Learning while Respecting Privacy and Robustness to Adversarial Distributed Datasets

Alireza Sadeghi and Georgios B. Giannakis
Dept. of Electrical and Computer Engineering, University of Minnesota, USA

***Abstract*—Massive datasets are typically distributed geographically across multiple sites, where scalability, data privacy and integrity, as well as bandwidth scarcity typically discourage uploading these data to a central server. This has propelled the so-called federated learning framework where multiple workers exchange information with a server to learn a "centralized" model using data locally generated and/or stored across workers. This learning framework necessitates workers to communicate iteratively with the server. Although appealing for its scalability, one needs to carefully address the various data distribution shifts across workers, which degrades the performance of the learnt model. In this context, the distributionally robust optimization framework is considered here. The objective is to endow the trained model with robustness against adversarially manipulated input data, or, distributional uncertainties, such as mismatches between training and testing data distributions, or among datasets stored at different workers. To this aim, the data distribution is assumed unknown, and to land within a Wasserstein ball centered around the empirical data distribution. This robust learning task entails an infinite-dimensional optimization problem, which is challenging. Leveraging a strong duality result, a surrogate is obtained, for which a primal-dual algorithm is developed. Compared to classical methods, the proposed algorithm offers robustness with little computational overhead. Numerical tests using image datasets showcase the merits of the proposed algorithm under several existing adversarial attacks and distributional uncertainties.**

I. Introduction

Machine learning models and tasks hinge on the premise that the training data are trustworthy, reliable, and representative of the testing data [8], [26], [28]. In practice however, data are usually generated and stored at geographically distributed devices (a.k.a., workers) each equipped with limited computing capability, and adhering to privacy, confidentiality, and possibly cost constraints [14]. Furthermore, the data quality is not guaranteed due to adversarially generated examples and distribution drifts across workers or from the training to testing phases [16]. Visually imperceptible perturbations to a dermatoscopic image of a benign mole can render the first-ever artificial intelligence (AI) diagnostic system approved by the U.S. Food and Drug Administration in 2018, to classify it as cancerous with 100% confidence [7]. A stranger wearing pixelated sunglasses can fool even the most advanced facial recognition software in a home security system to mistake it for the homeowner [4], [23]. Hackers indeed manipulated readings of field devices and control centers of the Ukrainian supervisory control and data acquisition system to cause the first ever cyberattack-caused power outage in 2015 [2], [18], [31]. Examples of such failures in widely used AI-enabled safety- and security-critical systems today could put national infrastructure and even lives at risk.

Recent research efforts have focused on devising defense strategies against adversarial attacks. These strategies fall under two groups: attack detection, and attack recovery. The former identifies whether a given input is adversarially perturbed [9], while the latter trains a model to gain robustness against such adversarial inputs [10], which is also the theme of the present contribution. To robustify learning models against adversarial data, a multitude of data pre-processing schemes have been devised [24], to identify anomalies not adhering to postulated or nominal data. Adversarial training on the other hand, adds imperceptible well-crafted noise to clean input data to gain robustness [8]; see also e.g., [19], [21], [22], and [3] for a recent survey. In these contributions, optimization tasks are formulated to craft adversarial perturbations. Despite their empirical success, solving the resultant optimization problem is challenging. Furthermore, analytical properties of these approaches have not been well understood, which hinders explainability of the obtained models. In addition, one needs to judiciously tune hyper parameters of the attack model, which tends to be cumbersome in practice.

On the other hand, data are typically generated and/or stored at geographically distributed sites, each having subsets of data with different distributions. While keeping data localized to e.g., respect privacy, as well as reduce communication- and computation-overhead, the federated learning (FL) paradigm targets a global model, whereby multiple devices are coordinated by a central parameter server [14]. Existing FL approaches have mainly focused on the communicating versus computing tradeoff by aggregating model updates from the learners; see e.g., [15], [20], [25], [29] and references therein. From the few works dealing with robust FL, [17] learns from dependent data through e.g., sparsification, while [12] entails an ensemble of untrusted sources. These methods are rather heuristic, and rely on aggregation to gain robustness. This context, motivates well a principled approach that accounts for the uncertainties associated with the underlying data distributions.

