Federated Composite Optimization

(arXiv: 2011.08474)

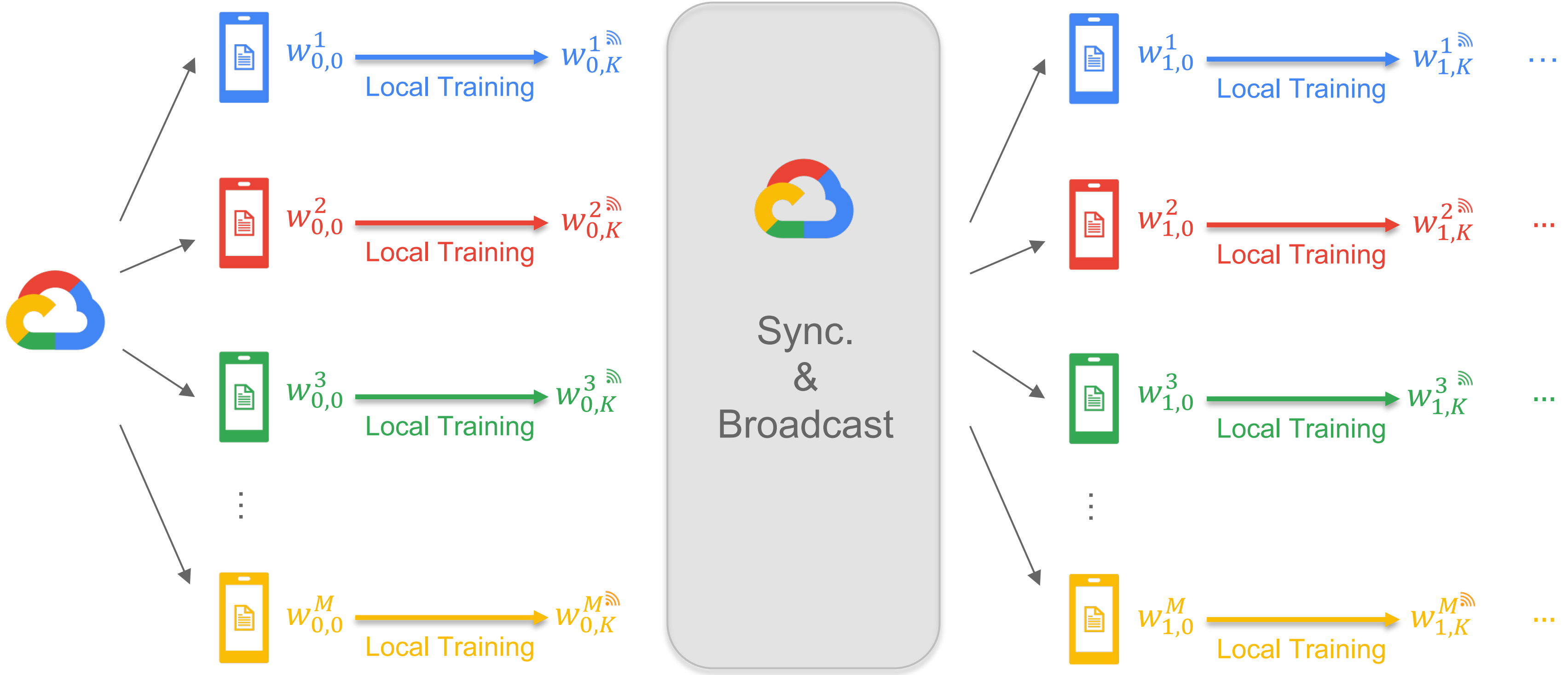
Honglin Yuan, Manzil Zaheer, Sashank Reddi

Thanks to: Zachary Charles, Zheng Xu, Andrew Hard, Ehsan Amid, Amr Ahmed, Aranyak Mehta, TensorFlow Federated team

Federated Learning

[Konečný et al., '15]

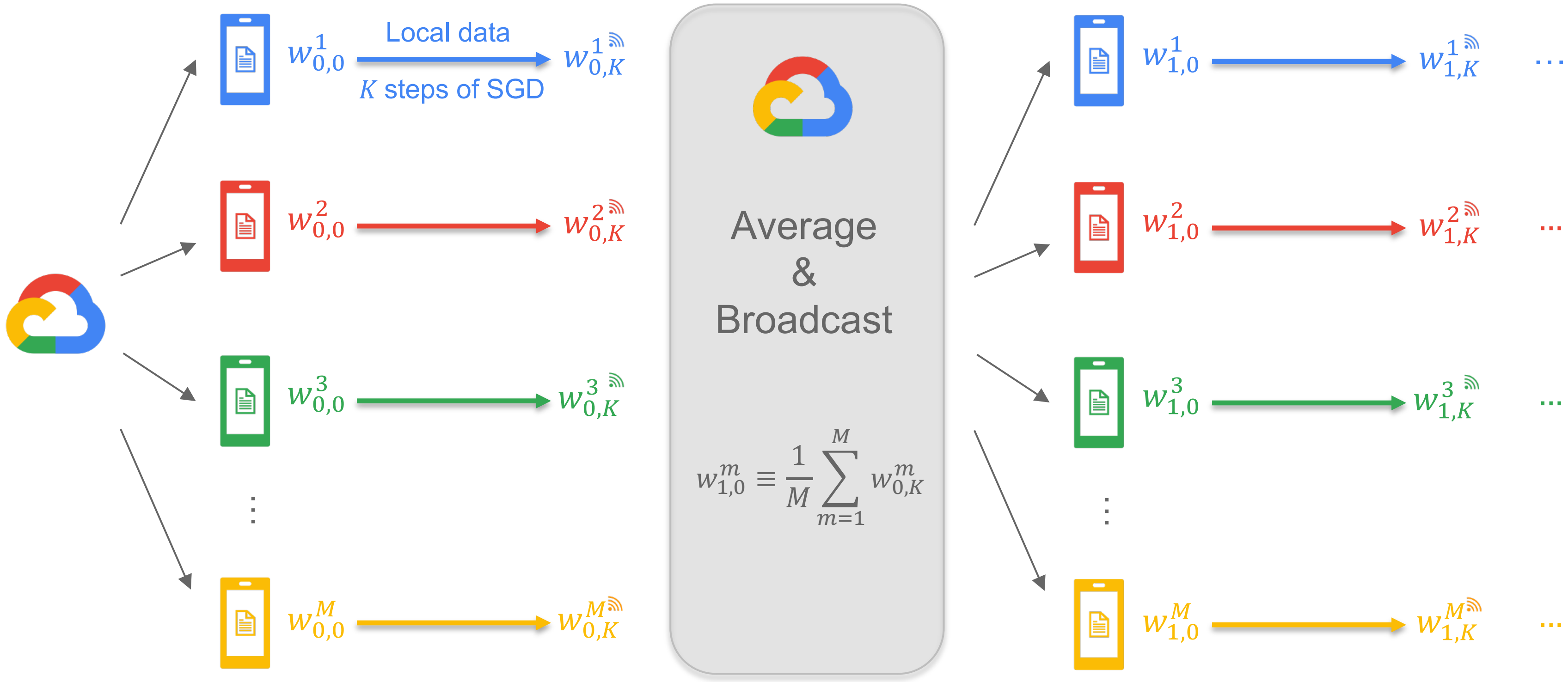
Notation: $w^{m}_{r,k}$ → m^{th} client; r^{th} round, k^{th} local iteration



FedAvg: the *de facto* standard of FL

[Konečný et al., '15]

Notation: $w^{m}_{r,k}$ → m^{th} client; r^{th} round, k^{th} local iteration



FedAvg: Generalized Formulation

[Karimireddy et al., ICML'20, Reddi et al., '20, etc]

Notation: $w^m_{r,k}$ → m^{th} client; r^{th} round, k^{th} local iteration

Algorithm 1 Federated Averaging (FedAvg)

procedure FedAvg(w_0, η_c, η_s)
  for $r = 0, \ldots, R-1$ **do**
    sample a subset of clients $\mathcal{S}_r \subseteq [M]$
    on client $m \in \mathcal{S}_r$ **in parallel do**
      client initialization $w^m_{r,0} \leftarrow w_r$
      for $k = 0, \ldots, K-1$ **do**
        $g^m_{r,k} \leftarrow \nabla f(w^m_{r,k}; \xi^m_{r,k})$
        $w^m_{r,k+1} \leftarrow w^m_{r,k} - \eta_c \cdot g^m_{r,k}$
    $\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} (w^m_{r,K} - w^m_{r,0})$
    $w_{r+1} \leftarrow w_r + \eta_s \cdot \Delta_r$

Client sampling

Client update

Average client deltas (as pseudo anti-gradient)

Server update with server learning rate η_s

Introducing Federated Composite Optimization (FCO)

