FLOW

Federated Learning of a Mixture of Global and Local Models: Local SGD and Optimal Algorithms

Filip Hanzely (KAUST)

Co-authors

Federated Learning of a Mixture of Global and Local Models
FH, P Richtárik, 2020

Lower Bounds and Optimal Algorithms for Personalized Federated
FH, S Hanzely, S Horváth, P Richtárik, 2020

Slavomír Hanzely
(KAUST)

Samuel Horváth
(KAUST)

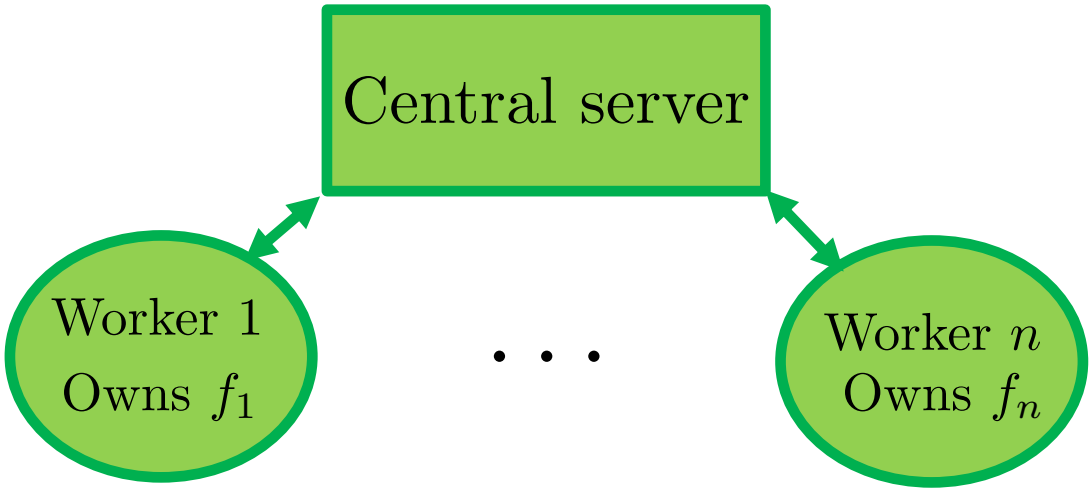
Peter Richtárik
(KAUST)

1. From classical FL to the mixture FL formulation

Distributed optimization in a datacenter

$$\min_{z \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(z)$$

number of workers not too big



Control over construction of $f_1, \ldots, f_n$
possibly shared data

Expensive communication

Communication/computation tradeoff

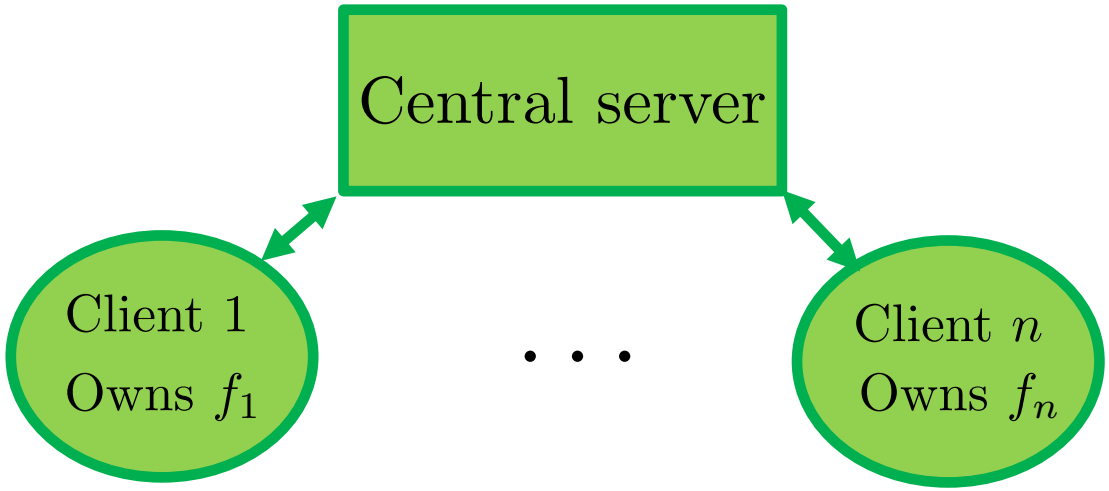
Goal: Find solution z^*

Classical federated learning

$$\min_{z \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(z)$$

number of clients
e.g. mobile phones

n is big



f_i constructed from data of i-th client only
No control over construction of $f_1, \ldots, f_n$
Local data never revealed

