# Zeroth-Order Methods for Nondifferentiable Nonconvex and Hierarchical Federated Optimization

Yuyang Qiu

Industrial & Systems Engineering
Rutgers University

Joint work with Dr. Uday V. Shanbhag & Dr. Farzad Yousefian

NeurIPS 2023



Funding acknowledgement:

- DOE (#DE-SC0023303)
- ONR (#N00014-22-1-2757 and #N00014-22-1-2589)

# Federated learning (FL)

Federated learning is a framework for learning predictive models from datasets that are

- distributed
- privacy-sensitive
- heterogeneous
- massive

This is accomplished through the use of efficiently devised periodic communications between a central server and clients.

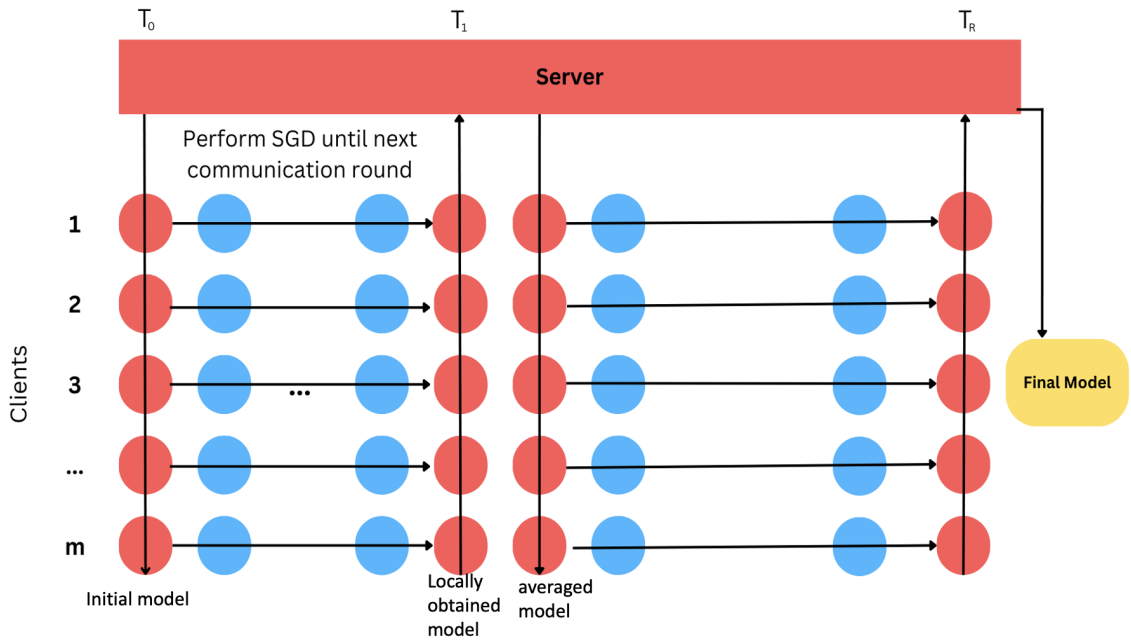


Figure: Simple example of Federated learning

# Motivation

- Federated averaging (FedAvg) [McMahan et al., 2017] is proposed to train deep networks, which could be nondifferentiable and nonconvex (e.g., ReLU).
  - [Stich, 2019], [Stich & Karimireddy, 2019], [Zhou & Cong, 2018], [Wang & Joshi, 2021], [Li et al., 2020], [Khaled et al., 2020]
- Hyperparameter optimization in ML and FL.
  - Bilevel models where the lower-level is a parameterized training model while the upper-level requires selecting the best configuration for the unknown hyperparameters. [Ghadimi & Wang, 2018], [Ji et al., 2021], [Tibshirani et al., 2005]
- Minimax FL problems. Recently, FL was extended to distributed minimax problems, but relatively little exists in nonsmooth nonconvex-strongly concave settings.
  - [Mohri et al., 2019], [Deng et al., 2020], [Sharma et al., 2022], [Tarzanagh et al., 2022]
- Research on such problems has relied on strong assumptions, including differentiability and L-smoothness of the local loss function or the implicit function. Such assumptions may fail to hold in practical settings.