- FedAvg (and other existing FL algorithms) solves **unconstrained** (smooth) problem only
  - $\min_{w \in \mathbb{R}^d} \frac{1}{M} \sum_{m=1}^{M} F_m(w)$, where $F_m(w) := \mathbb{E}_{\xi \sim \mathcal{D}_m}[f(w; \xi)]$

    [e.g., Woodworth et al., NeurIPS'20]

    → Data distribution of the m^{th} client
- We propose Federated **composite** optimization (FCO)
  - $\min_{w \in \mathbb{R}^d} \Phi(w) := \frac{1}{M} \sum_{m=1}^{M} [F_m(w) + \psi_m(w)]$, where ψ_m is convex composite functions

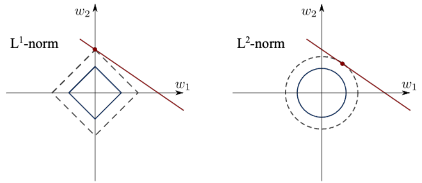
Example of ψ_m: FL with Regularization

$$\min_{w \in \mathbb{R}^d} \frac{1}{M} \sum_{m=1}^{M} [F_m(w) + \psi_m(w)]$$

- Let $\psi_m(w)$ be regularizers

- Federated Lasso for sparsity representations

$$\min_{w} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{(x,y)\sim\mathcal{D}_m} \|x^T w - y\|_2^2 + \lambda \|w\|_1$$



Potential application: cross-silo distributed biomedical data

- Federated matrix completion for recommendation system

$$\min_{W} \frac{1}{M} \sum_{m=1}^{M} F_m(W) + \lambda \|W\|_*$$

→ Matrix nuclear norm promotes low-rank

		-1		
			1	
1	1	-1	1	-1
1				-1
		-1		

1	1	-1	1	-1
1	1	-1	1	-1
1	1	-1	1	-1
1	1	-1	1	-1
1	1	-1	1	-1

Image source: Wikipedia

Example of ψ_m : FL with (Personalized) Constraints

- Let $\psi_m(w)$ be convex indicator $\begin{cases} 0 & \text{if } w \in \mathcal{C}, \\ +\infty & \text{if } w \notin \mathcal{C}. \end{cases}$

$$\min_{w \in \mathbb{R}^d} \frac{1}{M} \sum_{m=1}^{M} [F_m(w) + \psi_m(w)]$$

- Problem becomes

$$\min_{w} \frac{1}{M} \sum_{m=1}^{M} F_m(w) \quad \text{s.t.} \quad w \in \bigcap_{m=1}^{M} \mathcal{C}_m$$

→ Fulfill all constraints

- Budgeting, each customer has a budget constraint

- FL with monotonic constraints → Improve interpretability

- Inputs welcome!

Mix & Match of Setups

$$\min_{w \in \mathbb{R}^d} \frac{1}{M} \sum_{m=1}^{M} [F_m(w) + \psi_m(w)]$$

- **Homogeneous** vs **heterogeneous objective** F_m: standard "heterogeneity" in FL
  [e.g., Li et al., MLSys'20, Karimireddy et al., ICML'20, Woodworth et al., NeurIPS'20]
- **Homogeneous** vs **heterogeneous composite** ψ_m
- **Client** and/or **server** access to composite oracle ψ_m
  - Client-side oracle: better convergence? Privacy for personalized constraints?
  - Server-side oracle: computationally light
- **In this work, we focus on homogeneous $\psi_m \equiv \psi$ but allowing for heterogeneous F_m**

$$\min_{w \in \mathbb{R}^d} \Phi(w) := \frac{1}{M} \sum_{m=1}^{M} F_m(w) + \psi(w)$$

Composite 101: Proximal Gradient Descent

- Consider sequential $\min F(w) + \psi(w)$, where F smooth, ψ "simple" and convex
- Proximal Gradient Descent (PGD)

$$w_{t+1} \leftarrow \mathbf{prox}_{\eta\psi}\left(w_t - \eta \nabla F(w_t)\right)$$

$$:= \underset{w}{\operatorname{argmin}} \left\{ F(w_t) + \langle \nabla F(w_t), w - w_t \rangle + \frac{1}{2\eta}\|w - w_t\|_2^2 + \psi(w) \right\}$$

First-order Taylor expansion of F — *Smoothness estimation* — *Proximal additive*

- **prox** operator can often be computed analytically

$$\psi(w) = \chi_{\mathcal{C}}(w) := \begin{cases} 0 & \text{if } w \in \mathcal{C} \\ +\infty & \text{if } w \notin \mathcal{C} \end{cases}$$ → *Projected GD*

$$\psi(w) = \frac{1}{2}\lambda\|w\|_2^2$$ → *Weight decay (variant)*

$$\psi(w) = \lambda\|w\|_1$$ → *Soft-thresholding*

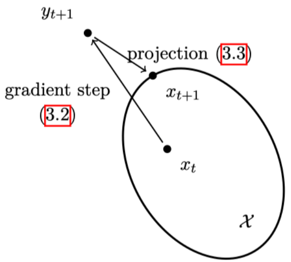


Image source: [Bubeck, 2015]

First Attempt: FedAvg + Proximal Gradient Descent

Algorithm 1 Federated Averaging (FedAvg)

procedure FedAvg(w_0, η_c, η_s)
  for $r = 0, \ldots, R-1$ **do**
    sample a subset of clients $\mathcal{S}_r \subseteq [M]$
    on client $m \in \mathcal{S}_r$ **in parallel do**
      client initialization $w_{r,0}^m \leftarrow w_r$
      for $k = 0, \ldots, K-1$ **do**
        $g_{r,k}^m \leftarrow \nabla f(w_{r,k}^m; \xi_{r,k}^m)$
        $w_{r,k+1}^m \leftarrow w_{r,k}^m - \eta_c \cdot g_{r,k}^m$
    $\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} (w_{r,K}^m - w_{r,0}^m)$
    $w_{r+1} \leftarrow w_r + \eta_s \cdot \Delta_r$

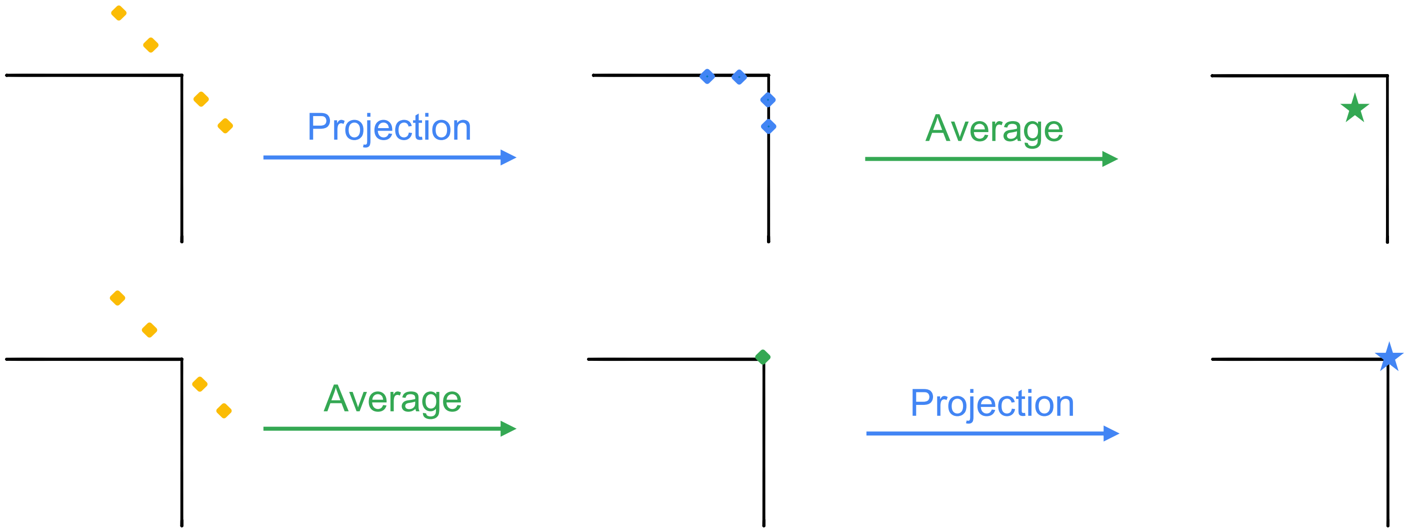
Algorithm 2 Federated PGD

procedure FedPGD(w_0, η_c, η_s)
  for $r = 0, \ldots, R-1$ **do**
    sample a subset of clients $\mathcal{S}_r \subseteq [M]$
    on client $m \in \mathcal{S}_r$ **in parallel do**
      client initialization $w_{r,0}^m \leftarrow w_r$
      for $k = 0, \ldots, K-1$ **do**
        $g_{r,k}^m \leftarrow \nabla f(w_{r,k}^m; \xi_{r,k}^m)$
        $w_{r,k+1}^m \leftarrow \mathbf{prox}_{\eta_c \psi}(w_{r,k}^m - \eta_c g_{r,k}^m)$
    $\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} (w_{r,K}^m - w_{r,0}^m)$
    $w_{r+1} \leftarrow \mathbf{prox}_{\eta_s \eta_c K \psi}(w_r + \eta_s \Delta_r)$