Communication is bottleneck

Goal: Find a model to be deployed on each client

Issue of classical FL: personalization

f_i constructed from data of i-th client only

$$\min_{z \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(z)$$

local data might differ greatly

Wrong objective

Goal: Find a model to be deployed on each client

Personalization

Next word prediction mobile keyboard

Federated learning for mobile keyboard prediction
A Hard et al, 2018

Issue of classical FL: FedAvg / Local SGD

Local gradient descent (LGD):

$$k \bmod T \neq 0$$
$$x_i^{k+1} = x_i^k - \beta \nabla f_i(x_i^k)$$

$$k \bmod T = 0$$
$$x_i^{k+1} = \bar{x}^k$$

$$\bar{x}^k := \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

f_i are L-smooth and μ-strongly convex

$T = 1$: LGD = GD $\Rightarrow$ linear rate

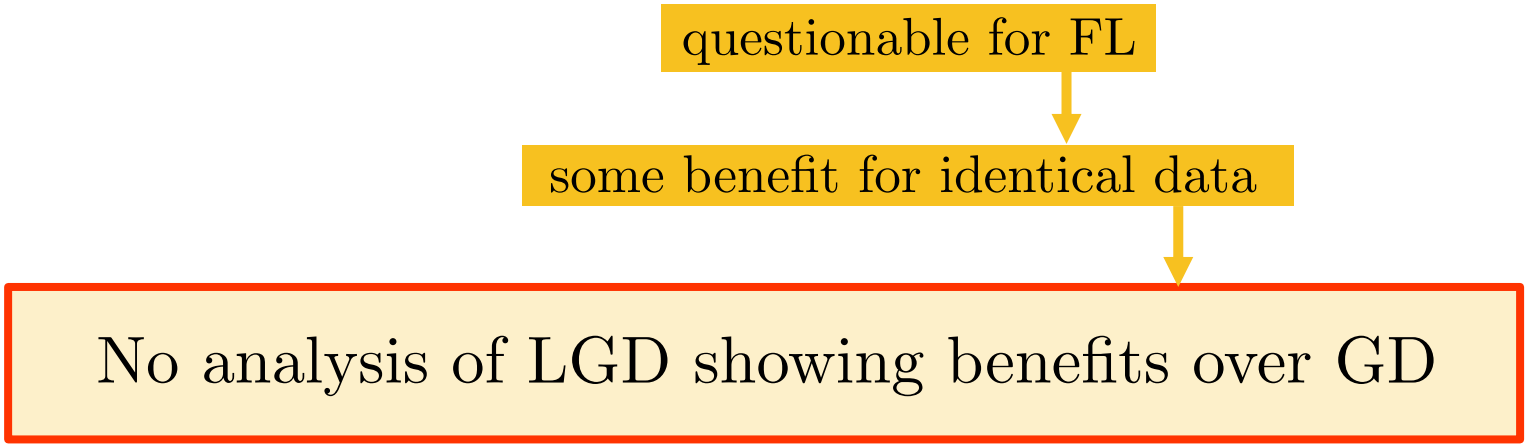
$T > 1$: sublinear rate of LGD

$T > 1$ + control variates

same communication as GD

T-times more computation

Issue of classical FL: FedAvg / Local SGD



No analysis of Local SGD showing benefits over minibatch SGD

Perhaps local methods are not meant to solve standard FL objective?

