Distributionally Robust Clustered Federated Learning: A Case Study in Healthcare

Xenia Konti[1] Hans Riess[2] Manos Giannopoulos[1] Yi Shen[3]
Michael J. Pencina[4] Nicoleta J Economou-Zavlanos[4] Michael M. Zavlanos[1,2,3]

***Abstract*— In this paper, we address the challenge of heterogeneous data distributions in cross-silo federated learning by introducing a novel algorithm, which we term Cross-silo Robust Clustered Federated Learning (CS-RCFL). Our approach leverages the Wasserstein distance to construct ambiguity sets around each client's empirical distribution that capture possible distribution shifts in the local data, enabling evaluation of worst-case model performance. We then propose a model-agnostic integer fractional program to determine the optimal distributionally robust clustering of clients into coalitions so that possible biases in the local models caused by statistically heterogeneous client datasets are avoided, and analyze our method for linear and logistic regression models. Finally, we discuss a federated learning protocol that ensures the privacy of client distributions, a critical consideration, for instance, when clients are healthcare institutions. We evaluate our algorithm on synthetic and real-world healthcare data.**

I. Introduction

The complex nature of healthcare systems and data presents significant regulatory [1] and ethical [2] challenges associated with the design and deployment of large-scale machine learning models that need to comply with, e.g., the Health Insurance Portability and Accountability Act (HIPAA) in the United States or similar laws in other nations, which generally prohibit the sharing of patient data across healthcare organizations. At the cutting-edge, are distributed and privacy-preserving machine learning systems, broadly known as *Federated Learning* (FL) systems [3], which tackle the regulatory problem of data-sharing by avoiding it altogether; they exploit accumulated statistical information on relevant patient populations or machine learning model parameters that carry no patient-specific information.

It is well known that the design of effective machine learning algorithms requires the availability of sufficient data. Depending on the characteristics of the model and the dimension of the feature space, a theoretical minimum number of samples, called the sample complexity, is required for a model to generalize to unseen data sampled from the same distribution on which it was trained [4]. In practice, however, such data are often unavailable. Federated learning algorithms overcome this challenge by instantiating a cross-silo collaboration across data centers, such as those housed within healthcare organizations, that allows sharing of data insights rather than raw data and uses this distributed knowledge to train better-performing models. These data centers and the organizations they are housed in are often referred to as *clients*; in healthcare systems they are simply the *hospitals*. Among the many possibilities, FL has promising applications in healthcare informatics [5] and smart healthcare technologies such as remote monitoring, disease detection, and medical imaging; see [6] for a survey.

An important challenge in the design of effective federated learning algorithms is that the participating organizations are often subject to statistically heterogeneous datasets. This is, e.g., the case in healthcare applications, where hospitals often attend to statistically different patient populations. In an effort to ensure model fairness, meaning that the learned local models are free of any possible biases that may arise by aggregating models trained on different data distributions, *personalized federated learning* algorithms have been developed [7], that take advantage of collaboration between hospitals while at the same time implementing ways to mitigate bias. In healthcare applications, bias can be caused by demographic, geographic, financial, educational, political, and environmental differences across different healthcare organizations and the patient populations they serve [8]. A common approach to personalized federated learning is, what is known as, *Clustered Federated Learning* (CFL) [9] that reduces bias by clustering together and restricting collaboration between hospitals that potentially have a common or similar data-generating distributions.

In this work, we propose a new clustered federated learning algorithm that accounts for both statistical heterogeneity in the data of different participating organizations and statistical sampling uncertainty (e.g., due to finite sampling) or other distributions shifts in the local data of each individual organization. For this, we employ ideas from Distributionally Robust Optimization (DRO) to construct an integer fractional program that we solve to determine the optimal distributionally robust clustering of clients into coalitions so that (i) the learned local models generalize to unseen data sampled from distributions that are different from, but close enough to, the training ones and (ii) possible biases in the local models caused by statistically heterogeneous hospital datasets are avoided. We term our proposed algorithm Cross-Silo Robust Clustered Federated Learning (CS-RCFL). We theoretically justify our algorithm for two different types of models: linear regression with absolute loss and logistic regression. We finally evaluate our algorithm on both synthetic and real-world healthcare

[1]Dept. of Computer Science, Duke University.
[2]Dept. of Electrical and Computer Engineering, Duke University.
[3]Dept. of Mechanical Engineering & Materials Science, Duke University.
[4]School of Medicine, Duke University.
Emails: {pek7, hmr14, eg261, ys267, mjp50, nje6, mz61}@duke.edu
This work is supported in part by The Duke Endowment (TDE) under grant #7262-SP and by the Onassis Foundation under award #ZT033-1/2023-2024.

data, and show that it outperforms related methods.

A. Related Work

1) Clustered Federated Learning (CFL): CFL methods often construct clusters in rounds, potentially training models for several epochs between consecutive rounds. This approach was employed in [10] that iteratively clusters users by bi-partitioning existing clusters at each round. To address the high communication overhead of typical CFL methods [11], more efficient algorithms have been recently developed. An example is IFCA [12], which randomly initializes cluster centers and assigns clients to the nearest one, though this method is sensitive to center initialization. Similarly, FeSEM [13] relies on an ℓ_2-distance -based expectation maximization (EM) algorithm. The convergence of these methods has been studied in [14]. There are also works that deviate from model similarity-based clustering, introducing alternative coalition formation objectives [15]–[17] or incorporating robustness notions to produce robust clustering schemes [18]–[20]. The literature discussed above focuses on cross-device federated learning, involving personal devices such as smartphones. Instead, in this paper, we focus on cross-silo federated learning that involves data centers affiliated with organizations [21]. In cross-silo FL, global models are typically trained iteratively with clients sharing locally fine-tuned versions. Clustered cross-silo federated learning methods construct coalitions during each round using model similarity measures [22], [23] or game-theoretic incentive compatibility [24]. Common in the methods discussed above is that they do not account for possible distribution shifts between the training and test data. In this paper, we address this challenge employing notions from DRO.