Goal: A unified FL framework accommodating nondifferentiable and nonconvex settings as well as allowing for bilevel or minimax interactions.

# Nondifferentiable nonconvex bilevel FL

Consider a bilevel FL problem of the form

$$\min_{x \in X \triangleq \bigcap_{i=1}^m X_i} \left\{ f(x) \triangleq \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{\xi_i \in \mathcal{D}_i} [\tilde{f}_i(x, y(x), \xi_i)] \right\}, \qquad (\mathbf{FL}_{bl})$$

where

$$y(x) \in \arg\min_{y \in \mathbb{R}^{\tilde{n}}} \quad \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{\zeta_i \in \tilde{\mathcal{D}}_i} [\tilde{h}_i(x, y, \zeta_i)].$$

$f$ is possibly nondifferentiable and nonconvex, $y(\bullet) : \mathbb{R}^n \to \mathbb{R}^{\tilde{n}}$ is a single-valued map returning the unique solution to the lower-level problem at $x$.

Even if the upper-level objective function is smooth and convex in (x,y), the implicit function $f(\bullet, y(\bullet))$ is often nondifferentiable nonconvex in $x$!

# Randomized smoothing

- For smoothening of the loss function $f$, we employ a *randomized smoothing* approach where the smoothing parameter is maintained as sufficiently small. This framework is rooted in the seminal work [Steklov, 1907], leading to progress in both convex and nonconvex regimes.
  - [Lakshmanan & Farias, 2008], [Yousefian et al., 2012], [Duchi et al., 2012]
  - [Nesterov & Spokoiny, 2017]

- We consider a smoothing of $f$, given by $f_\eta$ defined as

$$f^\eta(x) \triangleq \mathbb{E}_{u\in\mathbb{B}}[f(x+\eta u)],$$

  - where $u$ is a random vector in the unit ball $\mathbb{B}$, defined as $\mathbb{B} \triangleq \{u \in \mathbb{R}^n \mid \|u\| \leq 1\}$.
  - Further, $\mathbb{S}$ denotes the surface of the ball $\mathbb{B}$, i.e., $\mathbb{S} \triangleq \{v \in \mathbb{R}^n \mid \|v\| = 1\}$ and $\eta\mathbb{B}$ and $\eta\mathbb{S}$ denote ball with radius $\eta$ and its surface, respectively.
  - The gradient of the smoothed function is

$$\nabla f^\eta(x) = \frac{n}{\eta}\mathbb{E}_{v\in\eta\mathbb{S}}\left[(f(x+v)-f(x))\frac{v}{\|v\|}\right].$$

    [Cui et al., 2022]

# Construct a zeroth-order gradient

$$\mathbf{f}(x) \triangleq f(x) + \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}_{X_i}(x)$$

$$\Downarrow \text{(Randomized smoothing + Moreau smoothing)}$$

$$\mathbf{f}^{\eta}(x) \triangleq f^{\eta}(x) + \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}^{\eta}_{X_i}(x)$$

$$\mathbf{f}^{\eta}(x) \triangleq \frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{E}_{u_i\in\mathbb{B}}[\mathbb{E}_{\xi_i}[\tilde{f}_i(x+\eta u_i, y(x+\eta u_i), \xi_i)]]\right) + \frac{1}{2}\mathrm{dist}^2(x, X_i).$$

$$\Downarrow \text{(Implicit gradient)}$$

$$\nabla\mathbf{f}^{\eta}(x) \triangleq \frac{1}{m}\sum_{i=1}^{m}\left(\mathbb{E}_{v_i}[\mathbb{E}_{\xi_i}[\frac{n}{\eta}\frac{\tilde{f}_i(x+v_i, y(x+v_i), \xi_i) - \tilde{f}_i(x, y(x), \xi_i)}{\|v_i\|} v_i]] + \frac{1}{\eta}(x - \mathcal{P}_{X_i}(x))\right).$$

$$\Downarrow \text{(Inexact evaluation of stoch. zeroth-order local gradient)}$$

$$\tilde{\nabla}\mathbf{f}_i^{\eta}(x) \triangleq \frac{n}{\eta}\frac{\tilde{f}_i(x+v_i, y_\varepsilon(x+v_i), \xi_i) - \tilde{f}_i(x, y_\varepsilon(x), \xi_i)}{\|v_i\|} v_i + \frac{1}{\eta}(x - \mathcal{P}_{X_i}(x)).$$