Minibatch vs Local SGD for Heterogeneous Distributed Learning
B Woodworth, K Patel, N Srebro 2020

Our new FL formulation

Each client i has their own model $x_i \in \mathbb{R}^d$

$$x = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{nd}$$

Regularization parameter

$$\lambda > 0$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \{F(x) := f(x) + \lambda \cdot \psi(x)\}$$

$$f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x_i)$$

$$\psi(x) := \frac{1}{2n} \sum_{i=1}^{n} \|x_i - \bar{x}\|^2$$

Allow different local models, penalize dissimilarity

Local GD works well here

Both issues fixed!

What we do not do

No DL applications or generalization

Local methods are well studied in practice

explain why local methods might work

Enhancements: control variates, acceleration – not meant for DL

We do not aim to solve standard FL formulation

2. Mixture FL formulation

Effect of parameter λ

$$f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x_i)$$

$$\psi(x) := \frac{1}{2n}\sum_{i=1}^{n} \|x_i - \bar{x}\|^2$$

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \{F(x) := f(x) + \lambda \cdot \psi(x)\}$$

Solution is a function of λ:

$$x(\lambda) = (x_1(\lambda), \ldots, x_n(\lambda)) \in \mathbb{R}^d \times \cdots \times \mathbb{R}^d = \mathbb{R}^{nd}$$

Purely local scenario

$\lambda = 0$: $\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x_i)$ ⇒ $x_i(0) = \arg\min_{z \in \mathbb{R}^d} f_i(z)$

no communication required!

Purely global scenario

$\lambda = \infty$: $\min_{z \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(z)$ ⇒ $x_i(\infty) = \arg\min_{z \in \mathbb{R}^d} \frac{1}{n}\sum_{j=1}^{n} f_j(z)$

$$x_i(\infty) = x_j(\infty) \quad \forall i, j$$

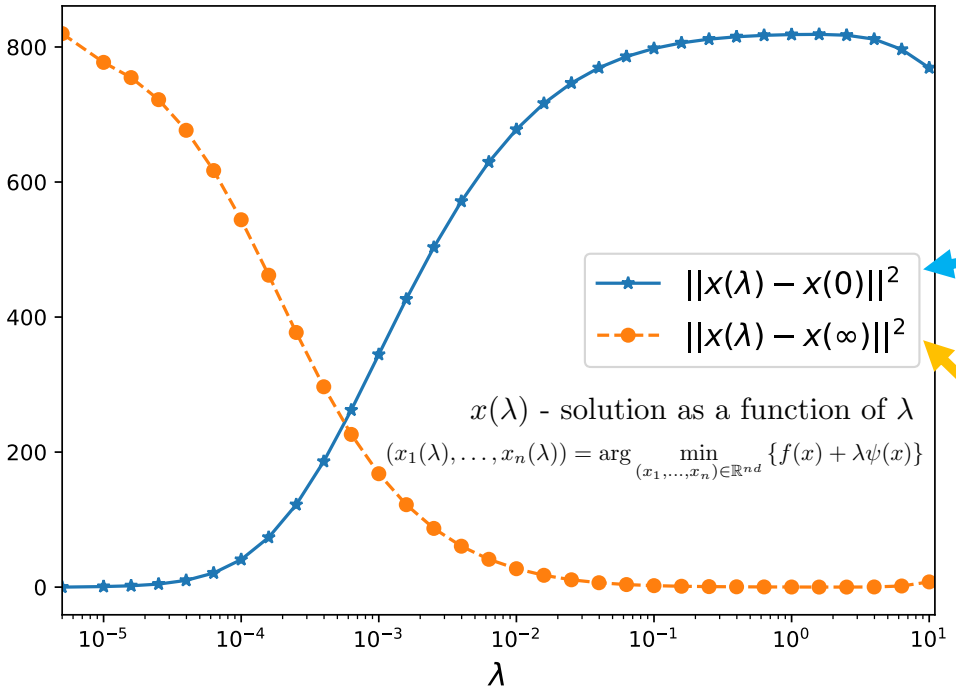
Effect of parameter λ

Purely local scenario ($\lambda = 0$)

$$\min_{x_1,\dots,x_n \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x_i)$$

