A Fast Douglas-Rachford Splitting Method for Solving Federated Minimax Problems

Anonymous Authors[1]

Abstract

Federated learning has recently been actively studied to train machine learning models collaboratively across devices/users without directly sharing data, which can address data hungry issues and also protect data privacy. Minimax problems arise in many machine learning tasks, such as adversarial training, GANs, fairness learning, and AUROC maximization. In this paper, we focus on designing fast learning algorithm to solve the federated minimax problems. To address the key challenges of heterogeneity and client drift in federated learning, we propose a new method to conduct efficient local training with tolerating client drift and improving performance in the face of data heterogeneity. We provide both asymptotic and non-asymptotic analyses for our proposed method to show that our method has a smaller sample complexity compared to existing approaches. Moreover, we demonstrate that the variables generated by our method converge finitely, linearly, or sublinearly under the Kurdyka-Łojasiewicz property. Unlike previous analyses for federated minimax methods, our analysis does not assume bounded heterogeneity among clients which could lead to large convergence error. We validate our federated learning method on AUC maximization tasks. The experimental results demonstrate that our method outperforms state-of-the-art federated learning methods when the distributions of local training data are non-IID.

1. Introduction

In recent years, federated learning (FL) has attracted increasing attentions in machine learning community due to many important real-world applications in finance, health, edge computing, AIoT, *etc.* Federated learning trains models across multiple devices or servers efficiently. It enables the training of machine learning models using data from different clients without requiring them to exchange the data with each other. Meanwhile, federated learning avoids transferring large datasets, thus reducing bandwidth requirements and associated costs.

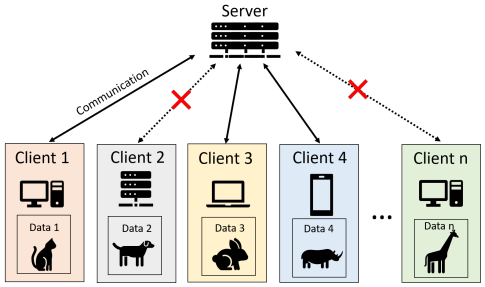


Figure 1. Illustration of (data and system) heterogeneity and client drift problems in federated learning.

Designing federated learning algorithm becomes challenging when we face problems with nested optimization structures. These structures are prevalent in a variety of domains, such as AUROC (area under the ROC curve) maximization (Lei & Ying, 2021), adversarial training (Tramèr et al., 2018; Bai et al., 2021), distributionally robust optimization (Levy et al., 2020; Gao & Kleywegt, 2023; Madras et al., 2018), and training generative adversarial networks (GANs) (Goodfellow et al., 2014). These problems often involve solving minimax optimization problems. While some methods have been developed to address these problems in a centralized manner, research on federated learning approaches for minimax problems is still in its early stages. In this work, we focus on studying federated methods for minimiax problem. Specifically, we consider:

$$\min_{x\in\mathbb{R}^l}\max_{y\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}f_i(x,y)+g(x), \tag{1}$$

where each $f_i(x,y)=\sum_{j\in\mathcal{D}_j}f(x,y;\xi_j)$ with $\mathcal{D}_j$ being the dataset on the j_{th} client, is nonconvex in x strongly concave in y, and g is a proper closed convex function. We assume the proximal operator of g is easy to compute

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Table 1. Local(L) SDGA (Sharma et al., 2022), Momentum Local(ML) SGDA (Sharma et al., 2023), FedSGDA (Wu et al., 2023), FEDNEST (Tarzanagh et al., 2022). BH=Bounded Heterogeneity, F/P=Partial/Full attendance.

	F/P	BH	Sample	Communication
LSDGA	F	✓	$O(\kappa^4 n^{-1}\epsilon^{-4})$	$O(\kappa^3\epsilon^{-3})$
MLSGDA	P	✓	$O(\kappa^4 n^{-1}\epsilon^{-4})$	$O(\kappa^3\epsilon^{-3})$
FedSGDA	F	✓	$O(\kappa^3 n^{-1}\epsilon^{-3})$	$O(\kappa^2\epsilon^{-2})$
FEDNEST	P	✓	$O(\kappa^3\epsilon^{-4})$	$O(\kappa^2\epsilon^{-4})$
FedSGDA	F	✓	$O(\kappa^3 n^{-1}\epsilon^{-3})$	$O(\kappa^2\epsilon^{-2})$
Our Stochastic Case	P	✗	$O(\kappa^2\log(\kappa) n^{-1}\epsilon^{-2})$	$O(\kappa^2\epsilon^{-2})$

through out this work. Examples of g in (1) include convex regularizers or indicator functions of convex constraints.

The challenge in solving federated minimax problems lies in how to deal with the max problem and its nested relation with the min problem when we are only allowed to train the model locally. A classical method for centralized minimax problems is the Gradient Descent Ascent method (GDA). One federated learning approach is to combine the classic federated learning method FedAvg, used for minimization problems, with GDA, which results in LocalSGDA (Deng et al., 2020). Sharma et al. (2022) proposed Momentum Local SGDA, which accelerates LocalSGDA via adding momentum to the local updates. Wu et al. (2023) presented FedSGDA+, which further reduces the complexity of Momentum Local SGDA.

However, these methods require all clients to participate in training during each round. Consequently, they cannot address the challenge of client drift, *i.e.* due to the device limitations, not all clients are able to participate in the training. Sharma et al. (2023) proposed a federated minimax optimization framework that allows the partial attendance of clients. In this work, we propose a more efficient method that also permits the partial attendance of clients.

Another critical challenge in federated learning is the data heterogeneity, which means the distribution of clients' datasets can vary significantly as shown in Figure 1. The existence of heterogeneity slows down federated learning methods and could make the trained model less efficient. Sharma et al. (2023; 2022); Wu et al. (2023) investigated how the heterogeneity affects the convergence of their proposed methods by assuming a bound on the heterogeneity. However, the heterogeneity bound could be very large in real-world scenarios, thus the error in the convergence upper bound associated with this heterogeneity bound could also be large. To address this challenging issue, in this work, we introduce a new method whose convergence guarantees do not rely on the heterogeneity bound.

