PERSONALIZED FEDERATED LEARNING: A UNIFIED FRAMEWORK AND UNIVERSAL OPTIMIZATION TECHNIQUES

Filip Hanzely [*]
Toyota Technological Institute at Chicago
Chicago, IL, USA
filip@ttic.edu

Boxin Zhao [*] **& Mladen Kolar**
Booth School of Business
The University of Chicago
Chicago, IL, USA
boxinz@uchicago.edu
mkolar@chicagobooth.edu

ABSTRACT

We study the optimization aspects of personalized Federated Learning (FL). We develop a universal optimization theory applicable to all strongly convex personalized FL models in the literature. In particular, we propose a general personalized objective capable of recovering essentially any existing personalized FL objective as a special case. We design several optimization techniques to minimize the general objective, namely a tailored variant of Local SGD and variants of accelerated coordinate descent/accelerated SVRCD. We demonstrate the practicality and/or optimality of our methods both in terms of communication and local computation. Surprisingly enough, our general optimization theory is capable of recovering best-known communication and computation guarantees for solving specific personalized FL objectives.

1 INTRODUCTION

Federated Learning (FL) (McMahan et al., 2017; Kairouz et al., 2019) is a novel paradigm for training machine learning models on individual devices rather than revealing their data while communicating the model updates using private and secure protocols. The original goal of FL was to search for a single model to be deployed on all devices, which has been questioned recently. As the user data distribution can vary greatly across the devices, a single model might not serve all the devices simultaneously (Hard et al., 2018). Thus, data heterogeneity becomes the main challenge in the search for efficient federated learning models. Recently, a range of personalized FL approaches has been proposed to deal with data heterogeneity (Kulkarni et al., 2020), where different local models are used to fit user-specific data, but also capture the common knowledge distilled from data of other devices.

Since the motivation and the goal of each of these personalized approaches varies greatly; examining them separately can only provide us with an understanding of a given model. Fortunately, all personalized FL models from the literature are trained by minimizing a specifically structured optimization program. In this paper, we analyze the general properties of such an optimization program which in turn provides us with high-level principles for training personalized FL models. We aim to solve the following optimization problem

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

where $w \in \mathbb{R}^{d_0}$ corresponds to the shared parameters, $\beta = (\beta_1, \dots, \beta_M)$ with $\beta_m \in \mathbb{R}^{d_m}$, $\forall m \in [M]$ corresponds to the local parameters, M is the number of devices, and $f_m : \mathbb{R}^{d_0 + d_m} \to \mathbb{R}$ is the objective that depends on the local data at the m-th client.

[*] Equal contribution

By carefully designing the local loss $f_m(w, \beta_m)$, the objective (1) can recover essentially any existing personalized FL approach as a special case. Note that the local objective f_m does not need to correspond to the empirical loss of a given model on the m-th device's data. See Section 2 for details. Therefore, (1) serves as a unified framework that includes all existing personalized FL approaches as special cases. The main goal of our work is to explore the problem (1) from the optimization perspective. Doing so, we provide universal optimization theory that applies to essentially all personalized FL approaches.

A longer version of this paper is attached as the appendix. Due to space limitations, we omit some details and definitions of notations. The purpose of this paper is to highlight the main points of our work, and we encourage interested readers to read the appendix for complete treatment.

1.1 Contributions

We outline the main contributions of this work.

Single personalized FL objective. We propose a single objective (1) capable of recovering, to the best of our knowledge, all the existing personalized FL approaches by carefully constructing the local loss $f_m(w, \beta_m)$. Consequently, training different personalized FL models is equivalent to solving a particular instance of (1).

Recovering best-known complexity and novel guarantees. We develop a tight strongly convex optimization theory for solving (1). Our convergence theory covers both the communication and computation guarantees. Furthermore, the computational guarantees include both the complexity with respect to the number of w-gradients evaluated and the number of β-gradients evaluated. *Despite the generality of our approach, specializing our rates to the individual personalized FL objectives, we recover best-known optimization guarantees from the literature or advance over the state-of-the-art*[1] Therefore, our results often deem the optimization tailored to solve a specific personalized FL unnecessary.

Universal (convex) optimization theory for personalized FL. In order to develop optimization theory for solving (1), we impose particular assumptions on the objective: $\mu-$strong convexity of F and convexity and (L^w, ML^β)-smoothness of f_m for all $m \in [M]$ (see appendix for details). These assumptions are naturally satisfied for the vast majority of personalized FL objectives from the literature, with the exception of personalized FL approaches that are inherently nonconvex, such as MAML (Finn et al., 2017). Under these assumptions, 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). Our convergence theory covers both the communication and computation guarantees. Furthermore, the computational guarantees include both the complexity with respect to the number of w-gradients evaluated and the number of β-gradients evaluated, as presented in Table 1.

Minimax optimal rates. We provide lower complexity bounds for solving (1). Using the construction of Hendrikx et al. (2020), we show that to solve (1), one requires at least a certain number of communication rounds, a certain number of (stochastic) gradients with respect to w and a certain number of (stochastic) gradients with respect to β. Note that communication is often the bottleneck when training distributed and personalized FL models. Next, we show that 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.

Personalization and communication complexity. Given that a specific FL objective contains a parameter which determines the amount of personalization, we observe that the value of $\sqrt{L^w/\mu}$ is always non-increasing function of this parameter. Since the communication complexity of (1) is equal to $\sqrt{L^w/\mu}$ up to constant and log factors, we 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. While we study various (parametric)

[1] With a single exception: objective (11) of the appendix with $\lambda > L'$.

Table 1: Complexity guarantees of proposed methods when ignoring constant and log factors. ∇_w/∇_β : number of (stochastic) gradient calls with respect to the w/β-parameters. Symbol ❀ indicates minimax optimal complexity. Local Stochastic Gradient Descent (LSGD): Local access to B-minibatches 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.

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	$n+\sqrt{n\mathcal{L}^w/\mu}$	$n+\sqrt{n\mathcal{L}^w/\mu}$ ❀	$n+\sqrt{n\mathcal{L}^\beta/\mu}$ ❀

extensions of known models, we believe that the objective (1) can lead to easier development of brand new objectives too. We stress that proposing novel personalized FL models is not the main focus of our work, but rather a low-hanging fruit enabled by other contributions; the paper's main focus consists of providing universal optimization guarantees for personalized FL.

The price of generality. As we impose a very generic assumptions on the structure (1), one can not hope to recover the minimax optimal rates, that is, the rates that match the lower complexity bounds, for all individual personalized FL objectives as a special case of our general guarantees. Therefore, our convergence guarantees are optimal in the light of our assumptions only. Despite all of this, our general rates specialize surprisingly well for these objectives: our complexities are state-of-the-art in all of the scenarios with a single exception: the communication complexity of the mixture FL objective of Hanzely & Richtárik (2020).

2 Personalized FL Objectives

We recover a range of known personalized FL approaches as a special case of (1). In particular, we recover the traditional FL, fully personalized FL, multi-task FL of Li et al. (2020), Moreau envelope personalized FL (T Dinh et al., 2020), mixture FL objective (Hanzely & Richtárik, 2020), adaptive personalized FL (Deng et al., 2020), personalized FL with explicit weight sharing (Arivazhagan et al., 2019; Liang et al., 2020), federated residual learning (Agarwal et al., 2020), and MAML based approaches (Fallah et al., 2020).

Due to space limitations, we only give a quick glimpse of our results here. In particular, Table 2 presents the smoothness and strong convexity constants with respect to (1) for the special cases, these in turn determine the communication and computation complexity of our methods.

3 Algorithms

We briefly describe each of the three proposed algorithms. The complexities of these algoritums are summarized in Table 1, while more details can be found in the appendix.

LSGD-PFL. This algorithm is a mixture between Local SGD (LSGD) (McMahan et al., 2016; Stich, 2019) and SGD – one takes a local SGD step with respect to w-parameters, while taking a minibatch SGD step with respect to β-parameters. Admittedly, LSGD-PFL was already proposed by Arivazhagan et al. (2019) and Liang et al. (2020) to solve a particular instance of (1), however, no optimization guarantees were provided. In contrast, we provide convergence guarantees of LSGD-PFL that recover the convergence rate of LSGD when $d_1 = d_2 = \cdots = d_M = 0$ and the rate of SGD when $d_0 = 0$. Next, we demonstrate that LSGD-PFL works the best when applied to an objective with rescaled w-space, unlike what was proposed in the aforementioned papers.

Table 2: Smoothness and strong convexity parameters for personalized FL objectives as an instance od (1), with a note about the rate: we either recover the best known rate for given objective, or give a novel rate that is to the best of our knowledge best under given assumptions. ♣: Rate for novel personalized FL objective (extension of a known one). ♠: We recover best-known communication complexity only for $\lambda = \mathcal{O}(L')$. Parameter L' (or $\mathcal{L}'$) correspond to the smoothness of the (components of) traditional FL objective, while μ' corresponds to the strong convexity of the traditional FL.

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
Li et al. (2020)	$\frac{\lambda}{2M}$	$\frac{\Lambda L'+\lambda}{2M}$	$\frac{L'+\lambda}{2M}$	$\frac{\Lambda \mathcal{L}'+\lambda}{2M}$	$\frac{\mathcal{L}'+\lambda}{2M}$	new♣
T Dinh et al. (2020) Hanzely & Richtárik (2020)	$\frac{\mu'}{3M}$	$\frac{\lambda}{M}$	$\frac{L'+\lambda}{M}$	$\frac{\lambda}{M}$	$\frac{\mathcal{L}'+\lambda}{M}$	recovered♠
Deng et al. (2020)	$\frac{\mu'(1-\alpha_{\max})^2}{M}$	$\frac{(\Lambda+\alpha_{\max}^2)L'}{M}$	$\frac{(1-\alpha_{\min})^2 L'}{M}$	$\frac{(\Lambda+\alpha_{\max}^2)\mathcal{L}'}{M}$	$\frac{(1-\alpha_{\min})^2 \mathcal{L}'}{M}$	new♣
Arivazhagan et al. (2019) Liang et al. (2020)	μ'	L'	L'	$\mathcal{L}'$	$\mathcal{L}'$	new
Agarwal et al. (2020)	μ	L_R^w	L_R^β	$\mathcal{L}_R^w$	$\mathcal{L}_R^\beta$	new

ACD-PFL. The second method we propose is an instance of the accelerated block coordinate descent (Allen-Zhu et al., 2016; Nesterov & Stich, 2017) with a very specific non-uniform sampling of coordinate blocks corresponding to either the w-variables or β-variables.

ASVRCD-PFL. Lastly, we propose a carefully designed instance of ASVRCD (Hanzely et al., 2020). Besides subsampling the global and local parameters as ACD-PFL does, ASVRCD-PFL subsamples the local finite sum as well and employs variance reduction with respect to both sources of the randomness.

4 Experiments

We present two experiments to validate the theoretical contributions of our work. In the first experiment, we compare three different methods – LSGD-PFL, SCD-PFL (=ASVRCD-PFL without acceleration and without variance reduction) and SVRCD-PFL (=ASVRCD-PFL without acceleration) [2] across different datasets and objective functions. In the second experiment, we demonstrate the need for reparametrization of w-space for SVRCD-PFL.

Setup. We implemented three personalized FL objectives each applied to three different datasets: MNIST (LeCun & Cortes, 2010), KMINIST (Clanuwat et al., 2018), and FMINST (Xiao et al., 2017). As a model, we use a multiclass logistic regression (i.e., a single-layer fully connected neural network composed with softmax function and cross entropy loss). See appendix for details.

Comparison between different optimization methods. We compare the convergence of LSGD-PFL versus SCD-PFL and SVRCD-PFL. We plot the loss against the number of communication rounds for three methods across different objectives and datasets. For Local SGD we set the synchronization step to be 5, which means that all devices synchronize in every 5 iterations. For SCD and SVRCD, devices only synchronize when we update the global parameter. The result is presented in Figure 1. We see that the variants of coordinate descent outperform widely-used LSGD-PFL. The addition of variance reduction term helps slightly improve the performance.[3]

Effect of reparametrization in SVRCD. In this experiment, we demonstrate the importance of reparametrization of global parameter w (i.e., divided by $\sqrt{M}$). We run reparameterized and non-reparameterized SVRCD-PFL across different objectives and datasets. Figure 2 shows the result.

[2] We drop the term as the condition number rather small for the acceleration to matter.

[3] We expect a more significant improvement if a closer neighborhood of the optimum was reached.

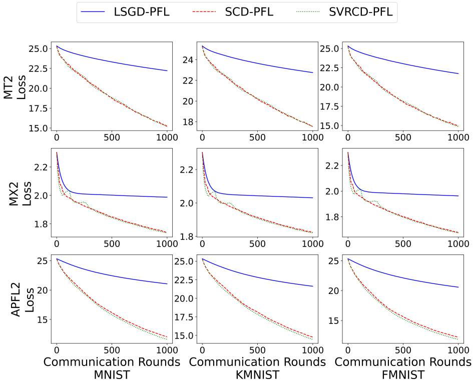

Figure 1: Comparison for three algorithms: LSGD-PFL, SCD-PFL, and SVRCD-PFL. Different rows correspond to different objective functions and columns correspond to different datasets.

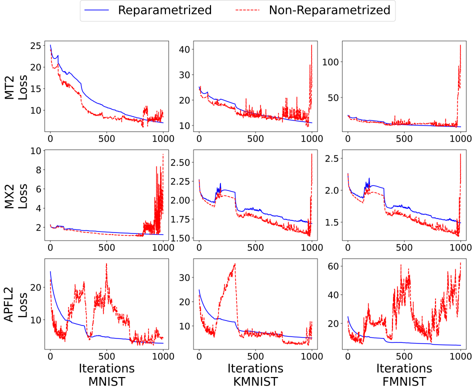

Figure 2: Effect of reparametrization of global space in SVRCD-PFL. Reparametrization helps SVRCD-PFL converge more smoothly, especially when it gets close to the optimum.

Indeed, we see that reparametrization improves the convergence of SVRCD-PFL. While the non-reparametrized variant might converge faster initially, soon enough, it becomes extremely unstable. This experiment confirms the necessity of reparametrization so that the scale of the learning rate is right for both global and local parameters.

REFERENCES

Alekh Agarwal, John Langford, and Chen-Yu Wei. Federated residual learning. *arXiv preprint arXiv:2003.12880*, 2020.

Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pp. 1110–1119, 2016.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature, 2018.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

Filip Hanzely, Dmitry Kovalev, and Peter Richtárik. Variance reduced coordinate descent with acceleration: New method with a surprising application to finite-sum problems. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4039–4048. PMLR, 13–18 Jul 2020.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Hadrien Hendrikx, Francis Bach, and Laurent Massoulie. An optimal algorithm for decentralized finite sum optimization. *arXiv preprint arXiv:2005.10675*, 2020.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pp. 794–797. IEEE, 2020.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Federated multi-task learning for competing constraints. *arXiv preprint arXiv:2012.04221*, 2020.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

H Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2016.

Yurii Nesterov and Sebastian U Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

Canh T Dinh, Nguyen Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017.

Personalized Federated Learning: A Unified Framework and Universal Optimization Techniques

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

Abstract

We study the optimization aspects of personalized Federated Learning (FL). We develop a universal optimization theory applicable to all strongly convex personalized FL models in the literature. In particular, we propose a general personalized objective capable of recovering essentially any existing personalized FL objective as a special case. We design several optimization techniques to minimize the general objective, namely a tailored variant of Local SGD and variants of accelerated coordinate descent/accelerated SVRCD. We demonstrate the practicality and/or optimality of our methods both in terms of communication and local computation. Surprisingly enough, our general optimization theory is capable of recovering best-known communication and computation guarantees for solving specific personalized FL objectives.

1. Introduction

Modern personal electronic devices such as mobile phones, wearable devices and home assistants can collectively generate and store vast amounts of user data. Such data are crucial for training and improving state-of-the-art machine learning models for tasks ranging from natural language processing to computer vision. Traditionally, the training process was performed by first collecting all the data into a datacenter (Dean et al., 2012), raising serious concerns about the user's privacy and bringing a huge burden on the storage ability of server suppliers. To address these issues, a novel paradigm – Federated Learning (FL) (McMahan et al., 2017; Kairouz et al., 2019) – has been proposed. Informally, the main idea of FL is to train a model locally on an individual's device instead of revealing their data while communicating the model updates using private and secure protocols.

While the original goal of FL was to search for a single model to be deployed on each device, such a goal has been questioned recently. As the user data distribution can vary greatly across the devices, a single model might not serve all the devices simultaneously (Hard et al., 2018). Thus, data heterogeneity becomes the main challenge in the search for efficient federated learning models. Recently, a range of personalized FL approaches has been proposed to deal with data heterogeneity (Kulkarni et al., 2020), where different local models are used to fit user-specific data, but also capture the common knowledge distilled from data of other devices.

Since the motivation and the goal of each of these personalized approaches varies greatly; examining them separately can only provide us with an understanding of a given model. Fortunately, all personalized FL models from the literature are trained by minimizing a specifically structured optimization program. In this paper, we analyze the general properties of such a optimization program which in turn provides us with high-level principles for training personalized FL models. We aim to solve the following optimization problem

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

where $w \in \mathbb{R}^{d_0}$ corresponds to the shared parameters, $\beta = (\beta_1, \ldots, \beta_M)$ with $\beta_m \in \mathbb{R}^{d_m}$, $\forall m \in [M]$ corresponds to the local parameters, M is the number of devices, and $f_m : \mathbb{R}^{d_0+d_m} \to \mathbb{R}$ is the objective that depends on the local data at the m-th client.

By carefully designing the local loss $f_m(w, \beta_m)$, the objective (1) can recover essentially any existing personalized FL approach as a special case. Note that the local objective f_m does not need to correspond to the empirical loss of a given model on the m-th device's data. See Section 2 for details. Therefore, (1) serves as a unified framework that includes all existing personalized FL approaches as special cases. The main goal of our work is to explore the problem (1) from the optimization perspective. Doing so, we provide universal optimization theory that applies to all personalized

[*] Equal contribution [1] Toyota Technological Institute at Chicago, Chicago, IL, USA [2] Booth School of Business, The University of Chicago, Chicago, IL, USA. Correspondence to: Filip Hanzely <filip@ttic.edu>, Boxin Zhao <boxinz@uchicago.edu>.

Preprint. Under review.

FL approaches.

1.1. Contributions

We outline the main contributions of this work.

Single personalized FL objective. We propose a single objective (1) capable of recovering, to the best of our knowledge, all the existing personalized FL approaches by carefully constructing the local loss $f_m(w, \beta_m)$. Consequently, training different personalized FL models is equivalent to solving a particular instance of (1).

Recovering best-known complexity and novel guarantees. We develop a tight strongly convex optimization theory for solving (1). Our convergence theory covers both the communication and computation guarantees. Furthermore, the computational guarantees include both the complexity with respect to the number of w-gradients evaluated and the number of β-gradients evaluated. *Despite the generality of our approach, specializing our rates to the individual personalized FL objectives, we recover best-known optimization guarantees from the literature or advance over the state-of-the-art* with a single exception: objective (11) with $\lambda > L'$. Therefore, our results often deem the optimization tailored to solve a specific personalized FL unnecessary.

Universal (convex) optimization theory for personalized FL. In order to develop optimization theory for solving (1), we impose particular assumptions on the objective: $\mu-$strong convexity of F and convexity and (L^w, ML^β)-smoothness of f_m for all $m \in [M]$ (see Assumptions 1.1, 1.2). These assumptions are naturally satisfied for the vast majority of personalized FL objectives from the literature, with the exception of personalized FL approaches that are inherently nonconvex, such as MAML (Finn et al., 2017). Under these assumptions, 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).

We briefly describe each of the three proposed algorithms.

LSGD-PFL. This algorithm is a mixture between Local SGD (LSGD) (McMahan et al., 2016; Stich, 2019) and SGD – one takes a local SGD step with respect to w-parameters, while taking a minibatch SGD step with respect to β-parameters. Admittedly, LSGD-PFL was already proposed by Arivazhagan et al. (2019) and Liang et al. (2020) to solve a particular instance of (1), however, no optimization guarantees were provided. In contrast, we provide convergence guarantees of LSGD-PFL that recover the convergence rate of LSGD when $d_1 = d_2 = \cdots = d_M = 0$ and the rate of SGD when $d_0 = 0$. Next, we demonstrate that LSGD-PFL works the best when applied to an objective with rescaled w-space, unlike what was proposed in the aforementioned papers.

ACD-PFL. The second method we propose is an instance of the accelerated block coordinate descent with non-uniform sampling (Allen-Zhu et al., 2016; Nesterov & Stich, 2017; Hanzely & Richtárik, 2019) that computes at each iteration the gradient with respect to w-parameters with probability $p_w = \frac{\sqrt{L^w}}{\sqrt{L^w}+\sqrt{L^\beta}}$ or the gradient with respect to β-parameters with probability $p_\beta = \frac{\sqrt{L^\beta}}{\sqrt{L^w}+\sqrt{L^\beta}}$.

ASVRCD-PFL. Lastly, we propose a carefully designed instance of ASVRCD (Hanzely et al., 2020b). Besides subsampling the global and local parameters as ACD-PFL does, ASVRCD-PFL subsamples the local finite sum as well and employs variance reduction with respect to both sources of the randomness.

Minimax optimal rates. We provide lower complexity bounds for solving (1). Using the construction of Hendrikx et al. (2020), we show that to solve (1), one requires at least $\mathcal{O}\left(\sqrt{L^w/\mu}\log \epsilon^{-1}\right)$ communication rounds. Note that communication is often the bottleneck when training distributed and personalized FL models. Furthermore, one needs at least $\mathcal{O}\left(\sqrt{L^w/\mu}\log \epsilon^{-1}\right)$ evaluations of $\nabla_w F$ and at least $\mathcal{O}\left(\sqrt{L^\beta/\mu}\log \epsilon^{-1}\right)$ evaluations of $\nabla_\beta F$. Given the n-finite sum structure of f_m with $(\mathcal{L}^w, M\mathcal{L}^\beta)$-smooth components, we show that one requires at least $\mathcal{O}\left(n + \sqrt{n\mathcal{L}^w/\mu}\log \epsilon^{-1}\right)$ stochastic gradient evaluations with respect to w-parameters and at least $\mathcal{O}\left(n + \sqrt{n\mathcal{L}^\beta/\mu}\log \epsilon^{-1}\right)$ stochastic gradient evaluations with respect to β-parameters. We show that 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.

Personalization and communication complexity. Given that a specific FL objective contains a parameter which determines the amount of personalization, we observe that the value of $\sqrt{L^w/\mu}$ is always non-increasing function of this parameter. Since the communication complexity of (1) is equal to $\sqrt{L^w/\mu}$ up to constant and log factors, we 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. While we study various (parametric) extensions of known models, we believe that the objective (1) can lead to easier development of brand new objectives too. We stress that proposing novel personalized FL models is not the main focus of our work, but rather a low-hanging fruit enabled by other contributions; the paper's main focus consists of providing universal optimization guarantees for personalized FL.

1.2. Assumptions and notations

Local Objective. We assume three different ways to access the local objective f_m. The first, and the most simple case, corresponds to having access to the full gradient of f_m with respect to either w or β_m for all $m \in [M]$ simultaneously. The second case corresponds to a situation where $f_m(w, \beta_m)$ is the expectation itself, i.e.,

$$f_m(w, \beta_m) = \mathbb{E}_{\xi \in \mathcal{D}_m} \left[\hat{f}_m(w, \beta_m, \xi) \right], \tag{2}$$

while having access to B stochastic gradients with respect to either w or β_m simultaneously for all $m \in M$. The third case corresponds to a finite sum f_m:

$$f_m(w, \beta_m) = \frac{1}{n} \sum_{i=1}^{n} f_{m,i}(w, \beta_m), \tag{3}$$

having an access to $\nabla_w f_{m,i}(w, \beta_m)$ or to $\nabla_\beta f_{m,i}(w, \beta_m)$ for all $m \in [M]$ and $i \in [n]$ selected uniformly at random.

