On Convergence of FedProx: Local Dissimilarity Invariant Bounds, Non-smoothness and Beyond

Xiao-Tong Yuan
Nanjing Univ. of Information Sci. & Tech.
219 Ningliu Rd, Nanjing, China
xtyuan1980@gmail.com

Ping Li
LinkedIn Ads
700 Bellevue Way NE, Bellevue, WA, USA
pinli@linkedin.com

Abstract

The `FedProx` algorithm is a simple yet powerful distributed proximal point optimization method widely used for federated learning (FL) over heterogeneous data. Despite its popularity and remarkable success witnessed in practice, the theoretical understanding of FedProx is largely underinvestigated: the appealing convergence behavior of `FedProx` is so far characterized under certain non-standard and unrealistic dissimilarity assumptions of local functions, and the results are limited to smooth optimization problems. In order to remedy these deficiencies, we develop a novel local dissimilarity invariant convergence theory for `FedProx` and its minibatch stochastic extension through the lens of algorithmic stability. As a result, we contribute to derive several new and deeper insights into `FedProx` for non-convex federated optimization including: 1) convergence guarantees invariant to certain stringent local dissimilarity conditions; 2) convergence guarantees for non-smooth FL problems; and 3) linear speedup with respect to size of minibatch and number of sampled devices. Our theory for the first time reveals that local dissimilarity and smoothness are not must-have for `FedProx` to get favorable complexity bounds.

1 Introduction

Federated Learning (FL) has recently emerged as a promising paradigm for communication-efficient distributed learning on remote devices, such as smartphones, internet of things, or agents (Konečnỳ et al., 2016; Yang et al., 2019). The goal of FL is to collaboratively train a shared model that works favorably for all the local data but without requiring the learners to transmit raw data across the network. The principle of optimizing a global model while keeping data localized can be beneficial for both computational efficiency and data privacy (Bhowmick et al., 2018). While resembling the classic distributed learning regimes, there are two most distinct features associated with FL: 1) large statistical heterogeneity of local data mainly due to the non-iid manner of data generalization and collection across the devices (Hard et al., 2020); and 2) partial participation of devices in the network mainly due to the massive number of devices. These fundamental challenges make FL highly demanding to tackle, both in terms of optimization algorithm design and in terms of theoretical understanding of convergence behavior (Li et al., 2020a).

FL is most conventionally formulated as the following problem of global population risk minimization averaged over a set of M devices:

$$\min_{w\in\mathbb{R}^p} \bar{R}(w) := \frac{1}{M}\sum_{m=1}^{M}\left\{R^{(m)}(w) := \mathbb{E}_{Z^{(m)}\sim\mathcal{D}^{(m)}}[\ell^{(m)}(w;Z^{(m)})]\right\}, \tag{1}$$

where $R^{(m)}$ is the local population risk on device m, $\ell^{(m)}:\mathbb{R}^p\times\mathcal{Z}^{(m)}\mapsto R^+$ is a non-negative loss function whose value $\ell^{(m)}(w;Z^{(m)})$ measures the loss over a random data point $Z^{(m)}\in\mathcal{Z}^{(m)}$ with parameter w, $\mathcal{D}^{(m)}$ represents an underlying random data distribution over $\mathcal{Z}^{(m)}$. Since the data distribution is typically unknown, the following empirical risk minimization (ERM) version of (1) is

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

often considered alternatively:

$$\min_{w\in\mathbb{R}^p} \bar{R}_{\texttt{erm}}(w) := \frac{1}{M}\sum_{m=1}^{M}\left\{R^{(m)}_{\texttt{erm}}(w) := \frac{1}{N_m}\sum_{i=1}^{N_m}\ell^{(m)}(w; z_i^{(m)})\right\}, \tag{2}$$

where $R^{(m)}_{\texttt{erm}}$ is the local empirical risk over the training sample $D^{(m)} = \{z_i^{(m)}\}_{i=1}^{N_m}$ on device m. The sample size N_m may vary significantly across devices, which can be regarded as another source of data heterogeneity. Federated optimization algorithms for solving (1) or (2) have attracted significant research interest from both academia and industry, with a rich body of efficient solutions developed that can flexibly adapt to the communication-computation tradeoffs and data/system heterogeneity. Several popularly used FL algorithms for this setting include `FedAvg` (McMahan et al., 2017), `FedProx` (Li et al., 2020b), `SCAFFOLD` (Karimireddy et al., 2020), and `FedPD` (Zhang et al., 2020), to name a few. A consensus among these methods on communication-efficient implementation is trying to extensively update the local models (e.g., with plenty epochs of local optimization) over subsets of devices so as to quickly find an optimal global model using a minimal number of inter-device communication rounds for model aggregation.

In this paper, we revisit the `FedProx` algorithm which is one of the most prominent frameworks for heterogeneous federated optimization. Reasons for the interests of `FedProx` include implementation simplicity, low communication cost, promise in dealing with data heterogeneity and tolerance to partial participation of devices (Li et al., 2020b). We analyze its convergence behavior, expose problems, and propose alternatives more suitable for scaling up and generalization. We contribute to derive several new and deeper theoretical insights into the algorithm from a novel perspective of algorithmic stability and generalization theory.

1.1 Review of `FedProx`

For solving FL problems in the presence of data heterogeneity, methods such as `FedAvg` based on local stochastic gradient descent (SGD) can fail to converge in practice when the selected devices perform too many local updates (Li et al., 2020b). To mitigate this issue, `FedProx` (Li et al., 2020b) was recently proposed for solving the empirical FL problem (2) using the (inexact) proximal point update for local optimization. The benefits of `FedProx` include: 1) it provides more stable local updates by explicitly enforcing the local optimization in the vicinity of the global model to date; 2) the method comes with convergence guarantees for both convex and non-convex functions, even under partial participation and very dissimilar amounts of local updates (Li et al., 2020a). More specifically, at each time instance t, `FedProx` uniformly randomly selects a subset $I_t \subseteq [M]$ of devices and introduces for each device $\xi \in I_t$ the following proximal point ERM sub-problem for local update around the previous global model w_{t-1}:

$$w_t^{(\xi)} \approx \operatorname*{arg\,min}_{w\in\mathbb{R}^p}\left\{Q^{(\xi)}_{\texttt{erm}}(w; w_{t-1}) := R^{(\xi)}_{\texttt{erm}}(w) + \frac{1}{2\eta_t}\|w - w_{t-1}\|^2\right\}, \tag{3}$$

where $\eta_t > 0$ is the learning rate that controls the impact of the proximal term. Then the global model is updated by uniformly aggregating those local updates from I_t as

$$w_t = \frac{1}{|I_t|}\sum_{\xi\in I_t} w_t^{(\xi)}.$$

In the extreme case of allowing $\eta_t \to +\infty$ in (3), `FedProx` reduces to the regime of `FedAvg` if using SGD for local optimization. Since its inception, `FedProx` and its variants have received significant interests in research (Pathak and Wainwright, 2020; Nguyen et al., 2020; Li et al., 2019) and become an algorithm of choice in application areas such as automatous driving (Donevski et al., 2021) and computer vision (He et al., 2021). Theoretically, `FedProx` comes with convergence guarantees under the following bounded *local gradient dissimilarity* assumption that captures the statistical heterogeneity of local objectives across the network:

Definition 1 ((B, H)-LGD)**.** *We say the local functions $R^{(m)}$ have (B, H)-local gradient dissimilarity (LGD) if the following holds for all $w \in \mathbb{R}^p$:*

$$\frac{1}{M}\sum_{m=1}^{M}\|\nabla R^{(m)}(w)\|^2 \le B^2\|\nabla \bar{R}(w)\|^2 + H^2.$$

The above definition naturally extends to the local empirical risks $\{R_{\texttt{erm}}^{(m)}\}_{m=1}^M$. Specially in the homogenous setting where $R^{(m)} \equiv \bar{R}, \forall m \in [M]$, we have $B = 1$ and $H = 0$. Under $(B, 0)$-LGD and some regularization conditions on the modulus B, it was shown that `FedProx` for non-convex problems requires $T = \mathcal{O}\left(\frac{1}{\epsilon}\right)$ rounds of inter-device communication to reach an ϵ-stationary solution in the sense of $\frac{1}{T}\sum_{t=1}^T \|\nabla \bar{R}_{\texttt{erm}}(w_t)\|^2 \le \epsilon$ (Li et al., 2020b). Similar guarantees have also been established for a variant of `FedProx` with non-uniform model aggregation (Nguyen et al., 2020).

Open issues and motivation. In spite of the remarkable success achieved by `FedProx` and its variants, there are still a number of important theoretical issues regarding the unrealistic assumptions, restrictive problem regimes and expensive local oracle cost that remain open for exploration, as specified below.

- **Local dissimilarity condition**. The appealing convergence behavior of `FedProx` is so far characterized under a key but non-standard $(B, 0)$-LGD condition (cf. Definition 1) with $B > 0$. Such a condition is obviously unrealistic in practice: it essentially requires the local objectives share the same stationary point as the global objective since $\|\nabla \bar{R}_{\texttt{erm}}(w)\| = 0$ implies $\|\nabla R_{\texttt{erm}}^{(m)}(w)\| = 0$ for all $m \in [M]$. However, if the optima of $R_{\texttt{erm}}^{(m)}$ are exactly (or even approximately) the same, there would be little point in distributing data across devices for federated learning. *It is thus desirable to understand the convergence behavior of `FedProx` for heterogeneous FL without imposing stringent local dissimilarity conditions like $(B, 0)$-LGD.*
- **Non-smooth optimization**. The existing convergence guarantees of `FedProx` are only available for FL with smooth losses. More often than not, however, FL applications involve non-smooth objectives due to the popularity of non-smooth losses (e.g., hinge loss and absolute loss) in machine learning, and training deep neural networks with non-smooth activation like ReLU. *Therefore, it is desirable to understand the convergence behavior of `FedProx` in non-smooth problem regimes.*
- **Local oracle complexity**. Unlike the (stochastic) first-order oracles such as SGD used by `FedAvg`, the proximal point oracle (3) for local update is by itself a full-batch ERM problem which tends to be expensive to solve even approximately per-iteration. Plus, due to the potentially imbalanced data distribution over devices, the computational overload of the proximal point oracle could vary significantly across the network. *Therefore, it is important to investigate whether using minibatch stochastic approximation to the proximal point oracle* (3) *can provably improve the computational efficiency of `FedProx`.*

Last but not least, existing convergence analysis of `FedProx` mainly focuses on the empirical FL problem (2). The optimality in terms of the population FL problem (1) is not yet clear for `FedProx`. The primary goal of this work is to remedy these theoretical issues simultaneously, so as to lay a more solid theoretical foundation for the popularly applied `FedProx` algorithm.

1.2 Our Contributions

In this paper, we make progress towards understanding the convergence behavior of `FedProx` for non-convex heterogenous FL under weaker yet more realistic conditions. The main results are a set of local dissimilarity invariant bounds for smooth or non-smooth problems.

Main results for the vanilla `FedProx`. As a starting point to address the restrictiveness of local dissimilarity assumption, we provide a novel convergence analysis for the vanilla `FedProx` algorithm invariant to the $(B, 0)$-LGD condition. For smooth and non-convex optimization problems, our result in Theorem 1 shows that the rate of convergence to a stationary point is upper bounded by

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla \bar{R}_{\texttt{erm}}(w_t)\right\|^2\right] \lesssim \max\left\{\frac{1}{T^{2/3}}, \frac{1}{\sqrt{TI}}\right\}, \tag{4}$$

where I is the number devices randomly selected for local update at each iteration. If all the devices participate in the local updates for every round, i.e. $I_t = [M]$, the rate of convergence can be improved to $\mathcal{O}(\frac{1}{T^{2/3}})$. For $T < I^3$, the rate in (4) is dominated by $\mathcal{O}(\frac{1}{T^{2/3}})$ which gives the communication complexity $\frac{1}{\epsilon^{3/2}}$ to achieve an ϵ-stationary solution. On the other hand when $T \ge I^3$, the rate is dominated by $\mathcal{O}(\frac{1}{\sqrt{TI}})$ which gives the communication complexity $\frac{1}{I\epsilon^2}$. Compared to the already known $\mathcal{O}(\frac{1}{\epsilon})$ complexity bound of `FedProx` under the unrealistic $(B, 0)$-LGD condition (Li et al., 2020b), our rate in (4) is slower but it holds without needing to impose stringent regularity conditions on the dissimilarity of local functions, and it reveals the benefit of device minibatch sampling for accelerating convergence. Further for *non-smooth* and weakly convex problems, we establish in Theorem 2 the following rate of convergence regarding proper ρ-Moreau-envelopes of $\bar{R}_{\texttt{erm}}$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla \bar{R}_{\texttt{erm},\rho}(w_t)\right\|^2\right] \lesssim \frac{1}{\sqrt{T}}, \tag{5}$$