Purely global scenario ($\lambda = +\infty$)

$$\min_{z \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(z)$$



Distance to $x(0)$
purely local solution

Distance to $x(\infty)$
purely global solution

Connection with MAML

Our global-local formulation:

$$(x_1(\lambda), \ldots, x_n(\lambda)) = \arg \min_{(x_1,\ldots,x_n)\in\mathbb{R}^{nd}} \{f(x) + \lambda\psi(x)\}$$

Optimality conditions:

$$x_i(\lambda) = \bar{x}(\lambda) - \tfrac{1}{\lambda}\nabla f_i\left(x_i(\lambda)\right)$$

Average of the personalized models:

$$\bar{x}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} x_i(\lambda)$$

MAML (Model Agnostic Meta Learning)

Stationary point, simplest scenario

$$z_i^* = z^* - \alpha\nabla f_i(z^*)$$

$$\nabla_{z^*} f_i(z_i^*(z^*)) = 0$$

$$z_i^*(z^*) = z^* - \alpha\nabla f_i(z^*)$$

Model-agnostic meta-learning for fast adaptation of deep networks
C Finn, P Abbeel, S Levine, 2017

3. FedAvg / Local SGD

Local GD as a special case of a cyclic GD

$$F(x) = f(x)+\lambda\psi(x) = \frac{1}{T+1}\left(\left(\sum_{i=1}^{T}\frac{T+1}{T}f(x)\right)+(T+1)\lambda\psi(x)\right) = \frac{1}{T+1}\sum_{i=1}^{T+1}\phi_i(x)$$

Apply cyclic GD

T times gradient step wrt f, then a single step wrt ψ

$$x_i^{k+1} = x_i^k - \beta\nabla f_i(x_i^k)$$

$$x_i^{k+1} = (1-\gamma)x_i^k + \gamma\bar{x}^k$$

$$(\nabla\psi(x))_i = \frac{1}{n}(x_i - \bar{x})$$

Step towards the average!

Local GD for a specific value of the stepsize

Local gradient methods

L2GD	random number of local steps
L2GD+	random number of local steps control variates
L2SGD+	random number of local steps control variates subsampling local objective

Simpler analysis
Better rate*

correct fixed point

local finite sum

Random Reshuffling: Simple Analysis with Vast Improvements
K Mishchenko, A Khaled, P Richtárik, 2020

Loopless Local GD (L2GD) = 2-sum SGD with importance sampling

$$F(x) = f(x) + \lambda\psi(x)$$

Apply non-uniform SGD to this 2-sum objective!

$$\mathbb{E}\left[G(x^k)\right] = \nabla F(x^k)$$

$$x^{k+1} = x^k - \alpha G(x^k)$$

$$G(x^k) := \begin{cases} \frac{\nabla f(x^k)}{1-p} & \text{with probability} \quad 1-p \\ \frac{\lambda \nabla \psi(x^k)}{p} & \text{with probability} \quad p \end{cases}$$

$$x_i^{k+1} = x_i^k - \beta \nabla f_i(x_i^k)$$

$$x_i^{k+1} = (1-\gamma)x_i^k + \gamma \bar{x}^k$$

Loopless Local GD (L2GD)

$$x^{k+1} = x^k - \alpha G(x^k)$$

$$G(x^k) := \begin{cases} \frac{\nabla f(x^k)}{1-p} & \text{with probability } 1-p \\ \frac{\lambda \nabla \psi(x^k)}{p} & \text{with probability } p \end{cases}$$

$$x_i^{k+1} = x_i^k - \beta \nabla f_i(x_i^k)$$

$$x_i^{k+1} = (1-\gamma) x_i^k + \gamma \bar{x}^k$$

step towards averaging
$\gamma \leq \frac{1}{2}$

geometric distribution

On average, $\frac{1-p}{p}$ local steps in between aggregations

On average, $p(1-p)k$ communications per k iterations

required only when switching

L2GD (Convergence)