Assumptions. We argue that the objective (1) is capable of recovering virtually any personalized FL objective. Since the structure of the individual personalized FL objectives varies greatly, it is important to impose reasonable assumptions on the problem (1) in order to obtain meaningful rates in the special cases.

Assumption 1.1. *Assume that function $F(w, \beta)$ is jointly μ-strongly convex for $\mu \geq 0$, while for all $m \in [M]$, function $f_m(w, \beta)$ is jointly convex, L^w-smooth w.r.t. parameter w and (ML^β)-smooth w.r.t. parameter β^m. In the case when $\mu = 0$, assume additionally that* (1) *has a unique solution.*

When f_m has a finite sum structure (3), we require the smoothness of the finite sum components as well.

Assumption 1.2. *Suppose that for all $m \in [M], i \in [n]$, function $f_{m,i}(w, \beta_m)$ is jointly convex, $\mathcal{L}^w$-smooth w.r.t. parameter w and $(M\mathcal{L}^\beta)$-smooth w.r.t. parameter β^m.*[1]

In Section 2 we justify Assumptions 1.1 and 1.2 and characterize the constants $\mu, L^w, L^\beta, \mathcal{L}^w, \mathcal{L}^\beta$ for special cases of (1). Table 2 provides a summary of what these parameters are for particular instances of (1) that we study.

The price of generality. Since Assumption 1.1 is the only structural assumption we impose on (1), one can not hope to recover the minimax optimal rates, that is, the rates that match the lower complexity bounds, for all individual personalized FL objectives as a special case of our general guarantees. Note that any given instance of (1) has a structure that is not covered by Assumption 1.1, but can be exploited by an optimization algorithm to improve either communication or local computation. Therefore, our convergence guarantees are optimal in the light of Assumption 1.1 only. Despite all of this, our general rates specialize surprisingly well for these objectives as we show in Section 2: our complexities are state-of-the-art in all of the scenarios with a single exception: the communication complexity of (11).

Individual treatment of w and β. Throughout this work, we allow different smoothness of the objective with respect to global parameters w and local parameters β. At the same time, our algorithm is allowed to exploit the separate access to gradients with respect to w and β, given that these gradients can be efficiently computed separately. Without such a distinction, one might not hope for the communication complexity better than $\Theta\left(\max\{L^w, \lambda\}/\mu \log \epsilon^{-1}\right)$, which is suboptimal in the special cases. Similarly, the computational guarantees would be suboptimal as well. See Section 2 for more details.

Data heterogeneity. We do not impose any similarity assumptions on different devices data. Our theory allows for an arbitrary dissimilarity among the individual clients.

2. Personalized FL Objectives

We recover a range of known personalized FL approaches as a special case of (1). In this section, we detail optimization challenges that arise in each one of the special cases. We discuss the relation to our results, particularly focusing on how Assumptions 1.1, 1.2 and our general rates (presented in Sections 3 and 4) behave in the special cases. Table 2 presents the smoothness and strong convexity constants with respect to (1) for the special cases, while Table C1 (in the appendix) provides the corresponding convergence rates for our methods when applied to these specific objectives.

Due to space limitations, we only present a subset of recovered personalized FL approaches here while the personalized FL with explicit weight sharing (Arivazhagan et al., 2019; Liang et al., 2020), federated residual learning (Agarwal et al., 2020), and MAML based approaches (Fallah et al., 2020) are discussed in the appendix.

[1] It is easy to see that $\mathcal{L}^w \geq L^w \geq \frac{\mathcal{L}^w}{n}$ and $\mathcal{L}^\beta \geq L^\beta \geq \frac{\mathcal{L}^\beta}{n}$.

Table 1. Complexity guarantees of proposed methods when ignoring constant and log factors. ∇_w / ∇_β : number of (stochastic) gradient calls with respect to the w/β-parameters. Symbol ❀ indicates minimax optimal complexity. Local Stochastic Gradient Descent (LSGD): Local access to B-minibatches 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.

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

Table 2. Parameters in Assumptions 1.1 and 1.2 for personalized FL objectives, with a note about the rate: we either recover the best known rate for given objective, or give a novel rate that is to the best of our knowledge best under given assumptions. ♣: Rate for novel personalized FL objective (extension of a known one).

Objective	μ	L^w	L^β	$\mathcal{L}^w$	$\mathcal{L}^\beta$	Rate?
(4)	μ'	L'	0	$\mathcal{L}'$	0	recovered
(5)	$\frac{\mu'}{M}$	0	$\frac{L'}{M}$	0	$\frac{\mathcal{L}'}{M}$	recovered
(8)	$\frac{\lambda}{2M}$	$\frac{\Lambda L'+\lambda}{2M}$	$\frac{L'+\lambda}{2M}$	$\frac{\Lambda\mathcal{L}'+\lambda}{2M}$	$\frac{\mathcal{L}'+\lambda}{2M}$	new♣
(11)	$\frac{\mu'}{3M}$	$\frac{\lambda}{M}$	$\frac{L'+\lambda}{M}$	$\frac{\lambda}{M}$	$\frac{\mathcal{L}'+\lambda}{M}$	recovered for $\lambda = \mathcal{O}(L')$
(14)	$\frac{\mu'(1-\alpha_{\max})^2}{M}$	$\frac{(\Lambda+\alpha_{\max}^2)L'}{M}$	$\frac{(1-\alpha_{\min})^2 L'}{M}$	$\frac{(\Lambda+\alpha_{\max}^2)\mathcal{L}'}{M}$	$\frac{(1-\alpha_{\min})^2\mathcal{L}'}{M}$	new♣
(16)	μ'	L'	L'	$\mathcal{L}'$	$\mathcal{L}'$	new
(18)	μ	L_R^w	L_R^β	$\mathcal{L}_R^w$	$\mathcal{L}_R^\beta$	new

2.1. Traditional FL

The traditional, non-personalized FL objective (McMahan et al., 2017) is given as

$$\min_{w\in\mathbb{R}^d} F'(w) := \frac{1}{M}\sum_{m=1}^{M} f'_m(w), \tag{4}$$

where f'_m corresponds to the loss on the m-th client's data. To properly compare the convergence rates, let us assume that f'_m is L'-smooth and $\mu'-$ strongly convex for all $m \in [M]$. The FL objective (4) is a special case of (1) with $d_1 = \cdots = d_M = 0$. It was shown that the minimax optimal communication to solve (4) up to ϵ-neighborhood of the optimum is $\tilde{\Theta}\left(\sqrt{L'/\mu'}\log\epsilon^{-1}\right)$ (Scaman et al., 2018). When $f'_m = \frac{1}{n}\sum_{j=1}^{n} f'_{m,j}(w)$ is a $n-$finite sum with convex and $\mathcal{L}'-$smooth components, the minimax optimal local stochastic gradient complexity is $\tilde{\Theta}\left(\left(n+\sqrt{n\mathcal{L}'/\mu}\right)\log\epsilon^{-1}\right)$ (Hendrikx et al., 2020). Both rates are recovered by our theory.

2.2. Fully personalized FL

At the other end of the spectrum lies the fully personalized FL where the m-th client trains their own model without any influence from other clients:

$$\min_{\beta_1,\dots,\beta_M\in\mathbb{R}^d} F_{full}(\beta) := \frac{1}{M}\sum_{m=1}^{M} f'_m(\beta_m). \tag{5}$$

The above objective is a special case of (1) with $d_0 = 0$. As the objective is separable in $\beta_1, \dots, \beta_M$, we do not require any communication to train it. At the same time, we need $\tilde{\Theta}\left(\left(n+\sqrt{n\mathcal{L}'/\mu}\right)\log\epsilon^{-1}\right)$ local stochastic oracle calls to solve it (Lan & Zhou, 2018) – which is what our algorithms achieve.

2.3. Multi-task FL of Li et al. (2020)

The objective is given as

$$\min_{\beta_1,\dots,\beta_M\in\mathbb{R}^d} F_{MT}(\beta) = \frac{1}{M}\sum_{i=1}^{M}\left(f'_m(\beta_m) + \frac{\lambda}{2}\|\beta_m - (w')^*\|^2\right) \tag{6}$$

where $(w')^*$ is a solution of the traditional FL in (4) and $\lambda \geq 0$. Assuming that $(w')^*$ is known (which Li et al. (2020) does), the problem (6) is a particular instance of (5); thus our approach achieves the optimal complexities.

A more challenging objective (in terms of the optimization) is the following relaxed version of (6):

$$\min_{w,\beta} \frac{1}{M}\sum_{m=1}^{M}\left(\Lambda f'_m(w) + f'_m(\beta_m) + \lambda\|w-\beta_m\|^2\right), \quad (7)$$

where $\Lambda \geq 0$ is the relaxation parameter, recovering the original objective for $\Lambda \to \infty$.

Next, we scale the global parameter w by a factor of $M^{-\frac{1}{2}}$ in order to obtain the right smoothness/strong convexity parameter (according to Assumption 1.1), arriving at the following objective:

$$\min_{w,\beta_1,\dots,\beta_M\in\mathbb{R}^d} F_{MT2}(w,\beta) = \frac{1}{M}\sum_{i=1}^{M} f_m(w,\beta_m), \quad (8)$$

where

$$f_m(w,\beta_m) = \Lambda f'_m(M^{-\frac{1}{2}}w)+f'_m(\beta_m)+\frac{\lambda}{2}\|\beta_m - M^{-\frac{1}{2}}w\|^2.$$

The next lemma determines parameters $\mu, L^w, L^\beta, \mathcal{L}^w, \mathcal{L}^\beta$ in Assumption 1.1.

Lemma 2.1. *Let $\Lambda \geq 3\lambda/(2\mu')$. Then, the objective (8) is jointly $(\lambda/(2M))$ $-$strongly convex, while f_m is jointly convex, $((\Lambda L' + \lambda)/M)$-smooth with respect to w and $(L'+\lambda)$-smooth with respect to β_m. Similarly, $f_{m,j}$ is jointly convex, $((\Lambda\mathcal{L}' + \lambda)/M)$-smooth with respect to w and $(\mathcal{L}'+\lambda)$-smooth with respect to β_m.*

Evaluating gradients. The nice thing about objective (8) is that evaluating $\nabla_w f_m(x,\beta_m)$ can be perfectly decoupled from evaluating $\nabla_\beta f_m(x,\beta_m)$ and vice versa. Therefore, we can make a full use of our theory and take advantage of different complexities with respect to ∇_w and ∇_β.

The resulting communication and computation complexities of solving (8) are presented in Table C1.

2.4. Multi-task personalized FL and implicit MAML

In its simplest form, the multi-task personalized objective (Smith et al., 2017; Wang et al., 2018) is given as (Hanzely & Richtárik, 2020)

$$\min_{\beta_1,\dots,\beta_M\in\mathbb{R}^d} F_{MX}(\beta) = \frac{1}{M}\sum_{m=1}^{M} f'_m(\beta_m) + \frac{\lambda}{2M}\sum_{m=1}^{M}\|\bar{\beta}-\beta_m\|^2 \quad (9)$$

where $\bar{\beta} := \frac{1}{M}\sum_{m=1}^{M}\beta_m$ and $\lambda \geq 0$.

On the other hand, the goal of implicit MAML (Rajeswaran et al., 2019; T Dinh et al., 2020) is to minimize

$$\min_{w\in\mathbb{R}^d} F_{ME}(w) = \frac{1}{M}\sum_{i=1}^{M}\left(\min_{\beta_m\in\mathbb{R}^d}\left(f'_m(\beta_m) + \frac{\lambda}{2}\|w-\beta_m\|^2\right)\right) \quad (10)$$

where $\lambda \geq 0$.

While we can not recover (9) or (10) in its exact form (1), we can recover objective which is simultaneously equivalent to both of them. In particular, by setting $f_m(w,\beta_m) = f'_m(\beta_m) + \lambda\|M^{-\frac{1}{2}}w - \beta_m\|^2$, the objective (1) becomes

$$\min_{w,\beta_1,\dots,\beta_M\in\mathbb{R}^d} F_{MX2}(w,\beta) := \frac{1}{M}\sum_{m=1}^{M} f'_m(\beta_m) + \frac{\lambda}{2M}\sum_{m=1}^{M}\|M^{-\frac{1}{2}}w - \beta_m\|^2. \quad (11)$$

It is a simple exercise to notice the equivalence of (11) to both (10) and (9).[2] Indeed, we can always minimize (11) in w arriving at $w^* = M^{\frac{1}{2}}\bar{\beta}$ and thus recovering the solution of (9). Similarly, by minimizing (11) in β we arrive at (10).

Next, we establish parameters in Assumptions 1.1 and 1.2.

Lemma 2.2. *Let $\mu' \leq \lambda/2$. Then, the objective (11) is jointly $(\mu'/(3M))$- strongly convex, while f_m is (λ/M)-smooth with respect to w and $(L'+\lambda)$-smooth with respect to β. Furthermore, function $f_{m,i}(w,\beta_m) = f'_{m,i}(\beta_m) + (\lambda/2)\|M^{-\frac{1}{2}}w - \beta_m\|^2$ is jointly convex, (λ/M)-smooth with respect to w and $(\mathcal{L}'+\lambda)$-smooth with respect to β.*

Hanzely et al. (2020a) showed that the minimax optimal communication complexity to solve (9) (and therefore to solve (10) and (11)) is $\Theta\left(\sqrt{\min(L',\lambda)/\mu'}\log\epsilon^{-1}\right)$. Furthermore, they showed that the minimax optimal number of gradients with respect to f' is $\tilde{\Theta}\left(\left(\sqrt{L'/\mu'}\right)\log\epsilon^{-1}\right)$ and proposed a method with $\Theta\left(\left(n + \sqrt{n(\mathcal{L}'+\lambda)/\mu'}\right)\log\epsilon^{-1}\right)$ complexity with respect to the number of $f'_{m,j}$-gradients. We match all of the aforementioned guarantees when $\lambda = \mathcal{O}(L')$. Furthermore, when $\lambda = \mathcal{O}(L')$, our complexity guarantees are strictly better when compared to guarantees for solving the implicit MAML objective (10) directly (Rajeswaran et al., 2019; T Dinh et al., 2020).

2.5. Adaptive personalized FL (Deng et al., 2020)

The objective is given as

$$\min_{\beta_1,\dots,\beta_M} F_{APFL}(\beta) = \frac{1}{M}\sum_{i=1}^{M} f'_m((1-\alpha_m)\beta_m + \alpha_m(w'^*)), \quad (12)$$

where $(w')^* = \text{argmin}_{w\in\mathbb{R}^d} F'(w)$ is a solution to (4) and $0 < \alpha_1, \dots \alpha_M < 1$.

Similar to (6), assuming that $(w')^*$ is known (which Deng et al. (2020) does), the problem (6) is an instance of (5); thus our approach achieves the optimal complexities.

[2] To the best of our knowledge, we are the first to notice the equivalence of (9) and (10).

Again, a more interesting case (in terms of the optimization) is when considering a relaxed variant of (12)

$$\min_{w,\beta} \frac{1}{M}\sum_{m=1}^{M} \left(\Lambda f'_m(w) + f'_m((1-\alpha_m)\beta_m + \alpha_m w)\right) \tag{13}$$

where $\Lambda \geq 0$ is the relaxation parameter which allows recovering the original objective when $\Lambda \to \infty$.

Such a choice, alongside with the usual rescaling of the parameter w results in the following objective:

$$\min_{w,\beta_1,\dots,\beta_M \in \mathbb{R}^d} F_{APFL2}(w,\beta) := \frac{1}{M}\sum_{i=1}^{M} f(w,\beta_m), \tag{14}$$

where

$$f(w,\beta_m) = \Lambda f'_m(M^{-\frac{1}{2}}w) + f'_m((1-\alpha_m)\beta_m + \alpha_m M^{-\frac{1}{2}}w).$$

Lemma 2.3. *Suppose that* $\Lambda \geq \max_{1\leq m\leq M}(3\alpha_m^2 + (1-\alpha_m)^2/2)$ *and define* $\alpha_{\min} := \min_{1\leq m\leq M}\alpha_m$, $\alpha_{\max} := \max_{1\leq m\leq M}\alpha_m$. *Then, function* F_{APFL2} *is jointly* $\left(\mu'(1-\alpha_{\max})^2/M\right)$*-strongly convex,* $\left((\Lambda+\alpha_{\max}^2)L'/M\right)$*-smooth with respect to* w *and* $\left((1-\alpha_{\min})^2 L'/M\right)$*-smooth with respect to* β.

3. Local SGD

Since the most popular optimizer to train non-personalized FL models is Local SGD/FedAvg (McMahan et al., 2016; Stich, 2019), we devise a local SGD variant tailored to solve personalized FL (1) – LSGD-PFL. Specifically, LSGD-PFL can be seen as a local SGD applied on global parameters w combined with a SGD applied on local parameters β. In order to mimic the classical setup of local SGD for non-personalized FL, we assume an access to the local objective $f_m(w,\beta_m)$ in the form of an unbiased stochastic gradient $\nabla \hat{f}_m(w,\beta_m,\zeta); \mathbb{E}\left[\hat{f}_m(w,\beta_m,\zeta)\right] = f_m(w,\beta_m)$ with bounded variance.

Assumption 3.1. *Assume that stochastic gradients* $\nabla_w \hat{f}_m(w,\beta_m,\zeta)$*,* $\nabla_\beta \hat{f}_m(w,\beta_m,\zeta)$ *satisfy for all* $m \in [M]$*,* $w \in \mathbb{R}^{d_0}$*,* $\beta_m \in \mathbb{R}^{d_m}$*:*

$$\begin{aligned} \mathbb{E}\left[\|\nabla_w \hat{f}_m(w,\beta_m,\zeta) - \nabla_w f_m(w,\beta_m)\|^2\right] &\leq \quad \sigma^2, \\ \mathbb{E}\left[\|\nabla_\beta \hat{f}_m(w,\beta_m,\zeta) - \nabla_\beta f_m(w,\beta_m)\|^2\right] &\leq \ M\sigma^2. \end{aligned}$$

Next, we state the convergence rate of LSGD-PFL.

Algorithm 1 LSGD-PFL

```
input Stepsize $\eta \in \mathbb{R}$, starting point $w^0 \in \mathbb{R}^{d_0}$, $\beta_m^0 \in \mathbb{R}^{d_m}$
  for all $m \in [M]$, communication period $\tau$.
  for $k = 0, 1, 2, \dots$ do
    if $k \bmod \tau = 0$ then
      Send all $w_m^k$'s to server, let $w^k = \frac{1}{M}\sum_{m=1}^M w_m^k$
      Send $w^k$ to each device, set $w_m^k = w^k$, $\forall m \in [M]$
    end if
    for $m = 1, 2, \dots, M$ in parallel do
      Sample $\xi_{1,m}^k, \dots \xi_{B,m}^k \sim \mathcal{D}_m$ independently
      Compute $g_m^k = \frac{1}{B}\sum_{j=1}^B \nabla \hat{f}_m(w_m^k, \beta_m^k; \xi_{j,m}^k)$
      Update the iterates $(w_m^{k+1}, \beta_m^{k+1}) = (w_m^k, \beta_m^k) - \eta \cdot g_m^k$
    end for
  end for
```

Theorem 3.1. *Let* $\zeta_*^2 := \frac{1}{M}\sum_{m=1}^{M}\|\nabla f_m(w^*,\beta^*)\|^2$ *be the data heterogeneity parameter at the optimum. Iteration complexity of Algorithm 1 to achieve* $\mathbb{E}\left[f(\overline{w}^K, \overline{\beta}^K)\right] - f(w^*,\beta^*) \leq \epsilon$ *is*

$$\tilde{O}\left(\frac{\max\left(L^\beta, \tau L^w\right)}{\mu} + \frac{\sigma^2}{MB\mu\epsilon} + \frac{\tau}{\mu}\sqrt{\frac{L^w(\zeta_*^2 + \sigma^2 B^{-1})}{\epsilon}}\right)$$

when $\mu > 0$ *and*

$$\tilde{O}\left(\frac{\max\left(L^\beta, \tau L^w\right)}{\epsilon} + \frac{\sigma^2}{MB\epsilon^2} + \frac{\tau\sqrt{L^w(\zeta_*^2 + \sigma^2 B^{-1})}}{\epsilon^{\frac{3}{2}}}\right)$$

when $\mu = 0$.

The iteration complexity of LSGD-PFL can be seen as a sum of two complexities – the complexity of minibatch SGD to minimize a problem with a condition number L^β/μ and complexity of local SGD to minimize a problem with a condition number L^w/μ. Note that the key reason why we were able to obtain such a rate of LSGD-PFL is the rescaling of w-space by constant $M^{-\frac{1}{2}}$. Earlier works that first introduced LSGD-PFL (without optimization guarantees) did not consider such a reparametrization.

Non-convex theory. To demonstrate that our approach is applicable in the non-convex setup (i.e., MAML based approaches), we provide a non-convex convergence guarantees of LSGD-PFL in the Appendix. Note that we do not argue about any form of optimality in the non-convex case.

4. Minimax optimal methods

We discuss the complexity of solving (1) in terms of the number of communication rounds to reach ϵ-solution and the amount of local computation – both in terms of the number of (stochastic) gradients with respect to global w-parameters and local β-parameters. Let us start with the lower complexity bounds.

4.1. Lower complexity bounds

We provide lower complexity bounds for solving (1) given that f_m is of a finite sum structure (3). Informally, we show that any algorithm with access to the communication oracle, local (stochastic) gradient oracle with respect to the global parameters w and local (stochastic) gradient oracle with respect to the local parameters β requires at least a certain number of oracle calls to approximately solve (1).

Oracle. The considered oracle allows us at any iteration to compute either • $\nabla_w f_{m,i}(w_m, \beta_m)$ on each device for a randomly selected $i \in [n]$ and any $w_m \in \mathbb{R}^{d_0}, \beta_m \in \mathbb{R}^{d_m}$, or • $\nabla_\beta f_{m,i}(w_m, \beta_m)$ on each device for a randomly selected $i \in [n]$ and any $w_m \in \mathbb{R}^{d_0}$, $\beta_m \in \mathbb{R}^{d_m}$, or • average of w_m's alongside with broadcasting the average back to clients, i.e., the communication step.

Our lower bound is provided for iterative algorithms whose iterates lie in the span of historical oracle queries only – let us denote such a class of algorithms as $\mathcal{A}$ (see technical details in Appendix E.1). While such a restriction is widespread in the classical optimization literature (Nesterov et al., 2018; Scaman et al., 2018; Hendrikx et al., 2020; Hanzely et al., 2020a), it can be avoided by more complex arguments (Nemirovskij & Yudin, 1983; Woodworth & Srebro, 2016; Woodworth et al., 2018).

Theorem 4.1. *Let F satisfy Assumptions 1.1 and 1.2. Then, any algorithm from the class $\mathcal{A}$ requires at least $\Omega(\sqrt{L^w/\mu}\log\epsilon^{-1})$ communication rounds, $\Omega\left(n + \sqrt{n\mathcal{L}^w/\mu}\log\epsilon^{-1}\right)$ calls to ∇_w-oracle and $\Omega\left(n + \sqrt{n\mathcal{L}^\beta/\mu}\log\epsilon^{-1}\right)$ calls to ∇_β-oracle to reach ϵ-solution.*

In the special case where $n = 1$, Theorem 4.1 provides a lower complexity bounds for solving (1) having an access to the full gradient locally.

