Balancing Energy Efficiency and Distributional Robustness in Over-the-Air Federated Learning

Mohamed Badi, Chaouki Ben Issaid, Anis Elgabli* and Mehdi Bennis
Centre of Wireless Communications (CWC), University of Oulu, Finland
*Interdisciplinary Research Center for Communication Systems and Sensing (IRC-CSS),
Industrial and System Engineering Department, KFUPM, Dhahran
Email: {Mohamed.badi, chaouki.benissaid, mehdi.bennis}@oulu.fi, anis.elgabli@kfupm.edu.sa

***Abstract*—The growing number of wireless edge devices has magnified challenges concerning energy, bandwidth, latency, and data heterogeneity. These challenges have become bottlenecks for distributed learning. To address these issues, this paper presents a novel approach that ensures energy efficiency for distributionally robust federated learning (FL) with over air computation (AirComp). In this context, to effectively balance robustness with energy efficiency, we introduce a novel client selection method that integrates two complementary insights: a deterministic one that is designed for energy efficiency, and a probabilistic one designed for distributional robustness. Simulation results underscore the efficacy of the proposed algorithm, revealing its superior performance compared to baselines from both robustness and energy efficiency perspectives, achieving more than 3-fold energy savings compared to the considered baselines.**

***Index Terms*—AirComp, distributionally robust optimization (DRO), energy efficiency, Federated learning, agnostic learning.**

I. Introduction

In our rapidly digitizing world, a multitude of edge devices, from smartphones to Internet of Things (IoT) systems, are continuously generating vast amounts of private data. This deluge of raw data brings about challenges in storage, capacity, and processing. Introduced in [1], federated learning (FL) proposes a solution to these challenges by allowing devices to learn collaboratively. Instead of transferring large amounts of data to a central location, devices learn a shared global model by communicating model parameters or gradients, thereby ensuring adherence to stringent privacy constraints.

However, FL comes with its own set of challenges. With a diverse range of edge devices operating in different environments, assuming an identical data distribution is often unrealistic. This diversity can result in biases, potentially favoring certain devices over others, leading to fairness concerns. To address this, the authors in [2] introduced a robust min-max formulation, ensuring that the learned model performs effectively over the worst-case combination of local empirical distributions. Furthermore, a communication-efficient iterative descent-ascent algorithm was proposed to solve this problem. This approach uniquely selects only a subset of clients in each communication round. Many recent works have built upon the formulation presented in [2], adapted for both centralized and decentralized settings [3]–[5]. On the other hand, scalability and latency are among the major challenges when integrating FL into wireless edge-centric systems [6]. AirComp, with its ability to exploit the superposition property of the wireless channel, emerges as a significant solution, especially with the exponential growth of IoT devices. Several studies such as [7]–[9], have explored the use of AirComp, highlighting its communication efficiency and advantages over traditional digital schemes. Beyond scalability and latency, energy consumption is another significant concern in wireless communication systems. Transmitting data wirelessly is energy-intensive, especially for battery-powered devices, which makes efficiency in this area essential. Dynamic client scheduling has recently been considered in a number of related works. For example, [10] proposes to integrate both channel conditions and gradient norms in client selection decisions, under the mild assumptions that the maximum gradient norm and the maximum channel condition across all clients are known in advance. Yet, this approach has drawbacks, including its reliance on several tuning parameters together with the heuristic construction of its indicator. Moreover, its adaptability with scheduled clients brings about unpredictability. For a given set of tuning parameters, the expected number of scheduled clients cannot be determined because it is intrinsically tied to the characteristics of the training problem, which cannot be known in advance. The growing emphasis on edge learning highlights the need to maximize edge devices' performance while adapting to diverse data distributions in an energy-efficient manner. Towards this goal, this work proposes a distributed learning approach that balances between energy efficiency and distributional robustness in model training. The main contributions can be summarized as follows:

- To the best of our knowledge, this is the first work to jointly address the problem of energy efficiency, and robustness to data heterogeneity, in terms of communication costs.
- We theoretically show that at the extremes of our configurable energy-conservation tuning parameter, the algorithm defaults either to the conventional distributionally robust AFL algorithm, which lacks channel awareness, or into a fully energy-conservative client selection algo-

This work was supported by the European Commission through Grant no. 101095363 (Horizon Europe SNS JU ADROIT6G project).

rithm. This latter approach selects clients with the lowest energy cost without performance guarantees. For intermediate values, the proposed client selection mechanism results in a blend of the two aforementioned metrics, offering a seamless transition between established algorithms and striking a balance between energy efficiency and distributional robustness performance.