Tapping on a distributionally robust optimization perspective, this paper develops robust learning procedures that ensure robustness to distributional uncertainties and adversarial

This work was supported by University of Minnesota Doctoral Dissertation Fellowship (DDF) and NSF grants 1901134, 2126052, and ARO STIR W911NF2110297. The authors are with the Dept. of ECE, Univ. of Minnesota, Minneapolis, MN 55455, USA (e-mails: sadeghi@umn.edu; georgios@umn.edu).

attacks. Building on [26], the adversarial input perturbations are constrained to lie in a Wasserstein ball, and the sought robust model minimizes the worst-case expected loss over this ball. As the resulting formulation leads to a challenging infinite-dimensional optimization problem, we leverage strong duality to arrive at a tractable and equivalent unconstrained minimization problem, requiring solely the empirical data distribution. To accommodate communication constraints or possibly untrusted datasets distributed across multiple workers, we propose a distributionally robust federated learning (DRFL) algorithm. In a nutshell, the main contribution of this paper is to develop a distributionally robust distributed learning framework to account for untrusted and possibly anonymized data from distributed datasets.

Notation. Bold lowercase letters denote column vectors, while calligraphic uppercase fonts are reserved for sets; $\mathbb{E}[\cdot]$ represents expectation; ∇ denotes the gradient operator; $(\cdot)^\top$ denotes transposition, and $\|\boldsymbol{x}\|$ is the 2-norm of the vector $\boldsymbol{x}$.

II. Problem Formulation

Consider the standard regularized statistical learning task

$$\min_{\boldsymbol{\theta}\in\Theta} \ \mathbb{E}_{\boldsymbol{z}\sim P_0}\big[\ell(\boldsymbol{\theta};\boldsymbol{z})\big] + r(\boldsymbol{\theta}) \tag{1}$$

where $\ell(\boldsymbol{\theta};\boldsymbol{z})$ denotes the loss of a model parameterized by the unknown parameter vector $\boldsymbol{\theta}$ on a datum $\boldsymbol{z}=(\boldsymbol{x},y)\sim P_0$, with feature $\boldsymbol{x}$ and label y, drawn from some nominal distribution P_0. Here, Θ denotes the feasible set for model parameters. To prevent over fitting or incorporate prior information, regularization term $r(\boldsymbol{\theta})$ is oftentimes added to the expected loss. Popular regularizers include $r(\boldsymbol{\theta}):=\beta\|\boldsymbol{\theta}\|_1^2$ or $\beta\|\boldsymbol{\theta}\|_2^2$, where $\beta\geq 0$ is a hyper-parameter controlling the importance of the regularization term relative to the expected loss.

The nominal distribution P_0 is typically unknown. Instead some data samples $\{\boldsymbol{z}_n\}_{n=1}^N\sim\widehat{P}_0^{(N)}$ drawn i.i.d. from P_0 are given. Upon replacing P_0 with the empirical distribution $\widehat{P}_0^{(N)}$ in (1), we arrive at the empirical loss minimization

$$\min_{\boldsymbol{\theta}\in\Theta} \ \bar{\mathbb{E}}_{\boldsymbol{z}\sim\widehat{P}_0^{(N)}}\big[\ell(\boldsymbol{\theta};\boldsymbol{z})\big] + r(\boldsymbol{\theta}) \tag{2}$$

where $\bar{\mathbb{E}}_{\boldsymbol{z}\sim\widehat{P}_0^{(N)}}[\ell(\boldsymbol{\theta};\boldsymbol{z})]=N^{-1}\sum_{n=1}^N\ell(\boldsymbol{\theta};\boldsymbol{z}_n)$. Indeed, a variety of machine learning tasks can be cast as (2), including e.g., ridge and Lasso regression, logistic regression, and reinforcement learning. The resultant models obtained by solving (2) however, have been shown vulnerable to abnormally corrupted data in $\widehat{P}_0^{(N)}$. Furthermore, the testing data distribution often deviates from the available $\widehat{P}_0^{(N)}$. For this reason, targeting a robust model against a set of distributions corresponding to perturbations of the underlying data distribution, leads to [26]

$$\min_{\boldsymbol{\theta}\in\Theta} \ \sup_{P\in\mathcal{P}} \ \mathbb{E}_{\boldsymbol{z}\sim P}[\ell(\boldsymbol{\theta};\boldsymbol{z})] + r(\boldsymbol{\theta}) \tag{3}$$