2) DRO in Federated Learning: Interest in robust federated learning from a Distributionally Robust Optimization (DRO) perspective has grown recently. Agnostic FL [25] was the first framework to optimize over the worst-case mixture of local models, instead of assigning uniform or size-proportional weights. An improved algorithm with less communication overhead was introduced in [26], which uses periodic averaging and samples a subset of clients at each round. By relaxing the worst-case mixture of local models, [27] imposes constraints on the weights determined by the superquantile. Similar to our approach, but for non-clustered FL, [28], [29] consider distribution shifts and rely on Wasserstein-DRO to account for them. Here we extend this framework to clustered FL by designing client coalitions that are robust to possible uncertainties in the training data distributions.

II. Background

A. Federated Learning

Federated learning (FL) is a privacy-centric distributed learning framework. Its appeal is mostly attributed to its scalable and decentralized nature, allowing for simultaneous training of local models on local datasets that will, then, be aggregated into a global model that combines the predictive capabilities of its individual parts. At the same time, federated learning guards individual users' data by providing privacy guarantees [30]. In the standard terminology, several *clients* (e.g., hospitals) and one or more *servers* can form a FL system. Communication takes place between the clients and the server (vertical FL) and possibly between the clients directly (horizontal FL). For instance, clients may upload parameters to the server or download parameters or aggregations of parameters that the server has received. As discussed before, FL frameworks are also classified by the types of clients, including mobile devices (cross-device FL) or data silos managed by organizations (cross-silo FL). In this paper, we develop a variation of the well-known FedAve [31] algorithm, which aggregates individual models into global models by taking weighted averages of model parameters; see Eq. (6).

B. Distributionally Robust Optimization (DRO)

We first recall some technical definitions from optimal transport theory [32] in order to define distances between distributions, essential to our problem. Suppose $\mathbb{Q}_1, \mathbb{Q}_2$ are distributions supported on a subset Ξ of a euclidean space with a chosen metric δ (e.g., the ℓ_2-norm), and suppose $\Gamma(\mathbb{Q}_1, \mathbb{Q}_2)$ is the set of joint distributions on $\Xi \times \Xi$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$. Then, the 1-*Wasserstein distance* [33] between $\mathbb{Q}_1$ and $\mathbb{Q}_2$ is defined as

$$W_1(\mathbb{Q}_1, \mathbb{Q}_2) = \inf_{\mathbb{P}\in\Gamma(\mathbb{Q}_1,\mathbb{Q}_2)} \left\{ \int_{\Xi\times\Xi} \delta(\xi_1, \xi_2) d\mathbb{P}(\xi_1, \xi_2) \right\}.$$

We consider the problem of minimizing an objective function under uncertainty. Suppose $f : \mathbb{R}^p \times \Xi \to \mathbb{R}$ is an objective function with deterministic decision variables and uncertain parameters $x \in \mathbb{R}^p$ and $\xi \in \Xi$, respectively, and suppose $\hat{\mathbb{Q}}$ is an estimated distribution of $\xi \in \Xi$. Where robustness to distributional shifts is desired, a common practice is to form *ambiguity sets*, that is sets $\mathcal{P}$ of probability distributions, that effectively convert the objective function f to an objective function of the form $\sup_{\mathbb{Q}\in\mathcal{P}} \mathbb{E}_{\xi\sim\mathbb{Q}}\big[f(x,\xi)\big]$. A typical choice for $\mathcal{P}$ is a ball of radius $\varepsilon > 0$ centered around an estimated distribution $\hat{\mathbb{Q}}$ in the metric space of probability measures with the W_1 distance metric, denoted by $\mathcal{B}_\varepsilon(\hat{\mathbb{Q}}) = \{\mathbb{Q} : W_1(\mathbb{Q}, \hat{\mathbb{Q}}) \leqslant \varepsilon\}$, giving rise to the robust optimization problem

$$\min_x \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\hat{\mathbb{Q}})} \mathbb{E}_{\xi\sim\mathbb{Q}}\big[f(x,\xi)\big]. \tag{1}$$

While solving (1) is generically computationally expensive, a dual formulation can lead to tractable solutions.

Lemma 1 (Strong Duality [34]): Suppose $\Xi \subseteq \mathbb{R}^d$ and $f : \mathbb{R}^m \times \Xi \to \mathbb{R}$ is proper, convex, and lower semi-continuous. Suppose $\hat{\mathbb{Q}} \in \mathcal{M}(\Xi)$ is an estimated distribution on Ξ. Consider the worst-case expectation problem

$$\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\hat{\mathbb{Q}})} \mathbb{E}_{\xi\sim\mathbb{Q}}\left[f(\xi)\right]. \tag{2}$$

Then, the dual formulation

$$\inf_{\lambda\geqslant 0} \Big\{ \lambda\varepsilon + \mathbb{E}_{\xi'\sim\hat{\mathbb{Q}}}\big[\sup_{\xi\in\Xi}(f(\xi) - \lambda\delta(\xi',\xi))\big] \Big\} \tag{3}$$

has a zero duality-gap.

Thus, we can solve (3) to get a solution for (1). Finally, we recall the dual formulation for both linear regression with absolute loss [35] and logistic regression [36].