The bound is not dependent on the number of selected devices in each round. In the case of $I = \mathcal{O}(1)$, the bounds in (4) and (5) are comparable, which indicates that smoothness is not must-have for `FedProx` to get sharper convergence bound especially with low participation ratio. On the other end when $I = \mathcal{O}(M)$, the bound (5) for non-smooth problems is slower than the bound (4) for smooth functions in large-scale networks.

Main results for minibatch stochastic `FedProx`. Then as the chief contribution of the present work, we propose a minibatch stochastic extension of `FedProx` along with its population optimization performance analysis from a novel perspective of algorithmic stability theory. Inspired by the recent success of minibatch stochastic proximal point methods (MSPP) (Li et al., 2014; Wang et al., 2017; Asi et al., 2020; Deng and Gao, 2021), we propose to implement `FedProx` using MSPP as the local update oracle. The resulting method, which is referred to as `FedMSPP`, is expected to attain improved trade-off between computation, communication and memory efficiency for large-scale FL. In the case of imbalanced data distribution, minibatching is also beneficial for making the local computation more balanced across the devices. Based on some extended uniform stability arguments for gradients, we show in Theorem 3 the following rate of convergence for `FedMSPP` in terms of population optimality, which is also invariant to the $(B, 0)$-LGD condition:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla \bar{R}(w_t)\right\|^2\right] \lesssim \max\left\{\frac{1}{T^{2/3}}, \frac{1}{\sqrt{TbI}}\right\}, \tag{6}$$

where b is the minibatch size of local update. For empirical FL, identical bound holds under sampling according to empirical distribution. For $T < (bI)^3$, the rate in (6) is dominated by $\mathcal{O}(\frac{1}{T^{2/3}})$ which gives the communication complexity $\frac{1}{\epsilon^{3/2}}$, and it matches that of the vanilla `FedProx`. For sufficiently large $T \geq (bI)^3$, the rate is dominated by $\mathcal{O}(\frac{1}{\sqrt{TbI}})$ which gives the communication complexity $\frac{1}{bI\epsilon^2}$. This shows that local minibatching and device sampling are both beneficial for linearly speeding up communication. Further, when applied to non-smooth problems, we show in Theorem 4 that `FedMSPP` converges at the following rate with respect to proper ρ-Moreau-envelopes of $\bar{R}$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla \bar{R}_{\rho}(w_t)\right\|^2\right] \lesssim \frac{1}{\sqrt{T}},$$

which is comparable to that of (6) when $b = \mathcal{O}(1)$ and $I = \mathcal{O}(1)$, but without witnessing the effect of linear speedup with respect to b and I.

Comparison with prior results. In Table 1, we summarize our communication complexity bounds for `FedProx` (`FedMSPP`) and compare them with several related heterogeneous FL algorithms in terms of the conditions on local dissimilarity, applicability to non-smooth problems and tolerance to partial participation. A few observations are in order. *First*, regarding the local dissimilarity condition, all of our $\mathcal{O}(\frac{1}{\epsilon^2})$ bounds are not dependent on the (B, H)-LGD type conditions, and they are comparable to those of `SCAFFOLD` and `FCO` (for convex problems) which are also invariant to local dissimilarity conditions. *Second*, with regard to the applicability to non-smooth optimization, our convergence guarantees in Theorem 2 and Theorem 4 are established for non-smooth and weakly convex functions. While `FCO` is the only one in the other considered algorithms that can be applied to non-smooth problems, it is customized for federated convex composite optimization with potentially non-smooth regularizers (Yuan et al., 2021). *Third*, in terms of tolerance to partial participation, all of our results are robust to device sampling, and the $\mathcal{O}(\frac{1}{bI\epsilon^2})$ bound in Theorem 3 for `FedMSPP` is comparable to the best known results under partial participation as achieved by `FedAvg` and `SCAFFOLD`. If assuming that all the devices participate in local update for each communication round and using momentum acceleration techniques, substantially faster $\mathcal{O}(\frac{1}{\epsilon})$ bounds are possible for `STEM` and `FedPD`, while the $\mathcal{O}(\frac{1}{\epsilon^{3/2}})$ bounds can be achieved by `FedAvg` (Khanduri et al., 2021). To summarize the comparison, our (B, H)-LGD invariant convergence bounds for `FedProx` (`FedMSPP`) are comparable to the best-known rates in the identical setting, while covering the generic non-smooth and non-convex cases which to our knowledge has not been provably possible for other FL algorithms.

Method	Work	Commun. Complex.	LD Condition	NS	PP
`FedProx`	(Li et al., 2020b)	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$	$(B,0)$-LGD	✗	✓
	Theorem 1 (ours)	$\mathcal{O}\left(\frac{1}{I\epsilon^2}+\frac{1}{\epsilon^{3/2}}\right)$	–	✗	✓
	Theorem 2 (ours)	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	–	✓	✓
`FedMSPP`	Theorem 3 (ours)	$\mathcal{O}\left(\frac{1}{bI\epsilon^2}+\frac{1}{\epsilon^{3/2}}\right)$	–	✗	✓
	Theorem 4 (ours)	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	–	✓	✓
`FedAvg`	(Karimireddy et al., 2020)	$\mathcal{O}\left(\frac{1}{bI\epsilon^2}+\frac{1}{\epsilon^{3/2}}+\frac{1}{\epsilon}\right)$	(B,H)-LGD	✗	✓
	(Yu et al., 2019)	$\mathcal{O}\left(\frac{1}{bM\epsilon^2}+\frac{Mb}{\epsilon}\right)$	$(0,H)$-LGD	✗	✗
	(Khanduri et al., 2021)	$\mathcal{O}\left(\frac{1}{\epsilon^{3/2}}\right)$	–	✗	✗
`SCAFFOLD`	(Karimireddy et al., 2020)	$\mathcal{O}\left(\frac{1}{bI\epsilon^2}+\frac{(M/I)^{2/3}}{\epsilon}\right)$	–	✗	✓
`FedPD`	(Zhang et al., 2020)	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$	–	✗	✗
`STEM`	(Khanduri et al., 2021)	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$	–	✗	✗
`FCO`	(Yuan et al., 2021)	$\mathcal{O}\left(\frac{1}{bM\epsilon^2}+\frac{1}{\epsilon}\right)$ (convex composite)	–	✓	✗