where $\mathcal{P}$ is a set of distributions centered around the data generating distribution $\widehat{P}_0^{(N)}$. Compared with (1), the worst-case formulation (3), yields models with reasonable performance across a continuum of distributions characterized by $\mathcal{P}$. In practice, datasets are typically distributed across multiple sites, where scalability, data privacy and integrity, as well as bandwidth scarcity discourage uploading them to a central server. This has propelled the so-called federated learning framework, where multiple workers exchange information with a server to learn a centralized model using data locally generated and/or stored across workers [5], [14], [16], [20]. Workers in this learning framework communicate *iteratively* with the server. Albeit appealing for its scalability, one needs to carefully address the bandwidth bottleneck associated with server-worker links. Furthermore, the workers' data may have (slightly) different underlying distributions, which further challenges the learning task. To seek a model robust to distribution drifts across workers, we will tune robust learning in (3) to design a robust federated algorithm algorithm.

Ensuing section targets learning a robust global model using data that is distributed across multiple locations.

III. Robust learning from distributed datasets

Consider K workers with each worker $k\in\mathcal{K}$ collecting samples $\{\boldsymbol{z}_n(k)\}_{n=1}^N$, and a globally shared model parameterized by $\boldsymbol{\theta}$. Parameters are to be updated at the server by aggregating gradients computed locally per worker. For simplicity, we consider workers having the same number of samples N. The goal is to learn a single global model from stored data at all workers by minimizing the following objective function

$$\min_{\boldsymbol{\theta}\in\Theta} \ \bar{\mathbb{E}}_{\boldsymbol{z}\sim\widehat{P}}[\ell(\boldsymbol{\theta};\boldsymbol{z})] + r(\boldsymbol{\theta}) \tag{4}$$

where $\bar{\mathbb{E}}_{\boldsymbol{z}\sim\widehat{P}}[\ell(\boldsymbol{\theta};\boldsymbol{z})] := \frac{1}{NK}\sum_{n=1}^N\sum_{k=1}^K\ell(\boldsymbol{\theta},\boldsymbol{z}_n(k))$. To endow the learned model with robustness against distributional uncertainties, our novel formulation will solve the following problem in a distributed fashion

$$\min_{\boldsymbol{\theta}\in\Theta} \ \sup_{P\in\mathcal{P}} \ \mathbb{E}_{\boldsymbol{z}\sim P}[\ell(\boldsymbol{\theta};\boldsymbol{z})] + r(\boldsymbol{\theta}) \quad \text{s. to. } P\in\mathcal{P}. \tag{5}$$

There are different approaches to define ambiguity sets $\mathcal{P}$, including momentum [6], [30], KL divergence [11], statistical test [1], and Wasserstein distance-based ambiguity sets [1], [26]. It has been shown that the Wasserstein ambiguity set $\mathcal{P}$ results in a tractable solution, thanks to its strong duality results [1] [26].

To formalize this, let us first defined Wasserstein distance between two probability distribution functions (pdfs). To this end, consider two probability measures P and Q supported on some set $\mathcal{Z}$, and let $\Pi(P,Q)$ be the set of all joint measures supported on $\mathcal{Z}^2$, with marginals P and Q. Let $c:\mathcal{Z}\times\mathcal{Z}\to[0,\infty)$ measure the cost of transporting a unit of mass from $\boldsymbol{z}$ in P to another element $\boldsymbol{z}'$ in Q. The celebrated optimal transport problem is given by [27, page 111]

$$W_c(P,Q) := \inf_{\pi\in\Pi}\mathbb{E}_\pi\big[c(\boldsymbol{z},\boldsymbol{z}')\big]. \tag{6}$$

Remark 1. If $c(\cdot,\cdot)$ satisfies the axioms of distance, then W_c defines a distance on the space of probability measures. For instance, if P and Q are defined over a Polish space equipped

with metric d, then choosing $c(\boldsymbol{z},\boldsymbol{z}')=d^p(\boldsymbol{z},\boldsymbol{z}')$ for some $p\in[1,\infty)$ asserts that $W_c^{1/p}(P,Q)$ is the well-known Wasserstein distance of order p between probability measures P and Q [27, Definition 6.1].