Lemma 2 (Linear regression [35]): Suppose $l_\theta(x,y) = \left|y - x^\top\theta\right|$. Then, $\inf_\theta \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_N)} \mathbb{E}_{(x,y)\sim\mathbb{Q}}[l_\theta(x,y)]$ is equal the solution of the following optimization problem:

$$\begin{array}{ll} \min\limits_{\theta,\alpha,b} & \alpha\varepsilon + \frac{1}{N}\sum_{i=1}^N b_i \\ \text{s.t.} & \|\theta\|_2^2 + 1 \leqslant \alpha^2 \\ & y_i - x_i^\top\theta \leqslant b_i, \forall i \in N \\ & -(y_i - x_i^\top\theta) \leqslant b_i, \forall i \in N \\ & \alpha, b_i \geqslant 0, \forall i \in N \\ & \theta \in \mathcal{B}. \end{array} \tag{4}$$

Lemma 3 (Logistic regression [36]): Suppose $l_\theta(x,y) = \log(1+\exp(-y\cdot x^T\theta))$ and define a distance metric between two data points (x,y) and (x',y') as $d((x,y),(x',y')) = ||x-x'|| + \kappa\frac{|y-y'|}{2}$, where κ is a positive weight. Then, $\inf_\theta \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\hat{\mathbb{P}}_N)} \mathbb{E}_{(x,y)\sim\mathbb{Q}}[l_\theta(x,y)]$ is equal to the solution of the following optimization problem:

$$\begin{array}{ll} \min\limits_{\theta,\alpha,b} & \alpha\varepsilon + \frac{1}{N}\sum_{i=1}^N b_i \\ \text{s.t.} & l_\theta(x_i,y_i) \leqslant b_i, \forall i \in N \\ & l_\theta(x_i,-y_i) - \alpha\kappa \leqslant b_i, \forall i \in N \\ & ||\theta||_\star \leqslant \alpha, \forall i \in N \\ & \alpha, b_i \geqslant 0, \forall i \in N \\ & \theta \in \mathcal{B}. \end{array} \tag{5}$$

III. Problem Formulation

Consider N hospitals (the clients in FL) and assume that each hospital $i \in \{1,\dots,N\}$ has data collected in the set $\mathcal{D}_i = \left\{(x_i^{(p)}, y_i^{(p)})\right\}_{p=1}^{|\mathcal{D}_i|}$, where every patient $p \in \{1,\dots,|\mathcal{D}_i|\}$ in the sample population of hospital i has n features $x_i^{(p)} \in \mathbb{R}^n$, e.g., height, weight, gender, blood pressure, etc., and observations $y_i^{(p)} \in \mathbb{R}$, e.g., their A1C. For each hospital i, the patient population is described by a true underlying nominal distribution, denoted by $\mathbb{P}_i$, so that $(x_i^{(p)}, y_i^{(p)}) \sim \mathbb{P}_i$, and an empirical distribution, denoted by $\hat{\mathbb{P}}_i$, that is obtained from the data in $\mathcal{D}_i$. We assume that the features and observations are related via a common underlying parametric model f, so that $y_i^{(p)} = f(x_i^{(p)};\theta_i^*)$, where $\theta_i^* \in \mathbb{R}^n$ are the model parameters. Then, given a loss function $\mathcal{L}$, the goal of each hospital is to estimate the true parameters of the model θ_i^* that minimize the loss on the underlying true data distribution. To do so, we formulate the *empirical risk minimization (ERM)* problem

$$\inf_{\theta\in\mathbb{R}^n} \mathbb{E}_{(x_i,y_i)\sim\hat{\mathbb{P}}_i}\left[\mathcal{L}\left(f(x_i;\theta),y_i\right)\right].$$

We consider a federated setting, where hospitals voluntarily collaborate with each other in order to train more accurate local models. To mitigate possible biases in the local models that can be caused by differences in the patient populations served by the hospitals, the hospitals form coalitions so that knowledge is shared among hospitals with common or similar data-generating distributions. These coalitions are coordinated by a lead hospital. Specifically, the lead hospital aims to create a *coalition structure* $\pi : \{1,\dots,N\} \to \{S_1,\dots,S_K\}$, that is a map from hospitals to K coalitions $\{S_1,\dots,S_K\}$ that satisfy $S_1 \cup \dots \cup S_K = \{1,\dots,N\}$, $S_k \cap S_{k'} = \emptyset$ for all $k \neq k'$, and $S_k \neq \emptyset$ for all $k \in \{1\dots,K\}$. Then, $\pi(i) = S_k$ denotes the unique coalition S_k to which hospital i is assigned. The hospitals participating in this coalition structure first use their limited data to learn parameter estimates θ_i^{local}, which we collect in a tuple $\boldsymbol{\theta} = \left(\theta_1^{\text{local}},\dots,\theta_N^{\text{local}}\right)$. Then, they share these parameters with the lead hospital. The lead hospital aggregates the local parameters for each coalition and returns their mean to the member hospitals. Thus, every hospital i belonging to coalition $\pi(i) = S_k$ receives the common estimated parameters

$$\theta_{S_k} = \frac{1}{|S_k|}\sum_{i\in S_k} \theta_i^{\text{local}}. \tag{6}$$

Using the model parameters θ_{S_k}, the expected loss of hospital i in coalition $\pi(i) = S_k$ becomes

$$\ell_i(\boldsymbol{\theta},\pi) = \mathbb{E}_{(x_i,y_i)\sim\hat{\mathbb{P}}_i}\left[\mathcal{L}\big(f(x_i;\theta_{S_k}),y_i\big)\right]. \tag{7}$$

The difficulty in solving the ERM problem defined before lies in the fact that the lead hospital requires knowledge of the empirical distributions $\hat{\mathbb{P}}_i$ for all hospitals. Even more, due to finite-sample bias or other distributional shifts, whether the empirical distributions $\hat{\mathbb{P}}_i$ are a good representation of the underlying data-generating method also comes into question. To model this uncertainty, for each hospital i, we construct an ambiguity set that is a Wasserstein ball of radius $\varepsilon_i \geqslant 0$ around $\hat{\mathbb{P}}_i$, denoted by $\mathcal{B}_{\varepsilon_i}(\hat{\mathbb{P}}_i)$; see Section II. The radius of the ambiguity set ε_i is selected by each hospital individually, and represents how robust and conservative this hospital wants to be when computing its loss. We argue that the mechanism for dividing hospitals into coalitions implemented by the lead hospital should take into account these ambiguity sets. Recall that if a hospital i belongs to coalition $\pi(i) = S_k$, then it receives model parameters θ_{S_k}. Using these model parameters, hospital i can also compute an upper bound on its expected loss by calculating the worst-case loss over all the distributions in its ambiguity set, i.e.,

$$\ell_i^{\text{rob}}(\boldsymbol{\theta},\pi) = \sup_{\mathbb{Q}_i\in B_{\varepsilon_i}(\hat{\mathbb{P}}_i)} \mathbb{E}_{(x_i,y_i)\sim\mathbb{Q}_i}\left[\mathcal{L}(f(x_i;\theta_{S_k}),y_i)\right]. \tag{8}$$

Therefore, to address the aforementioned challenges, the lead hospital should form clusters that minimize the accumulated *worst possible* loss across all hospitals. We translate this objective to the following optimization problem.