4.2. Accelerated Coordinate Descent for PFL

We apply an Accelerated block Coordinate Descent (ACD) algorithm (Allen-Zhu et al., 2016; Hanzely & Richtárik, 2019) to solve (1). We separate the domain into two blocks of coordinates to sample from: the first one corresponding to w parameters and the second one corresponding to $\beta = [\beta_1, \beta_2, \ldots, \beta_M]$. Specifically, at every iteration, we toss an unfair coin. With probability $p_w = \sqrt{L^w}/(\sqrt{L^w} + \sqrt{L^\beta})$, we compute $\nabla_w F(w, \beta)$ and update block w. Alternatively, with probability $p_\beta = 1 - p_w$, we compute $\nabla_\beta F(w, \beta)$ and update block β. Plugging the described sampling of coordinate blocks into ACD (Allen-Zhu et al., 2016), [3] we arrive at Algorithm 2 (See appendix).

Next, we give the optimization guarantees for Algorithm 2.

Theorem 4.2. *Suppose that Assumption 1.1 holds. Let $\nu = \frac{\mu}{(\sqrt{L^w}+\sqrt{L^\beta})^2}$, $\theta = \frac{\sqrt{\nu^2+4\nu}-\nu}{2}$ and $\eta = \theta^{-1}$. Iteration complexity of ACD-PFL is $\mathcal{O}\left(\sqrt{(L^w + L^\beta)/\mu}\log\epsilon^{-1}\right)$.*

Since $\nabla_w F(w, \beta)$ is evaluated on average once every $1/p_w$ iterations only, ACD-PFL requires $\mathcal{O}\left(\sqrt{L^w/\mu}\log\epsilon^{-1}\right)$ communication rounds and $\mathcal{O}\left(\sqrt{L^w/\mu}\log\epsilon^{-1}\right)$ gradient evaluations with respect to w, thus matching the lower bound. Similarly, as $\nabla_\beta F(w, \beta)$ is evaluated on average once every $1/p_\beta$ iterations, we require $\mathcal{O}\left(\sqrt{L^\beta/\mu}\log\epsilon^{-1}\right)$ evaluations of $\nabla_\beta F(w, \beta)$ to reach an ϵ-solution; again matching the lower bound. Consequently, ACD-PFL is (minimax) optimal in terms of all three quantities of interest simultaneously.

Remark 4.1. *We are not the first to propose a variant of the coordinate descent (Nesterov, 2012) for personalized FL. Wu et al. (2020) introduced block coordinate descent to solve a variant of (11) formulated over a network. However, they do not argue about any form of optimality for their approach, which is also less general as it only covers a single personalized FL objective.*

4.3. Accelerated SVRCD for PFL

Despite being minimax optimal, the main drawbacks of ACD-PFL is the necessity of having an access to the full gradient of local loss f_m with respect to either w or β at each iteration. Specifically, computing the full gradient with respect to f_m might be very expensive when f_m is a finite sum (3). Ideally, one would desire to have an algorithm that is i) subsampling the global/local variables w and β just as ACD-PFL, ii) subsampling the local finite sum, iii) employing control variates to reduce the variance of the local stochastic gradient (Johnson & Zhang, 2013; Defazio et al., 2014), and iv) accelerated in the sense of (Nesterov, 1983).

We propose a method – ASVRCD-PFL – that satisfies all four conditions above. ASVRCD-PFL is a carefully designed instance of ASVRCD (Accelerated proximal Stochastic Variance Reduced Coordinate Descent) (Hanzely et al., 2020b) applied to solve (1) written in a rather non-intuitive form.[4] Specifically, we lift the problem (1) into a larger

[3] Admittedly, ACD from (Allen-Zhu et al., 2016) only allows for subsampling individual coordinates and does not allow for "blocks". A variant of ACD (Allen-Zhu et al., 2016) that provides the right convergence guarantees for block sampling was proposed in Nesterov & Stich (2017) and Hanzely & Richtárik (2019).

[4] We are not aware of any other algorithm capable of satisfying i)-iv) simultaneously and achieving a fast convergence rate.

space as follows:

$$\begin{aligned}&\min_{w\in\mathbb{R}_0^d,\beta_m\in\mathbb{R}^{d_m},\forall m\in[M]} F(w,\beta)\\ =&\min_{\substack{X,\, X[1,:,:,:]\,\in\,\mathbb{R}^{M\times n\times d_0}\\ X[2,m,:,:]\,\in\,\mathbb{R}^{n\times d_m},\,\forall m\,\in\, M}} \left\{\mathbf{P}(X) := \mathbf{F}(X)+\psi(X)\right\},\end{aligned}$$

where

$$\mathbf{F}(X) := \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{n}\sum_{j=1}^{n} f_{m,j}(X[1,m,j],X[2,m,j])\right)$$

and

$$\psi(X) := \begin{cases} 0 & \begin{array}{l}\text{if } m,m'\in[M], j,j'\in[n]:\\ X[1,m,j]=X[1,m',j'],\\ X[2,m,j]=X[2,m,j']\end{array}\\ \infty & \text{otherwise.}\end{cases}$$

We apply ASVRCD with a very carefully chosen (non-uniform) sampling of coordinate blocks to minimize $\mathbf{P}(X)$. We detail the algorithm in Section E.3 of the appendix and provide convergence guarantees here.

Theorem 4.3. *Let Assumptions 1.1 and 1.2 hold. Then, iteration complexity of ASVRCD-PFL is*

$$\mathcal{O}\left(\left(\rho^{-1}+\sqrt{(\mathcal{L}^w+\mathcal{L}^\beta)/(\rho\mu)}\right)\log\epsilon^{-1}\right),$$

where ρ is the frequency of updating the control variates. Setting $\rho = \mathcal{L}^w/((\mathcal{L}^w+\mathcal{L}^\beta)n)$ yields both the communication complexity and the local stochastic gradient complexity with respect to w-parameters of order $\mathcal{O}\left(\left(n+\sqrt{n\mathcal{L}^w/\mu}\right)\log\epsilon^{-1}\right)$. Analogously, setting $\rho = \mathcal{L}^\beta/((\mathcal{L}^w+\mathcal{L}^\beta)n)$ yields the local stochastic gradient complexity with respect to β-parameters of order $\mathcal{O}\left(\left(n+\sqrt{n\mathcal{L}^\beta/\mu}\right)\log\epsilon^{-1}\right)$.

Contrasting with Theorem 4.1, we show that ASVRCD-PFL can be optimal in terms of the local computation either with respect to β-variables or in terms of the w-variables. Unfortunately, these bounds are not achieved simultaneously unless $\mathcal{L}^w, \mathcal{L}^\beta$ are of a similar order.

5. Experiments

We present two experiments to validate the theoretical contributions of our work.[5] In the first experiment, we compare three different methods – LSGD-PFL, SCD-PFL (=ASVRCD-PFL without acceleration and without variance reduction) and SVRCD-PFL (=ASVRCD-PFL without acceleration) [6] across different datasets and objective functions. In the second experiment, we demonstrate the need for reparametrization of w-space for SVRCD-PFL. The detailed statement of SCD-PFL and SVRCD-PFL us presented in the appendix.

Setup. We implemented three personalized FL objectives: (8), (11), and (14), each applied to three different datasets: MNIST (LeCun & Cortes, 2010), KMINIST (Clanuwat et al., 2018), and FMINST (Xiao et al., 2017). As a model, we use a multiclass logistic regression (i.e., a single-layer fully connected neural network composed with softmax function and cross entropy loss). See appendix for details.

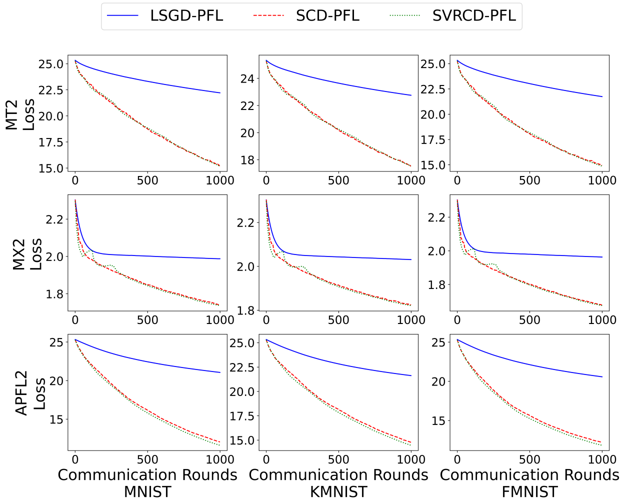

Figure 1. Comparison for three algorithms: LSGD-PFL, SCD-PFL, and SVRCD-PFL. Different rows correspond to different objective functions and columns correspond to different datasets.

Comparison between different optimization methods. We compare the convergence of LSGD-PFL versus SCD-PFL and SVRCD-PFL. We plot the loss against the number of communication rounds for three methods across different objectives and datasets. For Local SGD we set the synchronization step to be 5, which means that all devices synchronize in every 5 iterations. For SCD and SVRCD, devices only synchronize when we update the global parameter. The result is presented in Figure 1. We see that the variants of coordinate descent outperform widely-used LSGD-PFL. The addition of variance reduction term helps slightly improve the performance.[7]

Effect of reparametrization in SVRCD. In this experiment, we demonstrate the importance of reparametrization

[5] Our code is publicly available at https://github.com/boxinz17/PFL-Unified-Framework.

[6] We drop the acceleration term as f'_m may not have a large enough condition number for the acceleration to matter. Given that LSGD-PFL is not accelerated, considering the non-accelerated variant of coordinate descent results in a more fair comparison.

[7] We expect a more significant improvement if a closer neighborhood of the optimum was reached.

of global parameter w (i.e., divided by $\sqrt{M}$). We run reparameterized and non-reparameterized SVRCD-PFL across different objectives and datasets. Figure 2 shows the result. Indeed, we see that reparametrization improves the convergence of SVRCD-PFL. While the non-reparametrized variant might converge faster initially, soon enough, it becomes extremely unstable. This experiment confirms the necessity of reparametrization so that the scale of the learning rate is right for both global and local parameters.

6. Conclusions and extensions

In this work we proposed a general convex optimization theory for personalized FL. While our work answers a range of important questions, there are many directions in which our work can be extended in the future, in particular: partial participation, minimax optimal rates for specific personalized FL objectives, brand new personalized FL objectives and non-convex theory. See further detail in appendix.

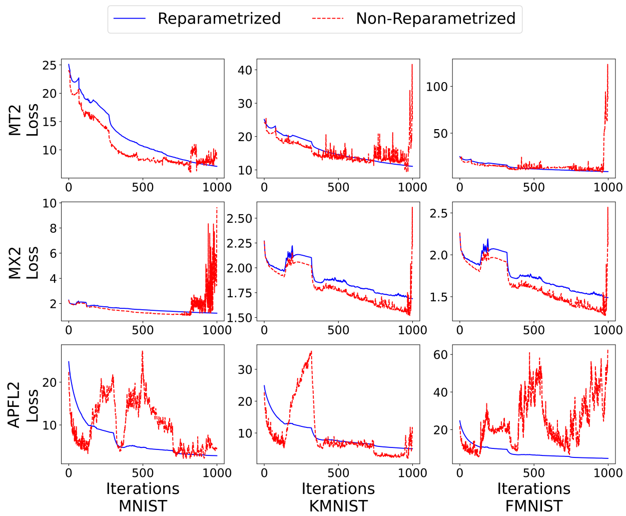

Figure 2. Effect of reparametrization of global space in SVRCD-PFL. Reparametrization helps SVRCD-PFL converge more smoothly, especially when it gets close to the optimum.

References

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

Allen-Zhu, Z., Qu, Z., Richtárik, P., and Yuan, Y. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pp. 1110–1119, 2016.

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

Chen, F., Luo, M., Dong, Z., Li, Z., and He, X. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*, 2018.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature, 2018.

Dean, J., Corrado, G. S., Monga, R., Chen, K., Devin, M., Le, Q. V., Mao, M. Z., Ranzato, M., Senior, A., Tucker, P., et al. Large scale distributed deep networks. 2012.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *arXiv preprint arXiv:1407.0202*, 2014.

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

Fallah, A., Mokhtari, A., and Ozdaglar, A. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Gorbunov, E., Hanzely, F., and Richtárik, P. Local sgd: Unified theory and new efficient methods. *arXiv preprint arXiv:2011.02828*, 2020.

Hanzely, F. and Richtárik, P. Accelerated coordinate descent with arbitrary sampling and best rates for minibatches. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 304–312. PMLR, 2019.

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

Hanzely, F., Hanzely, S., Horváth, S., and Richtárik, P. Lower bounds and optimal algorithms for personalized federated learning. *Advances in Neural Information Processing Systems*, 33, 2020a.

Hanzely, F., Kovalev, D., and Richtárik, P. Variance reduced coordinate descent with acceleration: New method with a surprising application to finite-sum problems. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4039–4048. PMLR, 13–18 Jul 2020b.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Hendrikx, H., Bach, F., and Massoulie, L. An optimal algorithm for decentralized finite sum optimization. *arXiv preprint arXiv:2005.10675*, 2020.

Jiang, Y., Konečný, J., Rush, K., and Kannan, S. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Khodak, M., Balcan, M.-F. F., and Talwalkar, A. S. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Kulkarni, V., Kulkarni, M., and Pant, A. Survey of personalization techniques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pp. 794–797. IEEE, 2020.

Lan, G. and Zhou, Y. An optimal randomized incremental gradient method. *Mathematical programming*, 171(1-2): 167–215, 2018.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

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

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

Lin, S., Yang, G., and Zhang, J. A collaborative learning framework via federated meta-learning. *arXiv preprint arXiv:2001.03229*, 3, 2020.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

McMahan, H. B., Moore, E., Ramage, D., and y Arcas, B. A. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2016.

Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.

Nesterov, Y. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pp. 543–547, 1983.

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Nesterov, Y. and Stich, S. U. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017.

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Rajeswaran, A., Finn, C., Kakade, S., and Levine, S. Meta-learning with implicit gradients. *arXiv preprint arXiv:1909.04630*, 2019.

Scaman, K., Bach, F., Bubeck, S., Massoulié, L., and Lee, Y. T. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems 31*, pp. 2740–2749. 2018.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. Federated multi-task learning. *arXiv preprint arXiv:1705.10467*, 2017.

Stich, S. U. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

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

Wang, W., Wang, J., Kolar, M., and Srebro, N. Distributed stochastic multi-task learning with graph regularization. *arXiv preprint arXiv:1802.03830*, 2018.

Woodworth, B. E. and Srebro, N. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3639–3647, 2016.

Woodworth, B. E., Wang, J., Smith, A., McMahan, B., and Srebro, N. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8496–8506, 2018.

Wu, R., Scaglione, A., Wai, H.-T., Karakoc, N., Hreinsson, K., and Ma, W.-K. Federated block coordinate descent scheme for learning global and personalized models. *arXiv preprint arXiv:2012.13900*, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017.

Appendix

A. Extensions and Future Work

Partial participation. An essential aspect of FL that is not covered in this work is the partial participation when one has access to a subset of devices at each iteration only. While we did not cover partial participation and focused on answering orthogonal questions, we believe that partial participation should be considered when extending our results in the future.

Minimax optimal rates for specific personalized FL objectives. As outlined in Section 1.2, one can not hope for the general optimization framework to be minimax optimal in every single special case. Consequently, there is a still need to keep exploring the optimization aspects of individual personalized FL objectives as one might come up with a more efficient optimizer that exploits the specific structure not covered by Assumptions 1.1 or 1.2.

Brand new personalized FL objectives. While in this work we propose a couple of novel personalized FL objectives obtained as an extension of known objectives, we believe that seeing the personalized FL as an instance of (1) might lead to the development of brand new approaches for personalized FL.

Non-convex theory. As mentioned, in this work, we propose a general convex optimization theory for personalized FL. Our convex rates are meaningful – they are minimax optimal and correspond to the empirical convergence. However, an inherent drawback of such an approach is the inability to cover non-convex FL approaches such as MAML (see Section C.4), or non-convex FL models. We believe that obtaining minimax optimal rates in the non-convex world would be very valuable.

B. Further Experimental Results and Details

B.1. Details of experiments in Section 5

Details on data preparation. We set the number of devices $M = 10$. We focus on a non-i.i.d. setting of McMahan et al. (2017) and Liang et al. (2020) by assigning two classes out of ten to each device. We then randomly select 100 samples for each device based on its class assignment. We first do normalization by feature (by column) for each dataset to make all feature columns have zero mean and unit variance; we then do normalization by sample (by row) to make every input vector have a unit norm.

Details on the model. More specifically, given a gray scale picture with label $y \in \{1, 2, \dots, C\}$, we unroll its pixel matrix into a vector $x \in \mathbb{R}^p$; then given a paramter matrix $\Theta \in \mathbb{R}^{p \times C}$, we have $f'_m(\cdot)$ in (8), (11) and (14) to be defined as

$$f'_m(\Theta) := l_{\text{CE}}\left(\varsigma(\Theta x)\,; y\right),$$

where $\varsigma(\cdot) : \mathbb{R}^K \to \mathbb{R}^K$ is softmax function and $l_{\text{CE}}(\cdot)$ is cross-entropy loss function. Note that under this setting, $f'_m(\cdot)$ is thus convex function.

Details on the personalized FL objectives. We consider three different objectives:

- Multitask FL objective (8) with $\Lambda = 10$ and $\lambda = 1$;
- Mixture FL objective (11), with $\lambda = 1$; and
- Adaptive personalized FL objective (14), with $\Lambda = 10$ and $\alpha_m = 0.2$ for all $m \in [M]$.

Details on optimization algorithms. For LSGD-PFL(Algorithm 1), we set the batch size to compute stochastic gradient $B = 1$. As for p_w in SCD-PFL (Algorithm 5) and SVRCD-PFL (Algorithm 6), we set it as $p_w = L^w/(L^\beta + L^w)$. For objective F_{MT2} in (8), we set $L^w = (\Lambda + \lambda)/M$ and $L^\beta = 1 + \lambda$; for objective F_{MX2} in (11), we set $L^w = 1/M$ and $L^\beta = 1 + \lambda$; for objective F_{APFL2} in (14), we set $L^w = (\Lambda + \max_{1 \le m \le M} \alpha_m^2)/M$ and $L^\beta = (1 - \max_{1 \le m \le M} \alpha_m)^2/M$. We set $\rho = 0.01$ for SVRCD-PFL. In the first experiment for comparison between different optimization methods, we set $\eta = 0.1$ for all experiments; in the second experiment for effect of reparametrization in SVRCD, we set $\eta = 0.45$ for objective F_{MT2}, $\eta = 0.5$ for objective F_{MX2} and $\eta = 1.1$ for objective F_{APFL2}, where the η's are chosen such that the desired effect is obvious within the first 1,000 iterations.

B.2. Subsampling of the Global and Local Parameters

In this section, we show that the choice of p_w based on Theorem E.1, that is, setting $p_w = \mathcal{L}^w/(\mathcal{L}^w + \mathcal{L}^\beta)$, can get both best communication complexity and best iteration complexity of SVRCD-PFL. More specifically, based on Theorem E.1, we set learning rate $\eta = 1/(4\mathcal{L})$, where $\mathcal{L} \coloneqq 2\max\left\{\mathcal{L}^w/p_w, \mathcal{L}^\beta/p_\beta\right\}$. The expressions of L^w and L^β for F_{MT2}, F_{MX2} and F_{APFL2} are stated in Lemma 2.1, Lemma 2.2 and Lemma 2.3, where $\mathcal{L}'$ is 1 after normalization. We set $\rho = 0.01$. We compare theoretically chose p_w with $p_w \in \{0.7, 0.5, 0.3, 0.1\}$.

To explore communication complexity, we plot loss against communication rounds. The result is shown in Figure 3. For completeness, Figure 4 shows the loss against the iteration so that we can better see the computational complexity. Our results show that by choosing p_w theoretically yields both the best communication complexity and the best iteration complexity (and consequently, the best computational complexity).

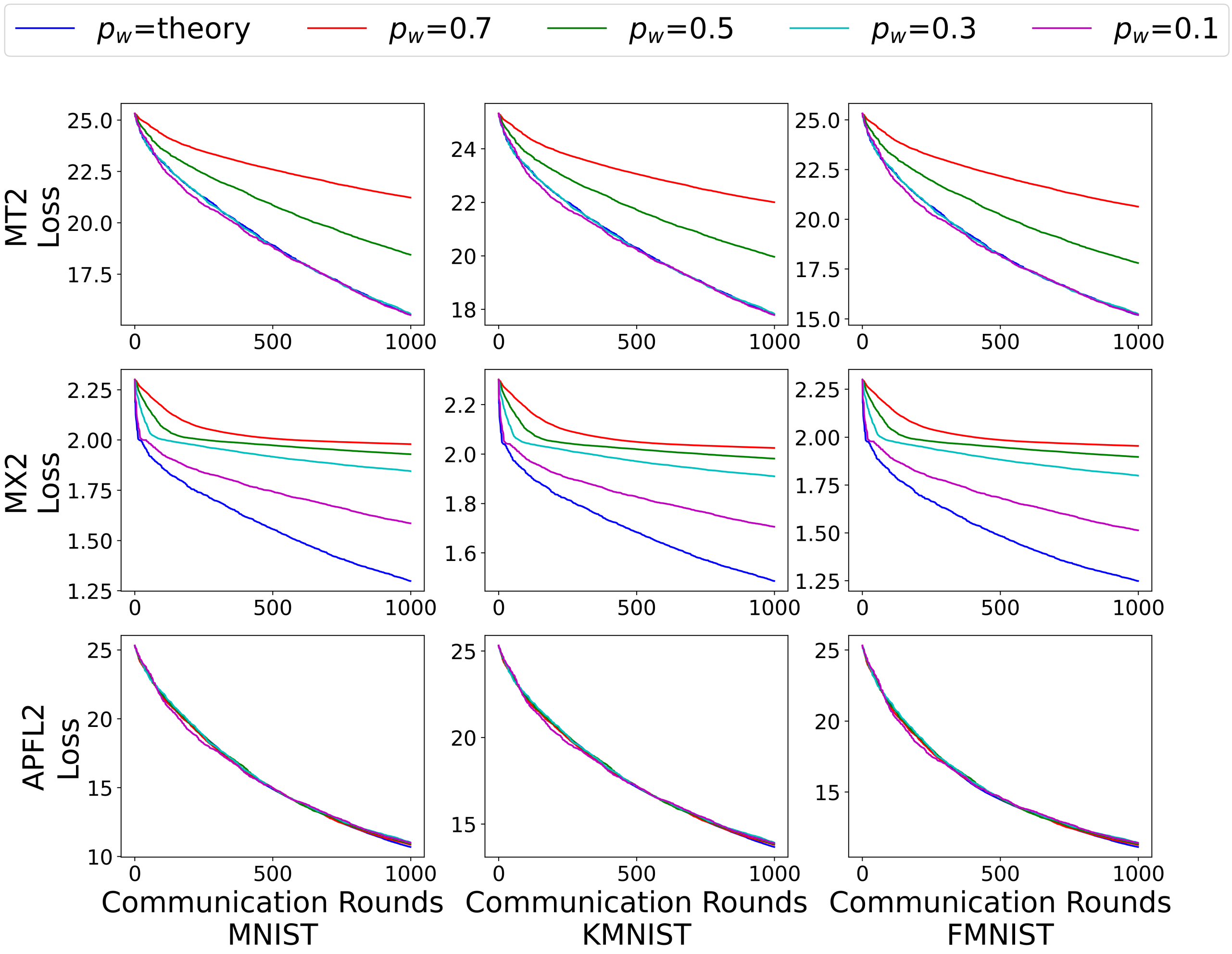

Figure 3. Communication complexity of SVRCD-PFL for different choices of p_w. Specifically, the theoretical choice of p_w gets $p_w = 0.35484$ for F_{MT2}, $p_w = 0.04762$ for F_{MX2} and $p_w = 0.94007$ for F_{APFL2}.