- Our numerical results show that our proposed algorithm, coined Channel-Aware Agnostic Federated Learning (CA-AFL), can achieve significant energy savings, up to one-third of the energy consumption, compared to the conventional AFL algorithm [2], with only a negligible reduction in performance. Moreover, these results demonstrate that our proposed algorithm can surpass the performance of the energy-efficient GCA algorithm [10] in terms of both worst client accuracy and global standard deviation (STD), even when the GCA algorithm's tuning parameters are experimentally configured for optimal performance.

The rest of the paper is organized as follows. Section II introduces the system model and the problem formulation. Section III details the proposed algorithm. In Section IV, we evaluate the average and worst test accuracy of our algorithm, CA-AFL, against several baselines in terms of communication rounds and total energy. Finally, the paper concludes with final remarks in Section V.

II. System Model and Problem Formulation

We consider a distributed learning system as depicted in Fig. 1. In this system model, a parameter server (PS) functions as a central coordinator serving a total of N edge devices. Unlike digital transmission systems, where devices are required to access communication resources orthogonally to ensure successful decoding at the PS, in AirComp devices exploit analog transmission and the superposition property of the wireless channel in the uplink. This enables efficient over-the-air model aggregation. To reduce the average energy consumption, at each communication round t, only K out of the N edge devices are selected to upload their models to the PS according to a probability distribution $\boldsymbol{\rho}^{(t)}$ defined over a combination of both energy and distributional robustness metrics. We assume that orthogonal frequency division multiplexing (OFDM) is utilized at the physical layer. Hence, leveraging AirComp the aggregated signal at the PS over the bth sub-carrier is expressed as

$$y_b = \sum_{i \in D^{(t)}} h_{i,b} \times x_{i,b} + z_b, \tag{1}$$

where $D^{(t)}$ is the set of sampled clients, $x_{i,b}$ represents the data stream of the ith client on the bth sub-carrier, z_b is the additive white Gaussian noise, and $h_{i,b}$ characterizes the channel of the bth sub-carrier between the ith client and the PS. We assume uplink-downlink channel symmetry throughout this paper.

As illustrated in Fig. 1, once local models are averaged at the PS, the global model is then broadcasted to all clients

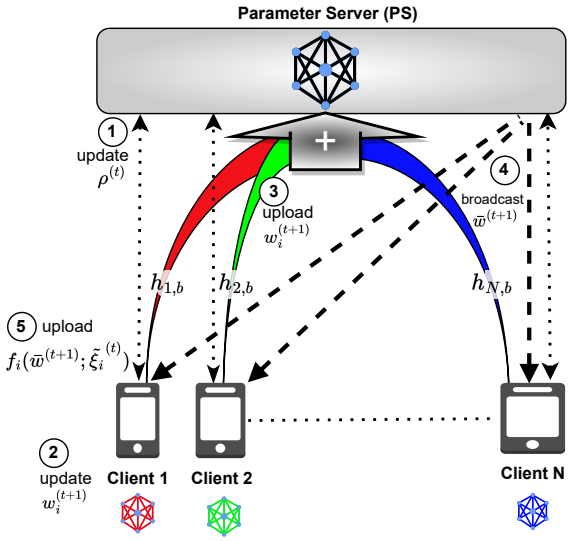


Fig. 1: Illustration of the AirComp System.

over a common broadcast channel. Additionally, the same figure highlights a dedicated bidirectional control channel, used to facilitate the exchange of control signals between the clients and the PS. The objective of distributionally robust optimization is to train a global model that is robust to data heterogeneity among clients. As initially discussed in [2], this optimization can be formulated as follows

$$\textbf{(P1)} \quad \min_{\boldsymbol{w}} \max_{\boldsymbol{\lambda} \in \Delta^{N-1}} \left(\sum_{i=1}^{N} \lambda_i f_i(\boldsymbol{w}) \right), \tag{2}$$

where $\Delta^{N-1} = \{\boldsymbol{\lambda} \in \mathbb{R}_+^N : \sum_{i=1}^N \lambda_i = 1\}$ is the $(N-1)$ standard simplex space. In this context, $\boldsymbol{w}$ signifies the shared global model and $f_i(\boldsymbol{w})$ is the local objective function for the ith client. It is presumed that the local data for the ith client is drawn from a distribution that may vary from that of other clients.