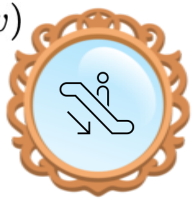
$$\min_{w \in \mathbb{R}^d} \Phi(w) := \frac{1}{M} \sum_{m=1}^{M} F_m(w) + \psi(w)$$

First Attempt: FedAvg + Proximal Gradient Descent

- Challenge: Averaging and proximal operations discord
  - Averaging and (nonlinear) proximal operators **do not commute**
  - Intuition: Averaging on post-projected points “blunt” the sharpness of projection



$$\min \ F(w) + \psi(w)$$



Composite 201: (composite) Mirror Descent

[Nemirovski et al., '83, Duchi et al., COLT'10]

$$\text{PGD}: \quad w_{t+1} = \underset{w}{\operatorname{argmin}} \left\{ F(w_t) + \langle \nabla F(w_t), w - w_t \rangle + \psi(w) + \frac{1}{2\eta} \|w - w_t\|_2^2 \right\}$$

Arbitrary distance-generating h

$$\text{MD}: \quad w_{t+1} = \underset{w}{\operatorname{argmin}} \left\{ F(w_t) + \langle \nabla F(w_t), w - w_t \rangle + \psi(w) + \frac{1}{\eta} D_h(w, w_t) \right\}$$

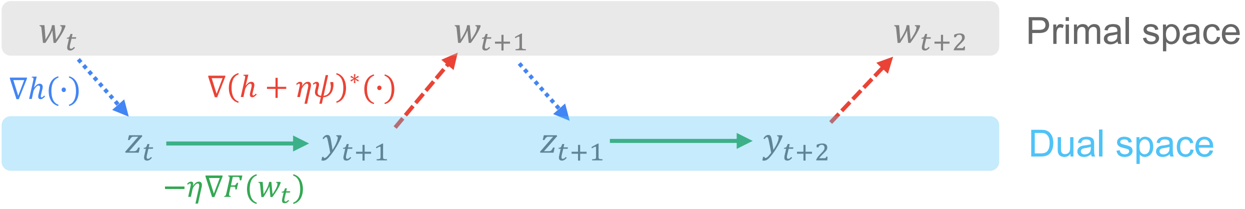
$$D_h(w, w_t) := h(w) - h(w_t) - \langle \nabla h(w_t), w - w_t \rangle$$

(reduces to PGD if $h(w) = \frac{1}{2}||w||^2$)

Primal-dual interpretation of MD

- $z_t = \nabla h(w_t)$ — Forward mirror (Primal -> Dual)
- $y_{t+1} = z_t - \eta \cdot \nabla F(w_t)$ — Gradient step (in dual space)
- $w_{t+1} = \nabla (h + \eta\psi)^* (y_{t+1})$ — Backward mirror (Dual -> Primal)

** indicates convex conjugate*



Federated Mirror Descent (FedMiD)

- Federated Mirror Descent (FedMiD) generalizes Federated PGD

Algorithm 2 Federated PGD

procedure FedPGD(w_0, η_c, η_s)
  for $r = 0, \ldots, R-1$ **do**
    sample a subset of clients $\mathcal{S}_r \subseteq [M]$
    on client $m \in \mathcal{S}_r$ **in parallel do**
      client initialization $w_{r,0}^m \leftarrow w_r$
      for $k = 0, \ldots, K-1$ **do**
        $g_{r,k}^m \leftarrow \nabla f(w_{r,k}^m; \xi_{r,k}^m)$
        $w_{r,k+1}^m \leftarrow \mathbf{prox}_{\eta_c \psi}(\nabla h(w_{r,k}^m) - \eta_c g_{r,k}^m)$
    $\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} (w_{r,K}^m - w_{r,0}^m)$
    $w_{r+1} \leftarrow \mathbf{prox}_{\eta_s \eta_c K \psi}(\nabla h(w_r) + \eta_s \Delta_r)$

Algorithm 2 Federated Mirror Descent (FedMiD)

procedure FedMiD(w_0, η_c, η_s)
  for $r = 0, \ldots, R-1$ **do**
    sample a subset of clients $\mathcal{S}_r \subseteq [M]$
    on client $m \in \mathcal{S}_r$ **in parallel do**
      client initialization $w_{r,0}^m \leftarrow w_r$
      for $k = 0, \ldots, K-1$ **do**
        $g_{r,k}^m \leftarrow \nabla f(w_{r,k}^m; \xi_{r,k}^m)$
        $w_{r,k+1}^m \leftarrow \nabla (h + \eta_c \psi)^* (\nabla h(w_{r,k}^m) - \eta_c g_{r,k}^m)$
    $\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} (w_{r,K}^m - w_{r,0}^m)$
    $w_{r+1} \leftarrow \nabla (h + \eta_s \eta_c K \psi)^* (\nabla h(w_r) + \eta_s \Delta_r)$

Composite 202: Dual Averaging

$$\min \; F(w) + \psi(w)$$

[Nesterov et al., '09, Xiao et al., '10, Flammarion et al., COLT'17]

Dual Averaging (a.k.a. Lazy Mirror Descent)

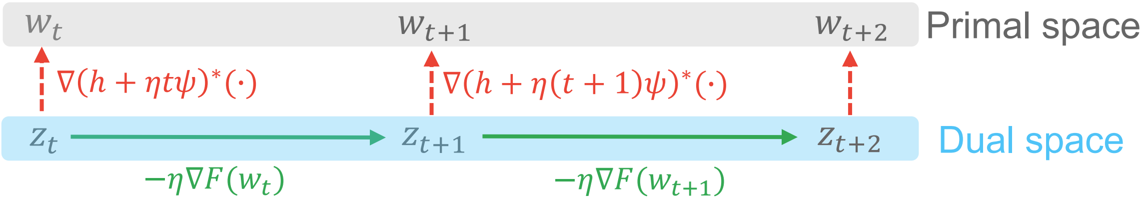
$$w_t = \nabla(h + \eta t \psi)^*(z_t)$$

$$= \arg\min_w \{\langle -z_t, w\rangle + \eta t \psi(w) + h(w)\}$$

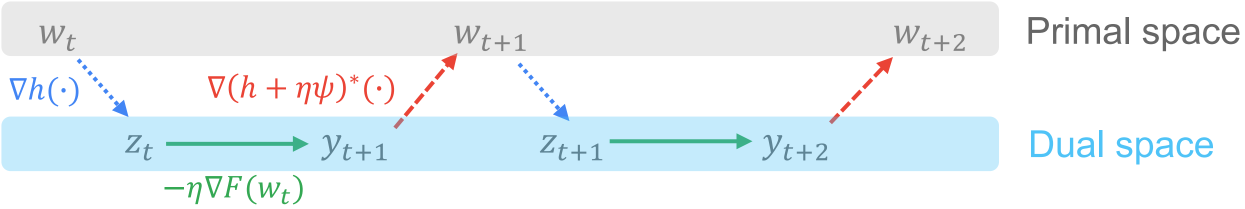
Backward mirror (Dual -> Primal) – retrieve primal

$$z_{t+1} = z_t - \eta \cdot \nabla F(w_t)$$

Gradient step (in dual space)

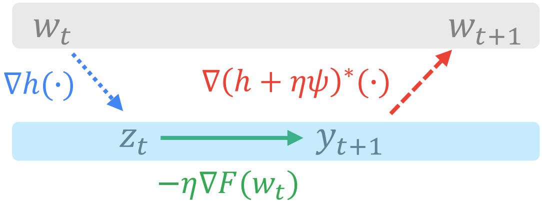


Recall MD:



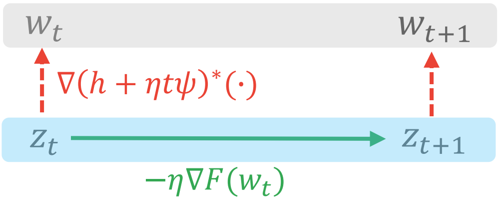
Mirror Descent vs Dual Averaging

Mirror Descent



- Forward **and** backward mirror
- Persistent **primal** states

Dual Averaging

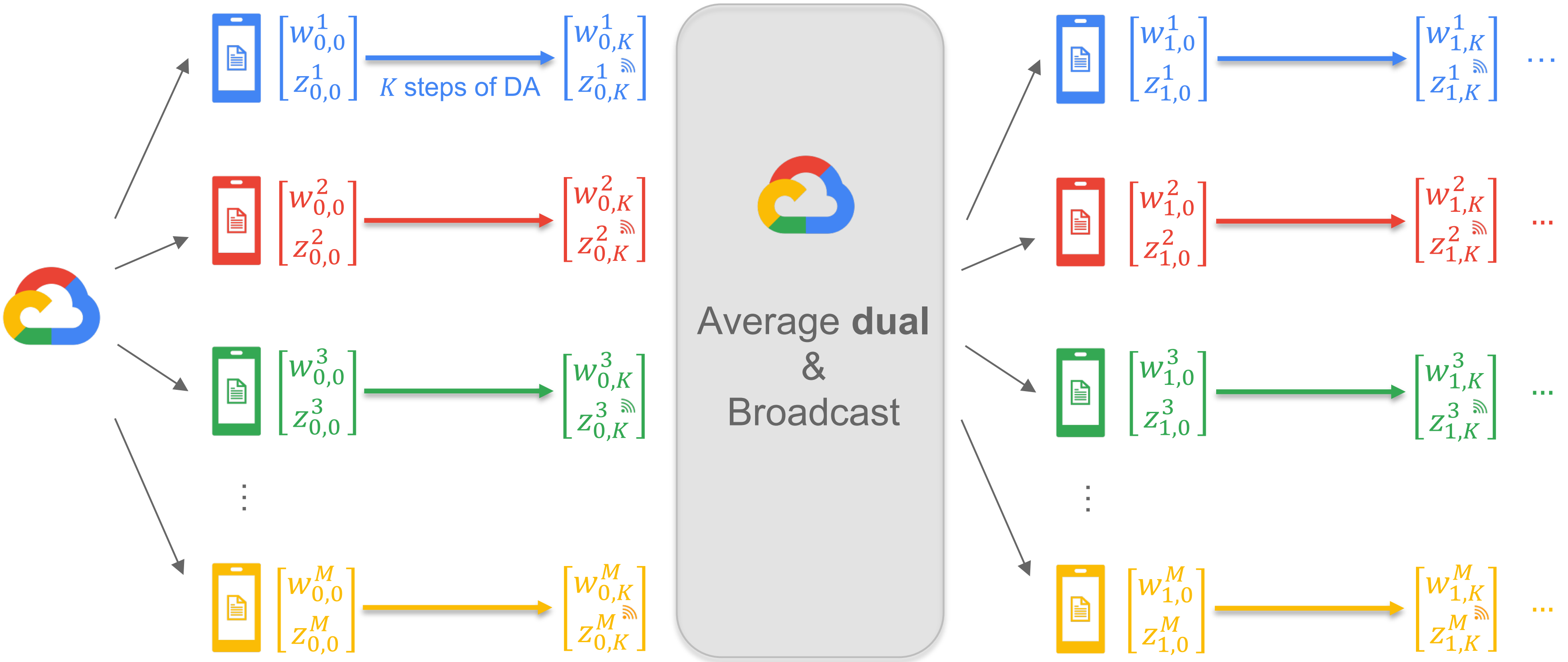


- Backward mirror **only**
- Persistent **dual** states

Google

Federated Dual Averaging (FedDualAvg)

Notation: $w_{r,k}^m$ → m^{th} client; r^{th} round, k^{th} local iteration



Federated Dual Averaging (FedDualAvg)

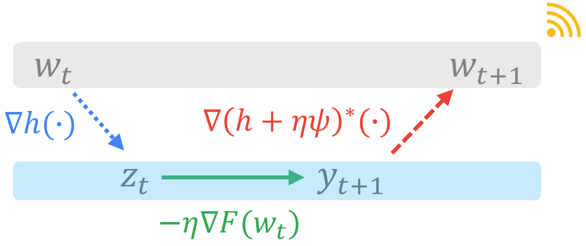
Notation: $w_{r,k}^{m}$ → m^{th} client; r^{th} round, k^{th} local iteration

Algorithm 3 Federated Dual Averaging

procedure FedDualAvg(w_0, η_c, η_s)
  server initialization $z_0 \leftarrow \nabla h(w_0)$
  for $r = 0, \ldots, R-1$ **do**
    sample a subset of clients $\mathcal{S}_r \subseteq [M]$
    on client $m \in \mathcal{S}_r$ **in parallel do**
      client initialization $z_{r,0}^{m} \leftarrow z_r$
      for $k = 0, \ldots, K-1$ **do**
        $\tilde{\eta}_{r,k} \leftarrow \eta_s \eta_c r K + \eta_c k$
        $w_{r,k}^{m} \leftarrow \nabla(h + \tilde{\eta}_{r,k}\psi)^{*}(z_{r,k}^{m})$ → Compute primal point
        $g_{r,k}^{m} \leftarrow \nabla f(w_{r,k}^{m}; \xi_{r,k}^{m})$
        $z_{r,k+1}^{m} \leftarrow z_{r,k}^{m} - \eta_c g_{r,k}^{m}$ → Client **dual** update
    $\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} (z_{r,K}^{m} - z_{r,0}^{m})$ → Average client **dual** deltas
    $z_{r+1} \leftarrow z_r + \eta_s \Delta_r$ → Server **dual** update
    $w_{r+1} \leftarrow \nabla(h + \eta_s \eta_c (r+1) K \psi)^{*}(z_{r+1})$ → (Optional) primal output

FedMiD (a.k.a. FedPGD) vs FedDualAvg

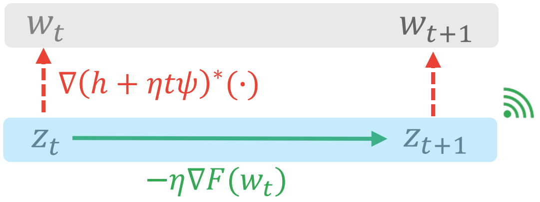
Mirror Descent



- Forward **and** backward mirror
- Persistent **primal** updates

- FedMiD: average the **primal**
- theoretically challenging due to the **nonlinearity of mirror map**.

Dual Averaging



- Backward mirror **only**
- Persistent **dual** updates

- FedDualAvg: average the **dual**
- Enjoys nice theoretical interpretation via **dual shadow sequence**.

- outperforms FedMiD empirically.

Theory: Blanket Assumptions

$$\min_{w \in \mathbb{R}^d} \quad \Phi(w) := \frac{1}{M} \sum_{m=1}^{M} F_m(w) + \psi(w)$$

$$\text{where} \quad F_m(w) := \mathbb{E}_{\xi \sim \mathcal{D}_m} [f(w;\xi)]$$

Assumption 1. *Let $\|\cdot\|$ be an arbitrary norm and $\|\cdot\|_*$ be its dual norm.*

(a) $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a closed convex function with closed $\mathbf{dom}\,\psi$. Assume $\Phi(w) = F(w) + \psi(w)$ attains a finite optimum at $\theta^\star \in \mathbf{dom}\,\psi$.

(b) $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a Legendre function that is 1-strongly convex with respect to $\|\cdot\|$. Assume $\mathbf{dom}\,h \supset \mathbf{dom}\,\psi$.

(a) & (b): standard regularity assumptions for composite setup

(c) $f(\cdot,\xi) : \mathbb{R}^d \to \mathbb{R}$ is a closed convex function that is differentiable on $\mathbf{dom}\,\psi$ for any fixed ξ. In addition, $f(\cdot,\xi)$ is L-smooth on $\mathbf{dom}\,\psi$, namely for any $u, w \in \mathbf{dom}\,\psi$,

$$f(u;\xi) \le f(w;\xi) + \langle \nabla f(w;\xi), u - w \rangle + \frac{1}{2} L \|u - w\|^2.$$

(c): smoothness of f

(d) ∇f has σ^2-bounded variance under $\|\cdot\|_*$ norm within $\mathbf{dom}\,\psi$, namely for any $w \in \mathbf{dom}\,\psi$,

$$\mathbb{E}_{\xi \sim \mathcal{D}_m} \|\nabla f(w,\xi) - \nabla F_m(w)\|_*^2 \le \sigma^2.$$

(d): additive bounded variance

(e) Assume all the M clients participate in client updates for every round, namely $\mathcal{S}_r = [M]$.