C. Missing Parts and Proofs for Section 2

For the sake of convenience, define

$$F_i(w, \beta) \coloneqq \frac{1}{M}\sum_{m=1}^{M} f_{m,i}(w, \beta_m).$$

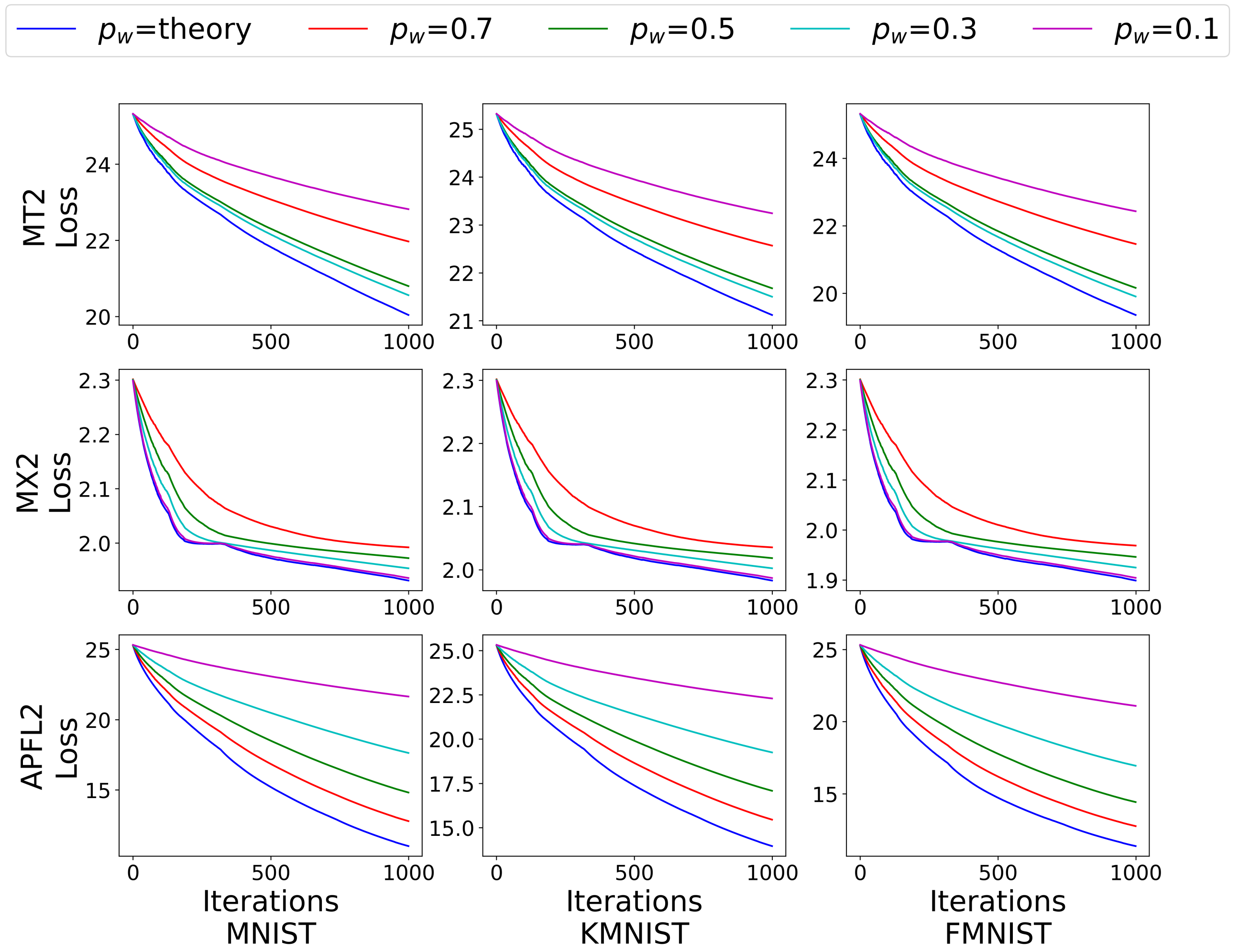

Figure 4. Iteration complexity of SVRCD-PFL for different choices of p_w. Specifically, the theoretical choice of p_w gets $p_w = 0.35484$ for F_{MT2}, $p_w = 0.04762$ for F_{MX2} and $p_w = 0.94007$ for F_{APFL2}.

in case when functions f_m are of a finite sum structure (3).

C.1. Table of Complexity Results for Special Cases

First, we present a table of complexity results for solving individual personalized FL objectives, which summarizes how our results apply in these special cases.

C.2. Personalized FL with explicit weight sharing

The most typical example of the weight sharing setting is when parameters w, β correspond to different layers of the same neural network: either $\beta_1, \ldots, \beta_M$ are the weights of first few layers of neural network while w are the weights of the remaining layers (Liang et al., 2020) or alternatively, each of $\beta_1, \ldots, \beta_M$ can correspond to the weights of last few layers while the remaining weights are included in the global parameter w (Arivazhagan et al., 2019). Overall, we can write the objective as follows:

$$\min_{\substack{w \in \mathbb{R}^{d_w}, \\ \beta_1, \ldots, \beta_M \in \mathbb{R}^{d_\beta}}} F_{WS}(w, \beta) = \frac{1}{M} \sum_{i=1}^{M} f'_m([w, \beta_m]), \tag{15}$$

Table C1. Complexity of solving personalized FL objectives by Algorithms 2 (second, third and fourth column) and 3 (fifth and sixth column). Constant and log factors ignored.

Objective	**# Comm**	# $\nabla_w F$	# $\nabla_\beta F$	# $\nabla_w F_i$	# $\nabla_\beta F_i$
(4)	$\sqrt{\frac{L'}{\mu'}}$	$\sqrt{\frac{L'}{\mu'}}$	0	$n+\sqrt{\frac{n\mathcal{L}'}{\mu'}}$	0
(5)	0	0	$\sqrt{\frac{L'}{\mu'}}$	0	$n+\sqrt{\frac{n\mathcal{L}'}{\mu'}}$
(8)	$\sqrt{\frac{\Lambda L'}{\lambda}}$	$\sqrt{\frac{\Lambda L'}{\lambda}}$	$\sqrt{\frac{L'}{\lambda}}$	$n+\sqrt{\frac{n\Lambda\mathcal{L}'}{\lambda}}$	$n+\sqrt{\frac{n\mathcal{L}'}{\lambda}}$
(11)	$\sqrt{\frac{\lambda}{\mu'}}$	$\sqrt{\frac{\lambda}{\mu'}}$	$\sqrt{\frac{L'+\lambda}{\mu'}}$	-	$n+\sqrt{\frac{n(\mathcal{L}'+\lambda)}{\mu'}}$
(14)	$\sqrt{\frac{(\Lambda+\alpha_{\max}^2)L'}{(1-\alpha_{\max})^2\mu'}}$	$\sqrt{\frac{(\Lambda+\alpha_{\max}^2)L'}{(1-\alpha_{\max})^2\mu'}}$	$\sqrt{\frac{(1-\alpha_{\min})^2 L'}{(1-\alpha_{\max})^2\mu'}}$	$n+\sqrt{\frac{n(\Lambda+\alpha_{\max}^2)\mathcal{L}'}{(1-\alpha_{\max})^2\mu'}}$	$n+\sqrt{\frac{n(1-\alpha_{\min})^2\mathcal{L}'}{(1-\alpha_{\max})^2\mu'}}$
(16)	$\sqrt{\frac{L'}{\mu'}}$	$\sqrt{\frac{L'}{\mu'}}$	$\sqrt{\frac{L'}{\mu'}}$	$n+\sqrt{\frac{n\mathcal{L}'}{\mu'}}$	$n+\sqrt{\frac{n\mathcal{L}'}{\mu'}}$
(18)	$\sqrt{\frac{L_R^w}{\mu}}$	$\sqrt{\frac{L_R^w}{\mu}}$	$\sqrt{\frac{L_R^\beta}{\mu}}$	$n+\sqrt{\frac{n\mathcal{L}_R^w}{\mu'}}$	$n+\sqrt{\frac{n\mathcal{L}_R^\beta}{\mu'}}$

where $d_w + d_\beta = d$. Using the same equivalent reparameterization of $w-$space we aim to minimize

$$\min_{\substack{w \in \mathbb{R}^{d_w}, \\ \beta_1,\dots,\beta_M \in \mathbb{R}^{d_\beta}}} F_{WS2}(w,\beta) = \frac{1}{M}\sum_{i=1}^{M} f'_m([M^{-\frac{1}{2}}w, \beta_m]), \tag{16}$$

which is an instance of (1) with $f_m(w,\beta_m) = f'_m([M^{-\frac{1}{2}}w, \beta_m])$.

Lemma C.1. *Function F_{WS2} is jointly μ'-strongly convex, $\left(\frac{L'}{M}\right)$-smooth with respect to w and L'-smooth with respect to β. Similarly, function f_m is jointly convex, $\left(\frac{\mathcal{L}'}{M}\right)$-smooth with respect to w and $\mathcal{L}'$-smooth with respect to β.*

A specific feature of the explicit weight sharing paradigm is that evaluating a gradient with respect to w-parameters automatically provides either free or very cheap access to the gradient with respect to β parameters (and vice versa).

C.3. Federated residual learning (Agarwal et al., 2020)

The last among the personalized FL models we mention is the federated residual learning from (Agarwal et al., 2020) given as:

$$\min_{\substack{w \in \mathbb{R}^{d_w}, \\ \beta_1,\dots,\beta_M \in \mathbb{R}^{d_\beta}}} F_R(w,\beta) = \frac{1}{M}\sum_{i=1}^{M} l_m(A^w(w, x_m^w), A^\beta(\beta_m, x_m^\beta)), \tag{17}$$

where (x_m^w, x_m^β) is a local feature vector (there might be an overlap between x_m^w and x_m^β), $A^w(w, x_m^w)$ is model prediction using global parameters/features, $A^\beta(\beta, x_m^\beta)$ is the model prediction using local parameters/features and $l(\cdot,\cdot)$ is a loss function. Clearly, we can recover (17) with

$$f_m(w,\beta_m) = l(A^w(M^{-\frac{1}{2}}w, x_m^w), A^\beta(\beta_m, x_m^\beta)). \tag{18}$$

Unlike in the for the other objectives, we can not relate constants $\mu', L', \mathcal{L}'$ to F_R since we do not write f_m as a function of f'_m. However, what seems natural is to assume L_R^w (or L_R^β)-smoothness of $l(A^w(w, x_m^w), a_m^\beta)$ (or $l(a_m^w, A^\beta(\beta_m, x_m^\beta))$) as a function of w (or β) for any $a_m^\beta, x_m^\beta, a_m^w, x_m^w$. Let us define $\mathcal{L}_R^w$, $\mathcal{L}_R^\beta$ analogously given that l is of a n-finite sum structure. Assuming further μ-strong convexity of F and convexity of f_m (for each $m \in M$), we can apply our theory.

Remark C.1. *The objective (18) allows us to properly make use of our theory as in different applications, functions A^w, A^β might have very different structure resulting in different L_R^w, L_R^β and different cost of accessing gradients w.r.t w and β.*

C.4. MAML based approaches

The last among the personalized FL models we mention is MAML (Finn et al., 2017) based personalized FL objective[8] given as

$$\min_{w\in\mathbb{R}^d} F_{MAML}(w) = \frac{1}{M}\sum_{i=1}^{M} f'_m(w - \alpha\nabla f'_m(w)). \tag{19}$$

While we can recover (19) as a special case of (1) setting $f_m(w, \beta_m) = f'_m(w - \alpha\nabla f'_m(w))$, our (convex) convergence theory does not apply due to the inherent non-convex structure of (19). In particular, objective F_{MAML} is non-convex even if function f'_m is convex. In this scenario, only our non-convex rates of Local SGD apply.

C.5. Proof of Lemma 2.1

Let us start by introducing some useful notation that will be useful throughout multiple proofs. In particular, set $\mathbf{I}_{d'}$ to be $d' \times d'$ identity matrix, $0_{d'_1\times d'_2}$ to be $d'_1 \times d'_2$ zero matrix and $\mathbf{1}'_d \in \mathbb{R}^{d'}$ to be vector of ones.

To show the strong convexity, we shall verify the positive definiteness of

$$\begin{aligned}
&\nabla^2 F_{MT2}(w,\beta) - \frac{\lambda}{2M}\mathbf{I}_{d(M+1)} \\
&= \begin{pmatrix} \frac{\Lambda}{M}\nabla F'(w) + \frac{\lambda}{M}I_d & -\frac{\lambda}{M^{\frac{3}{2}}}(\mathbf{1}_M^\top \otimes I_d) \\ -\frac{\lambda}{M^{\frac{3}{2}}}(\mathbf{1}_M \otimes I_d) & \frac{\lambda}{M}(I_m \otimes I_d) + \mathrm{Diag}(\nabla^2 f'_1(\beta_1), \dots, \nabla^2 f'_M(\beta_M)) \end{pmatrix} - \frac{\lambda}{2M}\mathbf{I}_{d(M+1)} \\
&\succeq \begin{pmatrix} \left(\frac{\Lambda\mu'}{M} + \frac{\lambda}{2M}\right) I_d & -\frac{\lambda}{M^{\frac{3}{2}}}(\mathbf{1}_M^\top \otimes I_d) \\ -\frac{\lambda}{M^{\frac{3}{2}}}(\mathbf{1}_M \otimes I_d) & \left(\frac{\lambda}{2M} + \frac{\mu'}{M}\right)(I_m \otimes I_d) \end{pmatrix} \\
&= \frac{1}{M}\underbrace{\begin{pmatrix} \Lambda\mu' + \frac{\lambda}{2} & -\frac{\lambda}{M^{\frac{1}{2}}}\mathbf{1}_M^\top \\ -\frac{\lambda}{M^{\frac{1}{2}}}\mathbf{1}_M & \left(\frac{\lambda}{2} + 2\mu'\right) I_m \end{pmatrix}}_{:=\mathbf{M}} \otimes I_d.
\end{aligned}$$

Note that $\mathbf{M}$ can be written as a sum of M matrices, each of them having

$$\mathbf{M}_m = \begin{pmatrix} \frac{\Lambda\mu' + \frac{\lambda}{2}}{M} & -\frac{\lambda}{M^{\frac{1}{2}}} \\ -\frac{\lambda}{M^{\frac{1}{2}}} & \left(\frac{\lambda}{2} + 2\mu'\right) \end{pmatrix}$$

as a $(1, m)$ submatrix and zeros everywhere else. To verify positive semidefiniteness of $\mathbf{M}_m$, we shall prove that the determinant is positive:

$$\begin{aligned}
\det(\mathbf{M}_m) &= \frac{1}{M}\left(\left(\Lambda\mu' + \frac{\lambda}{2}\right)\left(\frac{\lambda}{2} + 2\mu'\right) - \lambda^2\right) \\
&\geq \frac{1}{M}\left((2\lambda)\left(\frac{\lambda}{2} + 2\mu'\right) - \lambda^2\right) \geq 0
\end{aligned}$$

as desired. Verifying the smoothness constants is straightforward.

[8] Several recent papers study the meta-learning personalization approaches (Chen et al., 2018; Khodak et al., 2019; Jiang et al., 2019; Fallah et al., 2020; Lin et al., 2020).

C.6. Proof of Lemma 2.2

We have

$$\begin{aligned}
\nabla^2 F_{MFL2}(w,\beta) - \frac{\mu'}{3M}\mathbf{I}_{d(M+1)} &= \begin{pmatrix} \frac{\lambda}{M} I_d & -\frac{\lambda}{M^{\frac{3}{2}}}(\mathbf{1}_M^\top \otimes I_d) \\ -\frac{\lambda}{M^{\frac{3}{2}}}(\mathbf{1}_M \otimes I_d) & \frac{\lambda}{M}(I_m \otimes I_d) + \mathrm{Diag}(\nabla^2 f_1'(\beta_1), \dots, \nabla^2 f_M'(\beta_M)) \end{pmatrix} - \frac{\mu'}{3M}\mathbf{I}_{d(M+1)} \\
&\succeq \begin{pmatrix} \left(\frac{\lambda}{M} - \frac{\mu'}{3M}\right) I_d & -\frac{\lambda}{M^{\frac{3}{2}}}(\mathbf{1}_M^\top \otimes I_d) \\ -\frac{\lambda}{M^{\frac{3}{2}}}(\mathbf{1}_M \otimes I_d) & \left(\frac{\lambda}{M} + \frac{2\mu'}{3M}\right)(I_m \otimes I_d) \end{pmatrix} \\
&= \frac{1}{M} \underbrace{\begin{pmatrix} \lambda - \frac{\mu'}{3} & -\frac{\lambda}{M^{\frac{1}{2}}}\mathbf{1}_M^\top \\ -\frac{\lambda}{M^{\frac{1}{2}}}\mathbf{1}_M & \left(\lambda + \frac{2\mu'}{3}\right) I_m \end{pmatrix}}_{:=\mathbf{M}} \otimes I_d
\end{aligned}$$

Note that $\mathbf{M}$ can be written as a sum of M matrices, each of them having $\frac{\lambda}{M} - \frac{\mu'}{3M}$ at position $(1,1)$, $-\frac{\lambda}{M^{\frac{1}{2}}}$ at positions $(1,m), (m,1)$ and $\left(\frac{\lambda}{M} + \frac{2\mu'}{3M}\right)$ at position (m,m). Using the assumption $\mu' \leq \frac{\lambda}{2}$, it is easy to see that each of these matrices is positive semidefinite, and thus so is $\mathbf{M}$. Consequently, $\nabla F_{MFL2}(w,\beta) - \frac{\mu'}{3M}\mathbf{I}_{d(M+1)}$ is positive semidefinite and thus F_{MFL2} is jointly $\frac{\mu'}{3M}$- strongly convex. Verifying the smoothness constants is straightforward.

C.7. Proof of Lemma 2.3

Let $x_m = (1-\alpha_m)\beta_m + \alpha_m M^{-\frac{1}{2}} w$ for the notational simplicity. We have

$$\begin{aligned}
\nabla^2 f_m(w,\beta_m) &= \begin{pmatrix} \frac{\Lambda}{M}\nabla^2 f'(M^{-\frac{1}{2}}w) + \frac{\alpha_m^2}{M}\nabla^2 f_m'(x_m) & \frac{\alpha_m(1-\alpha_m)}{M^{\frac{1}{2}}}\nabla^2 f_m'(x_m) \\ \frac{\alpha_m(1-\alpha_m)}{M^{\frac{1}{2}}}\nabla^2 f_m'(x_m) & (1-\alpha_m)^2\nabla^2 f_m'(x_m) \end{pmatrix} \\
&= \begin{pmatrix} \frac{\Lambda}{M}\nabla^2 f'(M^{-\frac{1}{2}}w) & 0_{d\times d} \\ 0_{d\times d} & 0_{d\times d} \end{pmatrix} + \frac{1}{M}\begin{pmatrix} \frac{\alpha_m^2}{M} & \frac{\alpha_m(1-\alpha_m)}{M^{\frac{1}{2}}} \\ \frac{\alpha_m(1-\alpha_m)}{M^{\frac{1}{2}}} & (1-\alpha_m)^2 \end{pmatrix} \otimes \nabla^2 f_m'(x_m) \\
&\succeq \begin{pmatrix} \frac{\Lambda\mu'}{M}\mathbf{I}_d & 0_{d\times d} \\ 0_{d\times d} & 0_{d\times d} \end{pmatrix} + \begin{pmatrix} \frac{\alpha_m^2}{M} & \frac{\alpha_m(1-\alpha_m)}{M^{\frac{1}{2}}} \\ \frac{\alpha_m(1-\alpha_m)}{M^{\frac{1}{2}}} & (1-\alpha_m)^2 \end{pmatrix} \otimes (\mu'\mathbf{I}_d) \\
&= \mu' \underbrace{\begin{pmatrix} \frac{\Lambda+\alpha_m^2}{M} & \frac{\alpha_m(1-\alpha_m)}{M^{\frac{1}{2}}} \\ \frac{\alpha_m(1-\alpha_m)}{M^{\frac{1}{2}}} & (1-\alpha_m)^2 \end{pmatrix}}_{:=\mathbf{M}_m} \otimes \mathbf{I}_d.
\end{aligned}$$

Next, we show that

$$\mathbf{M}_m \succeq \begin{pmatrix} \frac{(1-\alpha_m)^2}{2M} & 0 \\ 0 & \frac{(1-\alpha_m)^2}{2} \end{pmatrix}. \tag{20}$$

For that, it suffices to show that

$$\det\left(\mathbf{M}_m - \begin{pmatrix} \frac{(1-\alpha_m)^2}{2M} & 0 \\ 0 & \frac{(1-\alpha_m)^2}{2} \end{pmatrix}\right) \geq 0,$$

which holds since

$$\begin{aligned}
\det\left(\mathbf{M}_m - \begin{pmatrix} \frac{(1-\alpha_m)^2}{2M} & 0 \\ 0 & \frac{(1-\alpha_m)^2}{2} \end{pmatrix}\right) &= \left(\frac{\Lambda + \alpha_m^2 - \frac{(1-\alpha_m)^2}{2}}{M} - \right)\frac{(1-\alpha_m)^2}{2} - \frac{\alpha_m^2(1-\alpha_m)^2}{M} \\
&\geq \left(2\frac{\alpha_m^2}{M}\right)\frac{(1-\alpha_m)^2}{2} - \frac{\alpha_m^2(1-\alpha_m)^2}{M} \\
&= 0.
\end{aligned}$$

Next, using (20) M times, it is easy to see that $\nabla^2 F_{APFL2}(w,\beta) \succeq \mu' \frac{(1-\alpha_{\max})^2}{M} \mathbf{I}_{d(M+1)}$ as desired. Verifying the smoothness constants is straightforward.

D. Missing Parts from Section 3

D.1. Proof of Theorem 3.1

The main idea consists of invoking the framework for analyzing local SGD methods from (Gorbunov et al., 2020) with several minor modifications. In particular, Algorithm 1 is an intriguing method that runs a local SGD on w-parameters and SGD on β-parameters, and therefore we shall treat both of these parameter sets differently. Define $V_k := \frac{1}{M}\sum_{m=1}^M \|w^k - w^k_m\|^2$ where $w^k := \frac{1}{M}\sum_{m=1}^M w^k_m$ is a sequence of so called virtual iterates.

The first step towards the convergence rate is to figure out parameters of Assumption 2.3 from (Gorbunov et al., 2020). In order to get these, let us show an analog of Lemma G.1 therein.