1.1. Contributions

In this work, we tackle the challenging heterogeneity and client drift issues in federated minimax optimization problems. We propose a new algorithm based on the Douglas-Rachford (DR) Splitting method, which allows partial client participation in each round. Meanwhile, our method addresses heterogeneity via integrating linear combinations and solving a strongly convex-strongly concave auxiliary minimax problem with an added quadratic term to balance the local and aggregated parameters. We propose a novel termination criterion in local training. This criterion enables the selected clients to terminate local updates after a fixed number of iterations while still maintaining the convergence guarantees. Our new method is named as the Fast Federated Minimax DR (FFMDR) method.

Table 2. Comparisons between Sample Complexity and Sequential Convergence of our method. z^* is the limiting point of $\{z^t\}$.

	Sample Complexity	Sequential Convergence
Focus	Loss function	Model parameters
Metric	$\text{dist}(0, \nabla\sum_{i=1}^n \frac{1}{n} f_i(z^t) + \partial g(z^t))$	$\|z^t - z^*\|$
A/N	Non-asymptotic	Asymptotic
Rate	Sublinear	Finite/Linear/Sublinear
Cases	Stochastic	Deterministic

Theoretically, we show that the proposed method has the following merits:

- Our method improves the sample complexity of existing federated minimax methods. The comparisons of our method and existing ones are summarized in Table 1. We show that the proposed method achieves a sample complexity of $O(\kappa^2\log(\kappa)/n\epsilon^{-2})$.
- Our method is the first one in federated learning to have sequential convergence guarantees beyond strong convexity, see Table 2 for the difference between complexity and sequential convergence. We analyze the convergence rate of variables, including the model parameters, generated by this method. We demonstrate that when all clients participate in training and the local solvers are deterministic, the sequences generated by our method converge, given the Kurdyka-Łojasiewicz (KL) exponent of a potential function. Specifically, we show that both x- and y-related updates generated by our method converge finitely, linearly, or sublinearly when the KL exponent is 0, $(0, \frac{1}{2}]$, or $(\frac{1}{2}, 1)$, respectively.
- We weaken the KL assumptions made on the potential function compared to the previous work on sequential analysis for the centralized minimax problem by Chen et al. (2021). In their work, the potential function relies on the maximizer $y(x) := \text{argmax}\, f(x, y)$ and the maximum function $f(x) := \max_y f(x, y)$. This makes verifying the KL assumption difficult. In our work, the potential function does not rely on $y(x) := \text{argmax}\, f(x, y)$, and we provide Proposition

4.7 to deduce the KL exponent of our potential function from the original objective function. Therefore, our analysis provides a weaker assumption for the sequential convergence analysis of the method for the minimax optimization problem.

We apply our method to AUC maximization problems. In the presence of data heterogeneity, the experiments show that our method outperforms other existing methods.

1.2. Related work

Federated learning for minimization problem Classical federated learning methods for minimization problem include FedAvg (McMahan et al., 2017) and FedDualAvg, (Yuan et al., 2021a) and SCAFFOLD (Karimireddy et al., 2020). In order to address the heterogeneity problem in FL, federated splitting methods are proposed, see (Yuan et al., 2021a; Li et al., 2020; Reddi et al., 2021; Pathak & Wainwright, 2020; Tran-Dinh et al., 2021) for examples.

Closely related work Our method is closely related to the FedDR method for the minimization problem in (Tran-Dinh et al., 2021). However, our work differs from ((Tran-Dinh et al., 2021)) in three perspectives:

1. We work on minimax problems. The existence of the maximization problem raises new challenges in theoretical analysis. To address this challenge, we propose a new potential function that is related to the variables in the maximization problem. Based on this new potential function, we are able to provide new analyses on complexity as well as sequential convergence.

2. We provide comprehensive sequential convergence analysis, the analysis of the behavior of the generated sequences, i.e. $\{(x^t, y^t)\}$. This result is not only new when our method degenerates to solve the minimization problems in federated learning, but also new when degenerated to the centralized minimax problem. In our sequential analysis, we provide the finte/linear/sublinear asymptotic convergence rate of the generated sequence.

3. We conducted further investigation into the KL assumption for sequential analysis. In fact, we demonstrate that we have weakened the KL assumptions made on the potential function compared to the previous work on sequential analysis for the centralized minimax problem by Chen et al. (2021). In their work, the potential function relies on the maximizer $y(x) := \operatorname{argmax} f(x, y)$ and the maximum function $f(x) := \max_y f(x, y)$. This makes verifying the KL assumption difficult. In our work, the potential function does not rely on $y(x) := \operatorname{argmax} f(x, y)$, and we provide Proposition 4.7 to deduce the KL exponent of our potential function from the original objective function. Therefore, our analysis provides a weaker assumption for the sequential convergence analysis of the method for the minimax optimization problem.

Federate methods for minimax The existing federated learning methods mainly focus on minimization problems such as minimizing the emprical loss in FL, (Kairouz et al., 2021; McMahan et al., 2017; Pathak & Wainwright, 2020). (Li et al., 2023; Deng et al., 2020; Peng et al., 2020) are among the early works that proposed federated minimax methods for adversarial training problems. (Sharma et al., 2022) investigated local stochastic gradient descent ascent in nonconvex-concave and nonconvex-nonconcave settings. Their analysis assumed an equal number of SGDA-like local updates with full client participation, whereas our method allows for different local updates and partial client participation. (Sharma et al., 2023) proposed a federated minimax optimization framework that includes local SGDA as a special case. They analyzed the convergence of the proposed algorithm under a global heterogeneity assumption that addresses inter-client data and system heterogeneity. In contrast to their work, we do not assume any specific assumptions on heterogeneity but still provide convergence guarantees. (Wu et al., 2023) analyzed the nonconvex-strongly-concave case and showed that their proposed method has a communication complexity of $O(\kappa^2\epsilon^{-2})$ and a gradient complexity of $O(\kappa^2 n^{-1}\epsilon^{-3})$. Their analysis also assumes bounded data heterogeneity. In our work, we remove this assumption. (Tarzanagh et al., 2022) proposed FEDNEST to address the general bilevel federated learning problem. When it degenerates to the nonconvex-strongly-concave problem, the sample complexity of their method is $O(\kappa^3\epsilon^{-4})$, and the communication complexity is $O(\kappa^2\epsilon^{-4})$.