(e): full participation (for simplicity of exposition)

Google

Theorem 1: Small Client Learning Rate η_c Regime

In small η_c regime, both FedMiD and FedDualAvg can match minibatch rate

Theorem 1. Assuming A1, for **sufficiently small** η_c, and appropriate η_s, both FedMiD and FedDualAvg can output $\hat{w}$ such that

$$\mathbb{E}\left[\Phi(\hat{w})\right] - \Phi(w^\star) \lesssim \frac{LB}{R} + \frac{\sigma B^{\frac{1}{2}}}{M^{\frac{1}{2}} K^{\frac{1}{2}} R^{\frac{1}{2}}}$$

where $B := D_h(w^\star, w_0)$ is the Bregman divergence distance between optimum $w^\star$ and initial w_0

L: smoothness
σ: variance bound
M: # of clients
K: # of local steps
R: # of rounds

Stronger Guarantee for FedDualAvg (bounded gradient)

We establish (possibly) stronger guarantee for **FedDualAvg** with larger η_c and unit $\eta_s = 1$

Theorem 2. Assuming A1, and in addition assume $\sup_{w \in \mathbf{dom} \psi} \|\nabla f(w, \xi)\|_* \leq G$, then for $\eta_s = 1$ and $\eta_c \leq \frac{1}{4L}$, FedDualAvg can output $\hat{w}$ such that

$$\mathbb{E}\left[\Phi\left(\hat{w}\right)\right] - \Phi(w^\star) \lesssim \frac{B}{\eta_c K R} + \frac{\eta_c \sigma^2}{M} + \eta_c^2 L K^2 G^2$$

Moreover for appropriate η_c

faster convergence (usefulness of client step)

Overhead for infrequent communication

$$\mathbb{E}\left[\Phi\left(\hat{w}\right)\right] - \Phi(w^\star) \lesssim \frac{LB}{KR} + \frac{\sigma B^{\frac{1}{2}}}{M^{\frac{1}{2}} K^{\frac{1}{2}} R^{\frac{1}{2}}} + \frac{L^{\frac{1}{3}} B^{\frac{2}{3}} G^{\frac{2}{3}}}{R^{\frac{2}{3}}}.$$

matches [Stich ICLR'19] bound on smooth unconstrained FedAvg

$B := D_h(w^\star, w_0)$
L: smoothness
σ: variance bound
M: # of clients
K: # of local steps
R: # of rounds

Google

Stronger Guarantee for FedDualAvg (quadratic F)

We can relax the bounded gradient assumption if F is quadratic, and heterogeneity is bounded.

Theorem 3. Assuming A1, and in addition assume $\sup_{w \in \mathbf{dom}\psi} \|\nabla F_m(w) - \nabla F(w)\|_* \leq \zeta^2$ and F is quadratic, then FedDualAvg can output $\hat{w}$ such that

$$\mathbb{E}\left[\Phi\left(\hat{w}\right)\right] - \Phi(w^\star) \lesssim \frac{B}{\eta_c K R} + \frac{\eta_c \sigma^2}{M} + \eta_c^2 L K \sigma^2 + \eta_c^2 L K^2 \zeta^2,$$

moreover for appropriate η_c

faster convergence (usefulness of client step)

Overhead for infrequent communication

$$\mathbb{E}\left[\Phi\left(\hat{w}\right)\right] - \Phi(w^\star) \lesssim \frac{LB}{KR} + \frac{\sigma B^{\frac{1}{2}}}{M^{\frac{1}{2}} K^{\frac{1}{2}} R^{\frac{1}{2}}} + \frac{L^{\frac{1}{3}} B^{\frac{2}{3}} \sigma^{\frac{2}{3}}}{K^{\frac{1}{3}} R^{\frac{2}{3}}} + \frac{L^{\frac{1}{3}} B^{\frac{2}{3}} \zeta^{\frac{2}{3}}}{R^{\frac{2}{3}}}.$$

matches best known bound on smooth unconstrained FedAvg [Khaled AISTATS'20, Woodworth NeurIPS'20 etc]

$B := D_h(w^\star, w_0)$
L: smoothness
σ: variance bound
M: # of clients
K: # of local steps
R: # of rounds

Summary of Theoretical Results

- FedMiD & FedDualAvg, small η_c: $$\frac{LB}{R} + \frac{\sigma B^{\frac{1}{2}}}{M^{\frac{1}{2}} K^{\frac{1}{2}} R^{\frac{1}{2}}}$$

- FedDualAvg, larger η_c: $$\frac{LB}{KR} + \frac{\sigma B^{\frac{1}{2}}}{M^{\frac{1}{2}} K^{\frac{1}{2}} R^{\frac{1}{2}}} + \frac{L^{\frac{1}{3}} B^{\frac{2}{3}} G^{\frac{2}{3}}}{R^{\frac{2}{3}}}$$
$$\frac{LB}{KR} + \frac{\sigma B^{\frac{1}{2}}}{M^{\frac{1}{2}} K^{\frac{1}{2}} R^{\frac{1}{2}}} + \frac{L^{\frac{1}{3}} B^{\frac{2}{3}} \sigma^{\frac{2}{3}}}{K^{\frac{1}{3}} R^{\frac{2}{3}}} + \frac{L^{\frac{1}{3}} B^{\frac{2}{3}} \zeta^{\frac{2}{3}}}{R^{\frac{2}{3}}}$$

$B \coloneqq D_h(w^\star, w_0)$
L: smoothness
σ: variance bound
M: # of clients
K: # of local steps
R: # of rounds
G: gradient bound
ζ: heterogeneity bound

Google

Proof Sketch -- FedDualAvg

Main observation: the averaged dual $\overline{z_{r,k}} := \frac{1}{M}\sum_{m=1}^{M} z_{r,k}^m$ "almost" does **centralized dual averaging**

$$\overline{z_{r,k+1}} = \overline{z_{r,k}} - \eta_c \cdot \frac{1}{M}\sum_{m=1}^{M} \nabla f(w_{r,k}^m; \xi_{r,k}^m)$$

***Variance-reduced** but **biased** stochastic gradient oracle*

Step 1: convergence of the averaged dual (a.k.a. perturbed iterate analysis)

$$\mathbb{E}\left[\Phi\left(\frac{1}{KR}\sum_{r=0}^{R-1}\sum_{k=1}^{K}\nabla\left(h + \tilde{\eta}_{r,k}\psi\right)^*\left(\overline{z_{r,k}}\right)\right)\right] - \Phi(w^\star) \leq \underbrace{\frac{B}{\eta_c KR} + \frac{\eta_c\sigma^2}{M}}_{\text{Rate if synchronize every iterations}} + \underbrace{\frac{L}{MKR}\left[\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\sum_{m=1}^{M}\mathbb{E}\left\|\overline{z_{r,k}} - z_{r,k}^m\right\|_*^2\right]}_{\text{Discrepancy overhead}}$$

Step 2: bound $\mathbb{E}\left\|\overline{z_{r,k}} - z_{r,k}^m\right\|_*^2$ by stability analysis

Experiments

- Platform setup: *TensorFlow/Federated* & *google-research/federated*

- We evaluate the following 4 algorithms:

  1. Federated Dual Averaging (FedDualAvg)
  2. Federated Mirror Descent (FedMiD)
  3. FedDualAvg-OSP (only-server-proximal)
  4. FedMiD-OSP (only-server-proximal)

potential light computation but less principled - for ablation study purpose

FedMiD vs FedMiD-OSP

Algorithm 2 Federated Mirror Descent (FedMiD)

procedure FedMiD(w_0, η_c, η_s)
  for $r = 0, \dots, R-1$ **do**
    sample a subset of clients $\mathcal{S}_r \subseteq [M]$
    on client $m \in \mathcal{S}_r$ **in parallel do**
      client initialization $w_{r,0}^m \leftarrow w_r$
      for $k = 0, \dots, K-1$ **do**
        $g_{r,k}^m \leftarrow \nabla f(w_{r,k}^m; \xi_{r,k}^m)$
        $w_{r,k+1}^m \leftarrow \nabla(h + \eta_c \psi)^* (\nabla h(w_{r,k}^m) - \eta_c g_{r,k}^m)$
    $\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} (w_{r,K}^m - w_{r,0}^m)$
    $w_{r+1} \leftarrow \nabla(h + \eta_s \eta_c K \psi)^* (\nabla h(w_r) + \eta_s \Delta_r)$

Algorithm 4 Federated Mirror Descent Only Server

procedure FedMiD-OSP(w_0, η_c, η_s)
  for $r = 0, \dots, R-1$ **do**
    sample a subset of clients $\mathcal{S}_r \subseteq [M]$
    on client $m \in \mathcal{S}_r$ **in parallel do**
      client initialization $w_{r,0}^m \leftarrow w_r$
      for $k = 0, \dots, K-1$ **do**
        $g_{r,k}^m \leftarrow \nabla f(w_{r,k}^m; \xi_{r,k}^m)$
        $w_{r,k+1}^m \leftarrow \nabla h^* (\nabla h(w_{r,k}^m) - \eta_c g_{r,k}^m)$ ▷
    $\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} (w_{r,K}^m - w_{r,0}^m)$
    $w_{r+1} \leftarrow \nabla(h + \eta_s \eta_c K \psi)^* (\nabla h(w_r) + \eta_s \Delta_r)$