The energy required for the ith client, employing the channel inversion technique to upload their model to the PS is expressed as

$$E_i^{(t)} = P_i^{(t)} \times t_{\text{trans}}, \tag{3}$$

where the transmission time t_{trans} is given by $t_{\text{trans}} = (M/N_{\text{sc}}) \times \tau$, with M and N_{sc} being the number of elements in the model and the number of sub-carriers, respectively, and τ representing the symbol period. The transmission power $P_i^{(t)}$ is defined as

$$P_i^{(t)} = \sum_{b=1}^{N_{\text{sc}}} \frac{|x_{i,b}^{(t)}|^2}{|h_{i,b}^{(t)}|^2}, \tag{4}$$

$x_{i,b}^{(t)} = \sqrt{\psi} \times s_{i,b}^{(t)}$, where ψ is a scaling factor that is assumed to be constant across carriers and clients and $s_{i,b}^{(t)}$ is the actual analog value that is transmitted at t. It is noteworthy that the transmission power is influenced by both the symbol power

and the power necessary for channel inversion. Given that the symbol power directly reflects the learning procedure and should therefore not be a factor in the scheduling decision, our focus in this work is primarily on minimizing the energy consumption attributed to channel inversion. To that end, we define

$$\tilde{P}_i^{(t)} = \psi \sum_{b=1}^{N_{\text{sc}}} \frac{1}{|h_{i,b}^{(t)}|^2} = \psi \frac{N_{\text{sc}}}{|h_i^{(t)}|^2}, \tag{5}$$

where the effective channel, $h_i^{(t)}$, is defined as

$$\frac{1}{|h_i^{(t)}|^2} = \frac{1}{N_{\text{sc}}} \sum_{b=1}^{N_{\text{sc}}} \frac{1}{|h_{i,b}^{(t)}|^2}. \tag{6}$$

Therefore, the energy attributed to the scaling and inversion operations at each transmission can be re-expressed as $\tilde{E}_i^{(t)} = \frac{\psi \times M \times \tau}{|h_i^{(t)}|^2}$. Furthermore, in this work, we denote $E^{(t)}$ as the cumulative energy consumed by the K clients selected in the tth round for model transmission to the PS. This energy is given by $E^{(t)} = \sum_{i \in D^{(t)}} \tilde{E}_i^{(t)}$.

III. Proposed Algorithm

The challenge of incorporating energy efficiency with distributional robustness into the client selection process arises from its inherent probabilistic nature, as proposed in the communication-efficient descent-ascent algorithms used to solve **(P1)** [2]–[4]. In the descent step, K clients are sampled according to the probabilities defined by $\boldsymbol{\lambda}^{(t)}$ to upload their models to the PS. However, this raises a potential issue: the selected clients might not have optimal communication channels, leading to intensive energy consumption. Hence, a refined client selection design that considers energy is crucial. With two metrics, namely the energy $E^{(t)}$ and the probability distribution $\boldsymbol{\lambda}^{(t)}$, our scheme transforms the deterministic energy-based client selection metric into a bias-configurable probability mass function (PMF). Governed by a tuning parameter, this allows for a smooth bias transition, which can either assign equal probabilities to all clients or give strong preference to those with the highest effective communication channels, who in turn require the least energy for model upload. In the following proposition, we provide a formal representation of this energy-conservative PMF.

Proposition 1. *The entries of the bias configurable PMF can be expressed as*

$$y_i^{(t)} = \frac{\left|h_i^{(t)}\right|^C}{\sum_{j=1}^N \left|h_j^{(t)}\right|^C}, \tag{7}$$

where $y_i^{(t)}$ is the i-th client's sampling probability at round t and C is denoted as the energy-conservation tuning factor.

Proof. The proof can be found in Appendix VI-A. □

For $C = 0$, the PMF (7) is unbiased; as C increases, it increasingly favors clients with better channels, becoming fully biased as $C \to \infty$.

