Nonconvex Federated Learning on Compact Smooth Submanifolds With Heterogeneous Data

Jiaojiao Zhang[1], **Jiang Hu**[2], Anthony Man-cho So[3], Mikael Johansson[1]

[1]KTH Royal Institute of Technology,
[2]University of California, Berkeley, [3]The Chinese University of Hong Kong

Introduction

- Consider federated learning (FL) on manifolds

$$\underset{x\in\mathcal{M}}{\text{minimize}}\ f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i(x) = \frac{1}{m_i}\sum_{l=1}^{m_i} f_{il}(x;\mathcal{D}_{il}) \tag{1}$$

  - Each client has local loss f_i that is smooth but nonconvex
  - Local datasets $\mathcal{D}_i$ across clients i are heterogeneous
  - $\mathcal{M}$ is a compact smooth submanifold embedded in $\mathbb{R}^{d\times k}$, with Euclidean metric serving as its Riemannian metric. E.g., Stiefel manifold: $\mathrm{St}(d,k) = \{x \in \mathbb{R}^{d\times k} : x^T x = I_k\}$

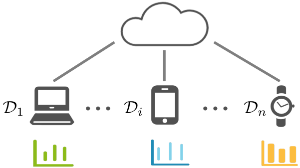

Figure: Federated learning

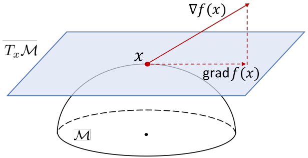

Figure: Optimization on manifolds

Challenges and Contributions

Challenges

- Single-machine optimization on $\mathcal{M}$ cannot be directly adapted to FL
  - Even if each local model lies on $\mathcal{M}$, their average typically does not
- Extending FL algorithms to manifold optimization is not straightforward
  - $\mathcal{M}$ is nonconvex
- FL algorithms with local updates need substantial modifications to accommodate $\mathcal{M}$
  - Client drift issue due to local updates and heterogeneous data persists

Contributions

- Propose a computation- and communication-efficient algorithm for solving (1)
- Establish sub-linear convergence to a neighborhood of a first-order optimal solution
- Demonstrate superior performance over alternative methods

Algorithm Intuition and Innovations

- The equivalent and compact form of our algorithm is

$$\begin{cases} \hat{\mathbf{z}}_{t+1}^r = \hat{\mathbf{z}}_t^r - \eta\Big(\underbrace{\operatorname{grad}\mathbf{f}\left(\mathbf{z}_t^r; \mathcal{B}_t^r\right)}_{\text{new}} + \underbrace{\frac{1}{\tau}\sum_{t=0}^{\tau-1} \overline{\operatorname{grad}\mathbf{f}}\left(\mathbf{z}_t^{r-1}; \mathcal{B}_t^{r-1}\right)}_{\text{average}} - \underbrace{\frac{1}{\tau}\sum_{t=0}^{\tau-1} \operatorname{grad}\mathbf{f}\left(\mathbf{z}_t^{r-1}; \mathcal{B}_t^{r-1}\right)}_{\text{old}} \Big) \\ \mathbf{z}_{t+1}^r = \mathcal{P}_{\mathcal{M}}\left(\hat{\mathbf{z}}_{t+1}^r\right) \\ \mathbf{x}^{r+1} = \mathcal{P}_{\mathcal{M}}(\mathbf{x}^r) - \eta_g \eta \sum_{t=0}^{\tau-1} \overline{\operatorname{grad}\mathbf{f}}\left(\mathbf{z}_t^r; \mathcal{B}_t^r\right) \end{cases}$$

- Mimic centralized projected Riemannian gradient descent
  - When $\tau = 1$ and $b = m_i$, we recover $\tilde{x}^{r+1} := \mathcal{P}_{\mathcal{M}}\Big(\mathcal{P}_{\mathcal{M}}(\overline{x}^r) - \tilde{\eta} \cdot \operatorname{grad} f(\mathcal{P}_{\mathcal{M}}(\overline{x}^r))\Big)$
- Feasibility of all iterates at a low computational cost by using $\mathcal{P}_{\mathcal{M}}$
  - Avoid exponential mapping, inverse exponential mapping, and parallel transport
- Overcome client drift
  - Correction employs "variance reduction" and does not incur extra communication