Table 1: Comparison of heterogeneous FL algorithms in terms of communication complexity bounds for reaching an ϵ-stationary solution, local dissimilarity (LD) condition, applicability to non-smooth (NS) functions and tolerance to partial participation (PP). Except for `FCO`, all the results listed are for non-convex functions. The involved quantities are M: total number of devices; I: number of chosen devices for partial participation; b: minibatch size for local stochastic optimization.

Highlight of contributions. The theoretical contributions of this work are highlighted as follows:

- From the perspective of algorithmic stability theory, we provide a set of novel local dissimilarity invariant convergence guarantees for the widely used `FedProx` algorithm for non-convex heterogeneous FL, with smooth or non-smooth local functions. Our theory for the first time reveals that local dissimilarity and smoothness are not necessary to guarantee the convergence of `FedProx` with reasonable rates.
- We present `FedMSPP` as a minibatch stochastic extension of `FedProx` and analyze its population optimization performance for both smooth and non-smooth FL problems, again without assuming local dissimilarity conditions. Particularly for smooth problems, our result provably shows that `FedMSPP` enjoys linear speedup in terms of minibatching size and partial participation ratio.

Paper organization. In Section 2 we present our local dissimilarity invariant convergence analysis for the vanilla `FedProx` with smooth or non-smooth loss functions. In Section 3 we propose `FedMSPP` as a minibatch stochastic extension of `FedProx` and analyze its convergence behavior through the lens of algorithmic stability theory. In Section 4, we discuss some additional related work on the topics covered by this paper. The concluding remarks are made in Section 5. Finally, all the technical proofs and some additional related work are relegated to the appendix sections.

2 Convergence of `FedProx`

We begin by providing an improved analysis for the vanilla `FedProx` which is not relying on the (B,H)-LGD type conditions. We first introduce notations that will be used in the analysis to follow.

Notations. Throughout the paper, we use $[n]$ to denote the set $\{1,...,n\}$, $\|\cdot\|$ to denote the Euclidean norm and $\langle\cdot,\cdot\rangle$ to denote the Euclidean inner product. We say a function f is G-Lipschitz continuous if $|f(w)-f(w')|\le G\|w-w'\|$ for all $w,w'\in\mathbb{R}^p$, and it is L-smooth if $|\nabla f(w)-\nabla f(w')|\le L\|w-w'\|$ for all $w,w'\in\mathbb{R}^p$. Moreover, we say f is ν-weakly convex if for any $w,w'\in\mathbb{R}^p$,

$$f(w)\ge f(w')+\langle\partial f(w'),w-w'\rangle-\frac{\nu}{2}\|w-w'\|^2,$$

where $\partial f(w')$ represents a subgradient of f evaluated at w'. We denote by

$$f_\eta(w):=\min_u\left\{f(u)+\frac{1}{2\eta}\|u-w\|^2\right\}$$

the η-Moreau-envelope of f, and by

$$\operatorname{prox}_{\eta f}(w) := \arg\min_u \left\{ f(u) + \frac{1}{2\eta}\|u - w\|^2 \right\}$$

the proximal mapping associated with f. We also need to access the following definition of inexact local update oracle for `FedProx`.

Definition 2 (Local inexact oracle of `FedProx`)**.** *Suppose that the local proximal point regularized objective $Q^{(m)}_{\texttt{erm}}(w; w_{t-1})$ (cf. (3)) admits a global minimizer. For each time instance t, we say that the local update oracle of `FedProx` is ε_t-inexactly solved with sub-optimality $\varepsilon_t \geq 0$ if*

$$Q^{(m)}_{\texttt{erm}}(w^{(m)}_t; w_{t-1}) \leq \min_w Q^{(m)}_{\texttt{erm}}(w; w_{t-1}) + \varepsilon_t.$$

Throughout our analysis, we focus on the case where the devices are sampled with replacement, while all the results extend well to the regime of sampling without replacement.

2.1 Results for Smooth Problems

The following theorem is our main result on the convergence rate of `FedProx` for smooth and non-convex federated optimization problems. A proof of this result is deferred to Appendix B.1. We assume that the initial sub-optimality $\bar{\Delta}^{(0)}_{\texttt{erm}} := \bar{R}_{\texttt{erm}}(w_0) - \min_{w\in\mathbb{R}^p} \bar{R}_{\texttt{erm}}(w)$ is bounded.

Theorem 1. *Assume that for each $m \in [M]$, the loss function $\ell^{(m)}$ is G-Lipschitz and L-smooth with respect to its first argument. Set $|I_t| \equiv I$ and $\eta_t \equiv \frac{1}{3L}\min\left\{\frac{1}{T^{1/3}}, \sqrt{\frac{I}{T}}\right\}$. Suppose that the local update oracle of `FedProx` is ε_t-inexactly solved with $\varepsilon_t \leq \min\left\{\frac{2L^2G^2\eta_t^3}{I^2(L\eta_t+1)}, \frac{G^2\eta_t}{2I(L\eta_t+1)}\right\}$. Let t^* be an index uniformly randomly chosen in $\{0, 1, ..., T-1\}$. Then it holds that*

$$\mathbb{E}\left[\left\|\nabla \bar{R}_{\texttt{erm}}(w_{t^*})\right\|^2\right] \lesssim \left(L\bar{\Delta}^{(0)}_{\texttt{erm}} + G^2\right)\max\left\{\frac{1}{T^{2/3}}, \frac{1}{\sqrt{TI}}\right\}.$$

A few remarks are in order.