Next, we introduce the probability distribution $\boldsymbol{\rho}^{(t)}$, derived from the product of two distinct PMFs. At round t, K clients are sampled according to

$$\rho_i^{(t)} = \frac{\lambda_i^{(t)} y_i^{(t)}}{\sum_{j=1}^N \lambda_j^{(t)} y_j^{(t)}}. \tag{8}$$

With this probability sampling, we aim to achieve consensus between two potentially conflicting insights. By multiplying the associated probabilities of these PMFs and subsequently normalizing the result by their summation, the weight assigned to clients evolves accordingly. Specifically, clients with elevated probabilities in both PMFs are prioritized, while clients dominant in only one PMF are given lower probabilities. Those who rank low in both PMFs are the least favored. This idea of multiplying different probability distributions and normalizing them is akin to the idea of the product of experts (PoE) in ML [11], in which multiple probabilistic models are combined to capture complex distributions by leveraging the strengths of each individual model. In our work, we regard each PMF as an "expert". One expert captures insights based on distributional robustness, whereas the other is geared towards energy conservation. The tuning parameter C serves a pivotal role in our algorithm, functioning as a gating mechanism. When the energy-conservative expert is unbiased, i.e., $C \to 0$, the robustness expert takes precedence, and our algorithm defaults to the non-channel-aware AFL [2]. On the other hand, when the energy efficiency expert is fully biased, i.e., $C \to \infty$, it completely overrides the distributional robustness expert. For values between the two extremes, the combined PMF will represent a blend of the two insights. Additionally, (8) can be simplified using (7) as

$$\rho_i^{(t)} = \frac{\lambda_i^{(t)} \left|h_i^{(t)}\right|^C}{\sum_{j=1}^N \lambda_j^{(t)} \left|h_j^{(t)}\right|^C}. \tag{9}$$

In the following proposition, we state the limiting behaviour of our algorithm as $C \to \infty$.

Proposition 2. *Our proposed algorithm, CA-AFL, reverts to the top-K greedy energy-efficient client selection as $C \to \infty$.*

Proof. The proof is deferred to Appendix VI-B. □

With the definition of the client selection probability given in (9), our proposed algorithm operates as follows: In the descent step, K clients are sampled according to the probabilities $\boldsymbol{\rho}^{(t)}$. These clients subsequently update and upload their models using a batch of size $|\xi_i^{(t)}|$. Following this, the PS broadcasts the aggregated model to all clients. On the other hand, during the ascent step, K clients are uniformly sampled to update the values of the probability simplex vector $\boldsymbol{\lambda}^{(t)}$, employing a batch of size $|\tilde{\xi}_i^{(t)}|$. Notably, because it is a scalar, this update does not demand a high communication cost and can be conveniently exchanged with the PS over a control channel as shown in Fig. 1. The detailed steps are summarized in Algorithm 1.

Algorithm 1 Channel-Aware Agnostic Federated Learning (CA-AFL)

Input: N Clients, K sampled clients, T total number of iterations, energy-conservation factor C , learning rates (η, γ), initial model $\bar{\boldsymbol{w}}^{(0)}$, and initial $\boldsymbol{\lambda}^{(0)}$.

Output: $\bar{\boldsymbol{w}}^{(T)}$, $\boldsymbol{\lambda}^{(T)}$

for $t = 0$ to $T-1$ **do**
  PS updates $\boldsymbol{\rho}^{(t)}$ according to (9)
  PS samples K clients $D^{(t)}$ according to $\boldsymbol{\rho}^{(t)}$
  for clients $i \in D^{(t)}$ **do** in parallel
    $\boldsymbol{w_i}^{(t+1)} = \boldsymbol{w_i}^{(t)} - \eta \nabla f_i(\boldsymbol{w_i}^{(t)}; \xi_i^{(t)})$
    client uploads $\boldsymbol{w_i}^{(t+1)}$ to the PS
  end for
  PS averages the AirComp aggregated models

$$\bar{\boldsymbol{w}}^{(t+1)} = \frac{\sum_{i \in D^{(t)}} \boldsymbol{w_i}^{(t+1)} + z^{(t)}}{K} \tag{10}$$

  PS broadcasts $\bar{\boldsymbol{w}}^{(t+1)}$ to all clients
  PS samples K clients uniformly $U^{(t)}$
  for clients $i \in U^{(t)}$ **do** in parallel
    compute and upload $f_i(\bar{\boldsymbol{w}}^{(t+1)}; \tilde{\xi}_i^{(t)})$ to the PS
    PS updates $\tilde{\boldsymbol{\lambda}}^{(t+1)}$ according to

$$\tilde{\lambda_i}^{(t+1)} = \lambda_i^{(t)} + \gamma f_i(\bar{\boldsymbol{w}}^{(t+1)}; \tilde{\xi}_i^{(t)})$$

  end for
  PS computes $\boldsymbol{\lambda}^{(t+1)} = \Pi_{\Delta}(\tilde{\boldsymbol{\lambda}}^{(t+1)})$
end for