2. Preliminaries

We denote $\mathbb{R}^n$ as the n-dimensional Euclidean space with inner product $\langle\cdot,\cdot\rangle$ and Euclidean norm $\|\cdot\|$. We denote the unit ball in $\mathbb{R}^n$ as $\mathcal{B}(0,1)$. We denote the set of positive real value as $\mathbb{R}_{++}$. Given a point $x \in \mathbb{R}^n$ and a set A, we denote the distance from x to A as $d(x, A)$.

An extended-real-valued function $f : \mathbb{R}^n \to [-\infty, \infty]$ is said to be proper if $\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < \infty\}$ is not empty and f never equals $-\infty$. We say a proper function f is closed if it is lower semicontinuous.

Following Definition 8.3 of (Rockafellar & Wets, 1998), the regular subdifferential of a proper function $f : \mathbb{R}^n \to [-\infty, \infty]$ at $x \in \operatorname{dom} f$ is defined as:

$$\hat{\partial} f(x) := \left\{ \xi \in \mathbb{R}^n : \liminf_{z \to x,\ z \neq x} \frac{f(z) - f(x) - \langle \xi, z - x\rangle}{\|z - x\|} \geq 0 \right\}.$$

The (limiting) subdifferential of f at $x \in \operatorname{dom} f$ is defined

as

$$\partial f(x) := \left\{\xi \in \mathbb{R}^n : \exists x^k \xrightarrow{f} x, \xi^k \to \xi \text{ with } \xi^k \in \hat{\partial} f(x^k), \forall k\right\},$$

where $x^k \xrightarrow{f} x$ means both $x^k \to x$ and $f(x^k) \to f(x)$. For $x \notin \text{dom } f$, we define $\hat{\partial} f(x) = \partial f(x) = \emptyset$. We denote $\text{dom } \partial f := \{x : \partial f(x) \neq \emptyset\}$. When f is convex, the limiting subdifferential reduces to the classical subdifferential in convex analysis.

For a proper function $f : \mathbb{R}^n \to [-\infty, \infty]$, we denote the proximal operator of f as

$$\text{Prox}_{\beta f}(x) := \underset{z \in \mathbb{R}^n}{\text{Arg min}} \left\{ f(z) + \frac{1}{2\beta}\|z - x\|^2 \right\}.$$

Next, we make a general assumption on (1).

Assumption 2.1. For (1), we assume the followings hold:

(i) Each f_i is strongly concave with modulus $\mu > 0$.

(ii) Each f_i is differentiable and ∇f_i is Lipschitz continuous with modulus L_f.

For the maximum of a strongly concave function, we have the following property, see (Lin et al., 2020; Huang et al., 2021; Chen et al., 2021) for examples.

Proposition 2.2. *Consider (1). Suppose Assumption 2.1 holds. Then for any x, there exists unique $y(x)$ such that $F_i(x) = f_i(x, y(x))$. In addition, F_i is continuously differentiable and $\nabla F_i(x) = \nabla_x f_i(x, y(x))$ is Lipschitz continuous with modulus $L := L_f(1 + \kappa)$, where $\kappa := \frac{L_f}{\mu}$.*

We say x is a stationary point of (1) if it satisfies $0 \in \nabla \sum_{i=1}^n \frac{1}{n} f_i(x) + \partial g(x)$. Thanks to Exercise 8.8 and Theorem 10.1 of (Rockafellar & Wets, 1998), we know that if x is a local minimizer of (1), it is a stationary point.

Now we give the definition of the KL property.

Definition 2.3 (Kurdyka-Łojasiewicz property and exponent). A proper closed function $f : \mathbb{R}^n \to (-\infty, \infty]$ is said to satisfy the Kurdyka-Łojasiewicz (KL) property at an $\hat{x} \in \text{dom } \partial f$ if there are $a \in (0, \infty]$, a neighborhood V of $\hat{x}$ and a continuous concave function $\varphi : [0, a) \to [0, \infty)$ with $\varphi(0) = 0$ such that

(i) φ is continuously differentiable on $(0, a)$ with $\varphi' > 0$ on $(0, a)$;

(ii) for any $x \in V$ with $f(\hat{x}) < f(x) < f(\hat{x}) + a$, it holds that $\varphi'(f(x) - f(\hat{x}))\text{dist}(0, \partial f(x)) \geq 1$.

If f satisfies the KL property at $\hat{x} \in \text{dom } \partial f$ and φ can be chosen as $\varphi(\nu) = a_0 \nu^{1-\alpha}$ for some $a_0 > 0$ and $\alpha \in [0, 1)$, then we say that f satisfies the KL property at $\hat{x}$ with exponent α. A proper closed function f satisfying the KL property at every point in $\text{dom } \partial f$ is called a KL function, and a proper closed function f satisfying the KL property with exponent $\alpha \in [0, 1)$ at every point in $\text{dom } \partial f$ is called a KL function with exponent α.

Algorithm 1 Fast Federated Minimax DR (FFMDR) method for (1)

1: Input: x_i^0, z_i^0, y_i^0, $\Upsilon_{i,0}$. Set $w_i^0 = z_i^0$. Set $\epsilon_{i,w} > 0$, $\beta \in (0, \frac{1}{L})$. Let $t = 0$.

2: Sample clients $\mathcal{S}^t \subseteq \{1, \ldots, n\}$ according to Assumption 3.2. For each client $i \in \mathcal{S}^t$:
Let
$$x_i^{t+1} = x_i^t + z^t - w_i^t \tag{2}$$
Find an approximate solution (w_i^{t+1}, y_i^{t+1}) to
$$\min_{w_i} \max_{y_i} r_{i,t+1}(w_i, y_i)$$
such that (10) is satisfied, where $r_{i,t+1}$ is defined in (7). Let $\tilde{z}_i^{t+1} = 2w_i^{t+1} - x_i^{t+1}$.

3: For the server:
Let
$$z^{t+1} = \text{Prox}_{\frac{\beta}{n} g}\left(\frac{1}{n}\sum_{i=1}^n \tilde{z}_i^{t+1}\right) \tag{3}$$

4: If a termination criterion is not met, let $t = t + 1$ and go to Step 2.