# Proposed algorithm for $\mathrm{FL}_{bl}$

**Algorithm** Randomized Implicit Zeroth-Order Federated Averaging ($\mathrm{FedRZO}_{bl}$)

**input:** Server chooses a random $\hat{x}_0 \in X$, stepsize $\gamma$, smoothing parameter $\eta$, synchronization indices $T_0 := 0$ and $T_r \geq 1$, where $r \geq 1$ is the upper-level communication round index

**for** $r = 0, 1, \ldots$ **do**

  Server generates a random replicate $v_{T_r} \in \eta\mathbb{S}$

  Server calls a lower-level federated algorithm to receive $y_{\varepsilon_r}(\hat{x}_r + v_{T_r})$ and $y_{\varepsilon_r}(\hat{x}_r)$, denoting the inexact evaluations of $y(\hat{x}_r + v_{T_r})$ and $y(\hat{x}_r)$, respectively.

  Server broadcasts $\hat{x}_r$, $\hat{x}_r + v_{T_r}$, $y_{\varepsilon_r}(\hat{x}_r)$, and $y_{\varepsilon_r}(\hat{x}_r + v_{T_r})$ to all clients; $x_{i,T_r} := \hat{x}_r$, $\forall i$

  **for** $k = T_r, \ldots, T_{r+1} - 1$ **in parallel by clients do**

    Client $i$ generates the random replicates $\xi_{i,k} \in \mathcal{D}_i$

    $$g_{i,k}^{\eta,\varepsilon_r} := \frac{n}{\eta^2}\left(\tilde{f}_i(x_{i,k} + v_{T_r}, y_{\varepsilon_r}(\hat{x}_r + v_{T_r}), \xi_{i,k}) - \tilde{f}_i(x_{i,k}, y_{\varepsilon_r}(\hat{x}_r), \xi_{i,k})\right) v_{T_r}$$

    (delayed inexact computation of $y(x)$ can reduce communications significantly!)

    Client $i$ does a local update as $x_{i,k+1} := x_{i,k} - \gamma\left(g_{i,k}^{\eta,\varepsilon_r} + \frac{1}{\eta}\left(x_{i,k} - \mathcal{P}_{X_i}(x_{i,k})\right)\right)$

  **end for**

  Server receives $x_{i,T_{r+1}}$ from all clients and aggregates, i.e., $\hat{x}_{r+1} := \frac{1}{m}\sum_{i=1}^{m} x_{i,T_{r+1}}$

**end for**

# Assumptions

Consider problem ($\mathbf{FL}_{bl}$). Let the following assumptions hold.

(i) For all $i \in [m]$, $\tilde{f}_i(\bullet, y, \xi_i)$ is $L^f_{0,x}(\xi_i)$-Lipschitz for any $y$ and $\tilde{f}_i(x, \bullet, \xi_i)$ is $L^f_{0,y}(\xi_i)$-Lipschitz for any $x$, where $L^f_{0,x} \triangleq \max_{i=1,\ldots,m} \sqrt{\mathbb{E}[(L^f_{0,x}(\xi_i))^2]} < \infty$ and $L^f_{0,y} \triangleq \max_{i=1,\ldots,m} \sqrt{\mathbb{E}[(L^f_{0,y}(\xi_i))^2]} < \infty$.

(ii) [Lower-level] For all $i \in [m]$, for any $x$, $h_i(x, \bullet)$ is $L^h_{1,y}$-smooth and $\mu_h$-strongly convex. Further, for any $y$, the map $\nabla_y h_i(\bullet, y)$ is Lipschitz continuous with parameter $L^{\nabla h}_{0,x}$.