IV. Numerical Results

A. Simulation Setup

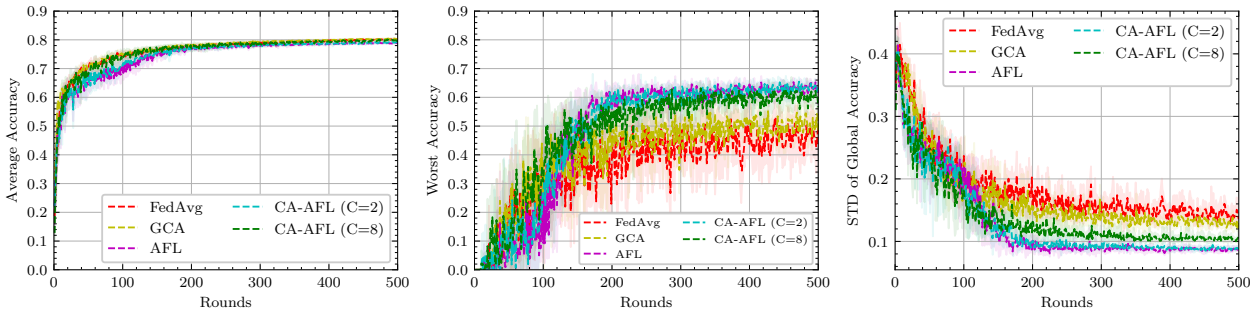
In this section, we detail the performance of our proposed algorithm, CA-AFL, and evaluate its effectiveness compared to other existing algorithms. In our experiments, we utilized the Fashion MNIST dataset and deployed a logistic regression model of size $M = 7850$. Our setup includes a total of $N = 100$ devices, out of which $K = 40$ actively participate in each communication round. We sorted the 60000 training data samples by label, then divided them into 100 equal-sized shards. This approach ensures that each client receives one shard, promoting a high degree of data heterogeneity (non-iidness) among clients. The training process spanned $T = 500$ rounds with a batch size of 50. For the descent step, we use an exponentially decaying learning rate with initial value $\eta^{(0)} = 0.1$ and a decaying rate of 0.998 per iteration. The ascent step size was set to $\gamma = 8 \times 10^{-3}$. Regarding communication, for a fair comparison against the non-channel-aware AFL algorithm in terms of energy consumption, we did not impose any power control mechanism. Instead, we adopted an independent and identically distributed (i.i.d.) truncated flat-fading Rayleigh distributed channel block $\mathcal{CN}(0,1)$, where $h \geq 0.05$. This setup favored the non-channel-aware baseline AFL. In the most challenging scenario, we assumed the channel remained coherent for only one communication round. The scaling factor was set at $\psi = 0.5$ mW, and the symbol period, τ, aligned with LTE standards at 1 ms. We averaged the results over five simulation runs. For the purpose of comparison, we consider the following baselines

- **FedAvg** [1]: The original algorithm that is neither distributionally robust nor channel-aware. In this approach, K clients are randomly sampled to upload their models to the parameter server for averaging.
- **AFL** [2]: A distributionally robust but non-channel-aware algorithm designed for solving the min-max formulation. In each round, K clients are sampled based on $\lambda^{(t)}$ for the descent steps, and a separate set of K clients is sampled uniformly for the ascent step update.
- **GCA** [10]: A non-robust, gradient and channel-aware algorithm that dynamically schedules clients based on a composite indicator consisting of energy, channel condition, and gradient norm. Its tuning parameters are $\lambda_E = 0.5$, $\lambda_V = 0.5$, $\rho_1 = 0.5$, $\rho_2 = 0.5$, following [10], while $\sigma_t = 1$ and $\alpha = 1500$ are experimentally optimized, resulting in an average of 42 scheduled clients.