3. Fast Federated Minimax DR method

The proposed Fast Federated Minimax DR (FFMDR) method is presented in Algorithm 1. The idea of this method is based on the Douglas-Rachford splitting method for the following reformation of (1):

$$\min_X \underbrace{\frac{1}{n}\sum_{i=1}^n F_i(x_i)}_{F(X)} + \underbrace{g(x_1) + \delta_{\mathcal{C}}(x_1, \ldots, x_n)}_{\tilde{g}(X)}, \tag{4}$$

where $F_i(x_i) := \max_{y_i \in \mathbb{R}^d} f_i(x_i, y_i)$, $X = (x_1, \ldots, x_n)$ and $\mathcal{C} = \{X : x_1 = x_2 = \cdots = x_n\}$. The Classic DR method (Lions & Mercier, 1979) to (4) is as follows: pick any X^0, let $Z^0 = X^0$ and $W^0 = \text{prox}_{\beta F}(X^0)$. Then for $t = 0, \ldots, T$, update:

$$\begin{aligned} X^{t+1} &= X^t + Z^t - W^t, \\ W^{t+1} &= \text{Prox}_{\beta F}(X^{t+1}), \\ Z^{t+1} &= \text{Prox}_{\beta \tilde{g}}(2W^{t+1} - X^{t+1}). \end{aligned} \tag{5}$$

Noting that F_i in (1) is a maximization function and F is

separable, the update of W^t in (5) is equivalent to

$$W^{t+1} = \min_W \max_Y \sum_i F(w_i, y_i) + \frac{1}{2\beta}\|w_i - x_i^{t+1}\|^2, \quad (6)$$

where $W = (w_1, \ldots, w_n)$ and $Y = (y_1, \ldots, y_n)$. The above problem is a minimax problem and cannot be solve exactly in the federated setting. This requires us to consider an efficient method that can find an good inexact solution to (6). We notice that (6) is a smooth strongly convex strongly concave (SC-SC) minimax problem. Since we let $\beta < \frac{1}{L}$, Proposition 2.2 guarantees the existence of the unique solution to the minimax subproblem.

Denote

$$r_{i,t+1}(w_i, y_i) := f_i(w_i, y_i) + \frac{1}{2\beta}\|w_i - x_i^{t+1}\|^2. \quad (7)$$

Then (6) is equivalent to

$$\min_{w_i} \max_{y_i} r_{i,t+1}(w_i, y_i), \quad (8)$$

for $i = 1, \ldots, n$. Then, we only need an inner solver to solve a SC-SC smooth minimax problem. Many methods such as those in (Benjamin et al., 2022; Fallah et al., 2020; Lin et al., 2020; Kovalev & Gasnikov, 2022; Palaniappan & Bach, 2016) can be applied as an inner solver for our subproblem. On the other hand, to have better convergence gurantees, we need an efficient termination criterion to terminate the inner solver. In the following lemma, we show how the SAGA in (Palaniappan & Bach, 2016) can be terminated in constant iterations when satisfying a termination criterion that depends on the current updates.

Proposition 3.1. *Suppose $r : \mathbb{R}^l \times \mathbb{R}^d \to \mathbb{R}$ is a μ_w-strongly convex μ_y strongly convex smooth function. Suppose ∇r is Lipschitz continuous with modulus l. Apply SAGA in (Palaniappan & Bach, 2016) to solve $\min_w \max_y r(w, y)$. Let (w^k, y^k) be the k_{th} iteration of SAGA. Let $(\bar{w}, \bar{y})$ satisfies $\nabla r(\bar{x}, \bar{y}) \neq 0$. Let $\epsilon_w > 0$. Then there exists $k = O(\max\{\frac{l}{m}, \log(\kappa)\})$ such that*

$$\begin{aligned} &\mathbb{E}\left\|(w^{k+1}, y^{k+1}) - (w_\star, y_\star)\right\|^2 \\ &\leq \epsilon_w \mathbb{E}\|(\bar{w}, \bar{y}) - (w^{k+1}, y^{k+1})\|^2, \end{aligned} \quad (9)$$

where (x^*, y^*) is the unique solution.

In inspired by (9), we propose to terminate the solver used in client i for solving (8) when[1]

$$\begin{aligned} &\mathbb{E}_t\left\|(w_i^{k+1}, y_i^{k+1}) - (w_{i,\star}^{t+1}, y_{i,\star}^{t+1})\right\|^2 \\ &\leq \epsilon_{i,w} \mathbb{E}_t \Upsilon_{i,t+1}, \end{aligned} \quad (10)$$

[1]We denote $E_t\xi$ as the expectation of the outputs ξ of local stochastic solver conditioned on $\{x_1^t, \ldots, x_n^t\}$,$\{y_1^t, \ldots, y_n^t\}$, $\{z^t\}$,$\{w_1^t, \ldots, w_n^t\}$.

where $(w_{i,\star}^{t+1}, y_{i,\star}^{t+1})$ is the exact solution to (8) and

$$\Upsilon_{i,t+1} := \|(w_i^t, y_i^t) - (w_i^{t+1}, y_i^{t+1})\|^2.$$

On the other hand, using the first-order optimality condition of the problem in the update of z^t in (5), Z^{t+1} in (5) is equivalent to $(\underbrace{z^{t+1}, \ldots, z^{t+1}}_{n's})$ with $z^{t+1} = \text{Prox}_{\frac{\beta}{n}g}(\frac{1}{n}\sum_i x_i^{t+1})$, see Appendix of A.1 in (Tran-Dinh et al., 2021) for more details.

Finally, considering the cliendt drift, we make the following assumption.

Assumption 3.2. At each round, the client i has the probability $p_i \in (0, 1]$ to attend the training.

Based on this fact, Assumption 3.2 and Proposition 3.1, we obtain Algorithm 1.

4. Convergence analysis

4.1. Sample Complexity of Algorithm 1

In this section, we analyze Algorithm 1 in a general stochastic case. We first present a recursive relation with respect to the $\Upsilon_{i,t}$.