2 differences wrt LGD

step towards averaging

random number of local steps

$$x^{k+1} = x^k - \alpha G(x^k)$$

$$G(x^k) := \begin{cases} \frac{\nabla f(x^k)}{1-p} & \text{with probability } 1-p \\ \frac{\lambda \nabla \psi(x^k)}{p} & \text{with probability } p \end{cases}$$

f_i is L-smooth

f_i is μ-strongly convex

f is $\frac{\mu}{n}$-strongly convex

$$\alpha \leq \frac{1}{2\mathcal{L}}$$

$$\mathcal{L} := \frac{1}{n} \max\left\{ \frac{L}{1-p}, \frac{\lambda}{p} \right\}$$

$$\mathbb{E}\left[\left\| x^k - x(\lambda) \right\|^2 \right] \leq \left(1 - \frac{\alpha\mu}{n}\right)^k \left\| x^0 - x(\lambda) \right\|^2 + \frac{2n\alpha\sigma^2}{\mu}$$

$$\sigma^2 := \frac{1}{n^2} \sum_{i=1}^{n} \left(\frac{1}{1-p} \left\| \nabla f_i\left(x_i(\lambda)\right) \right\|^2 + \frac{\lambda^2}{p} \left\| x_i\lambda - \bar{x}\lambda \right\|^2 \right)$$

SGD: General Analysis and Improved Rates

R Gower et al, 2019

L2GD+

$$x^{k+1} = x^k - \alpha G(x^k)$$

$$G(x^k) := \begin{cases} \frac{\nabla f(x^k)}{1-p} & \text{with probability} \quad 1-p \\ \frac{\lambda \nabla \psi(x^k)}{p} & \text{with probability} \quad p \end{cases}$$

+ control variates

$$\mathbb{E}\left[\left\|x^k - x(\lambda)\right\|^2\right] \leq \left(1 - \frac{\alpha\mu}{n}\right)^k \left\|x^0 - x(\lambda)\right\|^2 + \frac{2\alpha\sigma^2}{\mu}$$

Control variates

$$p^\star = \frac{\lambda}{L+\lambda}$$

Communication

$$\mathcal{O}\left(\frac{\min\{L,\lambda\}}{\mu} \log \frac{1}{\varepsilon}\right)$$

SCAFFOLD: Stochastic Controlled Averaging for Federated Learning
P Karimireddy et al, 2019

L2GD+ (Convergence)

$$x^{k+1} = x^k - \alpha G(x^k) \; + \text{control variates}$$

$$G(x^k) := \begin{cases} \frac{\nabla f(x^k)}{1-p} & \text{with probability} \quad 1-p \\ \frac{\lambda \nabla \psi(x^k)}{p} & \text{with probability} \quad p \end{cases}$$

$$p^\star = \frac{\lambda}{L+\lambda}$$

Communication

$$\mathcal{O}\left(\frac{\min\{L,\lambda\}}{\mu} \log \frac{1}{\varepsilon}\right)$$

$\lambda = 0$

0 communication!

Optimal # local steps: ∞

$\lambda = \mathcal{O}(L) \Rightarrow \mathcal{O}(1)$ local steps

The smaller λ the more local steps

L2SGD+

$$f_i(x_i) = \frac{1}{m}\sum_{j=1}^{m} \tilde{f}_{i,j}(x_i)$$

$$x_i^{k+1} = x_i^k - \alpha G_i(x_i^k)$$

+variance reduction

$$G_i(x_i^k) := \begin{cases} \frac{\frac{1}{n}\nabla \tilde{f}_{i,j}(x_i^k)}{1-p} & \text{w. p.} \quad 1-p \quad \text{for random } j \\ \frac{\lambda \nabla(\psi(x^k))_i}{p} & \text{w. p.} \quad p \end{cases}$$

$$p = \frac{4\lambda+\mu}{4\lambda+4\tilde{L}+(m+1)\mu}$$

Communication

$$\mathcal{O}\left(\frac{\min\{\tilde{L}+m\mu,\lambda\}}{\mu}\log\frac{1}{\varepsilon}\right)$$

