Federated Composite Optimization

Honglin Yuan (Stanford), Manzil Zaheer (Google), Sashank Reddi (Google)

Overview

New Setting: Federated Composite Optimization (FCO)

- FL with (possibly non-smooth) regularizers or constraints
- Arises naturally in FL applications that involve sparsity, low-rank, or constraints.
- Standard FL algorithms (e.g., FedAvg) are for smooth unconstrained settings.

[McMahan et al., 2017]

Straightforward Extension of FedAvg suffers from "curse of primal averaging"

Our proposal: Federated Dual Averaging

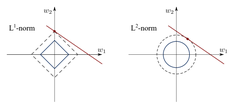
- Novel server dual averaging procedure
- Theoretical and empirical advantages

Problem definition and Examples

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

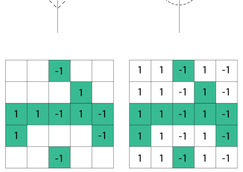
- $F_m(w) := \mathbb{E}_{\xi\sim\mathcal{D}_m}[f(w;\xi)]$ is the loss function of the m-th client
- ψ is a (possibly non-smooth, non-finite) convex regularizer
- **Federated Lasso** for sparsity representations

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



- **Federated matrix completion** via nuclear norm

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



- **FL with constraints:** let ψ be convex indicator $\begin{cases} 0 & \text{if } w \in \mathcal{C}. \\ +\infty & \text{if } w \notin \mathcal{C}. \end{cases}$

Background of (non-federated) composite optimization

Composite 101: **ProxGD** is the standard algorithm for solving non-federated CO:

$$w_{t+1} \leftarrow \mathbf{prox}_{\eta\psi}\left(w_t - \eta\nabla F(w_t)\right)$$
$$:= \operatorname*{argmin}_{w} \left\{ F(w_t) + \langle \nabla F(w_t), w - w_t\rangle + \frac{1}{2\eta}\|w - w_t\|_2^2 + \psi(w) \right\}$$

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

Composite 201: Mirror descent generalizes ProxGD to general Bregman divergence

$$w_{t+1} = \operatorname*{argmin}_{w} \left\{ F(w_t) + \langle \nabla F(w_t), w - w_t\rangle + \psi(w) + \frac{1}{\eta} D_h(w, w_t) \right\}$$

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

Primal-Dual interpretation of Mirror Descent *[Nemirovski and Yudin, 1983, Flammarion and Bach, 2017]*

- $z_t = \nabla h(w_t)$ 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)

FedMiD: a straightforward extension

Algorithm 1 Federated Averaging (FedAvg)

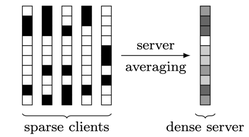
procedure FedAvg(w_0, η_c, η_s)
  for $r = 0, \dots, R-1$ **do**
    sample a subset of clients $\mathcal{S}_r \subseteq [M]$
    on client $m \in \mathcal{S}_r$ **in parallel do**
      client initialization $w_{r,0}^m \leftarrow w_r$
      for $k = 0, \dots, K-1$ **do**
        $g_{r,k}^m \leftarrow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

Algorithm 2 Federated Mirror Descent (FedMiD)

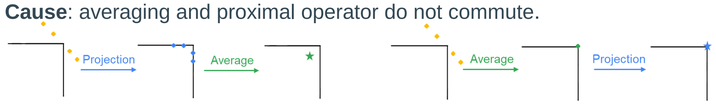
procedure FedMiD(w_0, η_c, η_s)
  for $r = 0, \dots, R-1$ **do**
    sample a subset of clients $\mathcal{S}_r \subseteq [M]$
    on client $m \in \mathcal{S}_r$ **in parallel do**
      client initialization $w_{r,0}^m \leftarrow w_r$
      for $k = 0, \dots, K-1$ **do**
        $g_{r,k}^m \leftarrow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

Issue of FedMiD: curse of primal averaging

While each client can locate a sparse solution, simply averaging yields a dense solution on the server.