For a given empirical distribution $\widehat{P}_0^{(N)}(k)$ per worker k, we resort to Wasserstein distance (6) to define the uncertainty set as $\mathcal{P}:=\Big\{P\Big|\sum_{k=1}^K W_c(P,\widehat{P}^{(N)}(k))\leq\rho\Big\}$, where $W_c(P,\widehat{P}^{(N)}(k))$ denotes the Wasserstein distance between distribution P and the local empirical data distribution $\widehat{P}^{(N)}(k)$, per worker k. Clearly, the constraint $P\in\mathcal{P}$, couples the optimization in (5) across all workers. Using this uncertainty set, we arrive at the distributionally robust federated learning formulation as

$$\min_{\boldsymbol{\theta}\in\Theta}\ \sup_{P\in\mathcal{P}}\ \mathbb{E}_{\boldsymbol{z}\sim P}[\ell(\boldsymbol{\theta};\boldsymbol{z})]+r(\boldsymbol{\theta})$$
$$\text{s. to. }\mathcal{P}:=\Big\{P\Big|\sum_{k=1}^K W_c(P,\widehat{P}^{(N)}(k))\leq\rho\Big\}\tag{7}$$

To offer distributed solvers for this learning problem, we resort to dual reformulation of the inner maximization in (7) (see [26] for strong duality details), to arrive at the following robust surrogate learning formulation

$$\min_{\boldsymbol{\theta}\in\Theta}\ \inf_{\gamma\in\Gamma}\ \sum_{k=1}^K\bar{\mathbb{E}}_{\boldsymbol{z}(k)\sim\widehat{P}^{(N)}(k)}\big[\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}\{\ell(\boldsymbol{\theta};\boldsymbol{\zeta})+\gamma(\rho-c(\boldsymbol{z}(k),\boldsymbol{\zeta}))\}\big]+r(\boldsymbol{\theta}).\tag{8}$$

Here γ denotes the dual variable associated with the uncertainty set constraint. To maximization over $\boldsymbol{\zeta}$ we need to rely on a strongly concave objective, thus we let the transportation cost $c(\cdot)$ to be strongly convex, and let $\gamma\in\Gamma:=\{\gamma|\gamma>\gamma_0\}$, where γ_0 is large enough. Since γ is the dual variable corresponding to the constraint in (7), having $\gamma\in\Gamma$ is tantamount to tuning ρ, which in turn *controls* the level of *robustness*. To this aim we assume that $\gamma\in\Gamma$, and our *robust learning model* is thus obtained as the solution of

$$\min_{\boldsymbol{\theta}\in\Theta}\ \inf_{\gamma\in\Gamma}\ \bar{\mathbb{E}}_{\boldsymbol{z}\sim\widehat{P}_0^{(T)}}\big[\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}\psi(\bar{\boldsymbol{\theta}},\boldsymbol{\zeta};\boldsymbol{z})\big]+r(\bar{\boldsymbol{\theta}})\tag{9}$$

where $\psi(\bar{\boldsymbol{\theta}},\boldsymbol{\zeta};\boldsymbol{z}):=\ell(\boldsymbol{\theta};\boldsymbol{\zeta})+\gamma(\rho-c(\boldsymbol{z},\boldsymbol{\zeta}))$. Intuitively, input $\boldsymbol{z}$ in (9) is pre-processed by maximizing ψ accounting for some perturbation. To iteratively solve our objective in (9), the ensuing section provides efficient solver under some mild conditions.

Specifically, our distributionally robust federated learning (DRFL) hinges on the fact that with fixed server parameters $\bar{\boldsymbol{\theta}}^t:=[{\boldsymbol{\theta}^t}^\top,\gamma^t]^\top$ per iteration t, the optimization problem becomes *separable* across all workers. Hence, upon receiving $\bar{\boldsymbol{\theta}}^t$ from the server, each worker $k\in\mathcal{K}$: i) samples a minibatch $\mathcal{B}^t(k)$ of data from $\widehat{P}^{(N)}(k)$; ii) forms the *perturbed* loss $\psi_k(\bar{\boldsymbol{\theta}}^t,\boldsymbol{\zeta};\boldsymbol{z}):=\ell(\boldsymbol{\theta}^t;\boldsymbol{\zeta})+\gamma^t(\rho-c(\boldsymbol{z},\boldsymbol{\zeta}))$ for each $\boldsymbol{z}\in\mathcal{B}^t(k)$; iii) lazily maximizes $\psi_k(\bar{\boldsymbol{\theta}}^t,\boldsymbol{\zeta};\boldsymbol{z})$ over $\boldsymbol{\zeta}$ using a single gradient ascent step to yield $\boldsymbol{\zeta}(\bar{\boldsymbol{\theta}}^t;\boldsymbol{z})=\boldsymbol{z}+\eta_t\nabla_{\boldsymbol{\zeta}}\psi_k(\bar{\boldsymbol{\theta}}^t,\boldsymbol{\zeta};\boldsymbol{z})|_{\boldsymbol{\zeta}=\boldsymbol{z}}$; and, iv) sends the stochastic gradient $|\mathcal{B}^t(k)|^{-1}\sum_{\boldsymbol{z}\in\mathcal{B}^t(k)}\nabla_{\bar{\boldsymbol{\theta}}}\psi_k(\bar{\boldsymbol{\theta}}^t,\boldsymbol{\zeta}(\bar{\boldsymbol{\theta}}^t;\boldsymbol{z});\boldsymbol{z})\big|_{\bar{\boldsymbol{\theta}}=\bar{\boldsymbol{\theta}}^t}$ back to the server. Upon receiving all local gradients, the server updates $\bar{\boldsymbol{\theta}}^t$ using a proximal gradient descent step to find $\bar{\boldsymbol{\theta}}^{t+1}$, that is