Problem 1: Given N hospitals, K coalitions, a loss function $\mathcal{L}$, and an ambiguity set $B_{\varepsilon_i}(\hat{\mathbb{P}}_i)$ for each hospital i, compute a coalition structure $\pi : \{1,\dots,N\} \to \{S_1,\dots,S_K\}$ that minimizes the expected robust loss of all hospitals, i.e.,

$$\begin{array}{ll} \min\limits_{\pi} & \sum_{k=1}^K \sum_{i\in S_k} \ell_i^{\text{rob}}(\boldsymbol{\theta},\pi) \\ \text{s.t.} & S_1 \cup S_2 \cup \dots S_K = \{1,\dots,N\} \\ & S_k \cap S_{k'} = \emptyset, \quad \forall k \neq k' \\ & S_k \neq \emptyset, \quad \forall k \in \{1,\dots,K\}. \end{array}$$

IV. Coalition Formation

In order to solve Problem 1, we introduce a binary variable $a_{i,k} \in \{0,1\}$ such that $a_{i,k} = 1$ if hospital i is assigned to coalition S_k and $a_{i,k} = 0$ otherwise. Then, the coalition formation Problem 1 can be reformulated into an integer program (IP) with binary decision variables $\{a_{i,k}\}_{N\times K}$ as

$$\begin{aligned} \min_{a_{i,k}} \quad & \textstyle\sum_{k=1}^{K}\sum_{i=1}^{N} a_{i,k}\ell_i^{\mathrm{rob}}(\boldsymbol{\theta},\pi) \\ \text{s.t.} \quad & \textstyle\sum_{k=1}^{K} a_{i,k} = 1, \quad i = 1,\dots,N \\ & \textstyle\sum_{i=1}^{N} a_{i,k} \geqslant 1, \quad k = 1,\dots,K \\ & a_{i,k} \in \{0,1\}, \quad i = 1,\dots,N,\ k = 1,\dots,K. \end{aligned} \tag{9}$$

A. Linear Integer Relaxation

In order to solve the optimization problem (9), we need to compute the robust loss $\ell_i^{\mathrm{rob}}(\boldsymbol{\theta},\pi)$ of hospital i for all possible aggregated models of all possible coalition structures π. This is a combinatorial problem that is very hard to solve. To address this challenge, we assume that the loss function $\mathcal{L}$ is convex with respect to the model parameters $\boldsymbol{\theta}$. As shown in the following result, this assumption allows us to relax problem (9) into a linear integer program that can be efficiently solved.

Lemma 4: Suppose a parametric model $f(x;\theta)$ defined by model parameters θ and suppose that $\mathcal{L}(f(x;\theta),y)$ is a loss function that is convex with respect to θ. Suppose also binary variables $a_{i,k} \in \{0,1\}$, for all $i \in \{1,\dots,N\},\ k \in \{1,\dots,K\}$, with $\sum_{k=1}^{K} a_{i,k} = 1$ for all $i \in \{1,\dots,N\}$, that denote whether hospital i belongs to coalition S_k. Then,

$$\ell_i^{rob}(\boldsymbol{\theta},\pi) \leqslant \textstyle\sum_{j=1}^{N} \frac{a_{j,k}}{\sum_{j=1}^{N} a_{j,k}} \sup_{\mathbb{Q}_i \in B_{\varepsilon_i}(\hat{\mathbb{P}}_i)} \mathbb{E}_{(x,y)\sim\mathbb{Q}_i}\big[\mathcal{L}(f(x;\theta_j),y)\big].$$

Proof: See Appendix. ∎

Define the loss

$$L_{i,j}(\boldsymbol{\theta}) = \sup_{\mathbb{Q}_i \in B_{\varepsilon_i}(\hat{\mathbb{P}}_i)} \mathbb{E}_{(x,y)\sim\mathbb{Q}_i}\big[\mathcal{L}(f(x;\theta_j),y)\big], \tag{10}$$

which we interpret as the wort-case loss obtained when hospital i evaluates the model parameters of hospital j over the ambiguity set centered at the empirical distribution of hospital i. In other words, $L_{i,j}(\boldsymbol{\theta})$ is the robust transfer loss. Then, we have the following result.

Theorem 1: As before, suppose a parametric model $f(x;\theta)$ defined by model parameters θ and suppose that $\mathcal{L}(f(x;\theta),y)$ is a loss function that is convex with respect to θ. Then, for the optimal value of problem (9) we have that it is upper bounded by:

$$\textstyle\sum_{i,j,k=1}^{N,N,K} \frac{a_{i,k}\cdot a_{j,k}}{\sum_{j=1}^{N} a_{j,k}} L_{i,j}(\boldsymbol{\theta}).$$

