Federated Learning with Proximal Stochastic Variance Reduced Gradient Algorithms

Canh T. Dinh
The University of Sydney

Nguyen H. Tran
The University of Sydney

Tuan Dung Nguyen
The University of Melbourne

Wei Bao
The University of Sydney

Albert Y. Zomaya
The University of Sydney

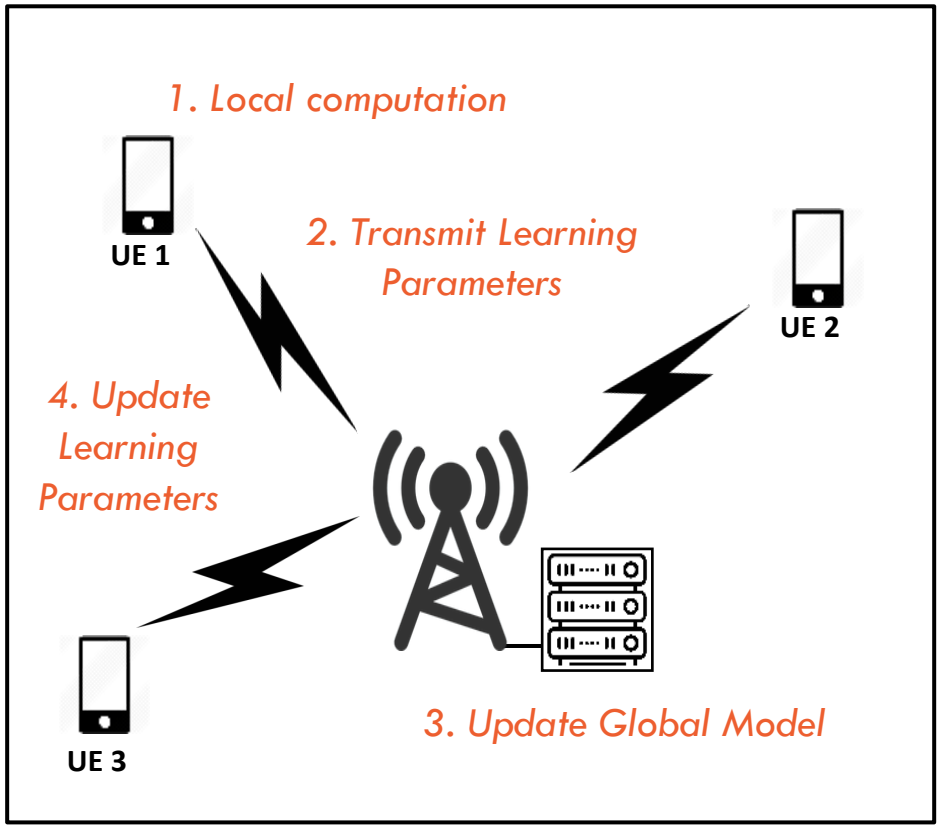
Bing B. Zhou
The University of Sydney

Outline

- Federated Learning
- System model
- Algorithm design
- Convergence analysis
- Experimental findings

Federated Learning* (FL)

- A fast-developing decentralized ML technique
- One global model, many local models in a network
- Pros: no need to send data, preserve privacy



Federated Learning Scheme.

* H. B. McMahan, E. Moore, D. Ramage, and S. Hampson, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, vol. 54, Fort Lauderdale, FL, USA, 2017

Challenges of FL

- **Systems heterogeneity**: differences in hardware (storage, computational power, connection) among users
- **Statistical heterogeneity**: devices' local data are non-identically distributed

Complicate algorithm design and convergence analysis

Our contributions

- FL algorithm using proximal stochastic variance reduced gradient (SVRG) method (FedProxVR): updating a model until some local accuracy threshold is achieved
- Convergence analysis: how to set the learning rate to achieve convergence
- Characterization of tradeoff between global and local convergence
- Method of minimizing the total training time

System Model

There are N users. Each user's dataset size is D_n. Total data size: $D = \sum_{n=1}^{N} D_n$.