(iii) The sets $X_i$, for $i \in [m]$, a *bounded set-dissimilarity* condition holds for all $x \in \mathbb{R}^n$ and some scalars $B_1$ and $B_2$.

$$\frac{1}{m}\sum_{i=1}^{m} \text{dist}^2(x, X_i) \leq B_1^2 + B_2^2 \ \text{dist}^2(x, \frac{1}{m}\sum_{i=1}^{m} \mathcal{P}_{X_i}(x)).$$

*It is when the bounded gradient dissimilarity assumption [Karimireddy et al., 2019] is written for the local functions.

## Theorem (FedRZO$_{bl}$ when using an arbitrary inexact FL method for lower-level)

*Let Assumption 2 hold. Let $k^*$ be chosen uniformly at random from $0,\dots,K := T_R - 1$ and let*

$\gamma \le \min\left\{ \frac{\max\{2,\sqrt{0.1\Theta_3},4B_2\sqrt{3\Theta_2},4B_2\sqrt{3\Theta_3}\}^{-1}}{4H}, \frac{\eta}{24(L_0^{imp}n+1)} \right\}$. *Let $\varepsilon_r$ denote the inexactness in obtaining the lower-level solution, i.e.,* $\mathbb{E}\left[\|y_{\varepsilon_r}(x)-y(x)\|^2\right] \le \varepsilon_r$ *for* $x \in \cup_{r=0}^{R}\{\hat{x}_r, \hat{x}_r + v_{T_r}\}$.

*(i)* **(Error bound)** *We have*

$$\mathbb{E}\left[\|\nabla \mathbf{f}^{\eta}(\bar{x}_{k^*})\|^2\right] \le 8(\gamma K)^{-1}(\mathbb{E}\left[\mathbf{f}^{\eta}(x_0)\right] - \mathbf{f}^{\eta,*}) + \frac{8\gamma\Theta_1}{m} + 8H^2\gamma^2\max\{\Theta_2,\Theta_3\}\Theta_5$$

$$+8\left(H^2\gamma^2\max\{\Theta_2,\Theta_3\}\Theta_4+\Theta_3\right)H\frac{\sum_{r=0}^{R-1}\varepsilon_r}{K}, \text{ where}$$

$$\Theta_1 := \frac{9(L_0^{imp}n+1)n^2}{2\eta}\left(\frac{2\nu_f^2}{\eta^2}+(L_0^{imp})^2\right),\ \Theta_2 := \frac{5(L_0^{imp}n+1)^2}{8\eta^2}, \Theta_3 := \left(\frac{L_{0,x}^{\nabla h}}{\mu_h}\right)^2\frac{60n^2}{\eta^2}\left(\frac{2\nu_f^2}{\eta^2}+(L_{0,y}^{f})^2\right),$$

$$\Theta_4 := \frac{144n^2}{\eta^2}\left(\frac{2\nu_f^2}{\eta^2}+(L_{0,y}^{f})^2\right), \Theta_5 := \left(36n^2\left(\frac{2\nu_f^2}{\eta^2}+(L_0^{imp})^2\right)+\frac{24B_1^2}{\eta^2}+48B_2^2(L_0^{imp})^2n^2\right).$$

*(ii)* **(Iteration complexity)** *Let* $\gamma := \sqrt{\frac{m}{K}}$ *and* $H := \left\lceil \sqrt[4]{\frac{K}{m^3}} \right\rceil$. *Let $K_\varepsilon$ denotes the number of iterations such that* $\mathbb{E}\left[\|\nabla \mathbf{f}^{\eta}(\bar{x}_{k^*})\|^2\right] \le \varepsilon$. *Suppose lower-level FL method has a linear speedup of the order* $\varepsilon_r := \tilde{\mathcal{O}}(\frac{1}{mr})$, *where the lower-level FL method is terminated after $r$ (communication round of upper-level) iterations. Then, the iteration complexity of FedRZO$_{bl}$ is*

$$K_\varepsilon := \tilde{\mathcal{O}}\left(\frac{\Theta_1^2+\max\{\Theta_2,\Theta_3\}^2\Theta_5^2}{m\varepsilon^2} + \frac{(\max\{\Theta_2,\Theta_3\}\Theta_4)^{0.8}}{m^{1.8}\varepsilon^{0.8}} + \frac{\Theta_3^{4/3}}{m^{7/3}\varepsilon^{4/3}}\right).$$