B. Results and Discussion

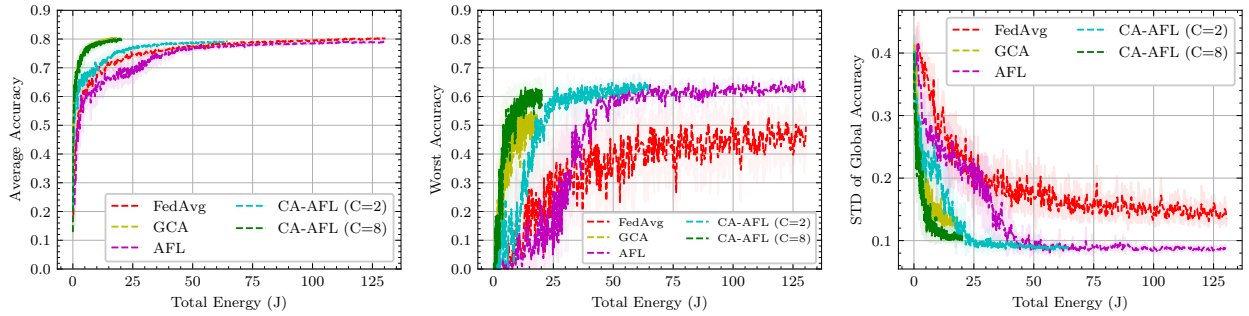
In Fig. 2, we depict the performance metrics in terms of average global accuracy, worst client accuracy, and the STD of the global accuracy versus the number of communication rounds. As shown in Fig. 2a, the proposed CA-AFL together with the baselines exhibit comparable average global accuracy with a convergence value of 80%. Fig. 2b illustrates that CA-AFL closely matches the performance of the non-channel-aware benchmark AFL algorithm in terms of worst client accuracy, with only a negligible degradation for $C = 8$. It delivers approximately 10% higher worst accuracy than both FedAvg and the GCA algorithm. Notably, CA-AFL attains the 50% worst accuracy, which is the maximum achieved by GCA and FedAvg, in less than half of the communication rounds compared to both FedAvg and GCA. Fig. 2c underscores that for varying values of the tuning parameter C, CA-AFL outperforms both GCA and FedAvg in terms of STD, which directly translates into more reliability in terms of fairness. CA-AFL aligns closely with AFL's STD for $C = 2$ and remains competitive for $C = 8$. Similar to the worst accuracy metric, CA-AFL reaches the convergence STD values of FedAvg and GCA baselines with significantly fewer communication rounds. The superior performance of GCA compared to FedAvg is attributed to incorporating user gradient norms in the client selection mechanism, which accelerates convergence. However, as expected, increasing the value of C leads to a degradation in terms of both worst-client accuracy and STD compared to the robust benchmark AFL.

In Fig. 3, we plot the accuracy performance metrics, including average accuracy, worst client accuracy, and the STD of global accuracy, against the total energy expenditure for uploading. As presented in Fig. 3a, FedAvg and AFL algorithms emerge as the most energy-intensive. This aligns with our expectations since neither of them is channel-aware. The figure further shows that increasing C for the proposed



(a) Average test accuracy (b) Worst test accuracy (c) STD

Fig. 2: Performance of CA-AFL compared to baselines in terms of the number of communication rounds.



(a) Average test accuracy (b) Worst test accuracy (c) STD

Fig. 3: Performance of CA-AFL compared to baselines in terms of the total energy.

CA-AFL leads to comparable average accuracy results but at considerably reduced energy costs. Notably, at $C = 8$, CA-AFL attains an energy efficiency that is on par with GCA algorithm without compromising performance. Fig. 3b underscores that CA-AFL delivers performance metrics in line with the robustness benchmark but with significantly reduced energy demands. Remarkably, at $C = 8$, CA-AFL not only matches the energy efficiency of GCA, it also achieves the best attainable worst client accuracy for GCA and FedAvg with substantially less energy consumption. Fig. 3c shows that CA-AFL outperforms the robustness benchmark AFL in energy efficiency, incurring only a minimal performance trade-off. At $C = 8$, CA-AFL surpasses GCA in terms of STD while consuming the same amount of energy, even with GCA operating at its optimal hyperparameter configuration. On the other hand, it is obvious that CA-AFL achieves the best attainable STD of GCA with much lower energy. An intriguing observation is that the proposed CA-AFL can match the performance of the benchmark AFL algorithm while consuming only one-third of the energy.

V. Conclusion

This work introduced a novel approach to incorporate energy efficiency considerations into the clients' selection process governed by the solution of the distributionally robust min-max formulation. Our proposed algorithm leverages a tuning configurable parameter, transitioning between two well-established algorithms—AFL and top-K greedy selection. Through extensive simulations, we have demonstrated that our algorithm outperforms existing baselines, in terms of both energy efficiency and distributional robustness.