Individual loss function on each device: $F_n(w) := \frac{1}{D_n} \sum_{i \in \mathcal{D}_n} f_i(w)$

Global minimization problem: $\min_{w \in \mathbb{R}^d} \bar{F}(w) := \sum_{n=1}^{N} \frac{D_n}{D} F_n(w).$

Assumptions:

$$\left\|\nabla f_i(w) - \nabla f_i(w')\right\| \leq L\left\|w - w'\right\| \tag{1}$$

$$F_n(w) + \left\langle \nabla F_n(w), w' - w \right\rangle \leq F_n(w') + \frac{\lambda}{2}\left\|w - w'\right\|^2 \tag{2}$$

$$\left\|\nabla F_n(w) - \nabla \bar{F}(w)\right\| \leq \sigma_n \left\|\nabla \bar{F}(w)\right\|. \tag{3}$$

Algorithm Design

ϵ-accurate solution: $\frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\left\|\nabla \bar{F}(\bar{w}^{(s)})\right\|^2 \leq \epsilon$

Local model update:

$$\min_{w\in\mathbb{R}^d}\left\{J_n(w) := F_n(w) + h_s(w)\right\}, \tag{1}$$

$$\text{where } h_s(w) := \frac{\mu}{2}\left\|w - \bar{w}^{(s-1)}\right\|^2, \tag{2}$$

(2) is a "soft" consensus constraint to penalize deviation from current global model

In each local iteration:

- Find a VR stochastic gradient estimator, $v_{n,s}^{(t)}$
- Update the parameters using the proximal operator
- Send the updated parameters to the server

Algorithm Design

Variance reduction stochastic gradient estimator:

SARAH $$v_{n,s}^{(t)} = \nabla f_{i_t}(w_{n,s}^{(t)}) - \nabla f_{i_t}(w_{n,s}^{(t-1)}) + v_{n,s}^{(t-1)}$$

SVRG $$v_{n,s}^{(t)} = \nabla f_{i_t}(w_{n,s}^{(t)}) - \nabla f_{i_t}(w_{n,s}^{(0)}) + v_{n,s}^{(0)}$$

SGD $$v_{n,s}^{(t)} = \nabla f_{i_t}(w_{n,s}^{(t)})$$

Find the proximal of the descent step, $w_{n,s}^{(t)} - \eta v_{n,s}^{(t)}$

Proximal operator: $$\text{prox}_{\eta h_s}(x) := arg\,min_{w \in \mathbb{R}^d} \left(h_s(w) + \frac{1}{2\eta} \|w - x\|^2 \right)$$

$$= \frac{\eta}{1 + \eta\mu} \left(\mu\, \bar{w}^{(s-1)} + \frac{1}{\eta} x \right)$$

Algorithm Design

Global iterations: $s = 1, \ldots, T$

Initial parameters

$$w_{n,s}^{(0)} = \bar{w}^{(s-1)}$$
$$v_{n,s}^{(0)} = \nabla F_n(w_{n,s}^{(0)})$$
$$w_{n,s}^{(1)} = \mathrm{prox}_{\eta h_s}(w_{n,s}^{(0)} - \eta v_{n,s}^{(0)})$$

$t = 1, \ldots, \tau$

Step 1. Randomly pick a batch

Step 2. Find $v_{n,s}^{(t)}$

Step 3. Update $w_{n,s}^{(t+1)} = \mathrm{prox}_{\eta h_s}(w_{n,s}^{(t)} - \eta v_{n,s}^{(t)})$,

Set $w_n^{(s)} = w_{n,s}^{(t')}$, where t' is chosen randomly in $1, \ldots, \tau$