*(iii)* **(Communication complexity)** *Suppose* $K_\varepsilon \ge m^3$. *Then, the number of communication rounds in FedRZO$_{bl}$ (upper-level only) to achieve the accuracy level in (ii) is* $R := \mathcal{O}\left((mK_\varepsilon)^{3/4}\right)$.

# Total communication complexity

Table: Communication complexity for nondifferentiable nonconvex, bilevel, and minimax FL.

| heterogeneous upper level | lower level (standard FL schemes are employed) | | total (**this work**) |
|---|---|---|---|
| $\mathcal{O}\left((mK)^{\frac{3}{4}}\right)$ (Prop. 1, Thm. 1) (**this work**) | Local SGD (i.i.d.) [Khaled et al., 20] | $\mathcal{O}(m)$ | $\mathcal{O}\left(m^{\frac{7}{4}} K^{\frac{3}{4}}\right)$ |
| | FedAc (i.i.d.) [Yuan & Ma, 20] | $\mathcal{O}\left(m^{\frac{1}{3}}\right)$ | $\mathcal{O}\left(m^{\frac{13}{12}} K^{\frac{3}{4}}\right)$ |
| | LFD (non-i.i.d.) [Haddadpour & Mahdavi, 19] | $\mathcal{O}\left(m^{\frac{1}{3}} r^{\frac{1}{3}}\right)$ | $\mathcal{O}\left(m^{\frac{4}{3}} K\right)$ |

In all cases, we assume heterogeneous data in the upper level. In the lower level, depending on what conventional FL scheme is employed, we obtain the communication complexity accordingly.

# Approximate Clarke stationarity

Recent findings on nonsmooth analysis [Zhang et al., 2020] shown that for a suitable class of nonsmooth functions, computing an $\varepsilon$-stationary point, is impossible in finite time.

As a weakening of $\varepsilon$-stationarity, a notion of $(\delta, \varepsilon)$-stationarity is introduced [Zhang et al., 2020] for a vector $\bar{x}$ when $\text{dist}(0_n, \partial_\delta \mathbf{f}(\bar{x})) \leq \varepsilon$, where the set

$$\partial_\delta \mathbf{f}(x) \triangleq \text{conv}\{\zeta : \zeta \in \partial \mathbf{f}(y), \|x - y\| \leq \delta\}$$

denotes the $\delta$-Clarke generalized gradient of $\mathbf{f}$ at $x$ [Goldstein, 1977];

i.e. if $x$ is $(\delta, \varepsilon)$-stationary, then there exists a convex combination of gradients in a $\delta$-neighborhood of $x$ that has a norm of at most $\varepsilon$ [Shamir, 2021].

Realation between original problem with its smoothed counterpart: if $\nabla \mathbf{f}^\eta(x) = 0$, then $0_n \in \partial_{2\eta} \mathbf{f}(x)$. [Mayne and Polak, 1984]

# Federated training of ReLU neural network

- We implement our method on a single-layer ReLU NN.

$$\min_{x:=(Z,w)\in\mathcal{X}} \quad \frac{1}{2m}\sum_{i=1}^{m}\sum_{\ell\in\mathcal{D}_i}(v_{i,\ell}-\sum_{q=1}^{N_1} w_q\sigma(Z_{\bullet,q}U_{i,\ell}))^2+\frac{\lambda}{2}\left(\|Z\|_F^2+\|w\|^2\right),$$

- $\sigma(x) := \max\{0, x\}$.