Remark 1. *Compared to the $\mathcal{O}\left(\frac{1}{T}\right)$ bound from Li et al. (2020b), our rate established in Theorem 1 is slower but it is valid without assuming the unrealistic $(B, 0)$-LGD conditions and imposing strong regularization conditions on I (see, e.g., Li et al., 2020b, Remark 5). Moreover, the dominant term $\frac{1}{\sqrt{TI}}$ in our bound reveals the benefit of device sampling for linear speedup which is not clear in the previous analysis by Li et al. (2020b).*

Remark 2. *In the extreme case of full device participation, i.e., $I_t \equiv [M]$, the terms related to I in Theorem 1 can be removed and thus the convergence rate becomes $\frac{1}{T^{2/3}}$ under $\eta_t = \mathcal{O}\left(\frac{1}{LT^{1/3}}\right)$. In this same setting, we comment that the rate can also be improved to $\mathcal{O}\left(\frac{1}{T}\right)$ using our proof augments if $(B, 0)$-LGD is additionally assumed.*

Remark 3. *The G-Lipschitz-loss assumption in Theorem 1 can be alternatively replaced by the bounded gradient condition as commonly used in the analysis of FL algorithms (Li et al., 2020b; Zhang et al., 2020). Despite that our analysis does not rely on the $(B, 0)$-LGD condition, the assumed G-Lipschitz (or bounded gradient) condition actually implies that the local objective gradients are not too dissimilar, which shares a close spirit to the typically assumed $(0, H)$-LGD condition (Karimireddy et al., 2020) and inter-client-variance condition (Khanduri et al., 2021). It is noteworthy that these mentioned client heterogeneity conditions are substantially milder than the $(B, 0)$-LGD condition as required in the original analysis of `FedProx`.*

2.2 Results for Non-smooth Problems

Now we turn to study the convergence of `FedProx` for weakly convex but not necessarily smooth problems. For the sake of presentation clarity, we work on the exact `FedProx` in which the local update oracle is assumed to be exactly solved, i.e. $\varepsilon_t \equiv 0$. Extension to the inexact case is more or less straightforward, though with somewhat more involved perturbation treatments. In the analysis to follow, we assume that the initial sub-optimality $\bar{\Delta}^{(0)}_{\texttt{erm},\rho} := \bar{R}_{\texttt{erm},\rho}(w_0) - \min_w \bar{R}_{\texttt{erm},\rho}(w)$ associated with ρ-Moreau-envelope of $\bar{R}_{\texttt{erm}}$ is bounded. The following is our main result on the convergence of `FedProx` for non-smooth and weakly convex problems.

Theorem 2. *Assume that for each $m \in [M]$, the loss function $\ell^{(m)}$ is G-Lipschitz and ν-weakly convex with respect to its first argument. Set $\eta_t \equiv \frac{\rho}{\sqrt{T}}$ for arbitrary $\rho < \frac{1}{2\nu}$. Suppose that the local update oracle of* `FedProx` *is exactly solved with $\varepsilon_t \equiv 0$. Let t^* be an index uniformly randomly chosen in $\{0, 1, ..., T-1\}$. Then it holds that*

$$\mathbb{E}\left[\left\|\nabla \bar{R}_{erm,\rho}(w_{t^*})\right\|^2\right] \lesssim \frac{\bar{\Delta}^{(0)}_{erm,\rho} + \rho G^2}{\rho\sqrt{T}}.$$

Proof. The proof technique is inspired by the arguments from Davis and Drusvyatskiy (2019) developed for analyzing stochastic model-based algorithms, with several new elements along developed for handling the challenges introduced by the model averaging and partial participation mechanisms associated with `FedProx`. A particular crux here is that due to the random subset model aggregation of $w_t = \frac{1}{|I_t|}\sum_{\xi \in I_t} w_t^{(\xi)}$, the local function values $R^{(\xi)}_{\texttt{erm}}(w_t)$ are no longer independent of each other though ξ is uniformly random. As a consequence, $\frac{1}{|I_t|}\sum_{\xi \in I_t} R^{(\xi)}_{\texttt{erm}}(w_t)$ is *not* an unbiased estimation of $\bar{R}_{\texttt{erm}}(w_t)$. To overcome this technical obstacle, we make use of a key observation that $w_t^{(m)}$ will be almost surely close enough to w_{t-1} if the learning rate η_t is small enough (which is the case in our choice of η_t), and thus we can replace the former with the latter whenever beneficial but without introducing too much approximation error. See Appendix B.2 for a full proof of this result. □

A few remarks are in order.

Remark 4. *To our best knowledge, Theorem 2 is the first convergence guarantee for FL algorithms applicable to generic non-smooth and weakly convex problems. This is in sharp contrast with* `FCO` *(Yuan et al., 2021) which focuses on composite convex and non-smooth problems such as ℓ_1-estimation, or* `Fed-HT` *(Tong et al., 2020) which is specially customized for cardinality-constrained sparse learning problems where the non-convexity essentially arises from the cardinality constraint.*

Remark 5. *Let us consider $\bar{w}_{t^*} := \mathrm{prox}_{\rho \bar{R}_{erm}}(w_{t^*})$, the proximal mapping of w_{t^*} associated with $\bar{R}_{erm}$. In view of a feature of Moreau envelope to characterize stationarity (Davis and Drusvyatskiy, 2019), if w_{t^*} has small gradient norm $\left\|\nabla \bar{R}_{erm,\rho}(w_{t^*})\right\|$, then $\bar{w}_{t^*}$ must be a near-stationary solution and w_{t^*} stays in the proximity of $\bar{w}_{t^*}$ due to the identity $\|w_{t^*} - \bar{w}_{t^*}\| = \rho\left\|\nabla \bar{R}_{erm,\rho}(w_{t^*})\right\|$. Therefore, the bound in Theorem 2 suggests that in expectation $\bar{w}_{t^*}$ converges to a stationary solution and w_{t^*} converges to $\bar{w}_{t^*}$, both at the rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$.*

