Personalized Federated Learning: A Unified Framework and Universal Optimization Techniques

Filip Hanzely [1] Boxin Zhao [2] Mladen Kolar [2]

[1]Toyota Technological Institute at Chicago (TTIC) [2]Booth School of Business, The University of Chicago

Main takeaway

- Optimization theory (= simple algorithms + convergence guarantees) applicable to all strongly convex personalized FL objectives.
- Tight convergence rates despite the generality: Matching best-known rates from the literature (in all but one case). Novel guarantees for new objectives.

A unified Personalized FL objective

Optimization problem of interest:

$$\min_{w,\beta}\left\{F(w,\beta) := \frac{1}{M}\sum_{m=1}^{M} f_m(w,\beta_m)\right\}. \tag{1}$$

Notation: $w \in \mathbb{R}^{d_0}$: shared parameters, $\beta = (\beta_1,\dots,\beta_M)$, $\beta_m \in \mathbb{R}^{d_m}, \forall m \in [M]$: local parameters, M: number of devices, $f_m : \mathbb{R}^{d_0+d_m} \to \mathbb{R}$: objective (not necessarily the local loss) that depends on the local data at the m-th client.

Main idea: Choose $f_m(w,\beta_m)$ to recover a particular personalized FL objective as an instance of (1) and apply our optimization theory.

Detailed contributions

- **Universal (convex) optimization theory for personalized FL**. We propose three algorithms for solving the general personalized FL objective (1): i) Local Stochastic Gradient Descent for Personalized FL (**LSGD-PFL**), ii) Accelerated block Coordinate Descent for Personalized FL (**ACD-PFL**), and iii) Accelerated Stochastic Variance Reduced Coordinate Descent for Personalized FL (**ASVRCD-PFL**).
- **Convergence rates**. We provide lower complexity bounds for solving (1). ACD-PFL is always optimal in terms of the communication and local computation when the full gradients are available, while ASVRCD-PFL can be optimal either in terms of the number of evaluations of the w-stochastic gradient or the β-stochastic gradient.
- **Single personalized FL objective**. We propose a single objective (1) capable of recovering all the existing personalized FL approaches by carefully constructing the local loss $f_m(w,\beta_m)$. Surprisingly, the optimization guarantees for (1) yield fast convergence for individual special cases.
- **Personalization and communication complexity** Our theory conclude that the personalization has positive effect on the communication complexity of training FL models.
- **New personalized FL objectives** The universal personalized FL objective (1) enables us to obtain a range of novel personalized FL formulations as a special case.

Algorithms

LSGD-PFL: Mixture between Local SGD and SGD. Local SGD step is taken wrt to w-parameters, minibatch SGD step taken wrt to β-parameters. Convergence guarantees of LSGD recovered when $d_1 = d_2 = \cdots = d_M = 0$. Convergence guarantees of SGD recovered when $d_0 = 0$.

ACD-PFL: An instance of the accelerated block coordinate descent with carefully designed non-uniform sampling of coordinate blocks (w-variables or β-variables).

ASVRCD-PFL: ACD-PFL that subsamples the local finite sum combined with the variance reduction.

Optimization guarantees for solving (1)

Alg.	Communication	# ∇_w	# ∇_β
LSGD-PFL	$\frac{\max\left(L^\beta\tau^{-1}, L^w\right)}{\mu} + \frac{\sigma^2}{MB\tau\mu\epsilon} + \frac{1}{\mu}\sqrt{\frac{L^w(\zeta_*^2+\sigma^2B^{-1})}{\epsilon}}$	$\frac{\max\left(L^\beta, \tau L^w\right)}{\mu} + \frac{\sigma^2}{MB\mu\epsilon} + \frac{\tau}{\mu}\sqrt{\frac{L^w(\zeta_*^2+\sigma^2B^{-1})}{\epsilon}}$	$\frac{\max\left(L^\beta, \tau L^w\right)}{\mu} + \frac{\sigma^2}{MB\mu\epsilon} + \frac{\tau}{\mu}\sqrt{\frac{L^w(\zeta_*^2+\sigma^2B^{-1})}{\epsilon}}$
ACD-PFL	$\sqrt{L^w/\mu}$ ❀	$\sqrt{L^w/\mu}$ ❀	$\sqrt{L^\beta/\mu}$ ❀
ASVRCD-PFL	$n + \sqrt{n\mathcal{L}^w/\mu}$	$n + \sqrt{n\mathcal{L}^w/\mu}$ ❀	$n + \sqrt{n\mathcal{L}^\beta/\mu}$ ❀

Table 1. Complexity guarantees for solving (1) ignoring constant and log factors. Assumptions: F is $\mu-$strongly convex, f_m is convex and L^w smooth wrt w and L^β-smooth wrt β. Symbol ❀ indicates minimax optimal complexity. Local Stochastic Gradient Descent (LSGD): Local access to B-minibatch of stochastic gradients, each with σ^2-bounded variance. Each device takes $(\tau-1)$ local steps in between of the communication rounds. Accelerated Coordinate Descent (ACD): access to the full local gradient, yielding both the optimal communication complexity and the optimal computational complexity (both in terms of ∇_w and ∇_β). ASVRCD: Assuming that f_i is n-finite sum, the oracle provides an access to a single stochastic gradient with respect to that sum. The corresponding local computation is either optimal with respect to ∇_w or with respect to ∇_β. Achieving both optimal rates simultaneously remains an open problem.