Lemma D.1. *Let Assumptions 1.1 and 3.1 hold. Let* $L = \max\{L^w, L^\beta\}$. *Then, we have:*

$$\frac{1}{M}\sum_{m=1}^M \|\nabla_w f_m(w^k_m, \beta^k_m)\|^2 \le 6L^w \left(f(w^k, \beta^k_m) - f(w^*, \beta^*)\right) + 3(L^w)^2 V_k + 3\zeta_*^2, \tag{21}$$

$$\left\|\frac{1}{M}\sum_{m=1}^M \nabla_w f_m(w^k_m, \beta^k_m)\right\|^2 + \frac{1}{M^2}\sum_{m=1}^M \left\|\nabla_\beta f_m(w^k_m, \beta^k_m)\right\|^2 \le 4L \left(f(w^k, \beta^k_m) - f(w^*, \beta^*)\right) + 2(L^w)^2 V_k \tag{22}$$

Proof. First, to show (21) we shall have

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^M \|\nabla_w f_m(w^k_m, \beta^k_m)\|^2 \quad &\le \quad \frac{3}{M}\sum_{m=1}^M \|\nabla_w f_m(w^k_m, \beta^k_m) - \nabla_w f_m(w^k, \beta^k_m)\|^2 \\
&\quad + \frac{3}{M}\sum_{m=1}^M \|\nabla_w f_m(w^k, \beta^k_m) - \nabla_w f_m(w^*, \beta^*)\|^2 \\
&\quad + \frac{3}{M}\sum_{m=1}^M \|\nabla_w f_m(w^*, \beta^*)\|^2 \\
&\overset{\text{As. } 1.1}{\le} \quad \frac{3L^2}{M}\sum_{m=1}^M \|w^k_m - w^k\|^2 + \frac{6L}{M}\sum_{m=1}^M D_{f_m}((w^k, \beta^k_m), (w^*, \beta^*)) + 3\zeta_*^2 \\
&= \quad 6L^w \left(f(w^k, \beta^k_m) - f(w^*, \beta^*)\right) + 3(L^w)^2 V_k + 3\zeta_*^2.
\end{aligned}$$

Next, to establish (22), we have

$$\begin{aligned}
& \left\|\frac{1}{M}\sum_{m=1}^{M} \nabla_w f_m(w_m^k,\beta_m^k)\right\|^2 + \frac{1}{M^2}\sum_{m=1}^{M} \left\|\nabla_\beta f_m(w_m^k,\beta_m^k)\right\|^2 \\
= \quad & \left\|\frac{1}{M}\sum_{m=1}^{M} \left(\nabla_w f_m(w_m^k,\beta_m^k) - \nabla_w f_m(w^*,\beta^*)\right)\right\|^2 + \frac{1}{M^2}\sum_{m=1}^{M} \left\|\nabla_\beta f_m(w_m^k,\beta_m^k) - \nabla_\beta f_m(w^*,\beta^*)\right\|^2 \\
\leq \quad & \frac{2}{M}\sum_{i=1}^{M} \|\nabla_w f_m(w_m^k,\beta_m^k) - \nabla_w f_m(w^k,\beta_m^k)\|^2 + \frac{2}{M}\sum_{m=1}^{M} \|\nabla_\beta f_m(w_m^k,\beta_m^k) - \nabla_\beta f_m(w^*,\beta^*)\|^2 \\
& \quad + \frac{2}{M^2}\sum_{m=1}^{M} \left\|\nabla_\beta f_m(w_m^k,\beta_m^k) - \nabla_\beta f_m(w^*,\beta^*)\right\|^2 \\
\overset{\text{As. 1.1}}{\leq} \quad & \frac{2(L^w)^2}{M}\sum_{m=1}^{M} \|w_m^k - w^k\|^2 + \frac{4L}{M}\sum_{m=1}^{M} D_{f_m}((w^k,\beta_m^k),(w^*,\beta^*)) \\
= \quad & 4L\left(f(w^k,\beta_m^k) - f(w^*,\beta^*)\right) + 2(L^w)^2 V_k.
\end{aligned}$$

□

The next lemma uses Lemma D.1 to recover a set of crucial parameters of Assumption 2.3 from (Gorbunov et al., 2020).

Lemma D.2. *Let* $g_{w,m}^k \coloneqq (g_m^k)_{1:d_0}$ *and* $g_{\beta,m}^k \coloneqq (g_m^k)_{(d_0+1):(d_0+d_m)}$. *Then we have*

$$\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[\|g_{w,m}^k\|^2\right] \leq 6L^w\left(f(w^k,\beta_m^k) - f(w^*,\beta^*)\right) + 3(L^w)^2 V_k + \frac{\sigma^2}{B} + 3\zeta_*^2, \tag{23}$$

$$\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} g_{w,m}^k\right\|^2 + \frac{1}{M^2}\sum_{m=1}^{M}\left\|g_{\beta,m}^k\right\|^2\right] \leq 4L\left(f(w^k,\beta_m^k) - f(w^*,\beta^*)\right) + 2(L^w)^2 V_k + \frac{2\sigma^2}{BM} \tag{24}$$

Proof. Let us start with (23):

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[\|g_{w,m}^k\|^2\right] \quad &= \quad \frac{1}{M}\sum_{m=1}^{M}\left(\mathbb{E}\left[\|g_{w,m}^k - \nabla_w f_m(w_m^k,\beta_m^k)\|^2\right] + \|\nabla_w f_m(w_m^k,\beta_m^k)\|^2\right) \\
&\leq \quad \frac{\sigma^2}{B} + \|\nabla_w f_m(w_m^k,\beta_m^k)\|^2.
\end{aligned}$$

It remains to apply (23). Similarly, to show (24), we have

$$\begin{aligned}
& \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} g_{w,m}^k\right\|^2 + \frac{1}{M^2}\sum_{m=1}^{M}\left\|g_{\beta,m}^k\right\|^2\right] \\
& \quad = \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}(g_{w,m}^k - \nabla_w f_m(w_m^k,\beta_m^k))\right\|^2\right] + \left\|\frac{1}{M}\sum_{m=1}^{M}\nabla_w f_m(w_m^k,\beta_m^k)\right\|^2 \\
& \qquad + \frac{1}{M^2}\sum_{m=1}^{M}\left(\mathbb{E}\left[\left\|g_{\beta,m}^k - \nabla_\beta f_m(w_m^k,\beta_m^k)\right\|^2\right] + \left\|\nabla_\beta f_m(w_m^k,\beta_m^k)\right\|^2\right) \\
& \quad \leq \frac{\sigma^2}{MB} + \left\|\frac{1}{M}\sum_{m=1}^{M}\nabla_w f_m(w_m^k,\beta_m^k)\right\|^2 \\
& \qquad + \frac{\sigma^2}{MB} + \frac{1}{M^2}\sum_{m=1}^{M}\left\|\nabla_\beta f_m(w_m^k,\beta_m^k)\right\|^2.
\end{aligned}$$

It remains to apply (22). □

To get the remaining parameters of Assumption 2.3 of Gorbunov et al. (2020), we shall prove an analog of Lemma E.1 therein.

Lemma D.3. *Suppose that Assumptions 1.1 and 3.1 hold and assume that*

$$\eta \leq \frac{1}{8\sqrt{3e}(\tau-1)L^w}$$

Then, we have

$$2L^w \sum_{k=0}^{K} (1-\eta\mu)^{-k-1}\mathbb{E}\left[V_k\right] \leq \frac{1}{2}\sum_{k=0}^{K}(1-\eta\mu)^{-k-1}\mathbb{E}\left[F(w^k,\beta^k) - F(w^*,\beta^*)\right] + 2L^w D\eta^2 \sum_{k=0}^{K}(1-\eta\mu)^{-k-1}, \tag{25}$$

where

$$D = 2e(\tau-1)\tau\left(3\zeta_*^2 + \frac{\sigma^2}{B}\right).$$

Proof. The proof is identical to the proof of Lemma E.1 from (Gorbunov et al., 2020) with a single difference – using inequality (23) instead of Assumption E.1 from (Gorbunov et al., 2020). □

Now we have all pieces together and we are ready to state the main convergence result for Algorithm 1.

Theorem D.1. *Let Assumptions 1.1, and 3.1 be satisfied and assume the stepsize η satisfies*

$$0 < \eta \leq \min\left\{\frac{1}{4L^\beta}, \frac{1}{8\sqrt{3e}(\tau-1)L^w}\right\}.$$

Define

$$(\overline{w}^K, \overline{\beta}^K) := \frac{\sum_{k=0}^{K}(1-\eta\mu)^{-k-1}(w^k,\beta^k)}{\sum\limits_{k=0}^{K}(1-\eta\mu)^{-k-1}}.$$

and $\Phi^0 := \frac{2\|w^0-w^*\|^2 + \sum_{m=1}^{M}\|\beta_m^0-\beta_m^*\|^2}{\eta}$ *and* $\Psi^0 := \frac{2\sigma^2}{BM} + 8L^w\eta e(\tau-1)\tau\left(3\zeta_*^2 + \frac{\sigma^2}{B}\right)$*. Then if $\mu > 0$, we have*

$$\mathbb{E}\left[f(\overline{w}^K,\overline{\beta}^K)\right] - f(w^*,\beta^*) \leq (1-\eta\mu)^K\,\Phi^0 + \eta\Psi^0, \tag{26}$$

and in the case when $\mu = 0$, we have

$$\mathbb{E}\left[f(\overline{w}^K,\overline{\beta}^K)\right] - f(w^*,\beta^*) \leq \frac{\Phi^0}{K} + \eta\Psi^0. \tag{27}$$

D.2. Nonconvex theory for LSGD-PFL

In order to demonstrate that our approaches work in the nonconvex setting too, we develop a non-convex theory for LSGD-PFL. Note that we do not claim optimality of our results and at the same time, we impose slightly different smoothness assumptions on the objective.

We set $k_p = p \cdot \tau$, where $\tau \in \mathbb{N}^+$ is the length of averaging period. Let $k_p = p\tau + \tau - 1 = k_{p+1} - 1 = v_p$. Denote the total number of iterations as K, and assume that $K = k_{\overline{p}}$ for some $\overline{p} \in \mathbb{N}^+$. The final result is set to be that $w = w^K$ and $\hat{\beta}_m = \beta_m^K$ for all $m \in [M]$. We assume that the solution to (1) is $w^*, \beta_1^*, \ldots, \beta_M^*$ and optimal value is f^*. Let $w^k = \frac{1}{M}\sum_{m=1}^{M} w_m^k$ for all k. Note that this quantity will not be actually computed in practice unless $k = k_p$ for some $p \in \mathbb{N}$, where we have $w^{k_p} = w_m^{k_p}$ for all $m \in [M]$. In addition, let $\xi_m^k = \{\xi_{1,m}^k, \xi_{2,m}^k, \ldots, \xi_{B,m}^k\}$, and $\xi^k = \{\xi_1^k, \xi_2^k, \ldots, \xi_M^k\}$.

Let $\theta_m = ((w_m)^\top, (\beta_m)^\top)^\top$, $\theta_m^k = ((w_m^k)^\top, (\beta_m^k)^\top)^\top$, $\theta_m^* = ((w^*)^\top, (\beta_m^*)^\top)^\top$ and $\hat{\theta}_m^k = ((w^k)^\top, (\beta_m^k)^\top)^\top$. Let

$$g_m^k = \frac{1}{B}\nabla \hat{f}_m(w_m^k, \beta_m^k; \xi_m^k), \tag{28}$$

where

$$\nabla \hat{f}_m(w_m^k, \beta_m^k; \xi_m^k) = \sum_{j=1}^{B} \nabla \hat{f}_m(w_m^k, \beta_m^k; \xi_{j,m}^k).$$

We assume that the gradient is unbiased, that is

$$\mathbb{E}\left[g_m^k\right] = \nabla f_m(w_m^k, \beta_m^k).$$

Besides, let

$$\begin{aligned} g_{m,1}^k &= \frac{1}{B}\nabla_w \hat{f}_m(w_m^k, \beta_m^k; \xi_m^k), \\ g_{m,2}^k &= \frac{1}{B}\nabla_{\beta_m} \hat{f}_m(w_m^k, \beta_m^k; \xi_m^k), \end{aligned} \tag{29}$$

then we have $g_m^k = ((g_{m,1}^k)^\top, (g_{m,2}^k)^\top)^\top$. We update parameters by $(w_m^{k+1}, \beta_m^{k+1}) = (w_m^k, \beta_m^k) - \eta_k g_m^k$.

In addition, we define

$$\begin{aligned} h^k &= \frac{1}{M}\sum_{m=1}^{M} g_{m,1}^k, \\ V^k &= \frac{1}{M}\sum_{m=1}^{M} \|w_m^k - w^k\|^2. \end{aligned}$$

Then we always have $w^{k+1} = w^k - \eta_k h^k$ for all k.

We denote the Bregman divergence associated with f_m for θ_m and $\bar{\theta}_m$ as

$$D_{f_m}(\theta_m, \bar{\theta}_m) \coloneqq f_m(\theta_m) - f(\bar{\theta}_m) - \langle \nabla f_m(\bar{\theta}_m), \theta_m - \bar{\theta}_m\rangle.$$

Furthermore, we define the sum of residuals of estimators as

$$r^k = \|w^k - w^*\|^2 + \frac{1}{M}\sum_{m=1}^{M}\|\beta_m^k - \beta_m^*\|^2 = \frac{1}{M}\sum_{m=1}^{M}\|\hat{\theta}_m^k - \theta_m^*\|^2. \tag{30}$$

Finally, let $\sigma_{\text{dif}}^2 = \frac{1}{M}\sum_{m=1}^{M}\|\nabla f_m(\theta_m^*)\|^2$.

Assumption D.1 (Smoothness). *The local objective function $f_m(\cdot)$ is differentiable and L-smooth, that is, $\|\nabla f_m(u) - f_m(v)\| \le L\|u - v\|$ for all u, v and $m \in [M]$.*

Assumption D.2 (Bounded Local Variance). *For local stochastic gradients defined in* (28) *and* (29)*, we assume that their variance are bounded as below:*

$$\begin{aligned} \mathbb{E}_{\xi_m^k}\left[\|g_{m,1}^k - \nabla_w f_m(\theta_m^k)\|^2\right] &\le C_1\|\nabla f_m(\theta_m^k)\|^2 + \frac{\sigma_1^2}{B} \\ \mathbb{E}_{\xi_m^k}\left[\|g_{m,2}^k - \nabla_{\beta_m} f_m(\theta_m^k)\|^2\right] &\le C_2\|\nabla f_m(\theta_m^k)\|^2 + \frac{\sigma_2^2}{B} \end{aligned}$$

for all $m \in [M]$, where $C_1, C_2, \sigma_1^2, \sigma_2^2$ are all positive constants.

Assumption D.3 (Bounded Dissimilarity). *There is a positive constant $\lambda > 0$ such that for all $\theta_m \in \mathbb{R}^{d_0 + d_m}$, $m \in [M]$, we have*

$$\frac{1}{M}\sum_{m=1}^{M}\|\nabla f_m(\theta_m)\|^2 \le \lambda\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m)\right\|^2 + \sigma_{dif}^2.$$

Assumption D.4 (μ-Polyak-Łojasiewicz (PL))**.** *Assume that there is a positive constant $\mu > 0$, such that for all $w \in \mathbb{R}^{d_0}$ and $\beta_m \in \mathbb{R}^{d_m}$, $m \in [M]$, we have*

$$\frac{1}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(w,\beta_m)\right\|^2 \geq \mu\left(\frac{1}{M}\sum_{m=1}^{M} f_m(w,\beta_m) - f^*\right).$$

Theorem D.2 (General Non-Convex Objectives)**.** *Under Assumptions D.1-D.3, let $\eta_k = \eta$ for all $k \geq 0$, and η small enough such that*

$$-1 + \eta L\lambda\left(\frac{C_1}{M} + C_2 + 1\right) + \lambda\eta^2 L^2(\tau-1)\tau(C_1+1) \leq 0,$$

we have

$$\begin{aligned}&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^k)\right\|^2\right]\\ \leq&\frac{2\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^0) - f^*\right]}{\eta K} + \eta L\lambda\left\{\left(\frac{C_1}{M} + C_2 + 1\right)\sigma_{dif}^2 + \frac{\sigma_1^2}{MB} + \frac{\sigma_2^2}{B}\right\}\\ &+\eta^2 L^2\sigma_{dif}^2(\tau-1)^2(C_1+1) + \frac{\eta^2 L^2\sigma_1^2(\tau-1)^2}{B}.\end{aligned}$$

Theorem D.3 (General Non-Convex Objectives under PL condition)**.** *Under Assumptions D.1-D.4, setting $\eta_k = 1/(\mu(k + \beta\tau + 1))$, where β is a positive constant satisfying*

$$\beta > \max\left\{\frac{2\lambda L}{\mu}\left(\frac{C_1}{M} + C_2 + 1\right) - 2, \frac{2L^2\lambda(C_1+1)}{\mu^2}, 1\right\},$$

when τ is large enough such that

$$\tau \geq \sqrt{\frac{\max\left\{(2L^2\lambda(C_1+1)/\mu^2)e^{1/\beta} - 4, 0\right\}}{\beta^2 - (2L^2\lambda(C_1+1)/\mu^2)e^{\frac{1}{\beta}}}},$$

we will have

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m\left(\hat{\theta}_m^K\right) - f^*\right]\\ \leq&\frac{b^3}{(K+\beta\tau)^3}\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m\left(\hat{\theta}_m^0\right) - f^*\right] + \frac{2L^2(\tau-1)^2 K}{\mu^3(K+\beta\tau)^3}\left\{\sigma_{dif}^2(C_1+1) + \frac{\sigma_1^2}{B}\right\}\\ &+\frac{LK(K+2\beta\tau+2)}{4\mu^2(K+\beta\tau)^3}\left\{\sigma_{dif}^2\left(\frac{C_1}{M} + C_2 + 1\right) + \frac{\sigma_1^2}{MB} + \frac{\sigma_2^2}{B}\right\}.\end{aligned}$$

D.3. Proof of nonconvex theory for LSGD-PFL

Proof of Theorem D.2. By Lemmas D.4-D.6 and under Assumptions D.1-D.3, given $\{\theta_m^k\}_{m\in[M]}$, taking conditional expectation with respect to ξ^k, we have

$$\begin{aligned}&\mathbb{E}_{\xi^k}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k+1})\right] - \frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^k)\\ \leq&-\frac{\eta}{2}\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^k)\right\|^2 - \frac{\eta}{2}\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\theta_m^k)\right\|^2 + \frac{\eta L^2}{2}V^k\\ &+\frac{1}{2}\eta^2 L\lambda\left(\frac{C_1}{M} + C_2 + 1\right)\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\theta_m^k)\right\|^2 + \frac{1}{2}\eta^2 L\lambda\left\{\left(\frac{C_1}{M} + C_2 + 1\right)\sigma_{\text{dif}}^2 + \frac{\sigma_1^2}{MB} + \frac{\sigma_2^2}{B}\right\}.\end{aligned}$$

Thus, taking unconditional expectation on both sides of the above equation, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k+1}) - \frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k})\right] \\
\leq &-\frac{\eta}{2}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k})\right\|^2\right] - \frac{\eta}{2}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\theta_m^{k})\right\|^2\right] + \frac{\eta L^2}{2}\mathbb{E}\left[V^k\right] \\
&+\frac{1}{2}\eta^2 L\lambda\left(\frac{C_1}{M}+C_2+1\right)\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\theta_m^{k})\right\|^2\right] + \frac{1}{2}\eta^2 L\lambda\left\{\left(\frac{C_1}{M}+C_2+1\right)\sigma_{\text{dif}}^2 + \frac{\sigma_1^2}{MB} + \frac{\sigma_2^2}{B}\right\},
\end{aligned}$$

which implies that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k_{p+1}}) - \frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k_p})\right] \\
=&\sum_{k=k_p}^{v_p}\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k+1}) - \frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k})\right] \\
\leq &-\frac{\eta}{2}\sum_{k=k_p}^{v_p}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k})\right\|^2\right] + \frac{\eta}{2}\left\{-1+\eta L\lambda\left(\frac{C_1}{M}+C_2+1\right)\right\}\sum_{k=k_p}^{v_p}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\theta_m^{k})\right\|^2\right] \\
&+\frac{\eta L^2}{2}\sum_{k=k_p}^{v_p}\mathbb{E}\left[V^k\right] + \frac{1}{2}\eta^2 L\lambda\left\{\left(\frac{C_1}{M}+C_2+1\right)\sigma_{\text{dif}}^2 + \frac{\sigma_1^2}{MB} + \frac{\sigma_2^2}{B}\right\}\sum_{k=k_p}^{v_p} 1.
\end{aligned} \tag{31}$$

By Lemma D.8, we have

$$\begin{aligned}
\mathbb{E}\left[V^k\right] \leq &\lambda\eta^2(\tau-1)(C_1+1)\sum_{k=k_p}^{k-1}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m^{k})\right\|^2\right] \\
&+\eta^2\sigma_{\text{dif}}^2(\tau-1)(C_1+1)(k-k_p) + \frac{\eta^2\sigma_1^2(\tau-1)}{B}(k-k_p) \\
\leq &\lambda\eta^2(\tau-1)(C_1+1)\sum_{k=k_p}^{v_p}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m^{k})\right\|^2\right] \\
&+\eta^2\sigma_{\text{dif}}^2(\tau-1)^2(C_1+1) + \frac{\eta^2\sigma_1^2(\tau-1)^2}{B}
\end{aligned}$$

hold for all $k_p \leq k \leq v_p$. We then have

$$\begin{aligned}
\frac{\eta L^2}{2}\sum_{k=k_p}^{v_p}\mathbb{E}\left[V^k\right] \leq &\frac{1}{2}\lambda\eta^3 L^2(\tau-1)\tau(C_1+1)\sum_{k=k_p}^{v_p}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m^{k})\right\|^2\right] \\
&+\frac{1}{2}\eta^3 L^2\sigma_{\text{dif}}^2(\tau-1)^2(C_1+1)\sum_{k=k_p}^{v_p} 1 + \frac{\eta^3 L^2\sigma_1^2(\tau-1)^2}{2B}\sum_{k=k_p}^{v_p} 1.
\end{aligned}$$

Plug the above equation into (31), we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k_{p+1}}) - \frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k_p})\right] \\
\leq & -\frac{\eta}{2}\sum_{k=k_p}^{v_p}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k})\right\|^2\right] \\
+ & \frac{\eta}{2}\left\{-1+\eta L\lambda\left(\frac{C_1}{M}+C_2+1\right)+\lambda\eta^2L^2(\tau-1)\tau(C_1+1)\right\}\sum_{k=k_p}^{v_p}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\theta_m^{k})\right\|^2\right] \\
+ & \frac{1}{2}\eta^2L\lambda\left\{\left(\frac{C_1}{M}+C_2+1\right)\sigma_{\text{dif}}^2+\frac{\sigma_1^2}{MB}+\frac{\sigma_2^2}{B}\right\}\sum_{k=k_p}^{v_p}1 \\
+ & \frac{1}{2}\eta^3L^2\sigma_{\text{dif}}^2(\tau-1)^2(C_1+1)\sum_{k=k_p}^{v_p}1+\frac{\eta^3L^2\sigma_1^2(\tau-1)^2}{2B}\sum_{k=k_p}^{v_p}1.
\end{aligned}
$$

Since we have already required that

$$
-1+\eta L\lambda\left(\frac{C_1}{M}+C_2+1\right)+\eta^2L^2(\tau-1)\tau(C_1+1)\leq 0,
$$

thus the above the equation implies that

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k_{p+1}}) - \frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k_p})\right] \\
\leq & -\frac{\eta}{2}\sum_{k=k_p}^{v_p}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k})\right\|^2\right]+\frac{1}{2}\eta^2L\lambda\left\{\left(\frac{C_1}{M}+C_2+1\right)\sigma_{\text{dif}}^2+\frac{\sigma_1^2}{MB}+\frac{\sigma_2^2}{B}\right\}\sum_{k=k_p}^{v_p}1 \\
+ & \frac{1}{2}\eta^3L^2\sigma_{\text{dif}}^2(\tau-1)^2(C_1+1)\sum_{k=k_p}^{v_p}1+\frac{\eta^3L^2\sigma_1^2(\tau-1)^2}{2B}\sum_{k=k_p}^{v_p}1.
\end{aligned}
$$