$$\bar{\boldsymbol{\theta}}^{t+1}=\text{prox}_{\alpha_t r}\Bigg[\bar{\boldsymbol{\theta}}^t-\frac{\alpha_t}{K}\sum_{k=1}^K\frac{1}{|\mathcal{B}^t(k)|}\times\sum_{\boldsymbol{z}\in\mathcal{B}^t(k)}\nabla_{\bar{\boldsymbol{\theta}}}\psi_k(\bar{\boldsymbol{\theta}}^t,\boldsymbol{\zeta}(\bar{\boldsymbol{\theta}}^t;\boldsymbol{z});\boldsymbol{z})\big|_{\bar{\boldsymbol{\theta}}=\bar{\boldsymbol{\theta}}^t}\Bigg]\tag{10}$$

which is then broadcasted to all workers to begin a new round of local updates. Our DRFL approach is tabulated in Alg. 1.

Algorithm 1: DRFL

Input : Initial guess $\bar{\boldsymbol{\theta}}^1$, a set of workers $\mathcal{K}$ with data samples $\{\boldsymbol{z}_n(k)\}_{n=1}^N$ per worker $k\in\mathcal{K}$, step size sequence $\{\alpha_t,\eta_t>0\}_{t=1}^T$

Output: $\bar{\boldsymbol{\theta}}^{T+1}$

for $t=1,\ldots,T$ **do**

  Each worker:

  Samples a minibatch $\mathcal{B}^t(k)$ of samples

  Given $\bar{\boldsymbol{\theta}}^t$ and $\boldsymbol{z}\in\mathcal{B}^t(k)$, forms local perturbed loss

$$\psi_k(\bar{\boldsymbol{\theta}}^t,\boldsymbol{\zeta};\boldsymbol{z}):=\ell(\bar{\boldsymbol{\theta}}^t;\boldsymbol{\zeta})+\gamma^t(\rho-c(\boldsymbol{z},\boldsymbol{\zeta}))$$

  Lazily maximizes $\psi_k(\bar{\boldsymbol{\theta}}^t,\boldsymbol{\zeta};\boldsymbol{z})$ over $\boldsymbol{\zeta}$ to find

$$\boldsymbol{\zeta}(\bar{\boldsymbol{\theta}}^t;\boldsymbol{z})=\boldsymbol{z}+\eta_t\nabla_{\boldsymbol{\zeta}}\psi_k(\bar{\boldsymbol{\theta}}^t,\boldsymbol{\zeta};\boldsymbol{z})|_{\boldsymbol{\zeta}=\boldsymbol{z}}$$

  Computes stochastic gradient

$$\frac{1}{|\mathcal{B}^t(k)|}\sum_{\boldsymbol{z}\in\mathcal{B}^t(k)}\nabla_{\bar{\boldsymbol{\theta}}}\psi_k(\bar{\boldsymbol{\theta}}^t,\boldsymbol{\zeta}(\bar{\boldsymbol{\theta}}^t;\boldsymbol{z});\boldsymbol{z})\big|_{\bar{\boldsymbol{\theta}}=\bar{\boldsymbol{\theta}}^t}$$

  and uploads to server

  Server:

  Updates $\bar{\boldsymbol{\theta}}^t$ according to (10)