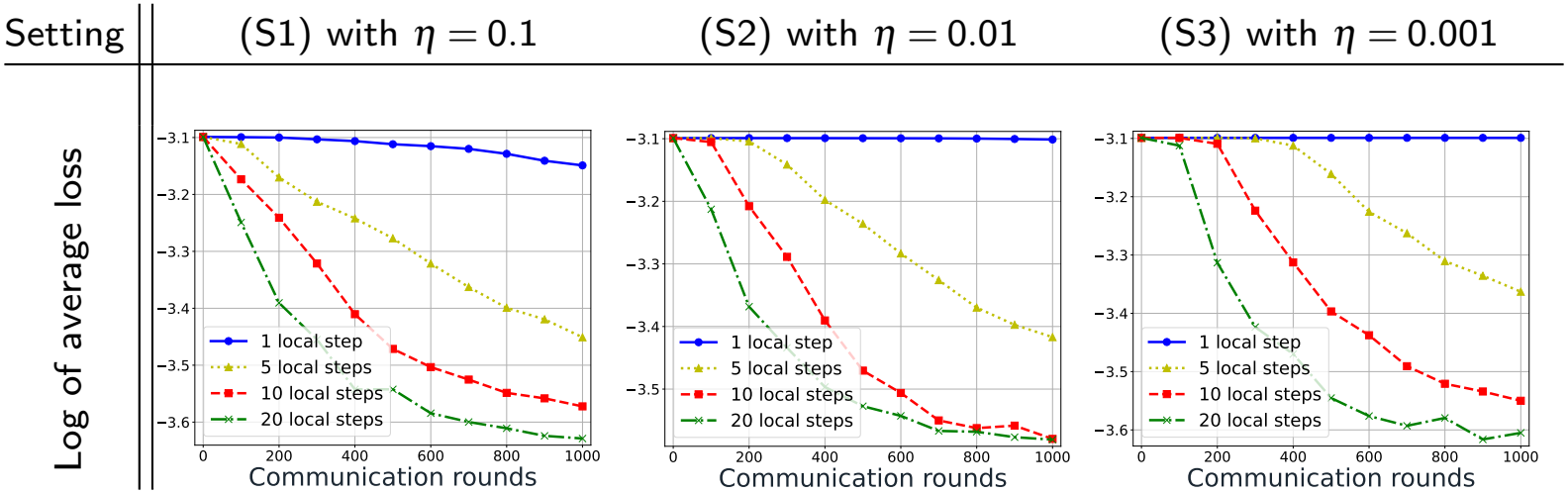


Figure: The proposed method improves with larger number of local steps H. Robustness of the scheme in terms of the original loss function, slight improvement in the empirical speed of convergence in early steps, as $\eta$ increases.

# Federated hyperparameter learning

We consider an FL hyperparameter learning problem for binary classification using logistic loss.

$$\min_{x\in X,\ y\in\mathbb{R}^n} \quad f(x,y) \triangleq \frac{1}{m}\sum_{i=1}^{m}\sum_{\ell\in\mathcal{D}_i}\log\left(1+\exp(-v_{i,\ell}U_{i,\ell}^T y)\right)$$

$$\text{subject to.} \quad y\in\arg\min_{y\in\mathbb{R}^n} \quad h(x,y)\triangleq\frac{1}{m}\sum_{i=1}^{m}\left(\sum_{\tilde{\ell}\in\tilde{D}_i}\log\left(1+\exp(-v_{i,\tilde{\ell}}U_{i,\tilde{\ell}}^T y)\right)+x_i\frac{\|y\|^2}{2}\right),$$

where $x_i$ denotes the regularization parameter (decision variable of the upper-level FL problem) for client $i$, $U_{i,\ell}\in\mathbb{R}^n/U_{i,\tilde{\ell}}\in\mathbb{R}^n$ and $v_{i,\ell}\in\{-1,1\}/v_{i,\tilde{\ell}}\in\{-1,1\}$ are the $\ell$th/$\tilde{\ell}$th input and output testing/training sample of client $i$, respectively.