Note that we have assumed that $K = k_{\bar{p}}$ for some $\bar{p} \in \mathbb{N}^+$. This way, we further have

$$\begin{aligned}
&\frac{1}{K}\mathbb{E}\left[\left(\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^K) - f^*\right) - \left(\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^0) - f^*\right)\right] \\
&\frac{1}{K}\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^K) - \frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^0)\right] \\
&= \frac{1}{K}\sum_{p=0}^{\bar{p}-1}\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k_{p+1}}) - \frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k_p})\right] \\
\leq& -\frac{\eta}{2K}\sum_{p=0}^{\bar{p}-1}\sum_{k=k_p}^{v_p}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^k)\right\|^2\right] + \frac{1}{2}\eta^2 L\lambda\left\{\left(\frac{C_1}{M} + C_2 + 1\right)\sigma_{\text{dif}}^2 + \frac{\sigma_1^2}{MB} + \frac{\sigma_2^2}{B}\right\}\frac{1}{K}\sum_{p=0}^{\bar{p}-1}\sum_{k=k_p}^{v_p} 1 \\
+&\frac{1}{2}\eta^3 L^2\sigma_{\text{dif}}^2(\tau-1)^2(C_1+1)\frac{1}{K}\sum_{p=0}^{\bar{p}-1}\sum_{k=k_p}^{v_p} 1 + \frac{\eta^3 L^2\sigma_1^2(\tau-1)^2}{2B}\frac{1}{K}\sum_{p=0}^{\bar{p}-1}\sum_{k=k_p}^{v_p} 1 \\
=& -\frac{\eta}{2K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^k)\right\|^2\right] + \frac{1}{2}\eta^2 L\lambda\left\{\left(\frac{C_1}{M} + C_2 + 1\right)\sigma_{\text{dif}}^2 + \frac{\sigma_1^2}{MB} + \frac{\sigma_2^2}{B}\right\} \\
+&\frac{1}{2}\eta^3 L^2\sigma_{\text{dif}}^2(\tau-1)^2(C_1+1) + \frac{\eta^3 L^2\sigma_1^2(\tau-1)^2}{2B},
\end{aligned}$$

which implies that

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^k)\right\|^2\right] \\
\leq&\frac{2\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^0) - f^*\right]}{\eta K} + \eta L\lambda\left\{\left(\frac{C_1}{M} + C_2 + 1\right)\sigma_{\text{dif}}^2 + \frac{\sigma_1^2}{MB} + \frac{\sigma_2^2}{B}\right\} \\
+&\eta^2 L^2\sigma_{\text{dif}}^2(\tau-1)^2(C_1+1) + \frac{\eta^2 L^2\sigma_1^2(\tau-1)^2}{B}.
\end{aligned}$$

□

Proof of Theorem D.3. By Lemmas D.4, D.5, D.7 and D.8, for $k_p + 1 \leq k \leq v_p$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^{k+1}) - f^*\right] \\
\leq&\Delta_k\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^k) - f^*\right] + \frac{\eta_k}{2}\left\{-1 + \eta_k\lambda L\left(\frac{C_1}{M} + C_2 + 1\right)\right\}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m^k)\right\|^2\right] \\
+&B_k\sum_{k=k_p}^{t-1}\eta_k^2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m^{(k)})\right\|^2\right] + c_t,
\end{aligned}$$

where

$$\begin{aligned}
\Delta_k =& 1-\eta_k\mu,\\
B_k =& \frac{1}{2}\eta_k L^2\lambda(\tau-1)(C_1+1),\\
c_k =& \frac{\eta_k L^2}{2}\left\{\sigma_{\text{dif}}^2(\tau-1)(C_1+1)\sum_{t=k_p}^{k-1}\eta_k^2+\frac{\sigma_1^2(\tau-1)}{B}\sum_{t=k_p}^{k-1}\eta_k^2\right\}\\
&+\frac{\eta_k^2 L}{2}\left\{\sigma_{\text{dif}}^2\left(\frac{C_1}{M}+C_2+1\right)+\frac{\sigma_1^2}{MB}+\frac{\sigma_2^2}{B}\right\}.
\end{aligned}$$

We also define

$$\begin{aligned}
a_k =& \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} f_m(\hat{\theta}_m^k)-f^*\right],\\
D =& \lambda L\left(\frac{C_1}{M}+C_2+1\right),\\
e_k =& \mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m^k)\right\|^2\right],
\end{aligned}$$

and denote

$$\begin{aligned}
&\sum_{k=k_p}^{k_p-1}\eta_k^2 e_t=0,\\
&c_{k_p}=\frac{\alpha_{k_p}^2 L}{2}\left\{\sigma_{\text{dif}}^2\left(\frac{C_1}{M}+C_2+1\right)+\frac{\sigma_1^2}{MB}+\frac{\sigma_2^2}{B}\right\},
\end{aligned}$$

then we have

$$a_{k+1}\le\Delta_k a_k+\frac{\eta_k}{2}(-1+D\eta_k)e_k+B_k\sum_{t=k_p}^{k-1}\eta_k^2 e_t+c_k$$

for all $k_p\le k\le v_p$. Under the conditions for β and τ, by Lemmas D.9 and D.10, we have

$$a_{v_p+1}\le\left(\prod_{k=k_p}^{v_p}\Delta_k\right)a_{k_p}+\sum_{k=k_p}^{v_p-1}\left(\prod_{i=k+1}^{v_p}\Delta_i\right)c_k+c_{v_p}. \tag{32}$$

Let $z_k=(k+b)^2$, where $b=\beta\tau+1$, we have

$$\begin{aligned}
\Delta_k\frac{z_k}{\eta_k} =& (1-\mu\eta_k)\mu(k+b)^3\\
=& (1-\frac{1}{k+b})\mu(k+b)^3\\
=& \mu(k+b-1)(k+b)^2\\
\le& \mu(k+b-1)^3\\
=& \frac{z_{k-1}}{\eta_{k-1}}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\frac{z_{v_p}}{\eta_{v_p}}\left(\prod_{i=k+1}^{v_p}\Delta_i\right)\\
=&\frac{z_{v_p}}{\eta_{v_p}}\Delta_{v_p}\left(\prod_{i=k+1}^{v_p-1}\Delta_i\right)\\
\le&\frac{z_{v_p-1}}{\eta_{v_p-1}}\left(\prod_{i=k+1}^{v_p-1}\Delta_i\right)\\
&\vdots\\
\le&\frac{z_k}{\eta_k}.
\end{aligned}$$

This way, note that $v_p+1=k_{p+1}$, and plug the above inequality into (32), we then get

$$\frac{z_{v_p}}{\eta_{v_p}}a_{k_{p+1}}\le\frac{z_{k_p}}{\eta_{k_p}}a_{k_p}+\sum_{k=k_p}^{v_p}\frac{z_k}{\eta_k}c_k.$$

Since we have assumed that $K=k_{\bar{p}}$, thus we have

$$\begin{aligned}
\frac{z_{K-1}}{\eta_{K-1}}a_K=&\frac{z_{v_{\bar{p}-1}}}{\eta_{v_{\bar{p}-1}}}a_{k_{\bar{p}}}\\
\le&\frac{z_{k_{\bar{p}-1}}}{\eta_{k_{\bar{p}-1}}}a_{k_{\bar{p}-1}}+\sum_{t=k_{\bar{p}-1}}^{v_{\bar{p}-1}}\frac{z_t}{\eta_t}c_t\\
&\vdots\\
\le&\frac{z_0}{\eta_0}a_0+\sum_{k=0}^{K-1}\frac{z_k}{\eta_k}c_k.
\end{aligned}\tag{33}$$

Recall that for $k_p\le k\le v_p$

$$\begin{aligned}
c_k=&\frac{\eta_k L^2}{2}\left\{\sigma^2_{\text{dif}}(\tau-1)(C_1+1)\sum_{k=k_p}^{t-1}\eta_k^2+\frac{\sigma_1^2(\tau-1)}{B}\sum_{k=k_p}^{t-1}\eta_k^2\right\}\\
&+\frac{\eta_k^2 L}{2}\left\{\sigma^2_{\text{dif}}\left(\frac{C_1}{M}+C_2+1\right)+\frac{\sigma_1^2}{MB}+\frac{\sigma_2^2}{B}\right\}\\
\le&\frac{\eta_k\eta^2_{\lfloor\frac{k}{\tau}\rfloor\tau}L^2(\tau-1)^2}{2}\left\{\sigma^2_{\text{dif}}(C_1+1)+\frac{\sigma_1^2}{B}\right\}\\
&+\frac{\eta_k^2 L}{2}\left\{\sigma^2_{\text{dif}}\left(\frac{C_1}{M}+C_2+1\right)+\frac{\sigma_1^2}{MB}+\frac{\sigma_2^2}{B}\right\}.
\end{aligned}$$

Then we have

$$\begin{aligned}
\sum_{k=0}^{K-1}\frac{z_k}{\eta_k}c_k\le&\frac{L^2(\tau-1)^2}{2}\left\{\sigma^2_{\text{dif}}(C_1+1)+\frac{\sigma_1^2}{B}\right\}\sum_{k=0}^{K-1}z_k\eta^2_{\lfloor\frac{k}{\tau}\rfloor\tau}\\
&+\frac{L}{2}\left\{\sigma^2_{\text{dif}}\left(\frac{C_1}{M}+C_2+1\right)+\frac{\sigma_1^2}{MB}+\frac{\sigma_2^2}{B}\right\}\sum_{k=0}^{K-1}z_k\eta_k.
\end{aligned}\tag{34}$$

First, assume that $k = p\tau + r$, where $0 \leq r \leq \tau - 1$, then we have

$$\left\lfloor \frac{t}{\tau} \right\rfloor \tau + b = p\tau + \beta\tau + 1 = (p+\beta)\tau + 1 \geq \beta\tau \geq r,$$

as we have assumed that $\beta > 1$, thus

$$2\left(\left\lfloor \frac{k}{\tau} \right\rfloor \tau + b \right) \geq (p+\beta)\tau + 1 + r = k + b.$$

This way, we have

$$\begin{aligned} \sum_{k=0}^{K-1} z_k \eta^2_{\lfloor \frac{k}{\tau} \rfloor \tau} =& \frac{1}{\mu^2} \sum_{k=0}^{K-1} \left(\frac{k+b}{\lfloor \frac{k}{\tau} \rfloor \tau + b} \right)^2 \\ \leq& \frac{4K}{\mu^2}. \end{aligned}$$

Next, note that

$$\sum_{k=0}^{K-1} z_k \eta_k = \frac{1}{\mu} \sum_{k=0}^{K-1} (k+b) \leq \frac{K(K+2b)}{2\mu}. \tag{35}$$

Combine equations (33)-(35), we have

$$\begin{aligned} \mathbb{E}\left[\frac{1}{M} \sum_{m=1}^{M} f_m\left(\hat{\theta}_m^K\right) - f^* \right] \leq& \frac{b^3}{(K+\beta\tau)^3} \mathbb{E}\left[\frac{1}{M} \sum_{m=1}^{M} f_m\left(\hat{\theta}_m^0\right) - f^* \right] \\ &+ \frac{2L^2(\tau-1)^2 K}{\mu^3 (K+\beta\tau)^3} \left\{ \sigma^2_{\text{dif}}(C_1+1) + \frac{\sigma_1^2}{B} \right\} \\ &+ \frac{LK(K+2\beta\tau+2)}{4\mu^2(K+\beta\tau)^3} \left\{ \sigma^2_{\text{dif}}\left(\frac{C_1}{M} + C_2 + 1 \right) + \frac{\sigma_1^2}{MB} + \frac{\sigma_2^2}{B} \right\}. \end{aligned}$$

□

We state some useful facts in the following proposition.

Proposition 1. *If f is differentiable and L-smooth, then we have*

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq \frac{L}{2}\|x-y\|^2. \tag{36}$$

If f is also convex, we then have

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2LD_f(x,y) \tag{37}$$

for all x, y.

Besides, for all vectors x, y, we have

$$2\langle x, y \rangle \leq \xi\|x\|^2 + \xi^{-1}\|y\|^2 \quad \forall \xi > 0 \tag{38}$$

$$-\langle x, y \rangle = -\frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2 + \frac{1}{2}\|x-y\|^2. \tag{39}$$

For vectors $v_1 . v_2, \ldots, v_n$, by Jensen's inequality and the convexity of map: $x \mapsto \|x\|^2$, we have

$$\left\| \frac{1}{n} \sum_{i=1}^{n} v_i \right\|^2 \leq \frac{1}{n} \sum_{i=1}^{n} \|v_i\|^2. \tag{40}$$

Lemma D.4. *Under Assumption D.1, given $\{\theta_m^k\}_{m\in[M]}$, taking conditional expectation with respect to ξ^k, we have*

$$\begin{aligned}&\mathbb{E}_{\xi^k}\left[\frac{1}{M}\sum_{m=1}^M f_m(\hat{\theta}_m^{k+1})\right]-\frac{1}{M}\sum_{m=1}^M f_m(\hat{\theta}_m^k)\\ \leq&-\eta_k\left\langle\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\hat{\theta}_m^k),\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\theta_m^k)\right\rangle-\frac{\eta_k}{M}\sum_{m=1}^M\langle\nabla_{\beta_m}f_m(\hat{\theta}_m^k),\nabla_{\beta_m}f_m(\theta_m^k)\rangle\\ &+\frac{\eta_k^2L}{2}\mathbb{E}_{\xi^k}\left[\|h^k\|^2\right]+\frac{\eta_k^2L}{2M}\sum_{m=1}^M\mathbb{E}_{\xi_m^k}\left[\|g_{m,2}^k\|^2\right]\end{aligned}$$

Proof. By L-smoothness assumption of $f_m(\cdot)$ and (36), we have

$$f_m(\hat{\theta}_m^{k+1})-f_m(\hat{\theta}_m^k)-\langle\nabla f_m(\hat{\theta}_m^k),\hat{\theta}_m^{k+1}-\hat{\theta}_m^k\rangle\leq\frac{L}{2}\|\hat{\theta}_m^{k+1}-\hat{\theta}_m^k\|^2.$$

Thus, we have

$$f_m(\hat{\theta}_m^{k+1})-f_m(\hat{\theta}_m^k)\leq-\eta_k\langle\nabla_w f_m(\hat{\theta}_m^k),h^k\rangle-\eta_k\langle\nabla_{\beta_m}f_m(\hat{\theta}_m^k),g_{m,2}^k\rangle+\frac{\eta_k^2L}{2}\|h^k\|^2+\frac{\eta_k^2L}{2}\|g_{m,2}^k\|^2,$$

which further implies that

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^M f_m(\hat{\theta}_m^{k+1})-\frac{1}{M}\sum_{m=1}^M f_m(\hat{\theta}_m^k)\leq&-\eta_k\left\langle\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\hat{\theta}_m^k),h^k\right\rangle-\frac{\eta_k}{M}\sum_{m=1}^M\langle\nabla_{\beta_m}f_m(\hat{\theta}_m^k),g_{m,2}^k\rangle\\ &+\frac{\eta_k^2L}{2}\|h^k\|^2+\frac{\eta_k^2L}{2M}\sum_{m=1}^M\|g_{m,2}^k\|^2.\end{aligned}$$

Finally, taking conditional expectation with respect to ξ^k, we have

$$\begin{aligned}&\mathbb{E}_{\xi^k}\left[\frac{1}{M}\sum_{m=1}^M f_m(\hat{\theta}_m^{k+1})\right]-\frac{1}{M}\sum_{m=1}^M f_m(\hat{\theta}_m^k)\\ \leq&-\eta_k\left\langle\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\hat{\theta}_m^k),\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\theta_m^k)\right\rangle-\frac{\eta_k}{M}\sum_{m=1}^M\langle\nabla_{\beta_m}f_m(\hat{\theta}_m^k),\nabla_{\beta_m}f_m(\theta_m^k)\rangle\\ &+\frac{\eta_k^2L}{2}\mathbb{E}_{\xi^k}\left[\|h^k\|^2\right]+\frac{\eta_k^2L}{2M}\sum_{m=1}^M\mathbb{E}_{\xi_m^k}\left[\|g_{m,2}^k\|^2\right]\end{aligned}$$

□

Lemma D.5. *Under Assumptions D.2 and D.3, given $\{\theta_m^k\}_{m\in[M]}$, taking conditional expectation with respect to ξ^k, we have*

$$\begin{aligned}&\mathbb{E}_{\xi^k}\left[\|h^k\|^2\right]+\frac{1}{M}\sum_{m=1}^M\mathbb{E}_{\xi_m^k}\left[\|g_{m,2}^k\|^2\right]\\ \leq&\left(\frac{C_1}{M}+C_2+1\right)\frac{1}{M}\sum_{m=1}^M\|\nabla f_m(\theta_m^k)\|^2+\frac{\sigma_1^2}{MB}+\frac{\sigma_2^2}{B}\\ \leq&\lambda\left(\frac{C_1}{M}+C_2+1\right)\left\|\frac{1}{M}\sum_{m=1}^M\nabla f_m(\theta_m^k)\right\|^2+\left(\frac{C_1}{M}+C_2+1\right)\sigma_{dif}^2+\frac{\sigma_1^2}{MB}+\frac{\sigma_2^2}{B}.\end{aligned}$$

Proof. Note that

$$\begin{aligned}
\mathbb{E}_{\xi^k}\left[\|h^k\|^2\right] =&\mathbb{E}_{\xi^k}\left[\left\|\frac{1}{M}\sum_{m=1}^M g_{m,1}^k\right\|^2\right]\\
\overset{(i)}{=}&\mathbb{E}_{\xi^k}\left[\left\|\frac{1}{M}\sum_{m=1}^M \left(g_{m,1}^k-\nabla_w f_m(\theta_m^k)\right)\right\|^2\right]+\left\|\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\theta_m^k)\right\|^2\\
\overset{(ii)}{=}&\frac{1}{M^2}\sum_{m=1}^M\mathbb{E}_{\xi_m^k}\left[\|g_{m,1}^k-\nabla_w f_m(\theta_m^k)\|^2\right]+\left\|\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\theta_m^k)\right\|^2\\
\overset{(iii)}{\le}&\frac{1}{M^2}\sum_{m=1}^M\left(C_1\|\nabla f_m(\theta_m^k)\|^2+\frac{\sigma_1^2}{B}\right)+\left\|\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\theta_m^k)\right\|^2\\
\overset{(iv)}{\le}&\frac{C_1}{M^2}\sum_{m=1}^M\|\nabla f_m(\theta_m^k)\|^2+\frac{\sigma_1^2}{MB}+\frac{1}{M}\sum_{m=1}^M\|\nabla_w f_m(\theta_m^k)\|^2,
\end{aligned}$$

where (i) is due to that $g_{m,1}^k$ is unbiased, (ii) is by the fact that $\xi_1^k,\xi_2^k,\ldots,\xi_M^k$ are independent, (iii) is by Assumption D.2, and (iv) is by (40).

Similarly, we have

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^M\mathbb{E}_{\xi_m^k}\left[\|g_{m,2}^k\|^2\right]=&\frac{1}{M}\sum_{m=1}^M\mathbb{E}_{\xi_m^k}\left[\|g_{m,2}^k-\nabla_{\beta_m}f_m(\theta_m^k)\|^2\right]+\frac{1}{M}\sum_{m=1}^M\|\nabla_{\beta_m}f_m(\theta_m^k)\|^2\\
\le&\frac{C_2}{M}\sum_{m=1}^M\|\nabla f_m(\theta_m^k)\|^2+\frac{\sigma_2^2}{B}+\frac{1}{M}\sum_{m=1}^M\|\nabla_{\beta_m}f_m(\theta_m^k)\|^2.
\end{aligned}$$

Combine the above equations, we have the final result. □

Lemma D.6. *Under Assumption D.1, we have*

$$\begin{aligned}
&-\eta_k\left\langle\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\hat\theta_m^k),\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\theta_m^k)\right\rangle-\frac{\eta_k}{M}\sum_{m=1}^M\langle\nabla_{\beta_m}f_m(\hat\theta^k),\nabla_{\beta_m}f_m(\theta_m^k)\rangle\\
\le&-\frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^M\nabla f_m(\hat\theta_m^k)\right\|^2-\frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^M\nabla f_m(\theta_m^k)\right\|^2+\frac{\eta_k L^2}{2}V^k.
\end{aligned}$$

Proof. By (39), we have

$$\begin{aligned}
&-\eta_k\left\langle\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\hat\theta_m^k),\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\theta_m^k)\right\rangle\\
=&-\frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\hat\theta_m^k)\right\|^2-\frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\theta_m^k)\right\|^2+\frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^M\left(\nabla_w f_m(\hat\theta_m^k)-\nabla_w f_m(\theta_m^k)\right)\right\|^2\\
\le&-\frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\hat\theta_m^k)\right\|^2-\frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^M\nabla_w f_m(\theta_m^k)\right\|^2+\frac{\eta_k}{2M}\sum_{m=1}^M\left\|\nabla_w f_m(\hat\theta_m^k)-\nabla_w f_m(\theta_m^k)\right\|^2,
\end{aligned}$$

where the last inequality is by (40). And we also have

$$-\eta_k\langle\nabla_{\beta_m}f_m(\hat\theta^k),\nabla_{\beta_m}f_m(\theta_m^k)\rangle=-\frac{\eta_k}{2}\|\nabla_{\beta_m}f_m(\hat\theta_m^k)\|^2-\frac{\eta_k}{2}\|\nabla_{\beta_m}f_m(\theta_m^k)\|^2+\frac{\eta_k}{2}\|\nabla_{\beta_m}f_m(\hat\theta_m^k)-\nabla_{\beta_m}f_m(\theta_m^k)\|^2,$$

thus

$$\begin{aligned}
-\frac{\eta_k}{M}\langle \nabla_{\beta_m} f_m(\hat{\theta}^k), \nabla_{\beta_m} f_m(\theta_m^k)\rangle =& -\frac{\eta_k}{2M}\sum_{m=1}^{M}\|\nabla_{\beta_m} f_m(\hat{\theta}_m^k)\|^2 - \frac{\eta_k}{2M}\sum_{m=1}^{M}\|\nabla_{\beta_m} f_m(\theta_m^k)\|^2 \\
&+\frac{\eta_k}{2M}\sum_{m=1}^{M}\|\nabla_{\beta_m} f_m(\hat{\theta}_m^k) - \nabla_{\beta_m} f_m(\theta_m^k)\|^2 \\
\le& -\frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla_{\beta_m} f_m(\hat{\theta}_m^k)\right\|^2 - \frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla_{\beta_m} f_m(\theta_m^k)\right\|^2 \\
&+\frac{\eta_k}{2M}\sum_{m=1}^{M}\|\nabla_{\beta_m} f_m(\hat{\theta}_m^k) - \nabla_{\beta_m} f_m(\theta_m^k)\|^2.
\end{aligned}$$

Combine the above equations, we have

$$\begin{aligned}
&-\eta_k\left\langle \frac{1}{M}\sum_{m=1}^{M}\nabla_w f_m(\hat{\theta}_m^k), \frac{1}{M}\sum_{m=1}^{M}\nabla_w f_m(\theta_m^k)\right\rangle - \frac{\eta_k}{M}\sum_{m=1}^{M}\langle \nabla_{\beta_m} f_m(\hat{\theta}_m^k), \nabla_{\beta_m} f_m(\theta_m^k)\rangle \\
\le& -\frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\hat{\theta}_m^k)\right\|^2 - \frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m^k)\right\|^2 + \frac{\eta_k}{2M}\sum_{m=1}^{M}\left\|\nabla f_m(\hat{\theta}_m^k) - \nabla f_m(\theta_m^k)\right\|^2 \\
\overset{(i)}{\le}& -\frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\hat{\theta}_m^k)\right\|^2 - \frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m^k)\right\|^2 + \frac{\eta_k L^2}{2M}\sum_{m=1}^{M}\left\|w_m^k - w^k\right\|^2 \\
=& -\frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\hat{\theta}_m^k)\right\|^2 - \frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m^k)\right\|^2 + \frac{\eta_k L^2}{2}V^k,
\end{aligned}$$

where (i) is by Assumption D.1. □

Lemma D.7. *Under Assumptions D.1 and D.4, we have*

$$\begin{aligned}
&-\eta_k\left\langle \frac{1}{M}\sum_{m=1}^{M}\nabla_w f_m(\hat{\theta}_m^k), \frac{1}{M}\sum_{m=1}^{M}\nabla_w f_m(\theta_m^k)\right\rangle - \frac{\eta_k}{M}\sum_{m=1}^{M}\langle \nabla_{\beta_m} f_m(\hat{\theta}^k), \nabla_{\beta_m} f_m(\theta_m^k)\rangle \\
\le& -\eta_k\mu\left(\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\hat{\theta}_m^k) - f^*\right) - \frac{\eta_k}{2}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m^k)\right\|^2 + \frac{\eta_k L^2}{2}V^k.
\end{aligned}$$

Proof. The proof follows directly from Lemma D.6 and Assumption D.4. □

Lemma D.8. *Under Assumptions D.2 and D.3, for $k_p + 1 \le k \le v_p$, taking unconditional expectation, we have*

$$\begin{aligned}
\mathbb{E}\left[V^k\right] \le& \lambda(\tau-1)(C_1+1)\sum_{t=k_p}^{k-1}\eta_t^2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m^t)\right\|^2\right] \\
&+\sigma_{dif}^2(\tau-1)(C_1+1)\sum_{t=k_p}^{k-1}\eta_t^2 + \frac{\sigma_1^2(\tau-1)}{B}\sum_{t=k_p}^{k-1}\eta_t^2.
\end{aligned}$$

Note that $V^{k_p} = 0$.