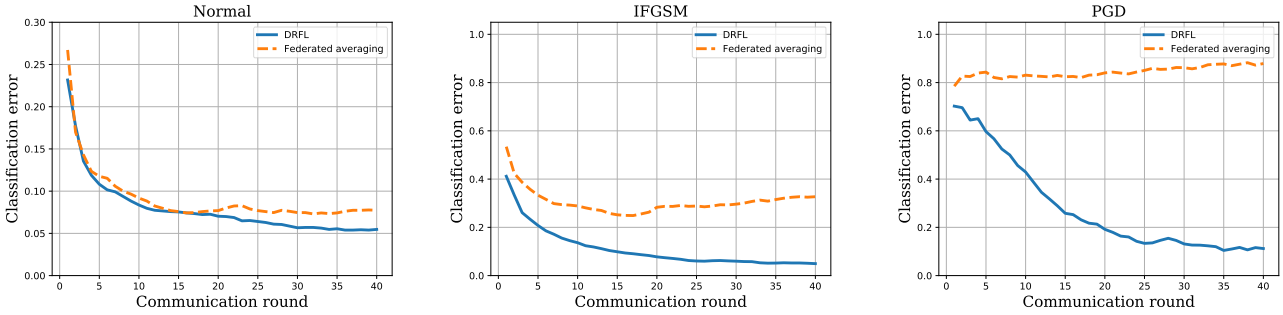
  Broadcasts $\bar{\boldsymbol{\theta}}^{t+1}$ to workers

end

IV. Numerical Tests

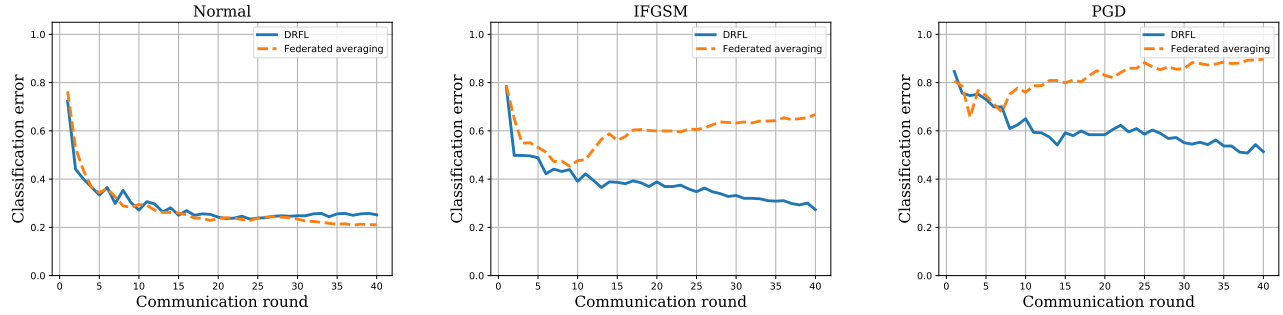
To assess the performance in the presence of distribution drifts and data perturbations, we will rely on empirical classification of standard MNIST and Fashion- (F-)MNIST datasets.

Specifically, we considered an FL environment consisting of a server and 10 workers, with local batch size 64, and assigned to every worker an equal-sized subset of training data containing i.i.d. samples from 10 different classes. All workers participated in each communication round. To benchmark the DRFL, we simulated the federated averaging method [20]. The testing accuracy on the MNIST dataset per communication round using clean (normal) images is depicted in Fig. 1a. Clearly, both DRFL and federated averaging algorithms exhibit reasonable performance when the data is not corrupted. The performance is further tested against adversarial samples generated according to th so-called iterative fast-



(a) No attack (b) IFGSM attack (c) PGD attack

Fig. 1: Federated learning for image classification using the MNIST dataset.



(a) No attack (b) IFGSM attack (c) PGD attack

Fig. 2: Distributionally robust federated learning for image classification using F-MNIST dataset.

gradient method (IFGSM) [13], and projected gradient descent (PGD) attack [19].

For IFGSM and PGD attacks we used a fixed adversarial budget $\epsilon_{\text{adv}} = 0.1$ during each communication round, and the corresponding misclassification error rates are shown in Figs. 1b and 1c, respectively. The classification performance using federated averaging does not improve in Fig. 1b, whereas the DRFL performance keeps improving across communication rounds. This is a direct consequence of accounting for the data uncertainties during the learning process. Moreover, Fig. 1c showcases that the federated averaging becomes even worse as the model gets progressively trained under the PGD attack. This indeed motivates our DRFL approach when data are from untrusted entities with possibly adversarial input perturbations. Similarly, Fig. 2 depicts the misclassification rate of the proposed DRFL method compared with federated averaging, when using the F-MNIST dataset.