Remark 6. *We comment that the bound in Theorem 2 is not dependent on I, the number of selected devices. On one hand, for $I = \mathcal{O}(1)$ and sufficiently large $T > \mathcal{O}(I^3)$, the bounds Theorem 1 and Theorem 2 are comparable to each other, which demonstrates that the smoothness is not must-have for* `FedProx` *to get sharper convergence bound with small device sampling rate. On the other hand, in the near-full participation setting where $I = \mathcal{O}(M)$, the bound in Theorem 2 for non-smooth problems will be slower when M is large. Extremely when $I_t = [M]$, the $\mathcal{O}(\frac{1}{\sqrt{T}})$ bound is substantially inferior to the smooth case which has improved rate of $\mathcal{O}(\frac{1}{T^{2/3}})$ as discussed in Remark 2.*

3 Convergence of `FedProx` with Stochastic Minibatching

When it comes to the implementation of `FedProx`, a notable challenge is that the local proximal point update oracle (3) is by itself a full-batch ERM problem which would be expensive to solve even approximately in large-scale settings. Moreover, in the settings where the data distribution over devices is highly imbalanced, the computational overload of local update could vary significantly across the network, which impairs communication efficiency. It is thus desirable to seek stochastic approximation schemes for hopefully improving the local oracle update efficiency and overload balance of `FedProx`. To this end, inspired by the recent success of minibatch stochastic proximal point methods (MSPP) (Asi et al., 2020; Deng and Gao, 2021), we propose to implement `FedProx` using MSPP as the local stochastic optimization oracle. More precisely, let $B_t^{(m)} = \{z_{i,t}^{(m)}\}_{i=1}^b \overset{\text{i.i.d.}}{\sim} (\mathcal{D}^{(m)})^b$ be a minibatch of b i.i.d. samples drawn from the distribution $\mathcal{D}^{(m)}$ at device m and time instance $t \geq 1$. We denote

$$R^{(m)}_{B_t^{(m)}}(w) := \frac{1}{b}\sum_{i=1}^{b} \ell^{(m)}(w; z_{i,t}^{(m)}) \tag{7}$$

Algorithm 1: `FedMSPP`: Federated Minibatch Stochastic Proximal Point

Input : Minibatch size b; learning rates $\{\gamma_t\}_{t\in[T]}$.
Output : w_T.
Initialization *Set w_0, e.g., typically as a zero vector.*
for $t = 1, 2, ..., T$ **do**
  `/* Device selection and model broadcast on the server */`
  Server uniformly randomly selects a subset $I_t \subseteq [M]$ of devices and sends w_{t-1} to all the selected devices;
  `/* Local model updates on the selected devices */`
  for $\xi \in I_t$ *in parallel* **do**
    Device ξ samples a minibatch $B_t^{(\xi)} = \{z_{i,t}^{(\xi)}\}_{i=1}^b \overset{\text{i.i.d.}}{\sim} (\mathcal{D}^{(\xi)})^b$.
    Device ξ inexactly updates the its local model as

$$w_t^{(\xi)} \approx \underset{w\in\mathbb{R}^p}{\arg\min} \left\{ Q_{B_t^{(\xi)}}^{(\xi)}(w; w_{t-1}) := R_{B_t^{(\xi)}}^{(\xi)}(w) + \frac{1}{2\eta_t}\|w - w_{t-1}\|^2 \right\}, \tag{8}$$

    where $R_{B_t^{(\xi)}}^{(\xi)}(w)$ is given by (7).
    Device ξ sends $w_t^{(\xi)}$ back to server.
  end
  `/* Model aggregation on the server */`
  Sever aggregates the local models received from I_t to update the global model as $w_t = \frac{1}{|I_t|}\sum_{\xi\in I_t} w_t^{(\xi)}$.
end

as the local minibatch empirical risk function over $B_t^{(m)}$. Here, the only modification we propose to make is to replace the empirical risk $R_{\mathtt{erm}}^{(m)}(w)$ in the original update form (3) with its minibatch counterpart $R_{B_t^{(m)}}^{(m)}(w)$. The resultant FL framework, which we refer to as `FedMSPP` (Federated MSPP), is outlined in Algorithm 1. Clearly, the vanilla `FedProx` is a special case of `FedMSPP` when applied to the federated ERM form (2) with full data batch $B_t^{(m)} \equiv D^{(m)}$.

3.1 Results for Smooth Problems

We first analyze the convergence rate of `FedMSPP` for smooth and non-convex problems using the tools borrowed from algorithmic stability theory. Analogous to the Definition 2, we introduce the following definition of inexact local update oracle for `FedMSPP`.

Definition 3 (Local inexact oracle of `FedMSPP`). *Suppose that the local proximal point regularized objective $Q_{B_t^{(m)}}^{(m)}(w; w_{t-1})$ (cf. (8)) admits a global minimizer. For each time instance t, we say that the local update oracle of `FedMSPP` is ε_t-inexactly solved with sub-optimality $\varepsilon_t \geq 0$ if*

$$Q_{B_t^{(m)}}^{(m)}(w_t^{(m)}; w_{t-1}) \leq \min_w Q_{B_t^{(m)}}^{(m)}(w; w_{t-1}) + \varepsilon_t.$$

We also assume that the initial population sub-optimality $\bar{\Delta}^{(0)} = \bar{R}(w^{(0)}) - \min_{w\in\mathbb{R}^p} \bar{R}(w)$ is bounded. The following theorem is our main result on `FedMSPP` for smooth and non-convex FL problems.