Send $w_n^{(s)}$ to the server for aggregation

For all devices in parallel

Server averages all $w_n^{(s)}$ to find updated global model $\bar{w}^{(s)}$

Convergence Analysis

Lemma 1. *Device n achieves θ-accurate solution* (11) *if β and τ satisfy the following conditions*

a) when SARAH update (8a) *is used:*

$$0 \leq \frac{3(\beta^2 L^2 + \mu^2)}{\theta^2 \tilde{\mu} L(\beta - 3)} \leq \tau \leq \frac{5\beta^2 - 4\beta}{8} \tag{13}$$

b) when SVRG update (8b) *is used:*

$$0 \leq \frac{3(\beta^2 L^2 + \mu^2)}{\theta^2 \tilde{\mu} L(\beta - 3)} \leq \tau \leq \frac{5\beta^2 - 4\beta}{8\,a} - 2 \tag{14}$$

where there exists $a > 0$ such that $a - 4 \geq 4\sqrt{a(\tau + 1)}$.

The following remarks are about relations between the *local accuracy θ, number of local iterations τ, and step size parameter β.*

ϵ-accurate solution:

$$\frac{1}{T}\sum_{s=1}^{T} \mathbb{E}\left\|\nabla \bar{F}\left(\bar{w}^{(s)}\right)\right\|^2 \leq \epsilon$$

Defining the cost gap of an arbitrary point $\bar{w}^{(0)}$ by $\Delta(\bar{w}^{(0)}) := \mathbb{E}\left[\bar{F}(\bar{w}^{(0)}) - \bar{F}(\bar{w}^*)\right]$, we next provide the convergence condition for the global model update of FedProxVR.

Theorem 1. *Consider FedProxVR with all devices satisfying conditions in Lemma 1, we have*

$$\frac{1}{T}\sum_{s=1}^{T} \mathbb{E}\left\|\nabla \bar{F}(\bar{w}^{(s)})\right\|^2 \leq \frac{\Delta(\bar{w}^{(0)})}{\Theta\, T}. \tag{17}$$

where

$$\Theta = \frac{1}{\mu}\left(1 - \theta\sqrt{2(1+\bar{\sigma}^2)} - \frac{2L}{\tilde{\mu}}\sqrt{(1+\theta^2)(1+\bar{\sigma}^2)} - \frac{2L\mu}{\tilde{\mu}^2}(1+\theta^2)(1+\bar{\sigma}^2)\right) > 0.$$

Corollary 1. *The number of global iterations required to achieve ϵ-accurate solution to* (2) *is*

$$T \geq \frac{\Delta(\bar{w}^{(0)})}{\Theta\,\epsilon}, \tag{18}$$

Convergence Analysis

Defining the cost gap of an arbitrary point $\bar{w}^{(0)}$ by $\Delta(\bar{w}^{(0)}) := \mathbb{E}\left[\bar{F}(\bar{w}^{(0)}) - \bar{F}(\bar{w}^{*})\right]$, we next provide the convergence condition for the global model update of FedProxVR.

Theorem 1. *Consider FedProxVR with all devices satisfying conditions in Lemma 1, we have*

$$\frac{1}{T}\sum_{s=1}^{T}\mathbb{E}\left\|\nabla\bar{F}(\bar{w}^{(s)})\right\|^2 \leq \frac{\Delta(\bar{w}^{(0)})}{\Theta\, T}. \tag{17}$$

where

$$\Theta = \frac{1}{\mu}\Bigg(1 - \theta\sqrt{2(1+\bar{\sigma}^2)} - \frac{2L}{\tilde{\mu}}\sqrt{(1+\theta^2)(1+\bar{\sigma}^2)} - \frac{2L\mu}{\tilde{\mu}^2}(1+\theta^2)(1+\bar{\sigma}^2)\Bigg) > 0.$$

Corollary 1. *The number of global iterations required to achieve ϵ-accurate solution to (2) is*

$$T \geq \frac{\Delta(\bar{w}^{(0)})}{\Theta\,\epsilon}, \tag{18}$$

θ and μ and vital control "knobs" for convergence

To ensure $\Theta > 0$:

- μ should be large
- $\theta < \left(2(1-\bar{\sigma}^2)\right)^{-1/2}$

Comparison:

- SVRG/SARAH: $O(1/\epsilon)$
- FedProxVR: $O(1/\Theta\epsilon)$

Θ is called the *federated factor*, which determines the iterations of FedProxVR

Parameter Optimization

Denoting the device's computation (i.e., steps 7 and 8 in Algorithm 1) and communication delays to send local model updates to the server by d_{cmp} and d_{com}, respectively, the total training time of FedProxVR is as follows

$$\mathcal{T} := T(d_{com} + d_{cmp}\tau). \tag{19}$$

Defining a weight factor $\gamma := \frac{d_{cmp}}{d_{com}}$ and $T = \frac{\Delta(\bar{w}^{(0)})}{\Theta\,\epsilon}$, we minimize $\mathcal{T}$ with convergence conditions as constraints:

$$\underset{\mu,\theta,\beta,\tau}{\text{minimize}} \quad \frac{1}{\Theta}\left(1+\gamma\,\tau\right) \tag{20}$$

$$\text{subject to} \quad (15), (16), \text{ and } \Theta > 0. \tag{21}$$

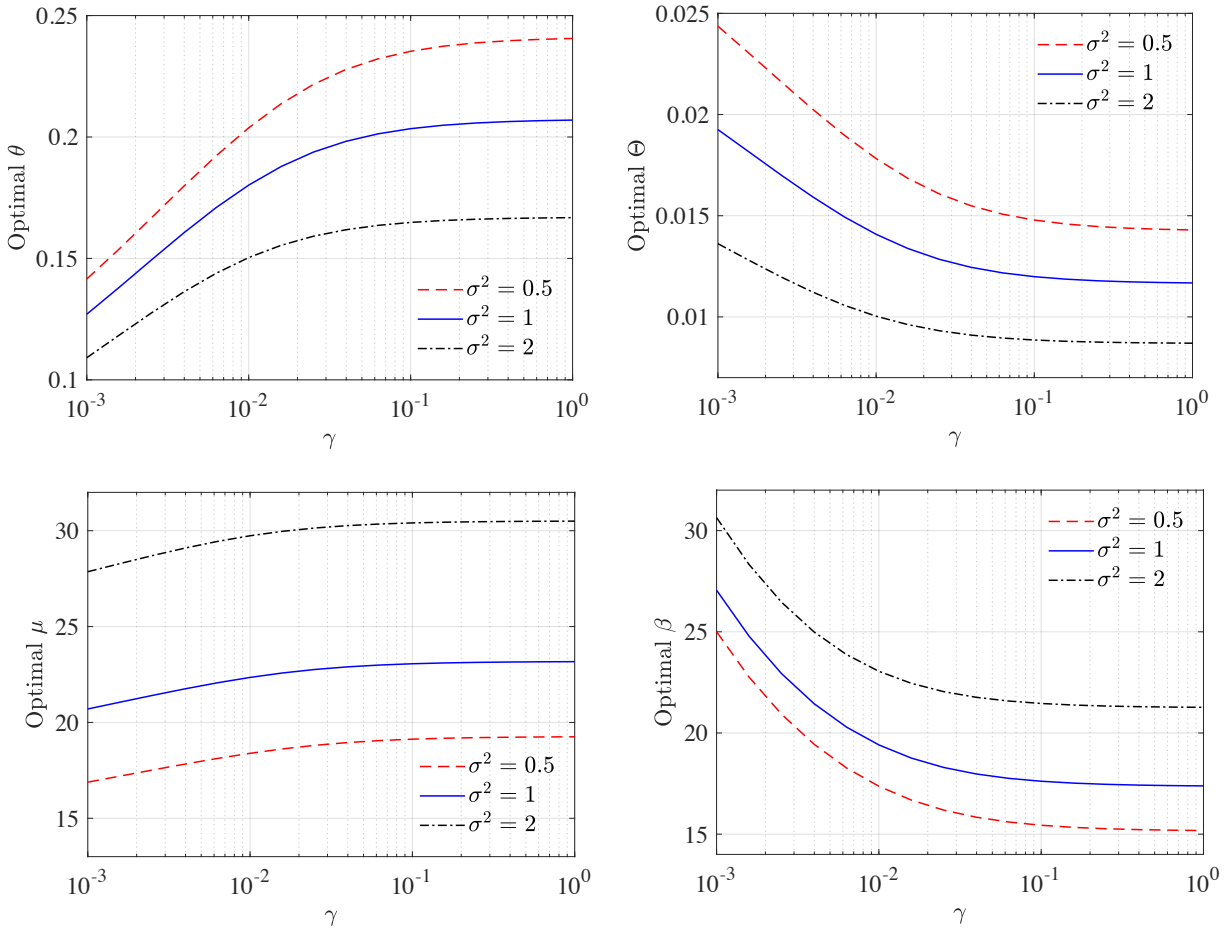
By removing constraint (15), (16) and substituting (with SARAH)