Proof: Apply Lemma 4 directly to (9). ∎

Replacing the objective function of the coalition formation problem (9) by the upper bound in Theorem 1 and introducing the binary variables $a_{i,j,k} \in \{0,1\}$, for all $i \in \{1,\cdots,N\}$, $j \in \{1,\cdots,N\}$, $k \in \{1,\cdots,K\}$, so that $a_{i,j,k} = 1$ if $a_{ik} = 1$ and $a_{jk} = 1$, the coalition formation problem (9) can be relaxed to the following linear integer fractional program:

$$\begin{aligned} \min_{a_{i,k}} \quad & \textstyle\sum_{i,j,k=1}^{N,N,K} \frac{a_{i,j,k}}{\sum_{j=1}^{N} a_{j,k}} L_{i,j}(\boldsymbol{\theta}) \\ \text{s.t.} \quad & \textstyle\sum_{k=1}^{K} a_{i,k} = 1, \ \ i \in \{1..N\} \\ & \textstyle\sum_{i=1}^{N} a_{i,k} \geqslant 1, \ \ k \in \{1..K\} \\ & a_{i,j,k} \leqslant a_{i,k}, \ \ i,j \in \{1..N\}, \ k \in \{1..K\} \\ & a_{i,j,k} \leqslant a_{j,k}, \ \ i,j \in \{1..N\}, \ k \in \{1..K\} \\ & a_{i,k} + a_{j,k} - a_{i,j,k} - 1 \leqslant 0, \ \ i,j \in \{1..N\}, k \in \{1..K\} \\ & a_{i,k} \in \{0,1\}, \ \ i \in \{1..N\}, \ k \in \{1..K\} \\ & a_{i,j,k} \in \{0,1\}, \ \ i,j \in \{1..N\}, \ k \in \{1..K\}. \end{aligned} \tag{11}$$

In what follows, we analyze how to compute the robust transfer loss in (10) for two specific model classes: linear models with ℓ_1-loss and logistic regression.

B. Example: ℓ_1-linear regression

In the case of linear models, the robust transfer loss in (10) can be written as

$$L_{i,j}(\boldsymbol{\theta}) = \sup_{\mathbb{Q}_i \in B_{\varepsilon_i}(\hat{\mathbb{P}}_i)} \mathbb{E}_{(x,y)\sim\mathbb{Q}_i}\big[|\theta_j^\top x - y|\big]. \tag{12}$$

To compute this loss, we can use the dual formulation for linear regression with ℓ_1-loss given in Lemma 2. More specifically, for fixed model parameters $\boldsymbol{\theta}$ we get

$$L_{i,j}(\boldsymbol{\theta}) = \varepsilon_i \cdot \alpha + \mathbb{E}_{(x,y)\sim\hat{\mathbb{P}}_i}\big[|\theta_j^\top x - y|\big], \tag{13}$$

where α is a decision variable in the dual formulation; see Eq. (4). Substituting the robust losses of every model θ_j and every hospital i into the coalition formation problem (11), we can obtain the desired optimal coalition structure.

Finally, recall that every hospital is responsible for selecting the value for the radius of its ambiguity set ε_i. If all hospitals select the same radius for their ambiguity sets, the following result holds true.

Proposition 1: Assume all hospitals use the same value ε for the radius of their ambiguity sets. Then, the coalition structure returned by the solution of the coalition formation problem (11) is independent of ε.

Proof: See Appendix ∎

C. Example: Logistic Regression

In the case of logistic regression models, the robust transfer loss in (10) can be written as

$$L_{i,j}(\boldsymbol{\theta}) = \sup_{\mathbb{Q}_i \in B_{\varepsilon_i}(\hat{\mathbb{P}}_i)} \mathbb{E}_{(x,y)\sim\mathbb{Q}_i}\big[\log(1+\exp(-y\cdot\theta_j^T x))\big]. \tag{14}$$

As in the case of linear models discussed above, here too we can use the dual formulation for logistic regression given in Lemma 3 to compute the robust loss in (14). More specifically, for fixed model parameters $\boldsymbol{\theta}$ we get

$$L_{i,j}(\boldsymbol{\theta}) = \varepsilon_i \cdot \alpha + \frac{1}{N}\textstyle\sum_{i=1}^{N} b_i, \tag{15}$$

where both α and b_i are decision variables in the dual formulation; see Eq. (5). Like in the linear case, substituting

the robust losses of every model θ_j and every hospital i, into the coalition formation problem (11), we can obtain the desired optimal coalition structure.

V. Federated Learning Protocol

In this section, we discuss a protocol that describes how the proposed clustered federated learning (CFL) method can be implemented.

1) *Training Phase*: Initially, the hospitals use their local data to learn estimates of their local model parameters θ_i^{local}. Then, they share their estimated local model parameters with the lead hospital and the coalition formation phase begins.
2) *Communication Phase*: To solve the coalition formation problem (11), the lead hospital requires a conservative estimate of the loss of every local model computed on the empirical distribution of every hospital, which is the robust transfer loss defined in (10). To obtain these robust losses, the lead hospital sends to all participating hospitals the local model parameters it has received, without revealing to which hospital each parameter set belongs to; this way information privacy is maintained as no hospital can determine which hospital trained which model. When the hospitals have received the model parameters of all other hospitals, they compute for each model the worst expected loss over all patient distributions in their local ambiguity set and send these values to the lead hospital.
3) *Coalition Phase*: When the lead hospital has received the worst-case expected losses of all models from all hospitals, it can compute the optimal coalition structure by solving the coalition formation problem (11). Then, the lead hospital can compute the aggregate model parameters for each coalition by (6) and send these aggregate parameters to the member hospitals of each coalition.