Proof. By noting that $w^{k_p} = w_m^{k_p}$ for all $m \in [M]$, thus for $k_p + 1 \leq k \leq v_p$, we have

$$\begin{aligned}\|w_m^k - w^k\|^2 =& \left\| w_m^{k_p} - \sum_{t=k_p}^{k-1} \eta_t g_{m,1}^t - w^{k_p} - \sum_{t=k_p}^{k-1} \eta_t h^t \right\|^2 \\ =& \left\| \sum_{t=k_p}^{k-1} \eta_t g_{m,1}^t - \sum_{t=k_p}^{k-1} \eta_t h^t \right\|^2 .\end{aligned}$$

Since

$$\frac{1}{M} \sum_{m=1}^{M} \sum_{t=k_p}^{k-1} \eta_t g_{m,1}^t = \sum_{t=k_p}^{k-1} \eta_t h^t,$$

we have

$$\begin{aligned}\frac{1}{M} \sum_{m=1}^{M} \|w_m^k - w^k\|^2 =& \frac{1}{M} \sum_{m=1}^{M} \left\| \sum_{t=k_p}^{k-1} \eta_t g_{m,1}^t - \sum_{t=k_p}^{k-1} \eta_t h^t \right\|^2 \\ =& \frac{1}{M} \sum_{m=1}^{M} \left\| \sum_{t=k_p}^{k-1} \eta_t g_{m,1}^t \right\|^2 - \left\| \sum_{t=k_p}^{k-1} \eta_t h^t \right\|^2 \\ \leq& \frac{1}{M} \sum_{m=1}^{M} \left\| \sum_{t=k_p}^{k-1} \eta_t g_{m,1}^t \right\|^2 \\ \leq& \frac{k - k_p}{M} \sum_{m=1}^{M} \sum_{t=k_p}^{k-1} \eta_t^2 \|g_{m,1}^t\|^2 \\ \leq& \frac{\tau - 1}{M} \sum_{m=1}^{M} \sum_{t=k_p}^{k-1} \eta_t^2 \|g_{m,1}^t\|^2 .\end{aligned} \tag{41}$$

Given $\{\theta_m^k\}_{m \in [M]}$, taking conditional expectation with respect to ξ^k, we have

$$\begin{aligned}\mathbb{E}_{\xi^k} \left[\frac{1}{M} \sum_{m=1}^{M} \|g_{m,1}^k\|^2 \right] =& \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{\xi_m^k} \left[\|g_{m,1}^k\|^2 \right] \\ =& \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{\xi_m^k} \left[\|g_{m,1}^k - \nabla_w f_m(\theta_m^k)\|^2 \right] + \frac{1}{M} \sum_{m=1}^{M} \|\nabla_w f_m(\theta_m^k)\|^2 \\ \leq& \frac{1}{M} \sum_{m=1}^{M} \left[(C_1 + 1) \nabla \|f_m(\theta_m^k)\|^2 + \frac{\sigma_1^2}{B} \right] + \frac{1}{M} \sum_{m=1}^{M} \|\nabla f_m(\theta_m^k)\|^2 \\ =& \frac{C_1 + 1}{M} \sum_{m=1}^{M} \|\nabla f_m(\theta_m^k)\|^2 + \frac{\sigma_1^2}{B} .\end{aligned}$$

Thus, taking unconditional expectation on both sides of (41), and by the independence of $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{k}$, we have

$$
\begin{aligned}
\mathbb{E}\left[V^k\right] =&(\tau-1)\sum_{t=k_p}^{k-1}\eta_t^2\mathbb{E}\left[\mathbb{E}_{\xi^t}\left[\frac{1}{M}\sum_{m=1}^{M}\|g_{m,1}^t\|^2\right]\right] \\
\leq&(\tau-1)(C_1+1)\sum_{t=k_p}^{k-1}\eta_t^2\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\|\nabla f_m(\theta_m^t)\|^2\right]+\frac{(\tau-1)\sigma_1^2}{B}\sum_{t=k_p}^{k-1}\eta_t^2 \\
\leq&\lambda(\tau-1)(C_1+1)\sum_{t=k_p}^{k-1}\eta_t^2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\theta_m^t)\right\|^2\right] \\
&+\sigma_{\text{dif}}^2(\tau-1)(C_1+1)\sum_{t=k_p}^{k-1}\eta_t^2+\frac{(\tau-1)\sigma_1^2}{B}\sum_{t=k_p}^{k-1}\eta_t^2,
\end{aligned}
$$

where the last inequality follows Assumption D.3. □

Lemma D.9. *For the sequence $\{a_k\}_{k_p\leq k\leq v_p}$ in the proof of Theorem D.3 that satisfies*

$$
a_{k+1}\leq\Delta_k a_k+\frac{\eta_k}{2}(-1+D\eta_k)e_k+B_k\sum_{t=k_p}^{k-1}\eta_t^2 e_t+c_k,
$$

when learning rates $\{\eta_k\}$ satisfy

$$
\begin{aligned}
\eta_{v_p}\leq&\frac{1}{D}, && (42)\\
\eta_{v_p-1}\leq&\frac{1}{D+\frac{2B_{v_p}}{\Delta_{v_p}}}, && (43)\\
&\vdots \\
\eta_{k_p}\leq&\frac{1}{D+\frac{2}{\prod_{i=k_p+1}^{v_p}\Delta_i}\left[\left(\prod_{i=k_p+2}^{v_p}\Delta_i\right)B_{k_p+1}+\left(\prod_{i=k_p+3}^{v_p}\Delta_i\right)B_{k_p+2}+\Delta_{v_p}B_{v_p-1}+B_{v_p}\right]}, && (44)
\end{aligned}
$$

we have

$$
a_{v_p+1}\leq\left(\prod_{k=k_p}^{v_p}\Delta_k\right)a_{k_p}+\sum_{k=k_p}^{v_p-1}\left(\prod_{i=k+1}^{v_p}\Delta_i\right)c_k+c_{v_p}.
$$

Proof. We start by noting that

$$
\begin{aligned}
a_{v_p+1}\leq&\Delta_{v_p}a_{v_p}+\frac{\eta_{v_p}}{2}(-1+D\eta_{v_p})e_{v_p}+B_{v_p}\sum_{k=k_p}^{v_p-1}\eta_k^2 e_k+c_{v_p} \\
\leq&\Delta_{v_p}a_{v_p}+B_{v_p}\sum_{k=k_p}^{v_p-1}\eta_k^2 e_k+c_{v_p},
\end{aligned}
$$

where the last inequality is due to (42). Thus, we have

$$\begin{aligned}
a_{v_p+1} \leq & \Delta_{v_p} a_{v_p} + B_{v_p} \sum_{k=k_p}^{v_p-1} \eta_k^2 e_k + c_{v_p} \\
= & \Delta_{v_p} a_{v_p} + B_{v_p} \left(\sum_{k=k_p}^{v_p-2} \eta_k^2 e_k + \eta_{v_p-1}^2 e_{v_p-1} \right) + c_{v_p} \\
\leq & \Delta_{v_p} \left(\Delta_{v_p-1} a_{v_p-1} + \frac{\eta_{v_p-1}}{2}(-1 + D\eta_{v_p-1}) e_{v_p-1} + B_{v_p-1} \sum_{k=k_p}^{v_p-2} \eta_k^2 e_k + c_{v_p-1} \right) \\
+ & B_{v_p} \left(\sum_{k=k_p}^{v_p-2} \eta_k^2 e_k + \eta_{v_p-1}^2 e_{v_p-1} \right) + c_{v_p} \\
= & \Delta_{v_p} \Delta_{v_{p-1}} a_{v_p-1} + \frac{\eta_{v_{p-1}} \Delta_{v_p}}{2} \left[-1 + D\eta_{v_p-1} + \frac{2B_{v_p} \eta_{v_p-1}}{\Delta_{v_p}} \right] e_{v_p-1} \\
+ & \left(\Delta_{v_p} B_{v_p-1} + B_{v_p} \right) \sum_{k=k_p}^{v_p-2} \eta_k^2 e_k + \left(\Delta_{v_p} c_{v_p-1} + c_{v_p} \right).
\end{aligned}$$

By (43), we have

$$-1 + D\eta_{v_p-1} + \frac{2B_{v_p} \eta_{v_p-1}}{\Delta_{v_p}} \leq 0,$$

thus we have

$$a_{v_p+1} \leq \Delta_{v_p} \Delta_{v_{p-1}} a_{v_p-1} + \left(\Delta_{v_p} B_{v_p-1} + B_{v_p} \right) \sum_{k=k_p}^{v_p-2} \eta_k^2 e_k + \left(\Delta_{v_p} c_{v_p-1} + c_{v_p} \right).$$

Repeat this process under corresponding assumptions about η_k, we get

$$\begin{aligned}
a_{v_p+1} \leq & \left(\prod_{i=k_p+1}^{v_p} \Delta_i \right) a_{k_p+1} + \left[\left(\prod_{i=k_p+2}^{v_p} \Delta_i \right) B_{k_p+1} + \left(\prod_{i=k_p+3}^{v_p} \Delta_i \right) B_{k_p+2} + \Delta_{v_p} B_{v_p-1} + B_{v_p} \right] \eta_{k_p}^2 e_{k_p} \\
& + \sum_{k=k_p}^{v_p-1} \left(\prod_{i=k+1}^{v_p} \Delta_i \right) c_k.
\end{aligned}$$

Note that

$$a_{k_{p+1}} \leq \Delta_{k_p} a_{k_p} + \frac{\eta_{k_p}}{2}(-1 + D\eta_{k_p}) e_{k_p} + c_{k_p},$$

and (44), we then have the final result. □

Lemma D.10. *Let*

$$\eta_k = \frac{1}{\mu(k + \beta\tau + 1)},$$

where β is a positive constant such that

$$\beta > \max \left\{ \frac{2\lambda L}{\mu} \left(\frac{C_1}{M} + C_2 + 1 \right) - 2, \frac{2L^2 \lambda (C_1 + 1)}{\mu^2} \right\}.$$

When τ is large enough such that

$$\tau \geq \sqrt{\frac{\max \left\{ (2L^2 \lambda (C_1 + 1)/\mu^2) e^{1/\beta} - 4, 0 \right\}}{\beta^2 - (2L^2 \lambda (C_1 + 1)/\mu^2) e^{\frac{1}{\beta}}}},$$

we will have conditions in Lemma D.9 to be satisfied for any η_k.

Proof. Let Δ_k and B_k be defined as in the proof of Theorem D.3, then as k increases, we have η_k decreases, Δ_k increases, and B_k decreases. Thus for $1 \leq k \leq K$, we have $\alpha_K \leq \alpha_{K-1} \leq \cdots \leq \alpha_1$. On the other hand, note that for the right hand side of (44), for $1 \leq k \leq K$, we have

$$\begin{aligned}
&\frac{1}{D + \frac{2}{\prod_{i=k_p+1}^{v_p} \Delta_i} \left[\left(\prod_{i=k_p+2}^{v_p} \Delta_i \right) B_{k_p+1} + \left(\prod_{i=k_p+3}^{v_p} \Delta_i \right) B_{k_p+2} + \Delta_{v_p} B_{v_p-1} + B_{v_p} \right]} \\
\geq &\frac{1}{D + \frac{2B_1}{\prod_{i=k_p+1}^{v_p} \Delta_1} \left[\left(\prod_{i=k_p+2}^{v_p} \Delta_K \right) + \left(\prod_{i=k_p+3}^{v_p} \Delta_K \right) + \Delta_K \right]} \\
\geq &\frac{1}{D + \frac{2B_1(\tau-1)}{\Delta_1^{\tau-1}}}.
\end{aligned}$$

Thus, when we let

$$\eta_1 \leq \frac{1}{D + \frac{2B_1(\tau-1)}{\Delta_1^{\tau-1}}},$$

we will have the conditions in Lemma D.9 to be satisfied for any η_k. For this purpose, we need to have

$$\begin{aligned}
&\left(D + \frac{2B_1(\tau-1)}{\Delta_1^{\tau-1}} \right) \tau_1 \leq 1 \\
\iff &\left(\lambda L \left(\frac{C_1}{M} + C_2 + 1 \right) + \frac{\eta_1 L^2 \lambda (\tau-1)^2 (C_1+1)}{(1-\eta_1\mu)^{\tau-1}} \right) \eta_1 \leq 1 \\
\iff &\lambda L \left(\frac{C_1}{M} + C_2 + 1 \right) (1-\eta_1\mu)^{\tau-1} + \eta_1 L^2 \lambda (\tau-1)^2 (C_1+1) \leq \frac{(1-\eta_1\mu)^{\tau-1}}{\eta_1}.
\end{aligned}$$

To satisfy the above equation, we only need

$$\begin{cases} \lambda L \left(\frac{C_1}{M} + C_2 + 1 \right) (1-\eta_1\mu)^{\tau-1} \leq \frac{(1-\eta_1\mu)^{\tau-1}}{2\eta_1} \\ \eta_1 L^2 \lambda (\tau-1)^2 (C_1+1) \leq \frac{(1-\eta_1\mu)^{\tau-1}}{2\eta_1}. \end{cases} \tag{45}$$

Note that $\eta_1 = 1/(\mu(\beta\tau+2))$, thus to satisfy the first inequality in (45), we need to have

$$2\lambda L \left(\frac{C_1}{M} + C_2 + 1 \right) \leq \frac{1}{\eta_1} = \mu(\beta\tau+2).$$

Since $\mu(\beta\tau+2) \geq \mu(\beta+2)$, we only need

$$\beta \geq \frac{2\lambda L}{\mu} \left(\frac{C_1}{M} + C_2 + 1 \right) - 2. \tag{46}$$

Next, to satisfy the second inequality in (45), we need

$$\begin{aligned}
&2\eta_1^2 L^2 \lambda (\tau-1)^2 (C_1+1) \leq (1-\eta_1\mu)^{\tau-1} \\
\iff &\frac{2L^2\lambda(C_1+1)}{\mu^2} \left(\frac{\tau-1}{\beta\tau+2} \right)^2 \left(\frac{\beta\tau+2}{\beta\tau+1} \right)^{\tau-1} \leq 1.
\end{aligned}$$

Since

$$\left(\frac{\beta\tau+2}{\beta\tau+1} \right)^{\tau-1} = \left(1 + \frac{1}{\beta\tau+1} \right)^{\tau-1} = \left(1 + \frac{(\tau-1)/(\beta\tau+1)}{\tau-1} \right)^{\tau-1} \leq \exp\left\{ \frac{\tau-1}{\beta\tau+1} \right\} \leq e^{\frac{1}{\beta}},$$

thus we only need

$$\frac{2L^2\lambda(C_1+1)}{\mu^2} \left(\frac{\tau-1}{\beta\tau+2} \right)^2 e^{\frac{1}{\beta}} \leq 1.$$

Let $\nu = 2L^2\lambda(C_1+1)/\mu^2$, then the above equation is equivalent to

$$(\beta^2 - \nu e^{\frac{1}{\beta}})\tau^2 + 2(\beta + \nu e^{\frac{1}{\beta}})\tau + (4 - \nu e^{\frac{1}{\beta}}) \geq 0.$$

First, we let $\beta^2 - \nu e^{\frac{1}{\beta}} > 0$, or equivalently

$$\frac{\beta^2}{e^{\frac{1}{\beta}}} > \frac{2L^2\lambda(C_1+1)}{\mu^2}, \tag{47}$$

then we need τ to be large enough such that

$$\tau \geq \frac{-2(\beta + \nu e^{\frac{1}{\beta}}) + \sqrt{4(\beta + \nu e^{\frac{1}{\beta}})^2 - \max\left\{4(\beta^2 - \nu e^{\frac{1}{\beta}})(4 - \nu e^{\frac{1}{\beta}}), 0\right\}}}{2(\beta^2 - \nu e^{\frac{1}{\beta}})}.$$

Since $\sqrt{a^2+b} \leq |a| + \sqrt{|b|}$ for any $a, b \in \mathbb{R}$, we have that the left hand side is smaller or equal than

$$\sqrt{\frac{\max\left\{\nu e^{1/\beta} - 4, 0\right\}}{\beta^2 - \nu e^{\frac{1}{\beta}}}} = \sqrt{\frac{\max\left\{(2L^2\lambda(C_1+1)/\mu^2)e^{1/\beta} - 4, 0\right\}}{\beta^2 - (2L^2\lambda(C_1+1)/\mu^2)e^{\frac{1}{\beta}}}},$$

thus we only need τ large enough to satisfy that

$$\tau \geq \sqrt{\frac{\max\left\{(2L^2\lambda(C_1+1)/\mu^2)e^{1/\beta} - 4, 0\right\}}{\beta^2 - (2L^2\lambda(C_1+1)/\mu^2)e^{\frac{1}{\beta}}}}. \tag{48}$$

The final result follows from the combination of (46)-(48). □

E. Missing parts from Section 4

E.1. Missing parts from Section 4.1

E.1.1. Formal definition of the oracle and algorithm restriction

As mentioned in the main body, our lower bound is provided for iterative algorithms whose iterates lie in the span of historical oracle queries only. In particular, for each m, k we must have

$$\beta_m^k \in \mathrm{Lin}\left(\beta_m^0, \nabla_\beta f_m(w_m^0, \beta_m^0), \dots, \nabla_\beta f_m(w_m^{k-1}, \beta_m^{k-1})\right)$$

Next, define

$$Q^k = \cup_{m=1}^M \left\{w^0, \nabla_w f_m(w_m^0, \beta_m^0), \dots, \nabla_w f_m(w_m^{l(k)}, \beta_m^{l(k)})\right\},$$

with $l(k)$ being the index of the last communication round until iteration k. Then, the span requirement on w-variables is given as

$$w_m^k \in \mathrm{Lin}\left(w^0, \nabla_w f_m(w_m^0, \beta_m^0), \dots, \nabla_w f_m(w_m^{k-1}, \beta_m^{k-1}), \mathrm{Q}_k\right).$$

E.1.2. Proof of Theorem 4.1

Nesterov's worst case objective. (Nesterov et al., 2018) Let $h' : \mathbb{R}^\infty \to \mathbb{R}$ be the Nesterov's worst case objective (see), i.e., $h'(y) = \frac{1}{2}y^\top Ay - e_1^\top y$ with tridiagonal A having diagonal elements equal to $2+c$ (for some $c > 0$) and offdiagonal elements equal to 1.[9] The proof rationale is to show that a k-th iterate of any first order method must satisfy $\|y^k\|_0 \leq k$ and consequently

$$\|y^k - y^*\|^2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|y^*\|^2 \tag{49}$$

where $y^* := \mathrm{argmin}_{y \in \mathbb{R}^\infty} h'(y)$, $\kappa := \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$.

[9] This is for the strongly convex case; one can do convex similarly.

Finite sum worst case objective. (**Lan & Zhou, 2018**) The construction of the worst case finite-sum objective[10] $h : \mathbb{R}^\infty \to \mathbb{R}, h(z) = \frac{1}{n}\sum_{j=1}^n h_j(z)$ is such that h_j corresponds only on a j-th block of the coordinates; in particular if $z = [z_1, z_2, \dots, z_n]; z_1, z_2, \dots, z_n \in \mathbb{R}^\infty$ we set $h_j(z) = h'(z_j)$. It was shown that to reach $\|z^k - z^*\|^2 \leq \epsilon$ one requires at least $\Omega\left(\left(n + \sqrt{\frac{n\mathcal{L}}{\mu}}\right)\log\frac{1}{\epsilon}\right)$ iterations for $\mathcal{L}$-smooth functions h_j and $\mu-$strongly convex h.

Distributed worst case objective. (**Scaman et al., 2018**) Define

$$g'_1(z) := \frac{1}{2}\left(c_1\|z\|^2 + c_2\left(e_1^\top z + z^\top \mathbf{M}_1 z\right)\right)$$
$$g'_2(z) = g'_3(z) = \dots = g'_M(z) := \frac{1}{2(M-1)}\left(c_1\|z\|^2 + c_2 z^\top \mathbf{M}_2 z\right)$$

where $\mathbf{M}_1$ is an infinite block diagonal matrix with blocks $\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ and $\mathbf{M}_2 := \begin{pmatrix} 1 & 0 & \\ 0 & 0 & \\ & & \mathbf{M}_1 \end{pmatrix}$ and $c_1, c_2 > 0$ are some constants determining the smoothness and strong convexity of the objective. The worst case objective of Scaman et al. (2018) is now $g(z) = \frac{1}{M}\sum_{m=1}^m g'_m(z)$.

Distributed worst case objective with local finite sum. (**Hendrikx et al., 2020**) The given construction is obtained from the one of Scaman et al. (2018) in the same way as the worst case finite sum objective (Lan & Zhou, 2018) was obtained from the construction of Nesterov et al. (2018). In particular, one would set $g_{m,j}(z) = g'_m(z_j)$ where $z = [z_1, z_2, \dots, z_n]$. Next, it was shown that such a construction with properly chosen c_1, c_2 yields a lower bound on the communication complexity of order $\Omega\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$ and the lower bound on the local computation of order $\Omega\left(\left(n + \sqrt{\frac{n\mathcal{L}}{\mu}}\right)\log\frac{1}{\epsilon}\right)$ where $\mathcal{L}$ is a smoothness constant of $g_{m,j}$, L is a smoothness constant of $g_m(z) = \frac{1}{n}\sum_{j=1}^n g_j(z)$ and μ is the strong convexity constant of $g(z) = \frac{1}{M}\sum_{m=1}^M g_m(z)$.