Proposition 4.1. *Consider (1). Suppose Assumptions 2.1 and 3.2 hold. Let $\{\Upsilon_{i,t}\}$ and $\{w_i^t\}$ be generated by Algorithm 1. Assume $\frac{1}{\beta} > L$, where L is defined as in Proposition 2.2. Then there exist ϵ_w such that for $t \geq 0$,*

$$\begin{aligned} \sum_i p_i \mathbb{E}\Upsilon_{i,t+1} \leq &\frac{1}{2}\left(\sum_i p_i \mathbb{E}\Upsilon_{i,t} - \sum_i p_i \mathbb{E}\Upsilon_{i,t+1}\right) \\ &+ 6L^2 \sum_i p_i \mathbb{E}\|w_i^t - w_i^{t+1}\|^2. \end{aligned}$$

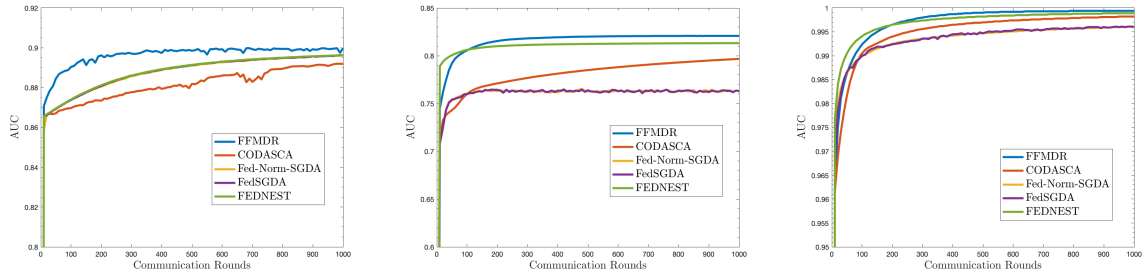
In the next theorem, we present a descent-type lemma.

Theorem 4.2. *Consider (1). Suppose the conditions in Proposition 4.1 hold. Let $\{(x_1^t, \ldots, x_n^t)\}$,$\{(y_1^t \ldots, y_n^t)\}$, $\{(w_1^t, \ldots, w_n^t)\}$, $\{z^t\}$ be generated by Algorithm 1. Let L be the one in Proposition 2.2. Given a $\delta > 0$, define*

$$\begin{aligned} &H(X, W, Z, Y, W', Y') := F(W) + \tilde{g}(Z) \\ &+ \frac{1}{2\beta}\left(\|X - W\|^2 - \|X - Z\|^2\right) + \frac{1}{\beta}\|W - Z\|^2 \\ &+ \frac{\delta}{\beta}\|W - W'\|^2 + \frac{1}{12L^2}\sum_i p_i \|(y_i, w_i) - (y_i', w_i')\|^2. \end{aligned} \quad (11)$$

where F and $\tilde{g}$ is defined in (4). Denote

$$\begin{aligned} X^t &= (x_1^t, \ldots, x_n^t), Y^t = (y_1^t, \ldots, y_n^t), \\ W^t &= (w_1^t, \ldots, w_n^t), Z^t = (z^t, \ldots, z^t) \end{aligned}$$



(a) a9a (b) covtype (c) gisette

(d) ijcnn1 (e) phishing (f) w8a

Figure 2. AUC values w.r.t. communication rounds on test dataset: a9a, covtype, gisette, ijcnn1, phishing and w8a.

and $H_t := \mathbb{E}H(X^t, W^t, Z^t, Y^t, W^{t-1}, Y^{t-1})$. *Let* $\delta_\beta \in (0, \frac{1}{2})$. *Let* $\beta \in (0, \frac{1}{L})$ *be such that*

$$(1+\beta L)^2 - \frac{3}{2} + \frac{5}{2}\beta L < -\delta_\beta.$$

Let $\delta' \in [0, \delta_\beta)$. *Let* $\iota > 0$ *and* $\tau \in (0,1)$ *be small enough such that*

$$\frac{1-L\beta}{2}\tau^2 + (1+\beta L)^2(2\iota+\iota^2) + (\beta L - 1)^2\iota < \delta'.$$

Denote $\delta := \delta_\beta - \delta'$. *Suppose that* ϵ_w *is small enough such that*

$$\left(\Gamma\frac{2}{(\frac{1}{\beta}-L)^2} + \frac{1}{\tau^2}\frac{1}{2(\frac{1}{\beta}-L)}\right) 6CL^2\epsilon_w \le \frac{\delta-\delta_\epsilon}{\beta},$$

for some $\delta_\epsilon > 0$, *where* $\Gamma := \frac{(1+\iota)^2}{\beta\iota} + \frac{2}{\beta}\left(\frac{1}{\iota} + \beta L - 1\right)$ *and* $C := 2\left(\frac{(L_f+\frac{1}{\beta})^2}{\mu^2} + 1\right)\left(L_f + \frac{1}{\beta}\right)^2$.

Then, for $t \ge 1$,

$$H_{t+1} \le H_t - \frac{\delta_\epsilon}{\beta}\|W^t - W^{t-1}\|^2. \tag{12}$$

Remark 4.3. The assumptions related to our method is on the β and ϵ_w. β and ϵ_w need to be small enough such that they satisfy the first and the third display before (12). The constants $\{\delta_\beta, \delta', \tau, \iota, \epsilon_w, \delta_\epsilon\}$ aims to make the first and the third display before (12) well defined.

Now we calculate the complexity of Algorithm 1.