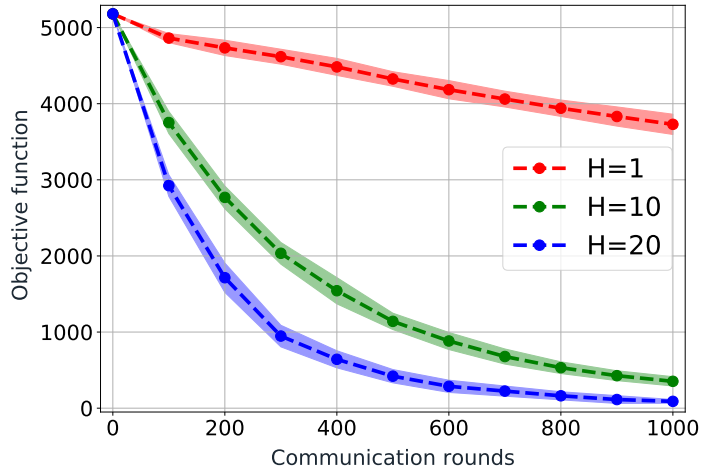


Figure: Convergence of FedRZO$_{\text{bl}}$ in hyperparameter FL for $\ell_2$ regularized logistic loss, where we plot the loss function on test data for different values of local steps with 95% CIs.

# Fair classification learning

Here, we study the performance of FedRZO$_{\text{bl}}$ in minimax FL. We consider solving an FL minimax formulation of the fair classification problem [Nouiehed et al., 2019].

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^C} \quad \frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{C}\sum_{\ell\in\mathcal{D}_{i,c}}\left(v_{i,\ell}-\sum_{q=1}^{N_1} w_q\sigma(Z_{\bullet,q}U_{i,\ell})\right)^2-\frac{\lambda}{2}\|y\|^2,$$

- where $c$ denotes the class index and $\mathcal{D}_{i,c}$ denotes the portion of local dataset associated with client $i$ that is comprised of class $c$ samples.
- This problem is nondifferentiable nonconvex-strongly concave, fitting well with the assumptions in our work in addressing minimax FL problems.

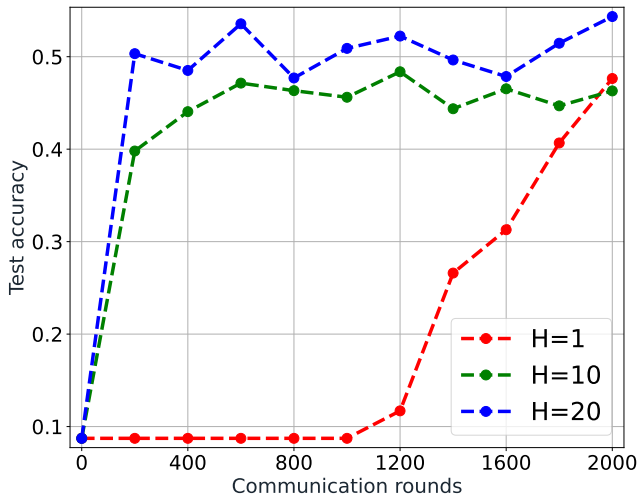


Figure: Test accuracy of FedRZO$_{bl}$ in minimax FL with different upper-level communication frequency. In terms of communication rounds, we observe that the performance of the method improves by using a larger number of local steps, motivating the need for the FL framework.

# Concluding remarks

- Federated learning has an important role in distributed ML. However, no existing FL scheme can provably address both nondifferentiability and nonconvexity.

- We resolve this gap via devising a unified, provably convergent, and communication-efficient randomized implicit zeroth-order method (with delayed inexact computation of $y(x)$) for addressing bilevel FL and minimax FL problems.

  - Our methods can contend with both nondifferentiability and nonconvexity for computing approximate Clarke-stationary points.
  - We derive iteration and communication complexity guarantees.

Thank you for your attention!

If you are interested in details, please see our full paper. We welcome any discussions!

Contact: yuyang.qiu@rutgers.edu