$$\theta^2 = \frac{24(\beta^2 L^2 + \mu^2)}{\tilde{\mu} L (5\beta^2 - 4\beta)(\beta - 3)} \tag{22}$$

into Θ, we further simplify this optimization problem as

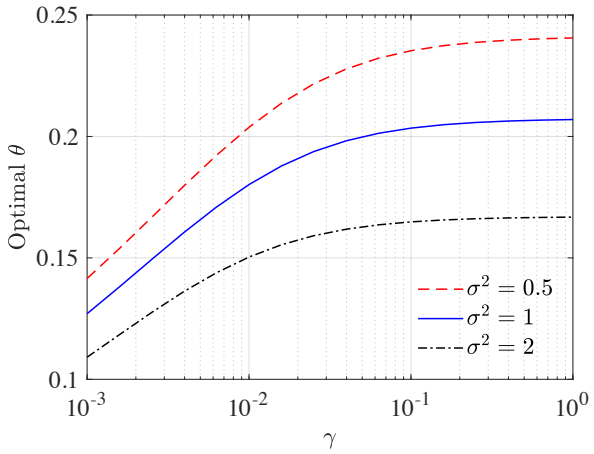
$$\underset{\mu,\beta}{\text{minimize}} \quad \frac{1}{\Theta}\left(1+\gamma\,\frac{5\beta^2-4\beta}{8}\right) \tag{23}$$

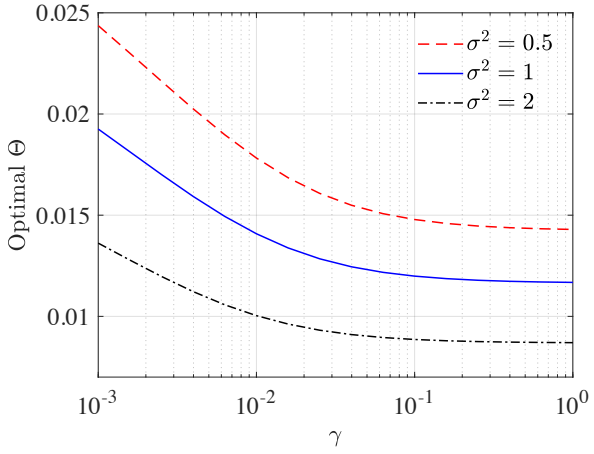
$$\text{subject to} \quad \beta > 3 \text{ and } \Theta > 0, \tag{24}$$