Theorem 4.4. *Let assumptions in Theorem 4.2 hold. Let* $\{(x_1^t, \dots, x_n^t)\}, \{(y_1^t \dots, y_n^t)\}$, $\{(w_1^t, \dots, w_n^t)\}$, $\{z^t\}$ *be generated by Algorithm 1. We further suppose* ϵ_w *and* β *are small enough such that* $\frac{1}{2(\frac{1}{\beta}-L)}C\epsilon_w + 6L^2\sum_i p_i \le \frac{\delta}{\beta}$, *where* C *is defined in Theorem 4.2. Then it holds that*

$$\frac{1}{T+1}\sum_{t=1}^{T+1}\mathbb{E}d^2(0, \nabla\sum_{i=1}^n f_i(z^t) + \partial g(z^t)) \le \frac{n}{p}\frac{1}{T+1}\left(D_1\bar{H}_0 + D_2\Upsilon_0 + D_3\|Y^0 - y(W^0)\|^2\right),$$

where $\bar{H}_0 := F(W^0) + \tilde{g}(Z^0) + \frac{1}{2\beta}\|X^0 - W^0\|^2 - \frac{1}{2\beta}\|X^0 - Z^0\|^2$, $D_1 := \frac{15L^2\beta}{\delta_\epsilon}$, $D_2 := 6\max\{1, L\}\epsilon_w + \frac{15L^2\beta}{\delta_\epsilon}C_u$, $D_3 := 3C_2 + \frac{15L^2\beta}{\delta_\epsilon}\frac{3}{2(\frac{1}{\beta}-L)}C\epsilon_w$, $C_u := 2\Gamma(\epsilon_w+1) + \frac{\frac{1}{\beta}-L}{2}(\frac{1}{\tau^2}-1)\epsilon_w + 6\max\{1, L\}\epsilon_w$ *and* (X^0, Y^0, W^0, Z^0) *are defined as in Theorem 4.2.*

Remark 4.5. This theorem indicates that the communication complexity of Algorithm 1 is $O(\kappa^2\epsilon^{-2})$. When the inner solver is chosen as SAGA, Theorem 4.4 together with Proposition 3.1 shows that the sample complexity of Algorithm 1 $O(\kappa^2\log(\kappa)n^{-1}\epsilon^2)$.

4.2. Sequential Convergence of Algorithm 1

In this section, we are devoted to analyze the convergence properties of the sequence generated by Algorithm 1 with (10). We make the following assumption.

Assumption 4.6. Suppose for all t, (10) is deterministic and all clients attend the training at each round.

Table 3. Maximum AUC values obtained by each algorithm after 1000 communication rounds.

Algorithm	a9a	covtype	gisette	ijcnn1	phishing	w8a
CODASCA (Yuan et al., 2021b)	0.8920	0.7967	0.9982	0.9264	0.9758	0.9007
Fed-Norm-SGDA (Sharma et al., 2023)	0.8961	0.7645	0.9961	0.9273	0.9786	0.8959
FedSGDA (Wu et al., 2023)	0.8963	0.7645	0.9962	0.9272	0.9786	0.8958
FEDNEST (Tarzanagh et al., 2022)	0.8963	0.8132	0.9989	0.9037	0.9714	0.9075
FFMDR (This Work)	**0.8998**	**0.8208**	**0.9994**	**0.9288**	**0.9797**	**0.9076**

Theorem 4.7. *Consider (1). Let $\{(X^t, W^t, Z^t, Y^t)\}$ as in Theorem 4.2. Suppose Assumption 4.6 holds. Suppose F and g are bounded from below and g is level-bounded. Suppose in addition that H is a KL function with exponent $\alpha \in [0,1)$. Then $\{(X^t, W^t, Z^t, Y^t)\}$ is convergent. In addition, denoting $(X^*, W^*, Z^*, Y^*) := \lim_t (X^t, W^t, Z^t, Y^t)$, it holds that*

(i) If $\alpha = 0$, then $\{(X^t, W^t, Z^t)\}$ converges finitely.

(ii) If $\alpha \in (0, \frac{1}{2}]$, then there exist $b > 0$, $t_1 \in \mathbb{N}$ and $\rho_1 \in (0,1)$ such that $\max\{\|W^t - W^*\|, \|X^t - X^*\|, \|Z^t - Z^*\|, \|Y^t - Y^*\|\} \le b\rho_1^t$ for $t \ge t_1$.

(iii) If $\alpha \in (\frac{1}{2}, 1)$, then there exist $t_2 \in \mathbb{N}$ and $c > 0$ such that $\max\{\|W^t - W^*\|, \|X^t - Y^*\|, \|Z^t - Z^*\|, \|Y^t - Y^*\|\} \le ct^{-\frac{1}{4\alpha - 2}}$ for $t \ge t_2$.

Finally, we elaborate on how to verify the KL assumption in Theorem 4.7. Note that the KL assumption is on H in (11). Since the F in H is a max function, H can be viewed as a max function, i.e.,

$$H(X, W, Z, Y, W', Y') := \max_{Y''} U(X, W, Z, Y, W', Y', Y''),$$

where $Y'' := (y_1'', \ldots, y_n'')$ and

$$U(X, W, Z, Y, W', Y', W') := \frac{1}{n}\sum_{i=1}^{n} f_i(w_i, y_i'') + \tilde{g}(Z) + \frac{1}{2\beta}\left(\|X - W\|^2 - \|X - Z\|^2\right) + \frac{1}{\beta}\|W - Z\|^2 + \frac{\delta}{\beta}\|W - W'\|^2 + \frac{1}{12L^2}\sum_i p_i \|(y_i, w_i) - (y_i', w_i')\|^2.$$

Therefore, it is hard to directly verify the KL property of H. However, it is easier to verify the KL property of U. For example, when U is a proper closed semi-algebraic function that has a closed domain and is continuous on their domains, U is a KL function (Attouch et al., 2010). Given this fact, it is natural to ask whether we can deduce the KL property of a max function like H from the KL property of the objective in the maximization like U. The following property provides a positive answer.