Proximal ψ skipped

Reduces to $w_{r,k+1}^m \leftarrow w_{r,k}^m - \eta_c g_{r,k}^m$ if $h = \frac{1}{2}\|\cdot\|^2$

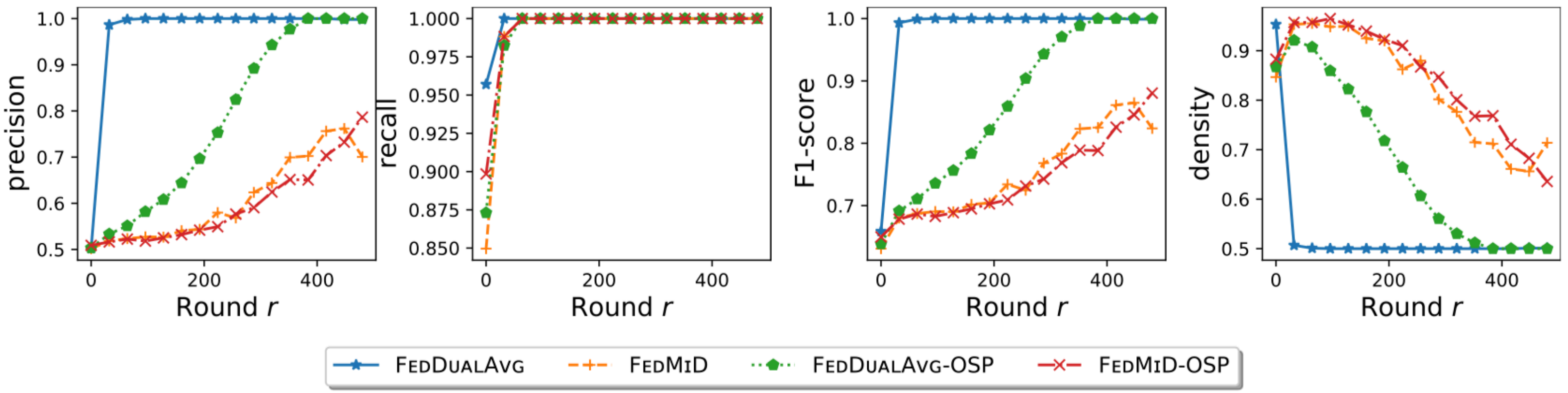
Experiment 1: Federated Lasso on Synthetic Dataset

- **Synthetic dataset**: $y = x^T w^\star + b^\star + \varepsilon$; known sparse ground truth w^*

  (64 clients, 128 samples per client, ground truth density 512/1024)

- **Problem**: $$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{(x,y)\sim\mathcal{D}_m}(x^\top w + b - y)_2^2 + \lambda\|w\|_1$$

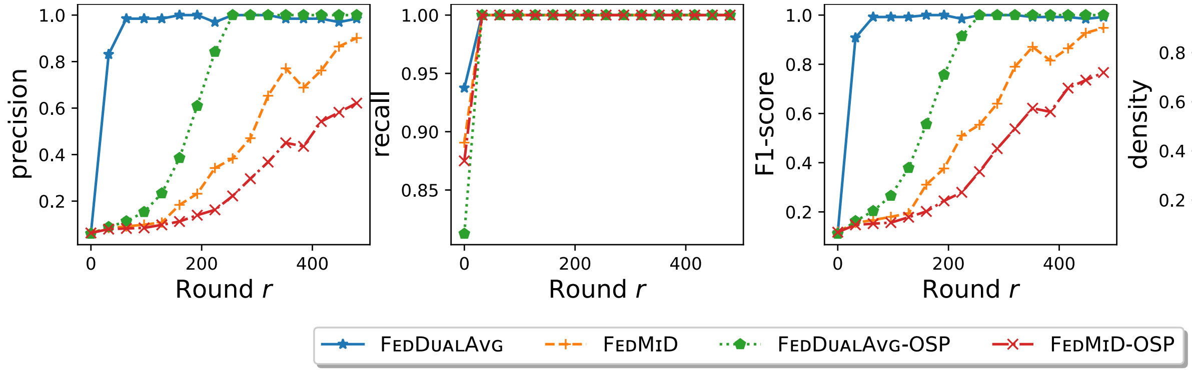
- **Metric**: F1-score of the estimated sparsity, precision, recall, density



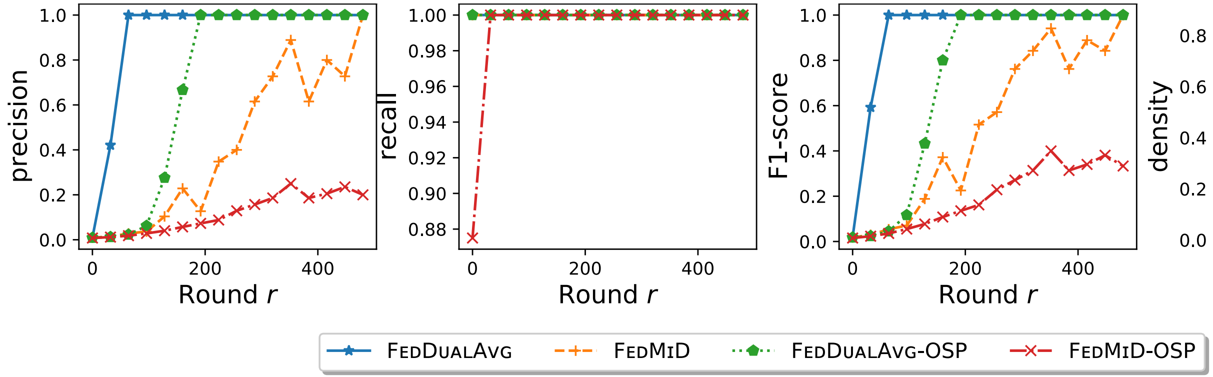
For all algorithms, we tune only η_s and η_c to attain the best F1-score

Experiment 1: Sparser Ground Truth

- **Sparser dataset:** *(64 clients, 128 samples per client, ground truth density **64**/1024)*

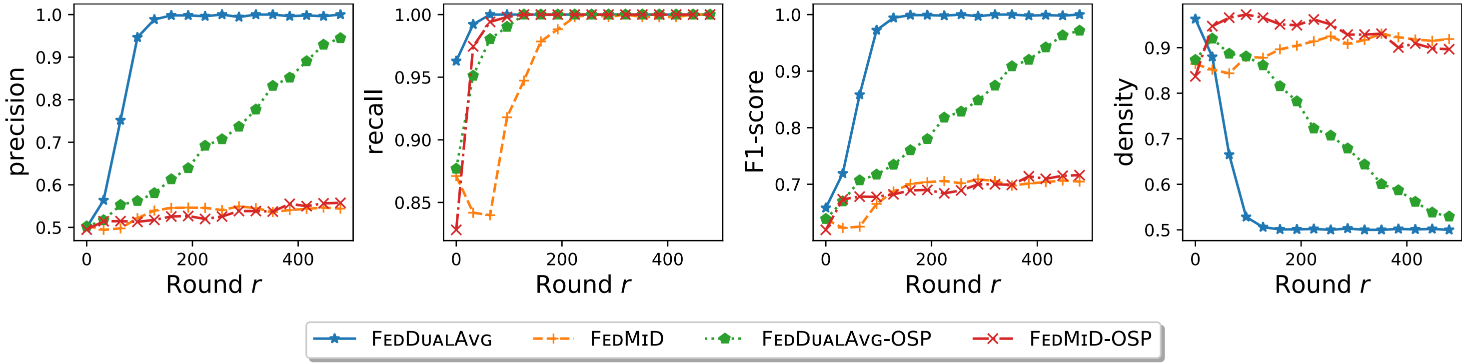


- **Even sparser dataset:** *(64 clients, 128 samples per client, ground truth density **8**/1024)*



Experiment 1: More Distributed Data

- **Even more distributed:** *(**256** clients, **32** samples per client, ground truth density 512/1024)*