References

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

[2] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 4615–4625.

[3] Y. Deng, M. M. Kamani, and M. Mahdavi, "Distributionally robust federated averaging," *Advances in neural information processing systems*, vol. 33, pp. 15 111–15 122, 2020.

[4] M. Zecchin, M. Kountouris, and D. Gesbert, "Communication-efficient distributionally robust decentralized learning," *Transactions on Machine Learning Research*, 2022.

[5] C. B. Issaid, A. Elgabli, and M. Bennis, "DR-DSGD: a distributionally robust decentralized learning algorithm over graphs," *Transactions on Machine Learning Research*, 2022.

[6] H. Hellström, J. M. B. da Silva Jr, M. M. Amiri, M. Chen, V. Fodor, H. V. Poor, C. Fischione *et al.*, "Wireless for machine learning: A survey," *Foundations and Trends® in Signal Processing*, vol. 15, no. 4, pp. 290–399, 2022.

[7] M. M. Amiri and D. Gündüz, "Over-the-air machine learning at the wireless edge," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2019, pp. 1–5.

[8] ——, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2155–2169, 2020.
[9] A. Elgabli, J. Park, C. B. Issaid, and M. Bennis, "Harnessing wireless channels for scalable and privacy-preserving federated learning," *IEEE Transactions on Communications*, vol. 69, no. 8, pp. 5194–5208, 2021.
[10] J. Du, B. Jiang, C. Jiang, Y. Shi, and Z. Han, "Gradient and channel aware dynamic scheduling for over-the-air computation in federated edge learning systems," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 4, pp. 1035–1050, 2023.
[11] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

VI. Appendices

A. Proof of Proposition 1

We define $G : \mathbb{R}_+^N \to \Delta^{N-1}$ as a transformation that projects an N-dimensional vector of non-negative real numbers onto an $(N-1)$-dimensional probability simplex. Let $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]$ be a vector in $\mathbb{R}_+^N$ and $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]$ be a vector in Δ^{N-1}. Additionally, we consider the symmetric group S_N, which represents all possible permutations of N distinct elements. The transformation G adheres to the following properties:

- **Permutation equivariance:** The transformation G is equivariant under the action of the symmetric group S_N on $\boldsymbol{x}$. In other words, any permutation of the inputs $\boldsymbol{x}$ will yield an equivalent permutation of the outputs $\boldsymbol{y}$.
- **Permutation invariance for non-corresponding inputs:** For any given output y_i, G is invariant under the action of the permutation group S_{N-1} on the set of inputs $\{x_j : j \in \{1, \ldots, N\} \text{ and } j \neq i\}$.

These properties allow us to express y_i as a function of x_i and a permutation-invariant function ζ acting on the vector $\boldsymbol{x}_{-i}$, where $\boldsymbol{x}_{-i}$ excludes the ith element of $\boldsymbol{x}$. The function is given by $y_i = \phi(x_i, \zeta(\boldsymbol{x}_{-i}))$. Furthermore, if $x_i > x_j$, then $\phi(x_i, \zeta(\boldsymbol{x}_{-i})) > \phi(x_j, \zeta(\boldsymbol{x}_{-j}))$ is also upheld, owing to the **order preservation** property. A potential representation of the function $\phi(x_i, \zeta(\boldsymbol{x}_{-i}))$, that satisfies the aforementioned properties, is

$$\phi(x_i, \zeta(\boldsymbol{x}_{-i})) = \frac{g(x_i)}{\sum_{j=1}^{N} g(x_j)}, \tag{11}$$

where $\zeta(\boldsymbol{x}_{-i}) = \sum_{j=1, j\neq i}^{N} g(x_j)$, and g is a monotonically increasing function, defined on the domain of non-negative real numbers and exhibits well-defined derivatives across its domain. The function g is characterized primarily by its non-zero Taylor series coefficients $g(x) = \sum_{k=1}^{M} a_k x^{C_k}$. Given this characterization, each y_i can be formulated as

$$y_i = \frac{\sum_{k=1}^{M} a_k x_i^{C_k}}{\sum_{j=1}^{N} \sum_{k=1}^{M} a_k x_j^{C_k}}. \tag{12}$$

An essential aspect of G is its reactivity to the elements in $\boldsymbol{x}$. We term G "unbiased" by the values of $\boldsymbol{x}$" when $y_i = y_j$, for all indices i and j. This unbiased behavior emerges when all exponents C_k are zero. In stark contrast, G is dubbed "fully biased" if $y_m = 1$, whenever x_m exceeds all other components of $\boldsymbol{x}$, i.e., $x_m > x_i \ \forall i \neq m$. This is satisfied when $\max_k\{C_k\}$ tends to infinity and all other C_k tend to zero. Consequently, we represent $g(x)$ as $g(x) = (L(x))^C$, designating it as our bias function where L is monotonic chosen as $L(x) = x$.