Proposition 4.8. *Let $F(x, y) : \mathbb{R}^m \times \mathbb{R}^n \to [-\infty, \infty]$ be a proper closed function that is continuous on its domain. Suppose $F(\cdot, y)$ is weakly convex for any y with modulus $\rho > 0$. Suppose for any y, $F(\cdot, y)$ has the KL property at x with exponent $\alpha \in [0,1)$ with constants $\epsilon(y)$, $c(y)$ and $a(y)$. Suppose $\epsilon(y)$, $c(y)$ and $a(y)$ are continuous in y. Let $G(x) = \max_y F(x, y)$. Let $x \in \operatorname{dom} \partial G$. Then G has KL property at x with exponent α.*

Remark 4.9. We provide an example where the assumptions in Proposition 4.8 is satisfied. For simplicity, we consider the following robust classification problem ((Sinha et al., 2017)):

$$\min_{\theta}\max_{\delta} F(\theta, \delta) := \underbrace{\log(1 + \exp(-y\theta(x + \delta)))}_{\ell(\theta, \delta)} - c|\delta|^2 + \lambda|\theta|, \tag{13}$$

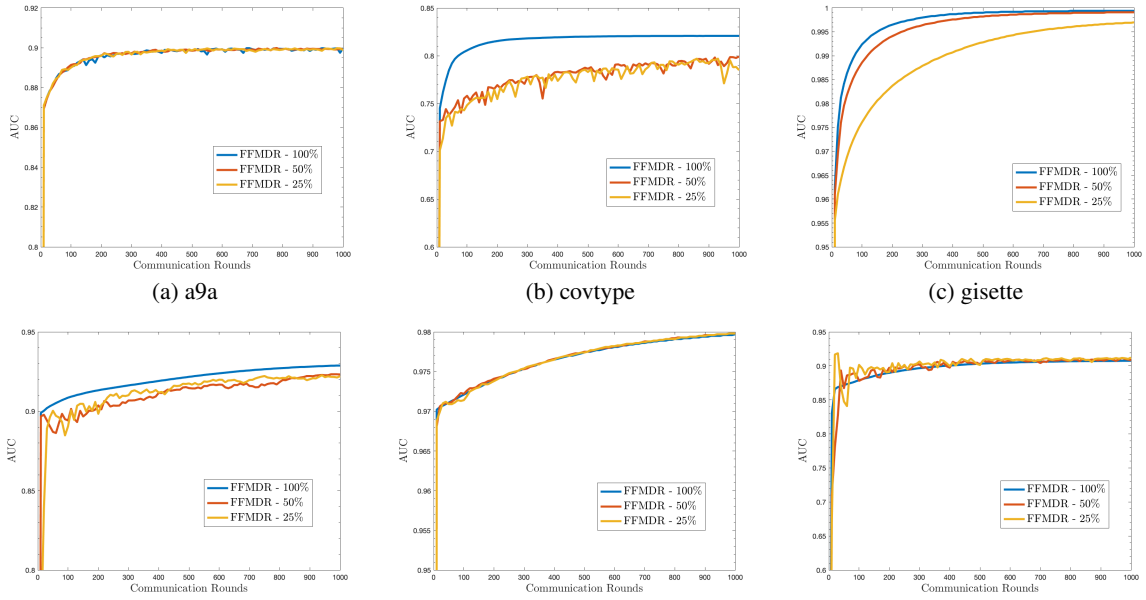
where $(x, y) \in \mathbb{R} \times \{-1, 1\}$ is a data point, $\theta \in \mathbb{R}$ is the weight, δ is a perturbation and $c, \lambda > 0$ are scalers. Now fix any δ. For any $\bar{\theta}$, there exists $\epsilon(\delta)$ continuous w.r.t. δ such that $F(\cdot, \delta)$ satisfies the KL property at $\bar{\theta}$ with exponent $\frac{1}{2}$ and constants $\epsilon(\delta)$, $c = 1$ and $a = 1$. More details can be found in the supplementary material.

5. Experiments

Learning task In this section, we apply our method to maximizing the Area under the ROC curve (AUC) problem (Natole et al., 2018) in the federated learning settings. This problem is formed as the following minimax problem:

$$\min_{\mathbf{w} \in \mathbb{R}^l, a \in \mathbb{R}, b \in \mathbb{R}} \max_{\alpha \in \mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}\sum_{\eta \in \mathcal{D}_i} [f_i(\mathbf{w}, a, b, \alpha; \eta)] + g(\mathbf{w}), \tag{14}$$

where, $\eta = (x, y)$ is a datapoint, n is the number of clients, $f_i(\mathbf{w}, a, b, \alpha; \eta) = p(1-p) + (1-p)(\mathbf{w}^T x - a)^2 \mathbb{I}_{[y=1]} + p(\mathbf{w}^T x - b)^2 \mathbb{I}_{[y=-1]} + 2(1+\alpha)\mathbf{w}^T x(p\mathbb{I}_{[y=-1]} - (1-p)\mathbb{I}_{[y=1]}) - p(1-p)\alpha^2$, $\mathbb{I}_A(x) = 1$ when $x \in A$ for any set A and $\mathbb{I}_A(x) = 0$ otherwise. Here p is the probability of $Pr(y = 1)$. The goal of AUC maximization tasks is to pursue a high AUC score for binary classification, which is defined by $Pr(\mathbf{w}^T x > \mathbf{w}^T x' | y = 1, y' = -1)$. This F is an equivalent formulation and it is strongly concave in α. The $g(\mathbf{w})$ in (14) is a convex regularization. In our



(a) a9a (b) covtype (c) gisette

(d) ijcnn1 (e) phishing (f) w8a

Figure 3. AUC values w.r.t. communication rounds on test dataset: a9a, covtype, gisette, ijcnn1, phishing and w8a.

experiments, we consider $g(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$ where $\lambda = 0.001$ is fixed during the experiment. In our experiment, the total number of clients is set to 20.

Dataset We perform our experiments on six real-world dataset for binary classification: a9a, covtype, gisette, ijcnn1, phishing and w8a, all of which can be downloaded from the LIBSVM repository (Chang & Lin, 2011). The training data is distributed to all clients heterogeneously where each client only owns the data from one class.

Compared methods We compare our stochastic method with CODASCA in (Yuan et al., 2021b), Fed-Norm-SGDA in (Sharma et al., 2023) and FedSGDA in (Wu et al., 2023). All these baselines are applicable to the AUC maximization problem in stochastic manner with a non-smooth regularization. CODASCA is an algorithm to solve federated AUC maximization problem for heterogeneous data. Other compared methods are general minimax algorithms which have been introduced in previous sections. In our experiments, the local solver of FFMDR is chosen as SGDA.

Parameters For FFMDR, we select the best value of $\frac{1}{2\beta}$ from $\{1, 0.1, 0.01, 0.001\}$, ϵ_w from $\{0.95, 0.75, 0.5, 0.25, 0.05\}$. For all methods, the stepsize is selected from $\{0.1, 0.01, 0.001, 0.0001, 0.00001\}$ so that it achieves the best experimental result. The batchsize is fixed to be 40. The local epoch is fixed to be 5.