VI. Experimental Results

A. Synthetic Data

We validate the effectiveness of the proposed CS-RCFL algorithm for linear regression models with absolute (ℓ_1) loss and for logistic regression models. We compare the performance of CS-RCFL to three benchmarks:

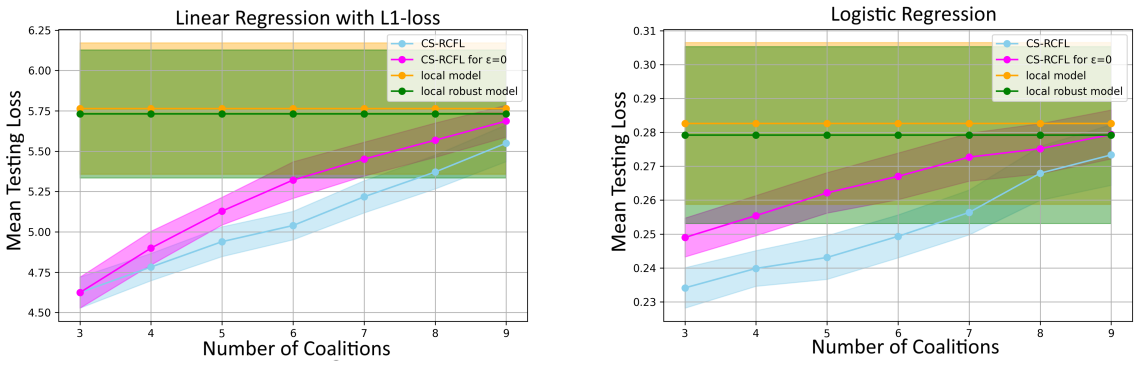
- The local models that the hospitals can learn from their local data, without considering distribution shifts or collaboration with other hospitals.
- The robust local models that the hospitals can learn by minimizing their robust local loss without collaborating with other hospitals.
- The non-robust clustered federated learning models that the hospitals can learn, without considering distribution shifts, i.e., using the non-robust losses to perform the clustering. This benchmark coincides with the proposed CS-RCFL algorithm for $\varepsilon = 0$.

Below we describe the synthetic datasets we generate for the linear regression and logistic regression models. These datasets are different because the first task is a regression problem whereas the second one is a classification problem.

- *Linear Regression Dataset.* The dataset consists of $N = 10$ hospitals. The number of samples n of each hospital ranges between 50 and 150, emulating hospitals with various sizes of patient populations. Every data sample has 50 features. For each hospital i, we generate the observations as $y_i^{(p)} = \tilde{w}_i^\top x_i^{(p)} + \eta$, where $x_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$ is sampled from a multi-variate normal distribution with random covariance matrix and mean, $\eta \sim \mathcal{N}(0, \sigma^2)$ is an i.i.d. noise with standard-deviation $\sigma = 5$, and $\tilde{w}_i$ is the true weight of the model. To simulate hospitals with similar joint distributions (that can be clustered together), we select 3 different values for $\tilde{w}$ and assign each hospital to one of the 3 true weights.
- *Logistic Regression Dataset.* Like before, this dataset consists of $N = 10$ hospitals and the number of samples n of each hospital ranges between 100 and 200. Every data sample has 50 features. For each hospital i, we generate the observations by the (log-loss) probability $P(y_i^{(p)} = 1 | x_i^{(p)}) = [1 + \exp(-\tilde{w}_i^\top x_i^{(p)})]^{-1}$. We generate labels based on a threshold on the log-loss; if $P(y_i^{(p)} = 1 | x_i^{(p)}) > 0.5$, then we set $y^{(p)} = 1$, otherwise $y^{(p)} = -1$.

In our proposed CS-RCFL method, we select the radii of the ambiguity sets of the hospitals depending on the number of samples each one of them has. If $n > 100$, we set $\varepsilon = 1$ for linear models and $\varepsilon = 0.5$ for logistic regression models. Otherwise, we use $\varepsilon = 2$ for linear models and $\varepsilon = 1$ for logistic regression models. For the robust local models method, we use cross-validation to select the radii of the ambiguity sets. Specifically, for linear models we get the values of ε that we also use in the CS-RCFL algorithm, while for logistic regression models we get $\varepsilon = 0.05$ if $n > 100$ and $\varepsilon = 0.1$, otherwise. Note that cross-validation is hard to implement in practice for the CS-RCFL algorithm as this would require coordination with the lead hospital. Instead, we selected conservative ambiguity sets (e.g., for logistic regression models) so that the coalitions returned by CS-RCFL are robust to possibly large distribution shifts.

In what follows, we examine the performance of our algorithm for different numbers of coalitions $K \in \{3, 4, \ldots, 10\}$. Specifically, we compare our proposed CS-RCFL method and the benchmarks described above on 10 testing datasets that contain 100 data-samples each, generated from the same distributions as the training data for each one of the $N = 10$ hospitals. In Fig. 1 we report the loss of each method averaged over the 10 different test datasets. We observe that even without federated learning, the use of DRO slightly improves the performance of the local models, which is expected due to the presence of distribution shifts between the training and test data. On the other hand, our CS-RCFL method consistently outperforms the other benchmarks by achieving lower average model loss and lower variance. It is also worth emphasizing that the models returned by CS-RCFL have lower



(a) (b)

Fig. 1: Loss of the CS-RCFL method and the benchmarks for (a) linear regression models with absolute (ℓ_1-) loss and (b) logistic regression models.

loss compared to those returned by the clustered federated learning method without robustness, showing that our method is more effective in addressing distribution shifts. Moreover, this improved performance is consistent for all values of K, meaning that our algorithm is robust to the choice of the numbers of coalitions, which is typically unknown. We finally note that our algorithm is able to consistently recover the true coalition structure in the case $K = 3$.

B. Real-world Healthcare Data

Next, we validate our algorithm on real-world healthcare data. We use the open access demo version of the eICU Collaborative Research Database, a multi-center database comprised of de-identified health data of over 200,000 admissions to ICUs across the United States between 2014-2015 [37]. Since not all patients in this dataset have the same number of attributes, we select a subset of attributes and construct a dataset with the patients that have those attributes in common. This way, we end up with data from 174 different hospitals with each hospital having between 3 and 18 patients admitted to ICU. For each one of these patients we have access to 57 different features, including demographic information (e.g., age, sex, ethnicity), vital sign measurements, severity of illness measures, and other diagnosis information (e.g., temperature, heart-rate, apachescore etc.). As seen in Fig. 2, the dataset consists of heterogeneous hospitals that serve different patient populations.