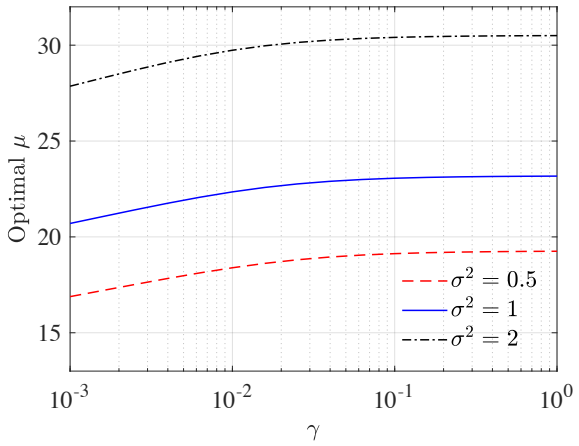


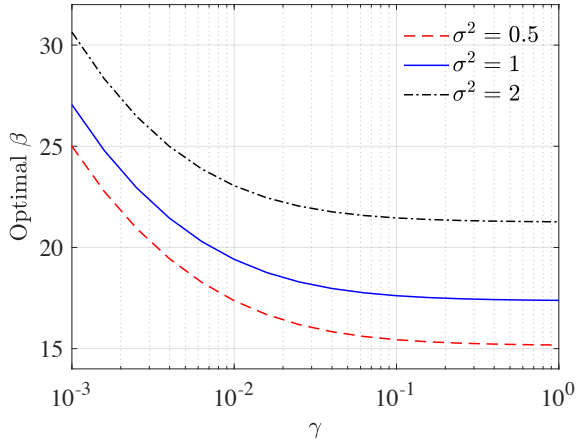
Solution using numerical methods

Parameter Optimization









When γ is small:

- Communication is more expensive
- Optimal β (thus τ) is large
- Devices are better off having more local computation than communication rounds

Large $\bar{\sigma}^2$ leads to

- higher β and μ
- lower θ and Θ

Devices will have to run more local iterations

Experiments

Datasets:

- Synthetic: logistic regression
- Real: MNIST and FASHION-MNIST
- 75% training, 25% test

Models: convex and non-convex

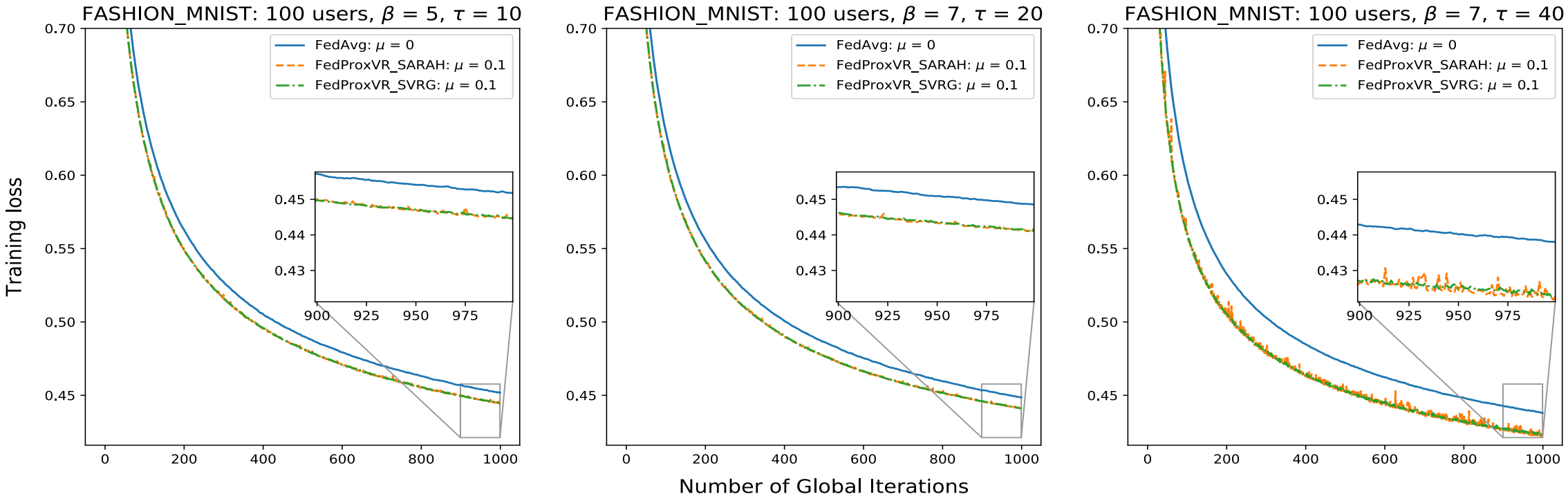
Federated setting:

- 100 users for convex task
- 10 users for non-convex task
- Data distributed by the power law*

*T. Li et al., “Federated Optimization in Heterogeneous Networks,” in Proceedings of the 3rd MLSys Conference, Austin, TX, USA, 2020.

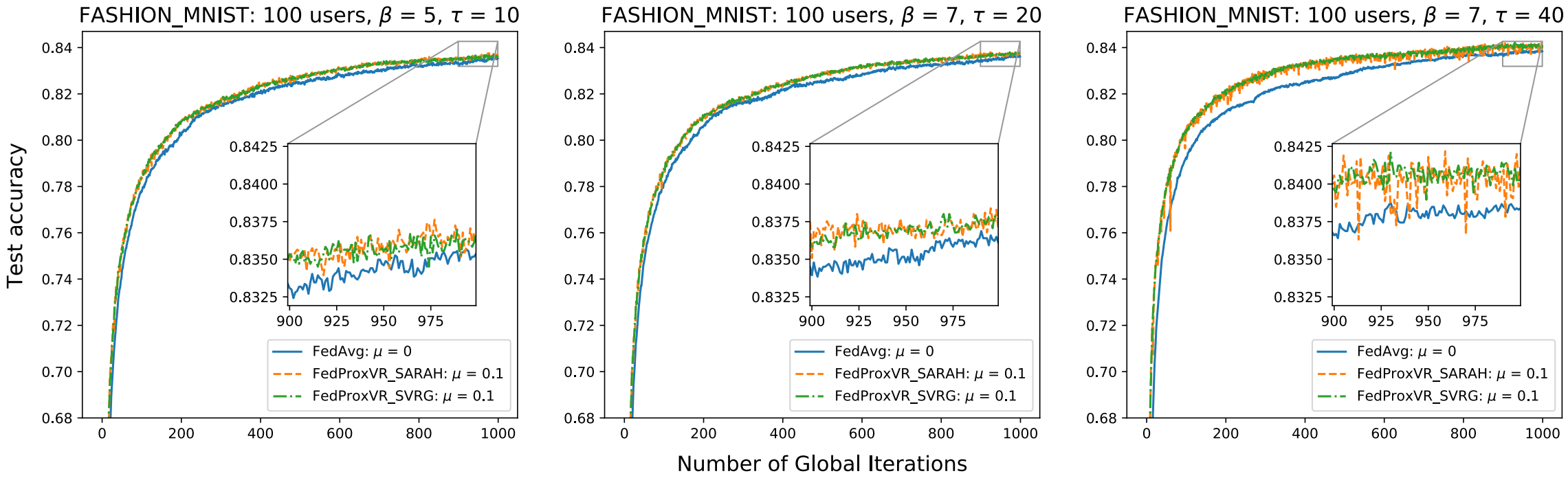
Experiments

Effects of step size parameter β and local iterations τ



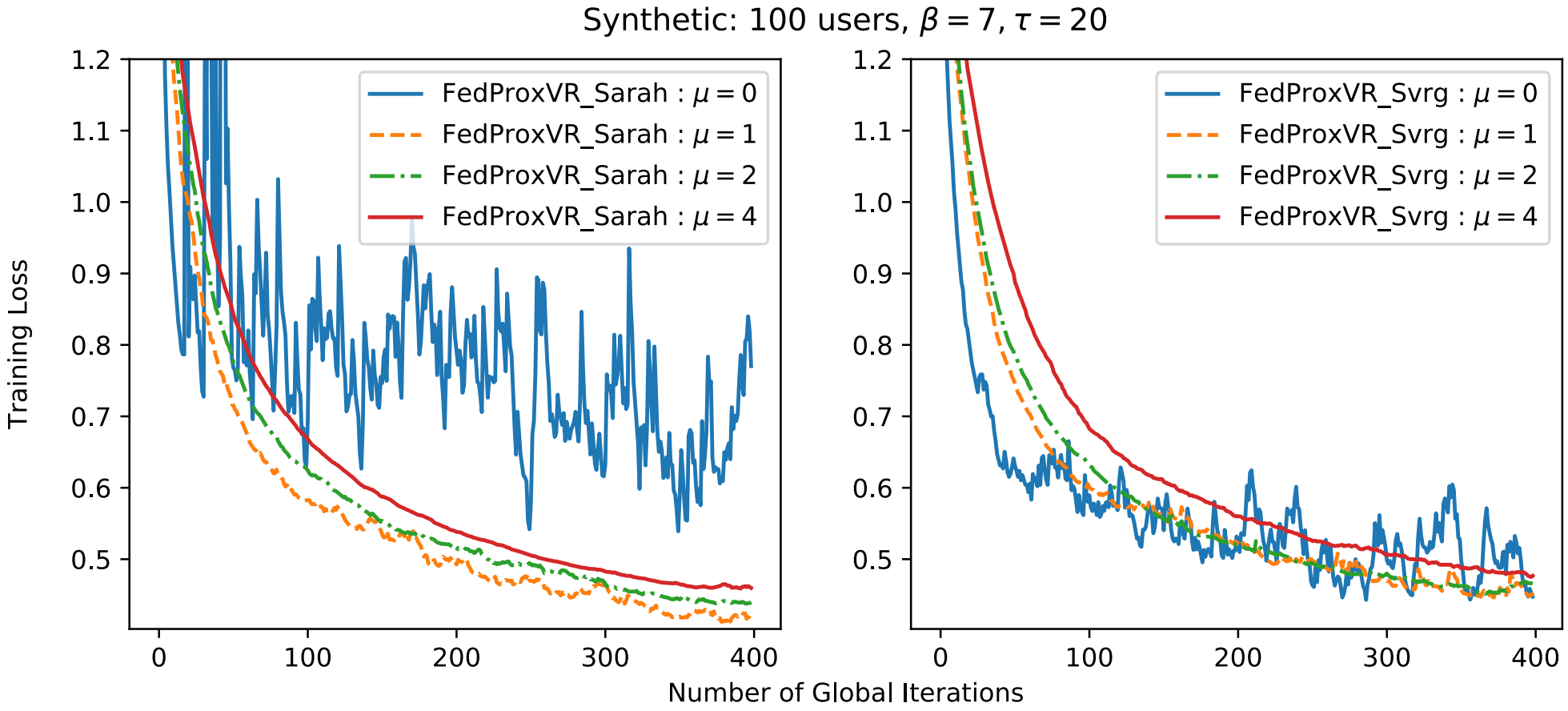
Experiments

Effects of step size parameter β and local iterations τ



Experiments

Effects of proximal penalty μ



Experiments

Comparisons with FedAvg

Table 1: Comparing the models' test accuracies using their best hyperparameters on a convex task.

Algorithm	τ	β	μ	B	T	Accuracy
FedAvg	10	10	0	16	983	**84.02%**
FedProxVR (SVRG)	20	10	0.1	32	895	**84.12%**
FedProxVR (SARAH)	20	5	0.1	32	965	**84.21%**

Table 2: Comparing the models' test accuracies using their best hyperparameters on a nonconvex task.

Algorithm	τ	β	μ	B	T	Accuracy
FedAvg	20	10	0	16	995	**93.52%**
FedProxVR (SVRG)	20	10	0.01	16	970	**94.06%**
FedProxVR (SARAH)	20	9	0.01	32	958	**93.75%**