Theorem 3. *Assume that for each $m \in [M]$, the loss function $\ell^{(m)}$ is G-Lipschitz and L-smooth with respect to its first argument. Set $|I_t| \equiv I$ and $\eta_t \equiv \frac{1}{8L}\min\left\{\frac{1}{T^{1/3}}, \sqrt{\frac{bI}{T}}\right\}$. Suppose that the local update oracle of `FedMSPP` is ε_t-inexactly solved with $\varepsilon_t \leq \min\left\{\frac{G^2\eta_t}{2(L\eta_t+1)}, \frac{G^2\eta_t}{8b^2}, \frac{L^2G^2\eta_t^3}{2bI(L\eta_t+1)}\right\}$. Let t^* be an index uniformly randomly chosen in $\{0, 1, ..., T-1\}$. Then it holds that*

$$\mathbb{E}\left[\left\|\nabla\bar{R}(w_{t^*})\right\|^2\right] \lesssim \left(L\bar{\Delta}^{(0)} + G^2\right)\max\left\{\frac{1}{T^{2/3}}, \frac{1}{\sqrt{TbI}}\right\}.$$

Proof. Let us consider $d_t^{(m)} = \nabla R^{(m)}_{B_t^{(m)}}(w_t^{(m)})$ which is roughly the local update direction on device m, in the sense that $w_t^{(m)} \approx w_{t-1} - \eta_t d_t^{(m)}$ given that the local update oracle is solved to sufficient accuracy. As a key ingredient of our proof, we show via some extended uniform stability arguments in terms of gradients (see Lemma 3) that the averaged directions $d_t := \frac{1}{|I_t|}\sum_{\xi \in I_t} d_t^{(\xi)}$ aligns well with the global gradient $\nabla \bar{R}(w_{t-1})$ in expectation (see Lemma 11). Therefore, in average it roughly holds that $w_t = \frac{1}{|I_t|}\sum_{\xi \in I_t} w_t^{(\xi)} \approx w_{t-1} - \eta_t d_t \approx w_{t-1} - \eta_t \nabla \bar{R}(w_{t-1})$, which suggests that w_t is updated roughly along the direction of global gradient descent and thus is expected to converge quickly. Based on this novel analysis, we are free of explicitly imposing local dissimilarity type conditions on local objectives. See Appendix C.1 for a full proof of this result. □

Remark 7. *For $T \geq (bI)^3$, the bound in Theorem 3 is dominated by $\mathcal{O}\left(\frac{1}{\sqrt{TbI}}\right)$ which gives the communication complexity $\frac{1}{bI\epsilon^2}$. This shows that* `FedMSPP` *enjoys linear speedup with respect to both local minibatching and device sampling sizes.*

Remark 8. *While the bound in Theorem 3 is derived for the population form of FL in (1), an identical bound naturally holds for the empirical form (2) under minibatch sampling according to local data empirical distribution.*

3.2 Results for Non-smooth Problems

Analogues to `FedProx` , we can further show that `FedMSPP` converges reasonably well when applied to weakly convex and non-smooth problems. In the analysis to follow, we assume that the initial sub-optimality $\bar{\Delta}_\rho^{(0)} := \bar{R}_\rho(w_0) - \min_{w \in \mathbb{R}^p} \bar{R}_\rho(w)$ associated with ρ-Moreau-envelope of $\bar{R}$ is bounded. The following is our main result in this line.

Theorem 4. *Assume that for each $m \in [M]$, the loss function $\ell^{(m)}$ is G-Lipschitz and ν-weakly convex with respect to its first argument. Set $\eta_t \equiv \frac{\rho}{\sqrt{T}}$ for arbitrary $\rho < \frac{1}{2\nu}$. Suppose that the local update oracle of* `FedMSPP` *is exactly solved with $\varepsilon_t \equiv 0$. Let t^* be an index uniformly randomly chosen in $\{0, 1, ..., T-1\}$. Then it holds that*

$$\mathbb{E}\left[\left\|\nabla \bar{R}_\rho(w_{t^*})\right\|^2\right] \lesssim \frac{\bar{\Delta}_\rho^{(0)} + \rho G^2}{\rho\sqrt{T}}.$$

Proof. The proof argument is a slight adaptation of that of Theorem 2 to the population FL setup (1) with `FedMSPP`. For the sake of completeness, a full proof is reproduced in Appendix C.2. □

We comment in passing that the discussions made in Remarks 4-6 extend directly to Theorem 4.

4 Additional Related Work

The present work is situated at the intersection of federated learning, stochastic proximal point optimization and algorithmic stability theory. We next briefly review some additional work in these lines of research that are closely related to ours.

Heterogenous federated learning. The presence of device heterogeneity features a key distinction between FL and classic distributed learning. The most commonly used FL method is `FedAvg` (McMahan et al., 2017), where the local update oracle is formed as multi-epoch SGD. `FedAvg` was early analyzed for identical functions (Stich, 2019; Stich and Karimireddy, 2020) under the name of local SGD. In heterogeneous setting, numerous recent studies have focused on the analysis of `FedAvg` and other variants under various notions of local dissimilarity (Li et al., 2020c; Woodworth et al., 2020; Chen et al., 2020; Khaled et al., 2020; Reddi et al., 2021; Khanduri et al., 2021; Li et al., 2022; Chen et al., 2022; Zhao et al., 2022). As another representative FL method, `FedProx` (Li et al., 2020b) has recently been proposed to apply averaged proximal point updates to solve heterogeneous federated minimization problems. The theoretical guarantees of `FedProx` have been established for both convex and non-convex problems, but under a fairly stringent assumption of gradient similarity (see Definition 1) to measure data heterogeneity (Li et al., 2020b; Pathak and Wainwright, 2020; Nguyen et al., 2021). This assumption was relaxed by `FedPD` (Zhang et al., 2020) inside a meta-framework of primal-dual optimization. The `SCAFFOLD` (Karimireddy et al., 2020) and

VRL-SGD (Liang et al., 2019) are two algorithms that utilize variance reduction techniques to correct the local update directions, achieving convergence guarantees independent of the data heterogeneity. For composite non-smooth FL problems, the FCO proposed in Yuan et al. (2021) employs a server dual averaging procedure to circumvent the curse of primal averaging suffered by FedAvg. In sharp contrast to these prior works which either require certain stringent local dissimilarity conditions, or require full device participation, or only applicable to smooth problems, we show through a novel analysis based on algorithmic stability theory that the well-known FedProx can elegantly overcome all these shortcomings in a simple algorithmic framework. Another research direction in FL is to adopt compression methods for efficient communication (e.g., Haddadpour et al. (2020); Li and Li (2022)), which could also be applied to FedProx, as a topic for future investigation.