An important feature contained in the dataset for every patient is their length-of-stay in the ICU. Knowledge of the length-of-stay in the ICU (similarly, knowledge of the utilization of other hospital resources, such as PACUs or step-down beds) can be used to streamline health system operations and improve delivery of care. Therefore, in this experiment, we focus on ICU length-of-stay prediction. Specifically, to simplify the problem, we consider a logistic regression model that can predict whether a patient will stay in the ICU longer than 1 day or not. Since the number of data samples at each hospital is very low, especially as it relates to the large number of patient features, we consider hospitals that have at least 10 data samples each. Among those, we randomly select 20 hospitals to participate in the proposed federation. Moreover, for these 20 hospitals, we reduce the number of features we use to train the logistic regression models from 57 to 6. These 6 features include both demographic information (gender, ethnicity, and age) and illness measurements (medicines, heart-rate, and apachescore).

To decide the radius ε of the ambiguity set of each hospital we use cross-validation. We get values of ε that range between $[0.001, 0.1]$ across the different hospitals. Small values (i.e., $\varepsilon = 0.001$) indicate that the training and validation sets are similar and thus there are small distribution shifts in the data. On the other hand, larger values (i.e., $\varepsilon = 0.1$) mean that the training distribution is significantly different compared to the validation one. To validate our method, we split the data at every hospital into a training set containing 70% of the data and a test set containing the rest. We then train our CS-RCFL model along with the other benchmarks on the training data and evaluate their loss on the corresponding test sets at the local hospitals. We construct 3 different experiments by randomly selecting 3 different training sets at each hospital and, in Fig. 3, we report the average loss of all models (averaged over the 3 experiments) for various coalition number values $K \in \{1, \ldots, 10\}$. We observe that the fact that hospitals have very few data affects the performance of the local models; both the local and local robust models suffer from high loss and high variance. Moreover, we observe that our proposed CS-RCFL model (with either $\varepsilon = 0$ or $\varepsilon > 0$) outperforms the models trained locally at each hospital both in average loss and in terms of variance. Finally, we see that incorporating robustness, i.e., letting $\varepsilon > 0$, can on average improve the loss of a clustered federated model.

VII. Conclusion

In this paper, we proposed a new clustered federated learning method that assigns hospitals to coalitions allowing hospitals in the same coalition to collaboratively train a common model. We assumed that the local data at each hospital may be subject to local distribution shifts and may also be statistically different across hospitals. Our proposed clustered federated learning method designs coalitions that are robust to distribution shifts in the local data and learns local models that are unbiased in the presence of statistically heterogeneous

Fig. 2: Distribution of patients' ethnicity at two different hospitals, an example that shows heterogeneity of patient populations across hospitals.

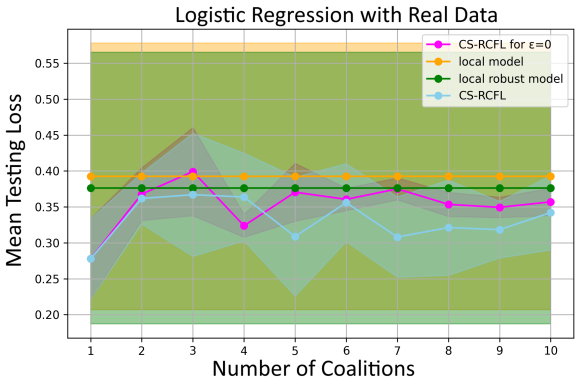


Fig. 3: Loss of the CS-RCFL method and the benchmarks for the logistic regression model, evaluated on the eICU Collaborative Research Dataset.

hospitals. We evaluated our method on synthetic and real healthcare data and showed that it outperforms models that are trained solely on local data or federated models that are not robust to distribution shifts. In future work, we will explore the use of a broader class of models, including multi-layer perceptrons, Neural Networks, etc.

APPENDIX

Proof of Lemma 4: Let $a_k = \sum_{i=1}^N a_{i,k}$, and let $\mathcal{P}_i = B_{\varepsilon_i}(\hat{\mathbb{P}}_i)$. Applying Jensen's inequality, along with the monotonicity of the supremum and the additive property of expectation yields $\ell_i^{rob}(\boldsymbol{\theta}, \pi) =$

$$\sup_{\mathbb{Q}_i \in \mathcal{P}_i} \mathbb{E}_{(x,y)\sim\mathbb{Q}_i}\left[\mathcal{L}\left(\sum_{j=1}^N \frac{a_{j,k}}{a_k}\theta_j^\top x, y\right)\right] \leqslant$$
$$\sup_{\mathbb{Q}_i \in \mathcal{P}_i} \mathbb{E}_{(x,y)\sim\mathbb{Q}_i}\left[\sum_{j=1}^N \frac{a_{j,k}}{a_k}\mathcal{L}\left(\theta_j^\top x, y\right)\right] =$$
$$\sup_{\mathbb{Q}_i \in \mathcal{P}_i} \sum_{j=1}^N \frac{a_{j,k}}{a_k}\mathbb{E}_{(x,y)\sim\mathbb{Q}_i}\left[\mathcal{L}\left(\theta_j^\top x, y\right)\right] =$$
$$\sum_{j=1}^N \frac{a_{j,k}}{a_k} \sup_{\mathbb{Q}_i \in \mathcal{P}_i} \mathbb{E}_{(x,y)\sim\mathbb{Q}_i}\left[\mathcal{L}\left(\theta_j^\top x, y\right)\right].$$

■

Proof of Proposition 1: From (13) and $\alpha = \sqrt{||\theta_i||_2^2 + 1}$ from Lemma 2, if $\varepsilon_i = \varepsilon\ \forall i \in \{1, \ldots, N\}$, then the objective becomes $\sum_{i=1}^N \varepsilon\sqrt{||\theta_i||_2^2 + 1} +$
$\sum_{k=1}^K \sum_{i=1}^N \sum_{j=1}^N \frac{a_{i,j,k}}{\sum_{j=1}^N a_{j,k}} \left(\mathbb{E}_{(x,y)\sim\hat{\mathbb{P}}_i}\left[|\theta_j^\top x - y|\right]\right)$.