B. Proof of Proposition 2

Let $\boldsymbol{\rho}^{[k]}$ be the probability vector for the kth client's sampling procedure at round t, where $\boldsymbol{\rho}^{[k]} \in \Delta^{N-1}$. Without loss of generality, let $\boldsymbol{\rho}^{[1]} = [\rho_m^{[1]}, \rho_{m-1}^{[1]}, \ldots, \rho_1^{[1]}]$ be sorted in descending order with respect to the values of the effective channels, i.e., $h_1 < \cdots < h_{m-1} < h_m$. For the particular case of $\rho_i^{[1]}$, we have

$$\lim_{C\to\infty} \rho_i^{[1]} = \lim_{C\to\infty} \frac{\lambda_i |h_i|^C}{\sum_{j=1}^{N} \lambda_j |h_j|^C}. \tag{13}$$

Furthermore, we can deduce

$$\lim_{C\to\infty} \frac{\rho_i^{[1]}}{\rho_j^{[1]}} = \lim_{C\to\infty} \frac{\lambda_i}{\lambda_j} \left(\frac{|h_i|}{|h_j|}\right)^C. \tag{14}$$

From these results, we derive

$$\lim_{C\to\infty} \frac{\rho_i^{[1]}}{\rho_j^{[1]}} = \begin{cases} 0, & \text{if } i < j, \\ 1, & \text{if } i = j. \end{cases} \tag{15}$$

Using (13), we further express

$$\lim_{C\to\infty} \rho_i^{[1]} = \lim_{C\to\infty} \frac{\rho_i^{[1]}/\rho_m^{[1]}}{\sum_{j=1}^{N} \rho_j^{[1]}/\rho_m^{[1]}} = \begin{cases} 0, & \text{if } i \neq m, \\ 1, & \text{if } i = m. \end{cases} \tag{16}$$

This suggests that as C increases, our algorithm tends to select the client that requires the lowest energy for model upload. Further exploration reveals

$$\rho_i^{[2]} = \sum_{j=1,\, j\neq i}^{N} \frac{\rho_j^{[1]}}{1-\rho_j^{[1]}} \rho_i^{[1]}. \tag{17}$$

Consequently, we have

$$\lim_{C\to\infty} \rho_i^{[2]} = \begin{cases} \lim\limits_{C\to\infty} \frac{\rho_m^{[1]}}{1-\rho_m^{[1]}} \rho_i^{[1]}, & \text{if } i \neq m, \\ 0, & \text{if } i = m. \end{cases} \tag{18}$$

These insights indicate a dimensionality reduction within the simplex space. This can be articulated by the following relationships for $i \neq m$

$$\lim_{C\to\infty} \rho_i^{[2]} = \lim_{C\to\infty} \frac{\rho_i^{[1]}}{\rho_{m-1}^{[1]} + \cdots + \rho_1^{[1]}}. \tag{19}$$

Furthermore, if $i \neq m$ and $j \neq m$, then we can write

$$\lim_{C\to\infty} \rho_i^{[2]} = \lim_{C\to\infty} \frac{\rho_i^{[1]}/\rho_{m-1}^{[1]}}{\sum_{j=1}^{N} \rho_j^{[1]}/\rho_{m-1}^{[1]}}. \tag{20}$$

Using (15) in (20), we deduce

$$\lim_{C\to\infty} \frac{\rho_i^{[2]}}{\rho_j^{[2]}} = \begin{cases} 0, & \text{if } i < j, i \neq m, j \neq m, \\ 1, & \text{if } i = j, i \neq m. \end{cases} \tag{21}$$

By upholding the prior conditions using mathematical induction, we establish that as $C \to \infty$, our algorithm simplifies to a greedy strategy, prioritizing the K clients who require the least energy to transmit their models to the PS at round t.