Results In Figure 2, we plot the AUC values of each algorithm with respect to the number of communication rounds. In Table 3, we report detailed AUC scores obtained by each algorithm after 1000 communication rounds. From these experimental results we can see our FFMDR algorithm achieves the best AUC scores on all of the six datasets. Also, our method converges faster than the compared methods in most cases. These experimental results verify the performance of our proposed method to solve federated minimax problems with data heterogeneity.

Additionally, we also test our FFMDR method in the case where only a fraction of clients can participate in the training process in each communication round. The result is shown in Figure 3, where the percentage of clients attending the training in each round is $100\%/50\%/25\%$. Figure 3 indicates that in most cases, our FFMDR method with partial attendance of the clients also works as well as FFMDR with full attendance of clients.

6. Conclusion

In this paper, we address the heterogeneity and client drift problems in federated learning for solving nonconvex, strongly concave, nonsmooth minimax problems. We propose a fast federated minimax Douglas-Rachford splitting method. We analyze the proposed method from two perspectives: sample complexity and sequential convergence. We demonstrate that our method exhibits smaller sample complexity compared to existing federated learning methods. Additionally, the proposed method provides global sequential convergence guarantees. Our analysis does not rely on the assumption that heterogeneity between clients is bounded. Empirically, we apply our method to the AUC maximization problem and find that it outperforms existing federated minimax methods in scenarios with high data heterogeneity.

Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here

References

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality. *Math. Oper. Res.*, 35(2):438–457, 2010.

Bai, T., Luo, J., Zhao, J., Wen, B., and Wang, Q. Recent advances in adversarial training for adversarial robustness. In Zhou, Z. (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 4312–4321. ijcai.org, 2021.

Benjamin, G., Haihao, L., Pratik, W., and Vahab, M. The landscape of the proximal point method for nonconvex–nonconcave minimax optimization. *To appear in Mathematical Programming*, 2022.

Chang, C. and Lin, C. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3): 27:1–27:27, 2011.

Chen, Z., Zhou, Y., Xu, T., and Liang, Y. Proximal gradient descent-ascent: Variable convergence under kł geometry. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7*, 2021.

Deng, Y., Kamani, M. M., and Mahdavi, M. Distributionally robust federated averaging. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Fallah, A., Ozdaglar, A. E., and Pattathil, S. An optimal multistage stochastic gradient method for minimax problems. In *59th IEEE Conference on Decision and Control, CDC 2020, Jeju Island, South Korea, December 14-18, 2020*, pp. 3573–3579. IEEE, 2020.

Gao, R. and Kleywegt, A. J. Distributionally robust stochastic optimization with wasserstein distance. *Math. Oper. Res.*, 48(2):603–655, 2023.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2672–2680, 2014.

Huang, F., Wu, X., and Huang, H. Efficient mirror descent ascent methods for nonsmooth minimax problems. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 10431–10443, 2021.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K. A., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2):1–210, 2021.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. SCAFFOLD: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 2020. URL http://proceedings.mlr.press/v119/karimireddy20a.html.

Kovalev, D. and Gasnikov, A. V. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. *CoRR*, abs/2205.05653, 2022.

Lei, Y. and Ying, Y. Stochastic proximal AUC maximization. *J. Mach. Learn. Res.*, 22:61:1–61:45, 2021.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. In Dhillon, I. S., Papailiopoulos, D. S., and Sze, V. (eds.), *Proceedings of Machine Learning and Systems 2020, MLSys 2020, Austin, TX, USA, March 2-4*, 2020.

Li, X., Song, Z., and Yang, J. Federated adversarial learning: A framework with convergence analysis. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19932–19959. PMLR, 2023.

Lin, T., Jin, C., and Jordan, M. I. Near-optimal algorithms for minimax optimization. In Abernethy, J. D. and Agarwal, S. (eds.), *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2738–2779. PMLR, 2020.

Lions, P. L. and Mercier, B. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979. ISSN 00361429. URL http://www.jstor.org/stable/2156649.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. S. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3381–3390, 2018.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April, Fort Lauderdale, FL, USA*, 2017.

Natole, M., Ying, Y., and Lyu, S. Stochastic proximal algorithms for AUC maximization. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3707–3716, 2018.

Palaniappan, B. and Bach, F. R. Stochastic variance reduction methods for saddle-point problems. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 1408–1416, 2016.

Pathak, R. and Wainwright, M. J. FedSplit: an algorithmic framework for fast federated optimization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12*, 2020.

Peng, X., Huang, Z., Zhu, Y., and Saenko, K. Federated adversarial domain adaptation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7*, 2021.

Rockafellar, R. T. and Wets, R. J. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1998.

Sharma, P., Panda, R., Joshi, G., and Varshney, P. K. Federated minimax optimization: Improved convergence analyses and algorithms. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19683–19730. PMLR, 2022.

Sharma, P., Panda, R., and Joshi, G. Federated minimax optimization with client heterogeneity. *CoRR*, abs/2302.04249, 2023.

Sinha, A., Namkoong, H., and Duchi, J. C. Certifiable distributional robustness with principled adversarial training. *CoRR*, abs/1710.10571, 2017. URL http://arxiv.org/abs/1710.10571.

Tarzanagh, D. A., Li, M., Thrampoulidis, C., and Oymak, S. Fednest: Federated bilevel, minimax, and compositional optimization. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21146–21179. PMLR, 2022.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I. J., Boneh, D., and McDaniel, P. D. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. FedDR - randomized douglas-rachford splitting algorithms for nonconvex federated composite optimization. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14*, 2021.

Wu, X., Sun, J., Hu, Z., Zhang, A., and Huang, H. Solving a class of non-convex minimax optimization in federated learning. *CoRR*, abs/2310.03613, 2023.

Yuan, H., Zaheer, M., and Reddi, S. J. Federated composite optimization. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July*, 2021a.

Yuan, Z., Guo, Z., Xu, Y., Ying, Y., and Yang, T. Federated deep auc maximization for hetergeneous data with a constant communication complexity. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12219–12229. PMLR, 18–24 Jul 2021b.