Algorithm Implementation

Input: R, τ, η, η_g, $\tilde{\eta} = \eta\eta_g\tau$, $\overline{x}^1$, and $c_i^1 = 0$ for all $i \in [n]$
for $r = 1, 2, \ldots, R$ **do**
  Client i
  Set $\hat{z}_{i,0}^r = \mathcal{P}_{\mathcal{M}}(\overline{x}^r)$ and $z_{i,0}^r = \mathcal{P}_{\mathcal{M}}(\overline{x}^r)$
  for $t = 0, 1, \ldots, \tau - 1$ **do**
    Sample a mini-batch dataset $\mathcal{B}_{i,t}^r \subseteq \mathrm{D}_i$ with $|\mathcal{B}_{i,t}^r| = b$
    Update $\operatorname{grad} f_i(z_{i,t}^r; \mathcal{B}_{i,t}^r) = \frac{1}{b}\sum_{\mathrm{D}_{il} \in \mathcal{B}_{i,t}^r} \operatorname{grad} f_{il}(z_{i,t}^r; \mathrm{D}_{il})$
    Update $\hat{z}_{i,t+1}^r = \hat{z}_{i,t}^r - \eta\left(\operatorname{grad} f_i(z_{i,t}^r; \mathcal{B}_{i,t}^r) + c_i^r\right)$
    Update $z_{i,t+1}^r = \mathcal{P}_{\mathcal{M}}\left(\hat{z}_{i,t+1}^r\right)$
  end for
  Send $\hat{z}_{i,\tau}^r$ to the server
  Server
  Update $\overline{x}^{r+1} = \mathcal{P}_{\mathcal{M}}(\overline{x}^r) + \eta_g\left(\frac{1}{n}\sum_{i=1}^n \hat{z}_{i,\tau}^r - \mathcal{P}_{\mathcal{M}}(\overline{x}^r)\right)$
  Broadcast $\overline{x}^{r+1}$ to all the clients
  Client i
  Update $c_i^{r+1} = \frac{1}{\eta_g\eta\tau}(\mathcal{P}_{\mathcal{M}}(\overline{x}^r) - \overline{x}^{r+1}) - \frac{1}{\tau}\sum_{t=0}^{\tau-1} \operatorname{grad} f_i(z_{i,t}^r; \mathcal{B}_{i,t}^r)$
end for
Output: $\mathcal{P}_{\mathcal{M}}(\overline{x}^{R+1})$

Convergence Analysis

Definitions

- (Riemannian gradient): The Riemannian gradient $\mathrm{grad} f(x)$ of a function f at the point $x \in \mathcal{M}$ is the unique tangent vector that satisfies

$$\langle \mathrm{grad} f(x), \xi \rangle_x = df(x)[\xi], \quad \forall \xi \in T_x\mathcal{M}$$

  - For a submanifold $\mathcal{M}$, $\mathrm{grad} f(x)$ can be computed as $\mathrm{grad} f(x) = \mathcal{P}_{T_x\mathcal{M}}(\nabla f(x))$
- ($\hat{\gamma}$-proximal smoothness of $\mathcal{M}$): The $\hat{\gamma}$-tube around $\mathcal{M}$ is $U_{\mathcal{M}}(\hat{\gamma}) := \{x : \mathrm{dist}(x, \mathcal{M}) < \hat{\gamma}\}$. We say that $\mathcal{M}$ is $\hat{\gamma}$-proximally smooth if the projection operator $\mathcal{P}_{\mathcal{M}}(x)$ is a singleton whenever $x \in U_{\mathcal{M}}(\hat{\gamma})$
  - Any compact smooth submanifold $\mathcal{M}$ embedded in $\mathbb{R}^{d\times k}$ is a proximally smooth set
  - Ensure not only the uniqueness of the projection but also the Lipschitz continuity of $\mathcal{P}_{\mathcal{M}}$

$$\|\mathcal{P}_{\mathcal{M}}(x) - \mathcal{P}_{\mathcal{M}}(y)\| \leq 2\|x - y\|, \quad \forall x, y \in \overline{U}_{\mathcal{M}}(\hat{\gamma}/2)$$

Convergence Analysis

Assumptions

- The proximal smoothness constant of $\mathcal{M}$ is 2γ
- L-smoothness: $\|\mathrm{grad} f_{il}(x;\mathcal{D}_{il}) - \mathrm{grad} f_{il}(y;\mathcal{D}_{il})\| \leq L\|x-y\|$
- Unbiasedness and bounded variance: $\mathbb{E}[\|\mathrm{grad} f_i(z_{i,t}^r;\mathcal{B}_{i,t}^r) - \mathrm{grad} f_i(z_{i,t}^r)\|^2 \mid \mathcal{F}_t^r] \leq \sigma^2/b$

(Theorem) Under some assumptions and conditions on $\tilde{\eta} := \eta\tau\eta_g$, we have

$$\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|\mathcal{G}_{\tilde{\eta}}(\mathcal{P}_{\mathcal{M}}(\overline{x}^r))\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{n}\tau R\eta} + \frac{\sigma^2}{n\tau b}\right)$$

where $\mathcal{G}_{\tilde{\eta}}(\mathcal{P}_{\mathcal{M}}(x^r)) := (\mathcal{P}_{\mathcal{M}}(x^r) - \tilde{x}^{r+1})/\tilde{\eta}$

Numerical Experiments

- kPCA problem with Mnist dataset

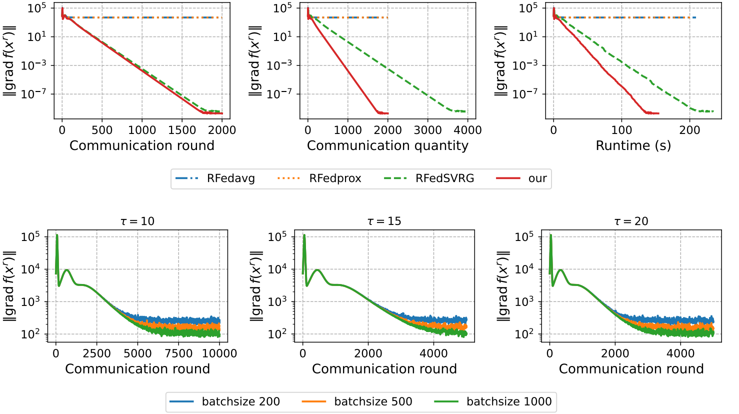

Figure: Comparison with alternative methods (1st row) and impacts of batch size (2nd row)