$\tilde{f}_{i,j}$ is $\tilde{L}$ -smooth

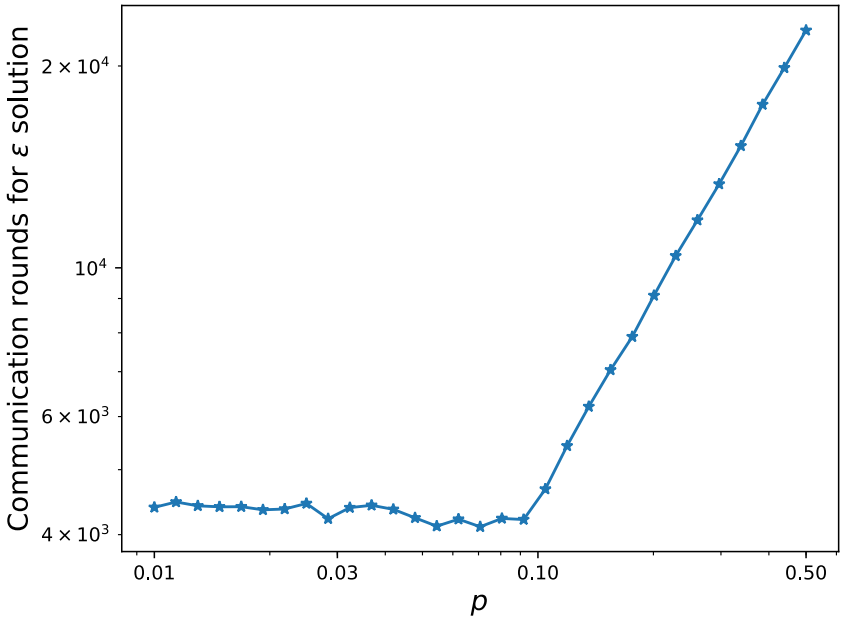
One Method to Rule Them All: Variance Reduction for Data, Parameters and Many New Methods
F Hanzely, P Richtárik, 2019

4. Experiments

L2SGD+: Usefulness of local steps

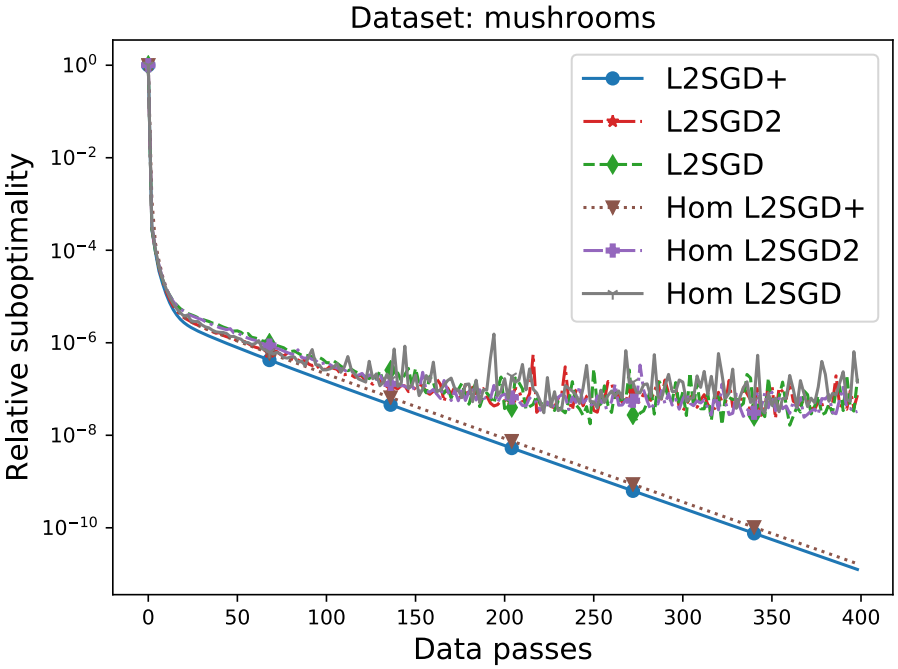
On average, $\frac{1-p}{p}$ local steps in between aggregations