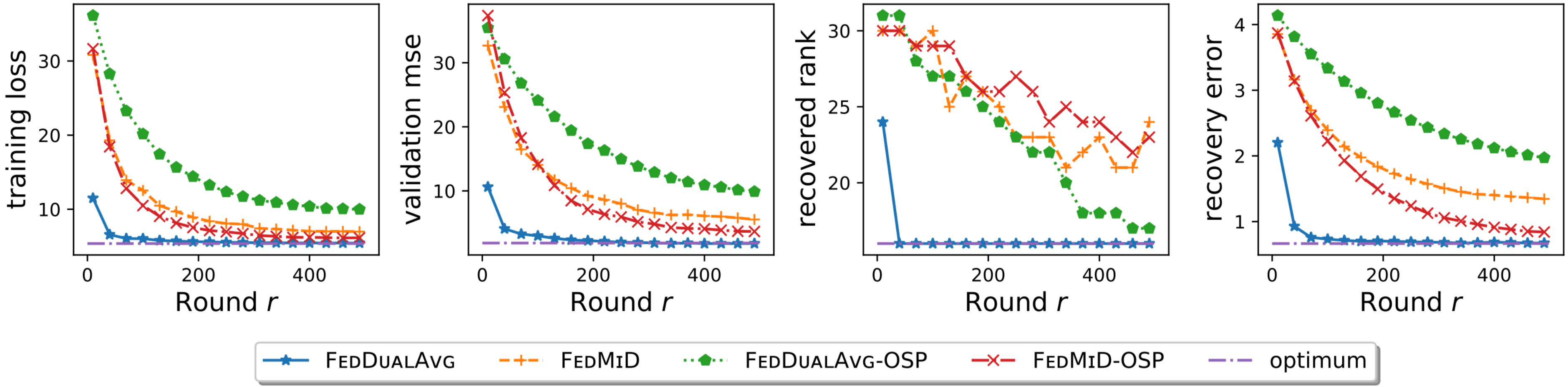


Experiment 2: Low-Rank Matrix Estimation

- **Synthetic dataset**: $y = \langle X, W^\star \rangle + b^\star + \varepsilon$; known **low-rank** ground truth W^*

  *(64 clients, 128 samples per client, ground truth rank **16**/32)*

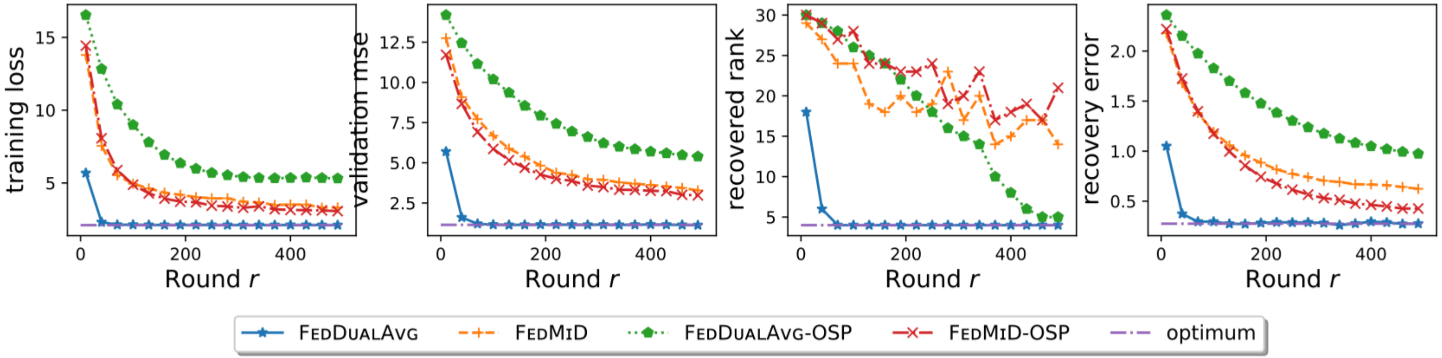
- **Problem**: $$\min_{W \in \mathbb{R}^{d_1 \times d_2}, b \in \mathbb{R}} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{(X,y)\sim\mathcal{D}_m} \left(\langle X, W \rangle + b - y\right)^2 + \lambda \|W\|_{\text{nuc}}$$
- **Metric**: training loss, validation mse, recovered rank, recovered error (in Frobenius norm)



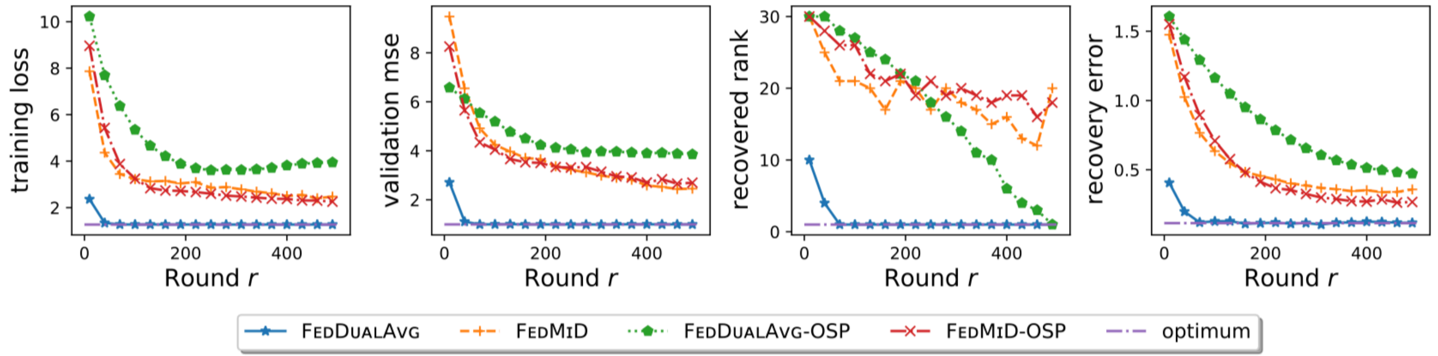
Google

Experiment 2: Sparser Ground Truth

- **Lower rank dataset:** *(64 clients, 128 samples per client, ground truth rank **4**/32)*

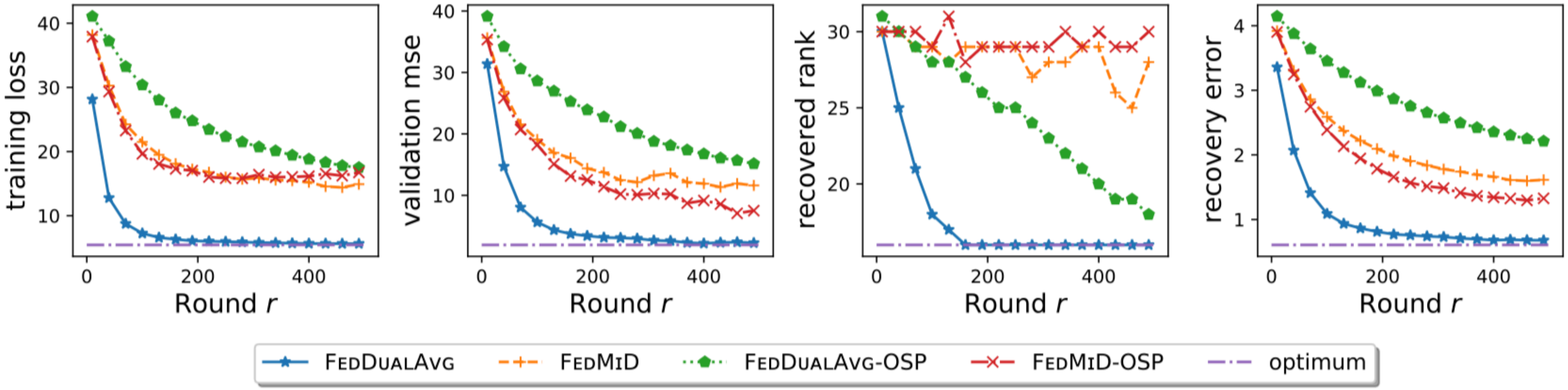


- **Even lower rank dataset:** *(64 clients, 128 samples per client, ground truth rank **1/32**)*



Experiment 2: More Distributed Data

- **More distributed:** *(**256** clients, **32** samples per client, ground truth density 512/1024)*



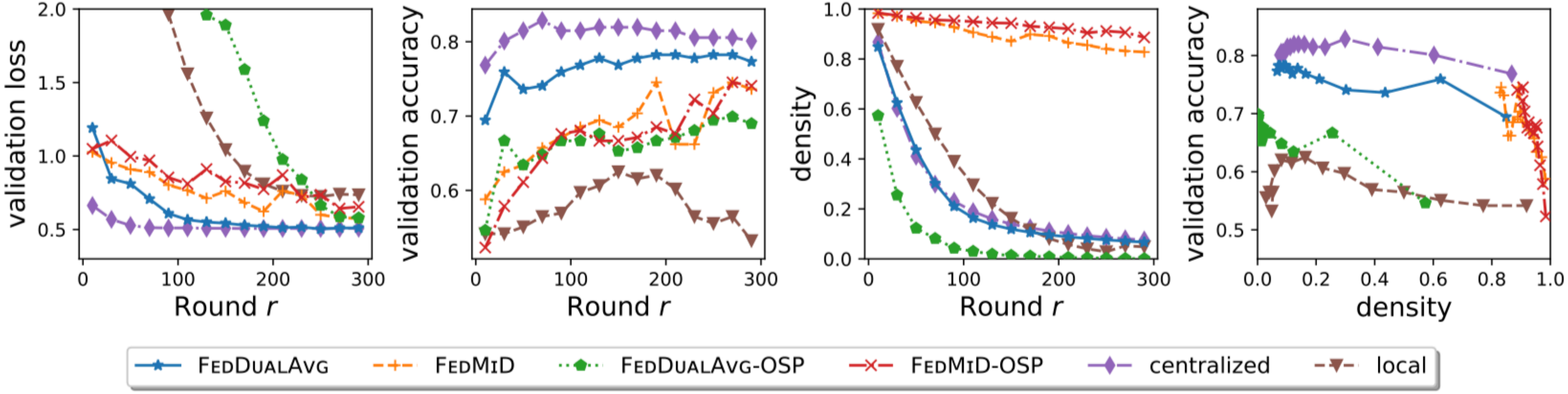
Google

Experiment 3: Sparse Logistic Regression for fMRI

- **Dataset**: fMRI scans on response to binary image recognition

  (6 subjects, 11-12 sessions per subject, 18 scans per session, 39,912 voxels)