Smoothness and strong convexity for special cases.

Objective / reference	μ	L^w	L^β	$\mathcal{L}^w$	$\mathcal{L}^\beta$	Rate?
Traditional	μ'	L'	0	$\mathcal{L}'$	0	recovered
Fully pers.	$\frac{\mu'}{M}$	0	$\frac{L'}{M}$	0	$\frac{\mathcal{L}'}{M}$	recovered
[5]	$\frac{\lambda}{2M}$	$\frac{\Lambda L'+\lambda}{2M}$	$\frac{L'+\lambda}{2M}$	$\frac{\Lambda\mathcal{L}'+\lambda}{2M}$	$\frac{\mathcal{L}'+\lambda}{2M}$	new♣
[7] [4]	$\frac{\mu'}{3M}$	$\frac{\lambda}{M}$	$\frac{L'+\lambda}{M}$	$\frac{\lambda}{M}$	$\frac{\mathcal{L}'+\lambda}{M}$	recovered♠
[3]	$\frac{\mu'(1-\alpha_{\max})^2}{M}$	$\frac{(\Lambda+\alpha_{\max}^2)L'}{M}$	$\frac{(1-\alpha_{\min})^2L'}{M}$	$\frac{(\Lambda+\alpha_{\max}^2)\mathcal{L}'}{M}$	$\frac{(1-\alpha_{\min})^2\mathcal{L}'}{M}$	new♣
[2] [6]	μ'	L'	L'	$\mathcal{L}'$	$\mathcal{L}'$	new
[1]	μ	L_R^w	L_R^β	$\mathcal{L}_R^w$	$\mathcal{L}_R^\beta$	new

Table 2. Smoothness and strong convexity parameters for personalized FL objectives as an instance of (1). ♣: Rate for novel personalized FL objective (extension of a known one). ♠: Best-known communication complexity recovered for $\lambda = \mathcal{O}(L')$. L' ($\mathcal{L}'$): smoothness of (components of) traditional FL objective, μ': strong convexity of the traditional FL.

References

[1] A. Agarwal, J. Langford, and C.-Y. Wei. Federated residual learning. *arXiv preprint arXiv:2003.12880*, 2020.

[2] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

[3] Y. Deng, M. M. Kamani, and M. Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

[4] F. Hanzely and P. Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

[5] T. Li, S. Hu, A. Beirami, and V. Smith. Federated multi-task learning for competing constraints. *arXiv preprint arXiv:2012.04221*, 2020.

[6] P. P. Liang, T. Liu, L. Ziyin, R. Salakhutdinov, and L.-P. Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

[7] C. T Dinh, N. Tran, and T. D. Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33, 2020.

Special cases

- Traditional FL: $\min_{w\in\mathbb{R}^d} F'(w) := \frac{1}{M}\sum_{m=1}^M f'_m(w)$,
- Fully personalized FL: $\min_{\beta_1,\dots,\beta_M\in\mathbb{R}^d} F_{full}(\beta) := \frac{1}{M}\sum_{m=1}^M f'_m(\beta_m)$.
- Multi-task personalized FL/implicit MAML [4, 7]:

$$\min_{w,\beta_1,\dots,\beta_M\in\mathbb{R}^d} F_{MX2}(w,\beta) := \frac{1}{M}\sum_{m=1}^M f'_m(\beta_m) + \frac{\lambda}{2M}\sum_{m=1}^M \|M^{-\frac{1}{2}}w - \beta_m\|^2.$$

- Multi-task FL [5] (generalization):

$$\min_{\beta_1,\dots,\beta_M\in\mathbb{R}^d} F_{MT2}(\beta) = \frac{1}{M}\sum_{i=1}^M\left(\Lambda f'_m(M^{-\frac{1}{2}}w) + f'_m(\beta_m) + \frac{\lambda}{2}\|\beta_m - M^{-\frac{1}{2}}w\|^2\right)$$

- Adaptive personalized FL [3] (generalization):

$$\min_{w,\beta}\frac{1}{M}\sum_{m=1}^M\left(\Lambda f'_m(M^{-\frac{1}{2}}w) + f'_m((1-\alpha_m)\beta_m + \alpha_m M^{-\frac{1}{2}}w)\right)$$

- Explicit parameter sharing [2, 6]: $\min_{w,\beta}\frac{1}{M}\sum_{m=1}^M f'_m(M^{-\frac{1}{2}}w,\beta_m)$.
- Federated residual learning [1]:

$$\min_{w,\beta} F_R(w,\beta) = \frac{1}{M}\sum_{i=1}^M l_m(A^w(w,x_m^w), A^\beta(\beta_m,x_m^\beta)),$$

Experiments

Setup: 3 personalized FL objectives, each applied to 3 datasets: MNIST, KMINIST, and FMINST. Model: multiclass logistic regression.

Goal of experiment: Demonstrate the effect of the $M^{-\frac{1}{2}}$ rescaling of the w-space.