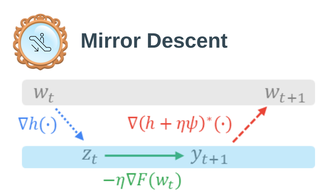
Cause: averaging and proximal operator do not commute.



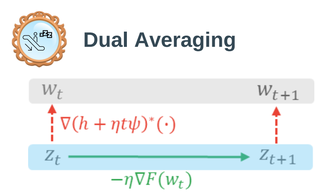
Background: Dual averaging

Composite 202: Dual Averaging (a.k.a. Lazy Mirror Descent) *[Nesterov, 2009]*

$$w_t = \nabla(h + \eta t\psi)^*(z_t)$$ Backward mirror (Dual -> Primal) – retrieve primal
$$= \arg\min_{w}\{\langle -z_t, w\rangle + \eta t\psi(w) + h(w)\}$$
$$z_{t+1} = z_t - \eta\cdot\nabla F(w_t)$$ Gradient step (in dual space)



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



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

Our main proposal: FedDualAvg

Algorithm 3 Federated Dual Averaging

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

Locally: each client runs dual averaging, tracking a pair of primal and dual states.

Communication: dual states are aggregated across clients.

- Compute primal point
- Client **dual** update
- Average client **dual** deltas
- Server **dual** update
- (Optional) primal output

Main theoretical results

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 $\theta^* \in \mathbf{dom}\,\psi$.

(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)

Theorem 4.3. Assuming A1, and in addition assume $\sup_{w\in\mathbf{dom}\,\psi}\|\nabla F_m(w) - \nabla F(w)\|_* \le \zeta^2$ and F is quadratic, then FedDualAvg can output $\hat{w}$ such that

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

moreover for appropriate η_c *faster convergence (usefulness of client step)* *Overhead for infrequent communication*

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

matches best known bound on smooth unconstrained FedAvg c.f. [Woodworth et al., 2020]

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

Experiments

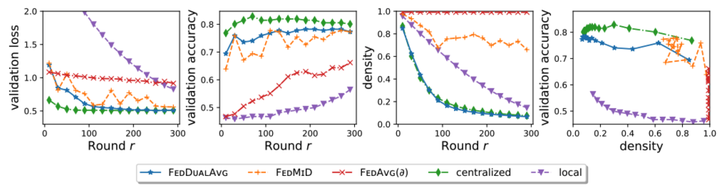


Figure 1: **Results on sparse (ℓ_1-regularized) logistic regression for a federated fMRI dataset based on (Haxby, 2001).** `centralized` corresponds to training on the centralized dataset gathered from **all** the training clients. `local` corresponds to training on the local data from only **one** training client without communication. FedAvg (∂) corresponds to running FedAvg algorithms with subgradient in lieu of SGD to handle the non-smooth ℓ_1-regularizer. FedMiD is another straightforward extension of FedAvg running local proximal gradient method (see Section 3.1 for details). We show that using our proposed algorithm FedDualAvg, one can 1) achieve performance comparable to the `centralized` baseline without the need to gather client data, and 2) significantly outperforms the `local` baseline on the isolated data and the FedAvg baseline. See Section 5.3 for details.

Main References

A.S. Nemirovski and D.B. Yudin "Problem complexity and method efficiency in optimization." 1983

Yurii Nesterov "Primal-dual subgradient methods for convex problems" In: Mathematical Programming, 120(1), 2009

Nicolas Flammarion and Francis Bach "Stochastic composite least-squares regression with convergence rate O(1/n)" In: COLT 2017.

Brendan McMahan et al. "Communicationefficient learning of deep networks from decentralized data" In: AISTATS 2017

Blake Woodworth et al "Minibatch vs Local SGD for Heterogeneous Distributed Learning" In: NeurIPS 2020.