- **Federated Setup:** Each client possesses the data of a **session**. (59 training clients in total)
- **Problem**: l1-regularized logistic regression
- **Metric**: density, validation accuracy

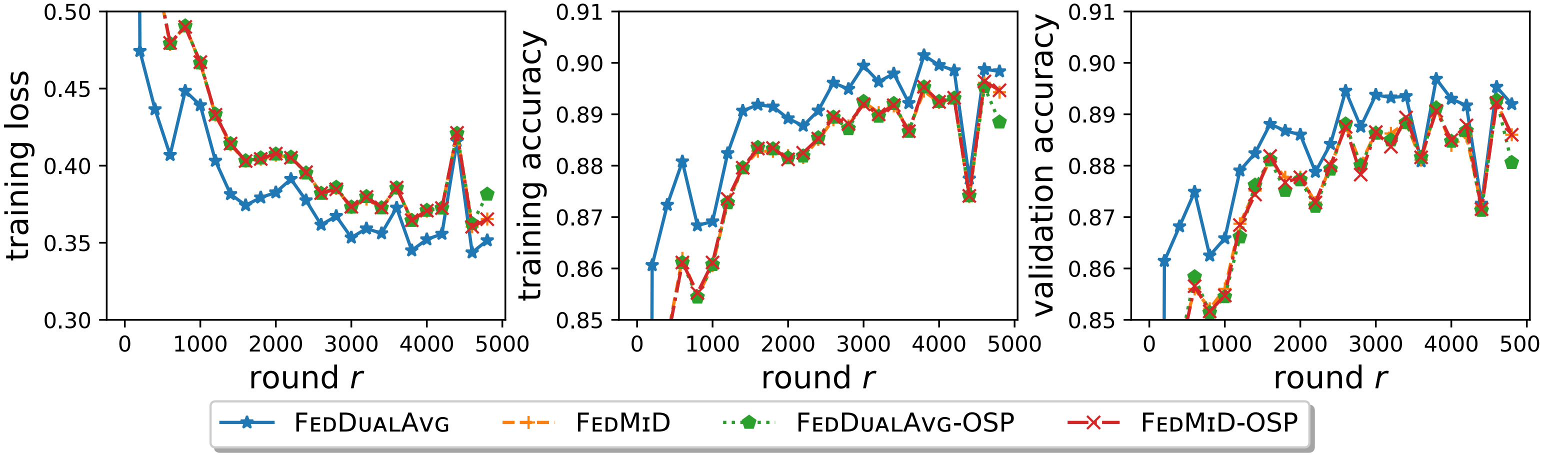


Google

Experiment 4: norm-ball constrained FL

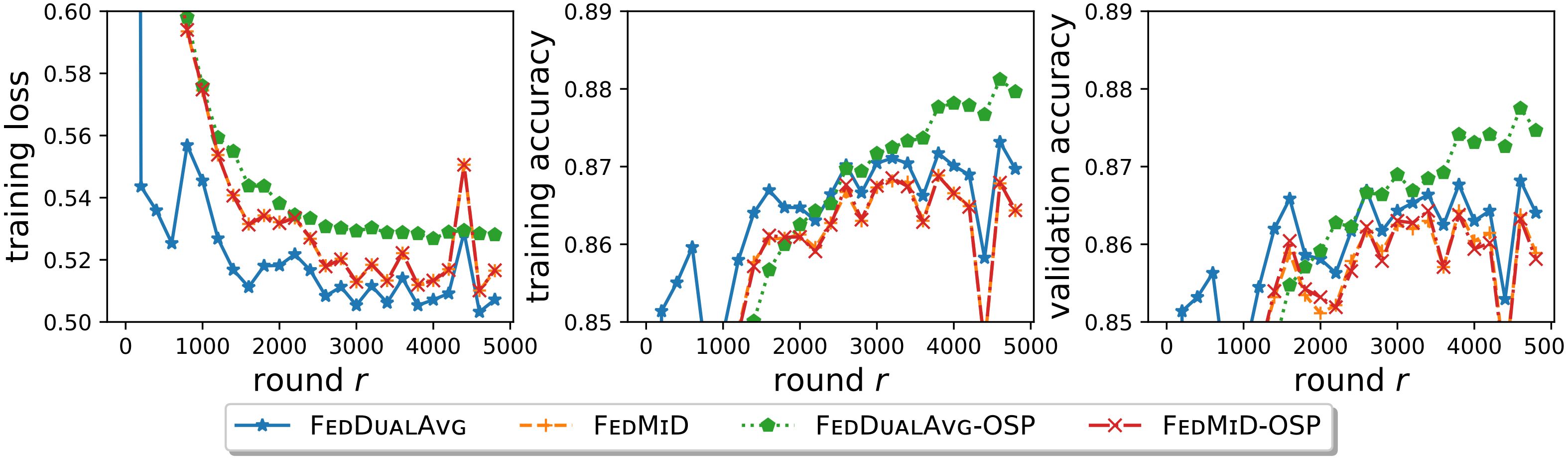
- **Dataset**: Federated EMNIST (10 classes or 62 classes)
- **Metric**: Training loss, training accuracy, test accuracy

- *L1-constrained logistic regression for EMNIST-10*



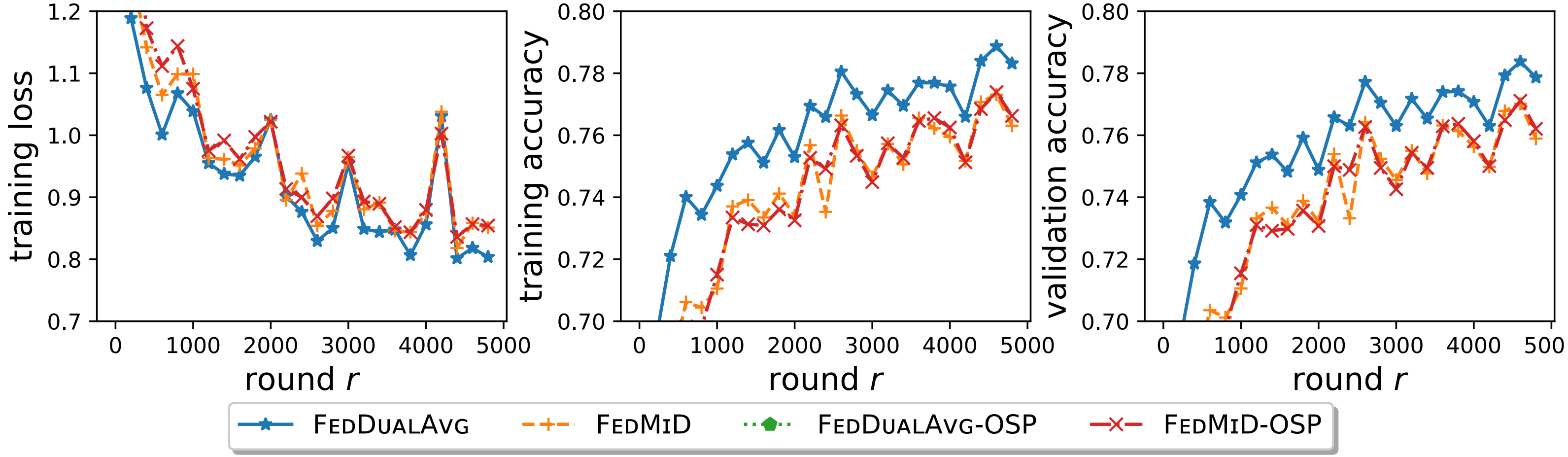
Experiment 4: norm-ball constrained FL

- *L2-constrained logistic regression for EMNIST-10*



Experiment 4: norm-ball constrained FL

- *L1-constrained 2-hidden-layer NN on EMNIST-62*



Google

Thank you!

Paper: https://arxiv.org/abs/2011.08474

Email: yuanhl@cs.stanford.edu

Google