," *IEEE Control Systems Letters*, vol. 7, pp. 2461–2466, 2023.

[16] Y. Lyu, W. Luo, and J. M. Dolan, "Risk-aware safe control for decentralized multi-agent systems via dynamic responsibility allocation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1–8.

[17] Z. Zhai, X. Yuan, and X. Wang, "Decentralized federated learning via MIMO Over-the-Air Computation: Consensus Analysis and Performance Optimization," *IEEE Transactions on Wireless Communications*, 2024.

[18] A. Shapiro et al., *Lectures on Stochastic Programming: Modeling and Theory*, ser. MOS-SIAM Series on Optimization. SIAM & Mathematical Optimization Society, Philadelphia, 2014.

[19] F. Beiser, B. Keith, S. Urbainczyk, and B. Wohlmuth, "Adaptive sampling strategies for risk-averse stochastic optimization with constraints," *IMA Journal of Numerical Analysis*, vol. 43, no. 6, pp. 3729–3765, 2023.

[20] N. Ghadban, P. Honeine, F. Mourad-Chehade, C. Francis, and J. Farah, "Diffusion strategies for in-network principal component analysis," in *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2014, pp. 1–6.

[21] D. S. Kalogerias, "Noisy linear convergence of stochastic gradient descent for CV@R statistical learning under Polyak-Lojasiewicz conditions," *arXiv preprint arXiv:2012.07785*, 2020.

[22] B. Le Bars, A. Bellet, M. Tommasi, E. Lavoie, and A.-M. Kermarrec, "Refined convergence and topology learning for decentralized sgd with heterogeneous data," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 1672–1702.

[23] Z. Shen, J. Cervino, H. Hassani, and A. Ribeiro, "An agnostic approach to federated learning with class imbalance," in *International Conference on Learning Representations*, 2021.

[24] P. Theodoropoulos, "Risk-aware dfl: Simulations," https://github.com/PeriklisTheodoropoulos/Risk-Aware-DFL, 2025.

[25] B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.