In this case, the coalition structure returned by this method is independent of the value of the radius of the ambiguity set ε, since all the decision variables $a_{i,j,k}, a_{j,k}$ are independent of ε. ■

REFERENCES

[1] L. O. Gostin, "National health information privacy: regulations under the health insurance portability and accountability act," *Journal of the American Medical Association*, vol. 285, no. 23, pp. 3015–3021, 2001.

[2] W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," *Nature medicine*, vol. 25, no. 1, pp. 37–43, 2019.

[3] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, and R. Cummings, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[4] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[5] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of healthcare informatics research*, vol. 5, pp. 1–19, 2021.

[6] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, and W.-J. Hwang, "Federated learning for smart healthcare: A survey," *ACM Computing Surveys (Csur)*, vol. 55, no. 3, pp. 1–37, 2022.

[7] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[8] A. Penman-Aguilar, M. Talih, D. Huang, R. Moonesinghe, K. Bouye, and G. Beckles, "Measurement of health disparities, health inequities, and social determinants of health to support the advancement of health equity," *J Public Health Manag Pract*, vol. 22 Suppl 1, pp. S33–42, Jan. 2016.

[9] Y. J. Cho, J. Wang, T. Chiruvolu, and G. Joshi, "Personalized federated learning for heterogeneous clients with clustered knowledge transfer," *arXiv preprint arXiv:2109.08119*, 2021.

[10] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 8, pp. 3710–3722, 2020.

[11] N. Shlezinger, S. Rini, and Y. C. Eldar, "The communication-aware clustered federated learning problem," in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2610–2615, 2020.

[12] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19586–19597, 2020.

[13] G. Long, M. Xie, T. Shen, T. Zhou, X. Wang, and J. Jiang, "Multi-center federated learning: clients clusteriuu for better personalization," *World Wide Web*, vol. 26, no. 1, pp. 481–500, 2023.

[14] J. Ma, G. Long, T. Zhou, J. Jiang, and C. Zhang, "On the convergence of clustered federated learning," *arXiv preprint arXiv:2202.06187*, 2022.

[15] S. Luo, Y. Xiao, and L. Song, "Personalized federated recommendation via joint representation learning, user clustering, and model adaptation," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 4289–4293, 2022.

[16] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, "Three approaches for personalization with applications to federated learning," *arXiv preprint arXiv:2002.10619*, 2020.

[17] M. Morafah, S. Vahidian, W. Wang, and B. Lin, "Flis: Clustered federated learning via inference similarity for non-iid data distribution," *IEEE Open Journal of the Computer Society*, vol. 4, pp. 109–120, 2023.

[18] M. Duan, D. Liu, X. Ji, Y. Wu, L. Liang, X. Chen, and Y. Tan, "Flexible clustered federated learning for client-level data distribution shift," 2021.

[19] M. Werner, L. He, M. Jordan, M. Jaggi, and S. P. Karimireddy, "Provably personalized and robust federated learning," 2023.

[20] Y. Guo, X. Tang, and T. Lin, "Fedrd: Tackling diverse distribution shifts challenge in federated learning by robust clustering," *arXiv preprint arXiv:2301.12379*, 2023.

[21] C. Huang, J. Huang, and X. Liu, "Cross-silo federated learning: Challenges and opportunities," *arXiv preprint arXiv:2206.12949*, 2022.

[22] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, "Personalized cross-silo federated learning on non-iid data," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 7865–7873, 2021.

[23] S. Jiang and J. Wu, "Coalition formation game in the cross-silo federated learning system," in *2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, pp. 49–57, IEEE, 2022.

[24] W. Bao, H. Wang, J. Wu, and J. He, "Optimizing the collaboration structure in cross-silo federated learning," in *International Conference on Machine Learning*, pp. 1718–1736, PMLR, 2023.

[25] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 4615–4625, PMLR, 09–15 Jun 2019.

[26] Y. Deng, M. M. Kamani, and M. Mahdavi, "Distributionally robust federated averaging," *Advances in neural information processing systems*, vol. 33, pp. 15111–15122, 2020.

[27] K. Pillutla, Y. Laguel, J. Malick, and Z. Harchaoui, "Federated learning with superquantile aggregation for heterogeneous data," *Machine Learning*, pp. 1–68, 2023.

[28] T.-A. Nguyen, T. D. Nguyen, L. T. Le, C. T. Dinh, and N. H. Tran, "On the generalization of wasserstein robust federated learning," *arXiv preprint arXiv:2206.01432*, 2022.

[29] A. Reisizadeh, F. Farnia, R. Pedarsani, and A. Jadbabaie, "Robust federated learning: The case of affine distribution shifts," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 21554–21565, Curran Associates, Inc., 2020.

[30] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, B. McMahan, T. Van Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," in *Proceedings of Machine Learning and Systems* (A. Talwalkar, V. Smith, and M. Zaharia, eds.), vol. 1, pp. 374–388, 2019.

[31] 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*, pp. 1273–1282, PMLR, 2017.

[32] M. Thorpe, "Introduction to optimal transport," *Notes of Course at University of Cambridge*, 2018.

[33] L. Rüschendorf, "The wasserstein distance and approximation theorems," *Probability Theory and Related Fields*, vol. 70, no. 1, pp. 117–129, 1985.

[34] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations," 2017.

[35] R. Chen and I. C. Paschalidis, "A robust learning approach for regression models based on distributionally robust optimization," *Journal of Machine Learning Research*, vol. 19, no. 13, pp. 1–48, 2018.

[36] S. Shafieezadeh Abadeh, P. M. Mohajerin Esfahani, and D. Kuhn, "Distributionally robust logistic regression," *Advances in neural information processing systems*, vol. 28, 2015.

[37] T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eicu collaborative research database, a freely available multi-center database for critical care research," *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.