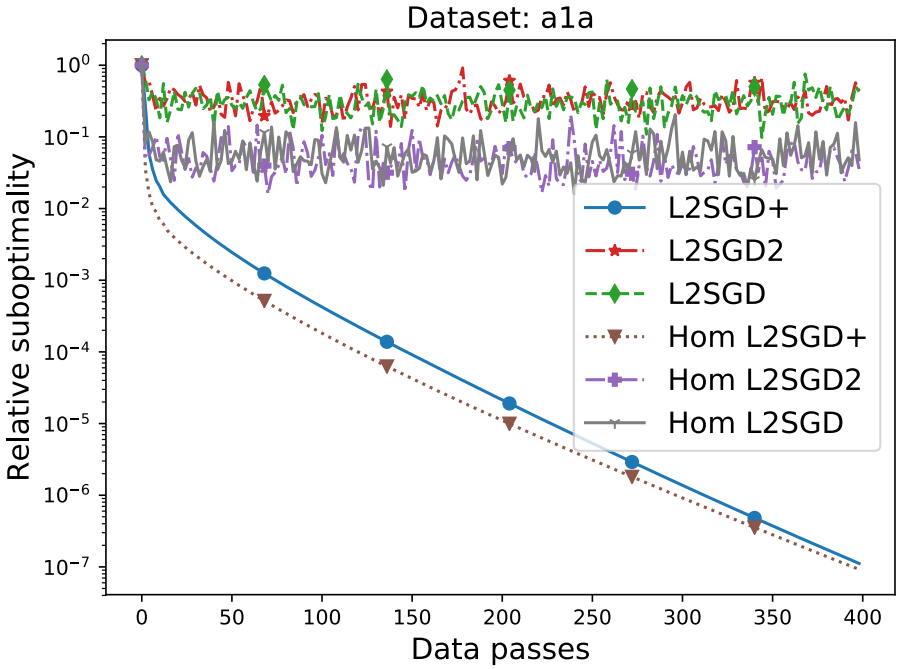
On average, $p(1-p)k$ communications per k iterations



Good communication for big number of local steps
Bad communication for small number of local steps

L2SGD+





L2SGD: no control variates at all

L2SGD2: control variates only for ψ, f (not for local finite sum)

Hom: data split uniformly

Bonus Material: Lower complexity bounds and optimal algorithms

Lower complexity bounds

$$f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x_i)$$

$$\psi(x) := \frac{1}{2n}\sum_{i=1}^{n} \|x_i - \bar{x}\|^2$$

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \{F(x) := f(x) + \lambda\psi(x)\}$$

Assume f_i is L-smooth and μ-strongly convex

Given L, μ, λ:

At least $\mathcal{O}\left(\sqrt{\frac{\min\{L,\lambda\}}{\mu}}\log\frac{1}{\varepsilon}\right)$ communications

local m-finite sum

At least $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$ local gradient calls

At least $\mathcal{O}\left(\sqrt{\frac{\min\{L,\lambda\}}{\mu}}\log\frac{1}{\varepsilon}\right)$ local prox calls

At least $\mathcal{O}\left(m + \sqrt{\frac{m\tilde{L}}{\mu}}\log\frac{1}{\varepsilon}\right)$ local stochastic gradients

Optimal algorithms at a glance

$$\text{Complexity} \in \{\text{Communication, Computation}\}$$

$$\times$$

$$\text{Local oracle} \in \{\text{Prox., Grad., Stoch. Grad.}\}$$

Optimal except [Computation, Stoch. Grad.]

Accelerated L2SGD+

APGD/Fista applied in 2 ways

Variant of FedProx

Accelerated local methods optimal for $\lambda < L$

Distributed Stochastic Multi-Task Learning with Graph Regularization
W Wang, J Wang, M Kolar, N Srebro, 2018

Summary

New FL formulation

$$\min_{x_1,\ldots,x_n \in \mathbb{R}^d} \{F(x) := f(x) + \lambda \psi(x)\}$$

$$f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x_i)$$

$$\psi(x) := \frac{1}{2n} \sum_{i=1}^{n} \|x_i - \bar{x}\|^2$$

Variants of Local SGD are beneficial on heterogeneous data

Lower complexity bounds ← communication, local computation

Discuss optimal algorithms

Locality helps for smaller λ