V. Conclusions

A robust learning framework was put forth here, where the objective was to endow parametric models robust against distributional uncertainties and possibly adversarial data. Specifically, we focused on federated learning setting to learn from unreliable datasets across geographically distributed workers. To this end, this paper proposed a distributionally robust federated learning (DRFL) algorithm, which ensures data privacy and integrity, while offering robustness with minimal computational and communication overhead. Numerical tests for classifying standard real images showcased the merits of the proposed algorithm against distributional uncertainties and adversaries. This work also opens up several interesting directions for future research, including distributionally robust deep reinforcement learning.

References

[1] C. Bandi and D. Bertsimas, "Robust option pricing," *Eur. J. Oper. Res.*, vol. 239, no. 3, pp. 842–853, 2014.

[2] D. U. Case, "Analysis of the cyber attack on the Ukrainian power grid," 2016.

[3] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *arXiv:1810.00069*, 2018.

[4] J. Chen, J. Sun, and G. Wang, "From unmanned systems to autonomous intelligent systems," *Eng.*, vol. 8, pp. 1–5, 2022.

[5] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *arXiv:1909.07972*, 2019.

[6] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Oper. Res.*, vol. 58, no. 3, pp. 595–612, 2010.

[7] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," Dec. 2015.

[9] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *Intl. Conf. Learn. Rep. Wrks*, Dec. 2015.

[10] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, "Countering adversarial images using input transformations," *Intl. Conf. Learn. Rep.*, Apr. 2018.

[11] Z. Hu and L. J. Hong, "Kullback-Leibler divergence constrained distributionally robust optimization," *Available at Optimization Online*, 2013.
[12] N. Konstantinov and C. Lampert, "Robust learning from untrusted sources," *Intl. Conf. Mach. Learn.*, June 2019.
[13] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.
[14] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.
[15] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *arXiv:1812.06127*, 2018.
[16] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surv. Tut.*, Apr. 8 2020.
[17] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *arXiv:2006.07242*, 2020.
[18] W. Liu, J. Sun, G. Wang, F. Bullo, and J. Chen, "Data-driven resilient predictive control under Denial-of-Service," *arXiv:2110.12766*, 2021.
[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *Intl. Conf. Learn. Rep.*, Apr. 2018.
[20] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficientlearning of deep networks from decentralized data," in *Proc. Intl. Conf. Artif. Intell. Stat.*, vol. 54, Fort Lauderdale, FL, USA, 20–22 Apr. 2017, pp. 1273–1282.
[21] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1765–1773.
[22] A. Sadeghi, M. Ma, B. Li, and G. B. Giannakis, "Distributionally robust semi-supervised learning over graphs," *Intl. Conf. on Learning Representations, Workshop on Responsible AI*, May 8, 2021.
[23] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *ACM SIGSAC Conf. on Comput. Commun. Security*, 2016, pp. 1528–1540.
[24] F. Sheikholeslami, S. Jain, and G. B. Giannakis, "Minimum uncertainty based detection of adversaries in deep neural networks," *arXiv:1904.02841*, 2019.
[25] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UVeQFed: Universal vector quantization for federated learning," *arXiv:2006.03262*, 2020.
[26] A. Sinha, H. Namkoong, and J. Duchi, "Certifying some distributional robustness with principled adversarial training," in *Intl. Conf. Learn. Rep.*, 2018.
[27] C. Villani, *Optimal Transport: Old and New*. Springer Science & Business Media, 2008, vol. 338.
[28] G. Wang, G. B. Giannakis, and J. Chen, "Learning ReLU networks on linearly separable data: Algorithm, optimality, and generalization," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2357–2370, 2019.
[29] K. Wei, J. Li, M. Ding, C. Ma, H. Su, B. Zhang, and H. V. Poor, "Performance analysis and optimization in privacy-preserving federated learning," *arXiv:2003.00229*, 2020.
[30] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Oper. Res.*, vol. 62, no. 6, pp. 1358–1376, 2014.
[31] G. Wu, G. Wang, J. Sun, and J. Chen, "Optimal partial feedback attacks in cyber-physical power systems," *IEEE Trans. Autom. Control*, vol. 65, no. 9, pp. 3919–3926, 2020.