Our construction and sketch of the proof. Now, our construction is straightforward – we set $f_m(w, \beta_m) = g(w) + h(\beta_m)$ with g, h scaled appropriately such that the strong convexity ratio is as per Assumption 1.1. Clearly, to minimize the global part $g(w)$, we require at least $\Omega\left(\sqrt{\frac{L^w}{\mu}}\log\frac{1}{\epsilon}\right)$ iterations and at least $\Omega\left(\left(n + \sqrt{\frac{n\mathcal{L}^w}{\mu}}\right)\log\frac{1}{\epsilon}\right)$ stochastic gradients of g. Similarly, to minimize h, we require at least $\Omega\left(\left(n + \sqrt{\frac{n\mathcal{L}^\beta}{\mu}}\right)\log\frac{1}{\epsilon}\right)$ stochastic gradients of h. Therefore, Theorem 4.1 is established.

E.2. Missing parts from Section 4.2

We state the ACD-PFL as Algorithm 2.

E.3. Missing parts from Section 4.3

E.3.1. Sampling of coordinate blocks

The key component of ASVRCD-PFL is the construction of the (unbiased) stochastic gradient estimator of $\nabla\mathbf{F}(X)$ which we describe here. We consider two independent sources of randomness:

• We toss an unfair coin ζ. With probability p_w we have $\zeta = 1$. In such a case, we ignore the local variables and update the global variables only (corresponding to w or $X[1]$ in our current notation). Alternatively, $\zeta = 2$ with probability $p_\beta := 1 - p_w$. In such a case, we ignore the global variables and update local variables only (corresponding to β or $X[2]$ in our current notation).

• Local subsampling. At each iteration the stochastic gradient is constructed using ∇F_j only, where $F_j(w, \beta) :=$

[10] We have lifted their construction to the infinite-dimensional space for the sake of simplicity. One can get a similar finite-dimensional results.

Algorithm 2 ACD-PFL

```
input $0 < \theta < 1$, $\eta, \nu > 0$, $w_y^0 = w_z^0 \in \mathbb{R}^{d_0}$, $\beta_{y,m}^0 = \beta_{z,m}^0 \in \mathbb{R}^{d_m}$ for $1 \leq m \leq M$.
  for $k = 0, 1, 2, \dots$ do
    $w_x^{k+1} = (1-\theta) w_y^k + \theta w_z^k$
    for $m = 1, \dots, M$ in parallel do
      $\beta_{x,m}^{k+1} = (1-\theta)\beta_{y,m}^k + \theta \beta_{z,m}^k$
    end for
    $\xi = \begin{cases} 1 & \text{w. p. } p_w = \frac{\sqrt{L^w}}{\sqrt{L^w}+\sqrt{L^\beta}} \\ 0 & \text{w. p. } p_\beta = \frac{\sqrt{L^\beta}}{\sqrt{L^w}+\sqrt{L^\beta}} \end{cases}$
    if $\xi = 0$ then
      $w_y^{k+1} = w_x^{k+1} - \frac{1}{L^w}\frac{1}{M}\sum_{m=1}^M \nabla_w f_m(w_x^{k+1}, \beta_{x,m}^{k+1})$
      $w_z^{k+1} = \frac{1}{1+\eta\nu}\left(w_z^k + \eta\nu w_x^{k+1} - \frac{\eta}{\sqrt{L^w}(\sqrt{L^w}+\sqrt{L^\beta})}\frac{1}{M}\sum_{m=1}^M \nabla_w f_m(w_x^{k+1}, \beta_{x,m}^{k+1})\right)$
      for $m = 1, \dots, M$ in parallel do
        $\beta_{z,m}^{k+1} = \frac{1}{1+\eta\nu}\left(\beta_{z,m}^k + \eta\nu\beta_{x,m}^{k+1}\right)$
      end for
    else
      for $m = 1, \dots, M$ in parallel do
        $\beta_{y,m}^{k+1} = \beta_{x,m}^{k+1} - \frac{1}{L^\beta}\nabla_\beta f_m(w_x^{k+1}, \beta_{x,m}^{k+1})$
      end for
      $\beta_{z,m}^{k+1} = \frac{1}{1+\eta\nu}\left(\beta_{z,m}^k + \eta\nu\beta_{x,m}^{k+1} - \frac{\eta}{\sqrt{L^\beta}(\sqrt{L^w}+\sqrt{L^\beta})}\nabla_\beta f_m(w_x^{k+1}, \beta_{x,m}^{k+1})\right)$
      $w_z^{k+1} = \frac{1}{1+\eta\nu}\left(w_z^k + \eta\nu w_x^{k+1}\right)$
    end if
  end for
```

$\frac{1}{M}\sum_{m=1}^M f_{m,j}(w, \beta_m)$ and j is selected uniformly at random from $[n]$.[11]

Overall, we arrive at the following construction of $\mathbf{G}(X)$ – an unbiased (stochastic) estimator of $\nabla\mathbf{F}(X)$:

$$\mathbf{G}(X)[1, m, j'] = \begin{cases} \frac{1}{p^w}\nabla_w f_{j',m}(X[1, m, j'], X[2, m, j']) & \begin{array}{c}\text{if } \zeta = 1 \\ \text{and } j' = j\end{array} \\ 0 \in \mathbb{R}^{d_0} & \text{otherwise} \end{cases}$$

$$\mathbf{G}(X)[2, m, j'] = \begin{cases} \frac{1}{p^\beta}\nabla_\beta f_{j',m}(X[1, m, j'], X[2, m, j']) & \begin{array}{c}\text{if } \zeta = 2 \\ \text{and } j' = j\end{array} \\ 0 \in \mathbb{R}^{d_m} & \text{otherwise} \end{cases}$$

Next, we enrich the stochastic gradient by control variates resulting in SVRG stochastic gradient estimator. In particular, the resulting stochastic gradient will take the form of $\mathbf{G}(X) - \mathbf{G}(Y) + \nabla\mathbf{F}(Y)$ where Y is another point that is updated upon a successful toss of a ρ-coin. The last ingredient of the method is to incorporate Nesterov's momentum. We state ASVRCD-PFL as Algorithm 3 and Algorithm 4 in the lifted notation and the notation consistent with the rest of the paper respectively.

E.3.2. Algorithm and convergence rate

Taking the stochastic gradient step followed by the proximal step with respect to ψ, both with stepsize η, is equivalent to (Hanzely et al., 2020b):

[11] We assume that all clients sample the same index, i.e., the randomness is synchronized. We do so only for the sake of simplicity; similar rate can be obtained without shared randomness.

Algorithm 3 ASVRCD-PFL (lifted notation)

```
input $0 < \theta_1, \theta_2 < 1$, $\eta, \nu, \gamma > 0$, $\rho \in (0,1)$, $Y^0 = Z^0 = X^0$.
  for $k = 0, 1, 2, \ldots$ do
    $X^k = \theta_1 Z^k + \theta_2 V^k + (1 - \theta_1 - \theta_2) Y^k$
    $g^k = \mathbf{G}(X^k) - \mathbf{G}(V^k) + \nabla \mathbf{F}(V^k)$
    $Y^{k+1} = \mathrm{prox}_{\eta\psi}(X^k - \eta g^k)$
    $Z^{k+1} = \nu Z^k + (1-\nu) X^k + \frac{\gamma}{\eta}(Y^{k+1} - Y^k)$
    $V^{k+1} = \begin{cases} Y^k, & \text{with probability } \rho \\ V^k, & \text{with probability } 1-\rho \end{cases}$
  end for
```

Algorithm 4 ASVRCD-PFL

```
input $0 < \theta_1, \theta_2 < 1$, $\eta, \nu, \gamma > 0$, $\rho \in (0,1)$, $w_y^0 = w_z^0 = w_v^0 \in \mathbb{R}^{d_0}$, $\beta_{y,m}^0 = \beta_{z,m}^0 = \beta_{v,m}^0 \in \mathbb{R}^{d_m}$ for $1 \le m \le M$.
  for $k = 0, 1, 2, \ldots$ do
    $w_x^k = \theta_1 w_z^k + \theta_2 w_v^k + (1 - \theta_1 - \theta_2) w_y^k$
    for $m = 1, \ldots, M$ in parallel do
      $\beta_{x,m}^k = \theta_1 \beta_{z,m}^k + \theta_2 \beta_{v,m}^k + (1 - \theta_1 - \theta_2) \beta_{y,m}^k$
    end for
    Sample random $j \in \{1, 2, \ldots, n\}$ and $\zeta = \begin{cases} 1 & \text{w.p.} p_w \\ 2 & \text{w.p.} p_\beta \end{cases}$
    $g_w^k = \begin{cases} \frac{1}{p_w}\left(\frac{1}{M}\sum_{m=1}^M \nabla_w f_{m,j}(w_x^k, \beta_{x,m}^k) - \frac{1}{M}\sum_{m=1}^M \nabla_w f_{m,j}(w_v^k, \beta_{v,m}^k)\right) + \nabla_w F(w_v^k, \beta_v^k) & \text{if } \zeta = 1 \\ \nabla_w F(w_v^k, \beta_v^k) & \text{if } \zeta = 2 \end{cases}$
    $w_y^{k+1} = w_x^k - \eta g_w^k$
    $w_z^{k+1} = \nu w_z^k + (1-\nu) w_x^k + \frac{\gamma}{\eta}(w_y^{k+1} - w_y^k)$
    $w_v^{k+1} = \begin{cases} w_y^k, & \text{with probability } \rho \\ w_v^k, & \text{with probability } 1-\rho \end{cases}$
    for $m = 1, \ldots, M$ in parallel do
      $g_{\beta,m}^k = \begin{cases} \frac{1}{M}\nabla_\beta f_m(w_v^k, \beta_{v,m}^k) & \text{if } \zeta = 1 \\ \frac{1}{p_\beta M}\left(\nabla_\beta f_{m,j}(w_x^k, \beta_{x,m}^k) - \nabla_\beta f_{m,j}(w_v^k, \beta_{v,m}^k)\right) + \frac{1}{M}\nabla_\beta f_m(w_v^k, \beta_{v,m}^k) & \text{if } \zeta = 2 \end{cases}$
      $\beta_{y,m}^{k+1} = \beta_{x,m}^k - \eta g_{\beta,m}^k$
      $\beta_{z,m}^{k+1} = \nu \beta_{z,m}^k + (1-\nu)\beta_{x,m}^k + \frac{\gamma}{\eta}(\beta_{y,m}^{k+1} - \beta_{y,m}^k)$
      $\beta_{v,m}^{k+1} = \begin{cases} \beta_{y,m}^k, & \text{with probability } \rho \\ \beta_{v,m}^k, & \text{with probability } 1-\rho \end{cases}$
    end for
  end for
```

$$
\begin{aligned}
&\text{w.p. } p_w: \quad \begin{cases} w^+ = w - \eta\left(\frac{1}{p_w}\left(\frac{1}{M}\sum_{m=1}^M \nabla_w f_{m,j}(w, \beta_m) - \frac{1}{M}\sum_{m=1}^M \nabla_w f_{m,j}(w', \beta'_m)\right) + \nabla_w F(w', \beta')\right), \\ \beta_m^+ = \beta_m - \frac{\eta}{M}\nabla f_m(w', \beta'_m) \end{cases} \\
&\text{w.p. } p_\beta: \quad \begin{cases} w^+ = w - \eta \nabla_w F(w', \beta'), \\ \beta_m^+ = \beta_m - \frac{\eta}{M}\left(\frac{1}{p_\beta}\left(\nabla_\beta f_{m,j}(w, \beta_m) - \nabla_\beta f_{m,j}(w', \beta'_m)\right) + \nabla_\beta f_m(w', \beta'_m)\right). \end{cases}
\end{aligned} \tag{50}
$$

Defining $x = [w, \beta_1, \ldots, \beta_M]$, $x' = [w', \beta'_1, \ldots, \beta'_M]$, update rule (50) can be rewritten as

$$
x^+ = x - \eta\left(g(x) - g(x') + \nabla F(x')\right)
$$

where $g(x)$ corresponds to the described unbiased stochastic gradient obtained by subsampling both the space and the finite sum simultaneously. In order to give the rate of aforementioned method, we shall determine the expected smoothness constant.

Lemma E.1. *Suppose that Assumptions 1.1 and 1.2 hold. Then, we have*

$$\mathbb{E}\left[\|(g(x)-g(x')+\nabla F(x'))-\nabla F(x)\|^2\right] \le 2\mathcal{L}D_F(x,y)$$

where $\mathcal{L} \coloneqq 2\max\left(\frac{\mathcal{L}^w}{p_w}, \frac{\mathcal{L}^\beta}{p_\beta}\right)$.

Proof. Let $d_\beta \coloneqq \sum_{m=1}^m d_m$. We have:

$$\begin{aligned}
&\mathbb{E}\left[\|(g(x)-g(x')+\nabla F(x'))-\nabla F(x)\|^2\right]\\
&\quad\le \mathbb{E}\left[\|g(x)-g(x')\|^2\right]\\
&\quad= p_w\mathbb{E}\left[\left\|p_w^{-1}\frac{1}{M}\sum_{m=1}^M\left(\nabla_w f_{m,j}(w,\beta_m)-\nabla_w f_{m,j}(w',\beta_m')\right)\right\|^2 \mid \zeta=1\right]\\
&\qquad+p_\beta\frac{1}{M^2}\sum_{m=1}^M\mathbb{E}\left[\|p_\beta^{-1}\nabla f_{m,j}(w,\beta_m)-p_\beta^{-1}\nabla f_{m,j}(w',\beta_m')\|^2 \mid \zeta=2\right]\\
&\quad= p_w^{-1}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{m=1}^M\left(\nabla_w f_{m,j}(w,\beta_m)-\nabla_w f_{m,j}(w',\beta_m')\right)\right\|^2 \mid \zeta=1\right]\\
&\qquad+p_\beta^{-1}\frac{1}{M^2}\sum_{m=1}^M\mathbb{E}\left[\|\nabla f_{m,j}(w,\beta_m)-\nabla f_{m,j}(w',\beta_m')\|^2 \mid \zeta=2\right]\\
&\quad= \mathbb{E}\left[(F_j(x)-\nabla F_j(x'))^\top\begin{pmatrix} p_w^{-1}I^{d_0\times d_0} & 0\\ 0 & p_\beta^{-1}I^{d_\beta\times d_\beta}\end{pmatrix}(F_j(x)-\nabla F_j(x'))\right]\\
&\quad\overset{(*)}{\le} \mathbb{E}\left[4\max\left(\frac{\mathcal{L}^w}{p_w},\frac{L^\beta}{p_\beta}\right)D_{F_j}(x,x')\right]\\
&\quad= 4\max\left(\frac{\mathcal{L}^w}{p_w},\frac{L^\beta}{p_\beta}\right)D_F(x,x'),
\end{aligned}$$

where $(*)$ holds due to the $(\mathcal{L}^w,\mathcal{L}^\beta)$-smoothness of F_j (from Assumption 1.2) and Lemma E.2.

□

Theorem E.1. *Iteration complexity of Algorithm 3 with*

$$\begin{aligned}
\eta &= \frac{1}{4\mathcal{L}},\\
\theta_2 &= \frac{1}{2},\\
\gamma &= \frac{1}{\max\{2\mu, 4\theta_1/\eta\}},\\
\nu &= 1-\gamma\mu \text{ and}\\
\theta_1 &= \min\left\{\frac{1}{2}, \sqrt{\eta\mu\max\left\{\frac{1}{2},\frac{\theta_2}{\rho}\right\}}\right\}
\end{aligned}$$

is $\mathcal{O}\left(\left(\frac{1}{\rho}+\sqrt{\frac{\max\left(\frac{\mathcal{L}^w}{p_w},\frac{L^\beta}{p_\beta}\right)}{\rho\mu}}\right)\log\frac{1}{\epsilon}\right)$. *Setting* $p_w = \frac{\mathcal{L}^w}{\mathcal{L}^\beta+\mathcal{L}^w}$ *yields the complexity* $\mathcal{O}\left(\left(\frac{1}{\rho}+\sqrt{\frac{\mathcal{L}^w+\mathcal{L}^\beta}{\rho\mu}}\right)\log\frac{1}{\epsilon}\right)$.

Overall, the algorithm requires

$$\mathcal{O}\left(\left(\frac{1}{\rho}+\sqrt{\frac{\mathcal{L}^w+\mathcal{L}^\beta}{\rho\mu}}\right)\left(\log\frac{1}{\epsilon}\right)(\rho n+p_w)\right)$$

communication rounds and the same amount of gradient calls w.r.t. parameter w. Set $\rho=\frac{p_w}{n}$. In such a case, we have

$$\begin{aligned}\left(\frac{1}{\rho}+\sqrt{\frac{\mathcal{L}^w+\mathcal{L}^\beta}{\rho\mu}}\right)\left(\log\frac{1}{\epsilon}\right)(\rho n+p_w) &= 2\left(\frac{1}{\rho}+\sqrt{\frac{\mathcal{L}^w+\mathcal{L}^\beta}{\rho\mu}}\right)\left(\log\frac{1}{\epsilon}\right)\rho n\\ &= 2\left(n+\sqrt{\frac{\rho n^2(\mathcal{L}^w+\mathcal{L}^\beta)}{\mu}}\right)\left(\log\frac{1}{\epsilon}\right)\\ &= 2\left(n+\sqrt{\frac{n\mathcal{L}^w}{\mu}}\right)\left(\log\frac{1}{\epsilon}\right).\end{aligned}$$

and thus Algorithm 3 enjoys both communication complexity and the global gradient complexity of order $\mathcal{O}\left(\left(n+\sqrt{\frac{n\mathcal{L}^w}{\mu}}\right)\log\frac{1}{\epsilon}\right)$. Analogously, setting $\rho=\frac{p_\beta}{n}$ yields personalized/local gradient complexity of order $\mathcal{O}\left(\left(n+\sqrt{\frac{n\mathcal{L}^\beta}{\mu}}\right)\log\frac{1}{\epsilon}\right)$.

Lemma E.2. *Let $H(x,y):\mathbb{R}^{d_x+d_y}\to\mathbb{R}$ be (jointly) convex function such that $\nabla^2_x H(x,y)\leq L_x\mathbf{I}$ and $\nabla^2_y H(x,y)\leq L_y\mathbf{I}$. Then we have*

$$\nabla^2 H(x,y)\leq 2\begin{pmatrix}L_x\mathbf{I} & 0\\ 0 & L_y\mathbf{I}\end{pmatrix} \tag{51}$$

and therefore,

$$D_H((x,y),(x'y'))\geq\frac{1}{2}\left(\nabla H(x,y)-\nabla H(x',y')\right)^\top\begin{pmatrix}\frac{1}{2}L_x^{-1}\mathbf{I} & 0\\ 0 & \frac{1}{2}L_y^{-1}\mathbf{I}\end{pmatrix}\left(\nabla H(x,y)-\nabla H(x',y')\right). \tag{52}$$

Proof. First, we show (51).

$$\begin{aligned}2\begin{pmatrix}L_x\mathbf{I} & 0\\ 0 & L_y\mathbf{I}\end{pmatrix}-\nabla^2 H(x,y) &= \begin{pmatrix}2L_x\mathbf{I}-\nabla^2_{x,x}H(x,y) & -\nabla^2_{x,y}H(x,y)\\ -\nabla^2_{y,x}H(x,y) & 2L_y\mathbf{I}-\nabla^2_{y,y}H(x,y)\end{pmatrix}\\ &\succeq \begin{pmatrix}\nabla^2_{x,x}H(x,y) & -\nabla^2_{x,y}H(x,y)\\ -\nabla^2_{y,x}H(x,y) & \nabla^2_{y,y}H(x,y)\end{pmatrix}\\ &\succeq \nabla^2_{x,x}H(x,-y)\\ &\succeq 0.\end{aligned}$$

It suffices to notice that (52) is a direct consequence of (51) and joint convexity of H. □

E.4. SCD-PFL and SVRCD-PFL

We state SCD-PFL in Algorithm 5 and SVRCD-PFL in Algorithm 6. These algorithms correspond to a simplified version of ASVRCD-PFL: SVRCD-PFL does not incorporate Nesterov's acceleration while ASVRCD-PFL does not incorporate the control variates or Nesterov's acceleration.

Algorithm 5 SCD-PFL

```
input $\eta > 0$, $p_w \in (0,1)$, $p_\beta = 1 - p_w$, $w^0 \in \mathbb{R}^d$, $\beta_m^0 \in \mathbb{R}^d$ for $1 \leq m \leq M$.
  for $k = 0, 1, 2, \dots K-1$ do
    Sample random $j_m \in \{1, 2, \dots, n_m\}$ for $1 \leq m \leq M$ and $\zeta = \begin{cases} 1 & \text{w.p. } p_w \\ 2 & \text{w.p. } p_\beta \end{cases}$
    $g_w^k = \begin{cases} \frac{1}{p_w M} \sum_{m=1}^M \nabla_w f_{m,j_m}(w^k, \beta_m^k) & \text{if } \zeta = 1 \\ 0 & \text{if } \zeta = 2 \end{cases}$
    $w^{k+1} = w^k - \eta g_w^k$
    for $m = 1, \dots, M$ in parallel do
      $g_{\beta,m}^k = \begin{cases} 0 & \text{if } \zeta = 1 \\ \frac{1}{p_\beta M} \nabla_\beta f_{m,j_m}(w^k, \beta_m^k) & \text{if } \zeta = 2 \end{cases}$
      $\beta_m^{k+1} = \beta_m^k - \eta g_{\beta,m}^k$
    end for
  end for
output $w^K$, $\beta_m^K$ for $1 \leq m \leq M$.
```

Algorithm 6 SVRCD-PFL

```
input $\eta > 0$, $p_w \in (0,1)$, $p_\beta = 1 - p_w$, $\rho \in (0,1)$, $w_y^0 = w_v^0 \in \mathbb{R}^d$, $\beta_{y,m}^0 = \beta_{v,m}^0 \in \mathbb{R}^d$ for $1 \leq m \leq M$.
  for $k = 0, 1, 2, \dots K-1$ do
    Sample random $j_m \in \{1, 2, \dots, n_m\}$ for $1 \leq m \leq M$ and $\zeta = \begin{cases} 1 & \text{w.p.} p_w \\ 2 & \text{w.p.} p_\beta \end{cases}$
    $g_w^k = \begin{cases} \frac{1}{p_w M} \sum_{m=1}^M \nabla_w f_{m,j_m}(w_y^k, \beta_{y,m}^k) + \nabla_w F(w_v^k, \beta_v^k) & \text{if } \zeta = 1 \\ \nabla_w F(w_v^k, \beta_v^k) & \text{if } \zeta = 2 \end{cases}$
    $w_y^{k+1} = w_y^k - \eta g_w^k$
    $w_v^{k+1} = \begin{cases} w_y^k, & \text{with probability } \rho \\ w_v^k, & \text{with probability } 1 - \rho \end{cases}$
    for $m = 1, \dots, M$ in parallel do
      $g_{\beta,m}^k = \begin{cases} \frac{1}{M} \nabla_\beta f_m(w_v^k, \beta_{v,m}^k) & \text{if } \zeta = 1 \\ \frac{1}{p_\beta M} \nabla_\beta f_{m,j_m}(w_y^k, \beta_{y,m}^k) + \frac{1}{M} \nabla_\beta f_m(w_v^k, \beta_{v,m}^k) & \text{if } \zeta = 2 \end{cases}$
      $\beta_{y,m}^{k+1} = \beta_{y,m}^k - \eta g_{\beta,m}^k$
      $\beta_{v,m}^{k+1} = \begin{cases} \beta_{y,m}^k, & \text{with probability } \rho \\ \beta_{v,m}^k, & \text{with probability } 1 - \rho \end{cases}$
    end for
  end for
output $w_y^K$, $\beta_{y,m}^K$ for $1 \leq m \leq M$.
```