Minibatch stochastic proximal point methods. The proposed FedMSPP algorithm is a variant of FedProx that simply replaces the local proximal point oracle with MSPP, which in each iteration updates the local model via (approximately) solving a proximal point estimator over a stochastic minibatch. The MSPP-type methods have been shown to attain a substantially improved iteration stability and adaptivity for large-scale machine learning, especially in non-smooth optimization settings (Li et al., 2014; Wang et al., 2017; Asi and Duchi, 2019; Deng and Gao, 2021). However, it is not yet known if FedProx or FedMSPP can achieve similar strong guarantees for non-smooth heterogenous FL problems.

Algorithmic stability. Our analysis for FedMSPP builds largely upon the classic algorithmic stability theory. Since the seminal work of Bousquet and Elisseeff (2002), algorithmic stability has been serving as a powerful proxy for establishing strong generalization bounds (Zhang, 2003; Mukherjee et al., 2006; Shalev-Shwartz et al., 2010). Particularly, the state-of-the-art risk bounds of strongly convex ERM are offered by approaches based on the notion of uniform stability (Feldman and Vondrák, 2018, 2019; Bousquet et al., 2020; Klochkov and Zhivotovskiy, 2021). It was shown by Hardt et al. (2016) that the solution obtained via SGD is stable for smooth convex or non-convex loss functions. For non-smooth convex losses, the stability induced generalization bounds have been established for SGD (Lei and Ying, 2020; Bassily et al., 2020; Yuan and Li, 2022). Through the lens of uniform algorithmic stability, convergence rates of MSPP have been studied for non-smooth and convex (Wang et al., 2017), or weakly convex (Deng and Gao, 2021) losses.

5 Conclusions

In this paper, we have exposed three shortcomings of the prior analysis for FedProx in unrealistic assumptions about local dissimilarity, inapplicability to non-smooth problems and expensive (and potentially imbalanced) computational cost of local update. In order to tackle these issues, we developed a novel convergence theory for the vanilla FedProx and its minibatch stochastic variant, FedMSPP, through the lens of algorithmic stability theory. In a nutshell, our results reveal that with minimal modifications, FedProx is able to kill three birds with one stone: it enjoys favorable rates of convergence which are simultaneously invariant to certain stringent local dissimilarity conditions, applicable to smooth or non-smooth problems, and scaling linearly with respect to local minibatch size and device sampling ratio for smooth problems. To the best of our knowledge, the present work is the first theoretical contribution that achieves all these appealing properties in a single FL framework.

Limitations. While our results in Theorems 2 and 4 for the first time guarantee the non-asymptotic convergence of FedProx and FedMSPP for non-smooth and weakly-convex problems, the corresponding rates of convergence so far cannot demonstrate any linear speedup effort w.r.t. device sampling ratio and local minibatching size. This is as opposed to what have been shown for smooth problems in Theorems 1 and 3; thus we view it as a potential limitation of the techniques used by our analysis. In the smooth-loss case, the comparison in Table 1 suggests that our results in Theorem 1 and Theorem 3 show no faster convergence rates than those of the existing FL methods based on local SGD update, despite that FedProx/FedMSPP requires a considerably more expensive oracle for local update.

Acknowledgments and Disclosure of Funding

The authors sincerely thank the anonymous reviewers for their insightful comments which are truly helpful for improving this paper. The research was conducted while both authors worked for the Baidu Cognitive Computing Lab. Xiao-Tong Yuan is also funded in part by the National Key Research and Development Program of China under Grant No. 2018AAA0100400, and in part by the Natural Science Foundation of China (NSFC) under Grant No.U21B2049, No.61876090 and No.61936005.

References

Hilal Asi and John C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM J. Optim.*, 29(3):2257–2290, 2019.

Hilal Asi, Karan N. Chadha, Gary Cheng, and John C. Duchi. Minibatch stochastic approximate proximal point methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2020.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2020.

Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 610–626, Virtual Event [Graz, Austria], 2020.

Xiangyi Chen, Xiaoyun Li, and Ping Li. Toward communication efficient adaptive gradient method. In *Proceedings of the ACM-IMS Foundations of Data Science Conference (FODS)*, pages 119–128, Virtual Event, 2020.

Xiangyi Chen, Belhal Karimi, Weijie Zhao, and Ping Li. On the convergence of decentralized adaptive gradient methods. In *Proceedings of The 14th Asian Conference on Machine Learning (ACML)*, Hyderabad, India, 2022.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. EMNIST: Extending MNIST to handwritten letters. In *International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926, Anchorage,AK, 2017.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29(1):207–239, 2019.

Qi Deng and Wenzhi Gao. Minibatch and momentum model-based methods for stochastic weakly convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 23115–23127, virtual, 2021.

Igor Donevski, Jimmy Jessen Nielsen, and Petar Popovski. On addressing heterogeneity in federated learning for autonomous vehicles connected to a drone orchestrator. *Frontiers in Communications and Networks*, 2:1–14, 2021.

André Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *J. Mach. Learn. Res.*, 6:55–79, 2005.

Vitaly Feldman and Jan Vondrák. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9770–9780, Montréal, Canada, 2018.

Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 1270–1279, Phoenix, AZ, 2019.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 2009.

Farzin Haddadpour, Belhal Karimi, Ping Li, and Xiaoyun Li. Fedsketch: Communication-efficient and private federated learning via sketching. *arXiv preprint arXiv:2008.04975*, 2020.

Andrew Hard, Kurt Partridge, Cameron Nguyen, Niranjan Subrahmanya, Aishanee Shah, Pai Zhu, Ignacio Lopez-Moreno, and Rajiv Mathews. Training keyword spotting models on non-iid data with federated learning. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech)*, pages 4343–4347, Virtual Event, Shanghai, China, 2020.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 1225–1234, New York City, NY, 2016.

Chaoyang He, Alay Dilipbhai Shah, Zhenheng Tang, Di Fan1Adarshan Naiynar Sivashunmugam, Keerti Bhogaraju, Mita Shimpi, Li Shen, Xiaowen Chu, Mahdi Soltanolkotabi, and Salman Avestimehr. FedCV: A federated learning framework for diverse computer vision tasks. *arXiv preprint arXiv:2111.11066*, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5132–5143, Virtual Event, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4519–4529, Online [Palermo, Sicily, Italy], 2020.

Prashant Khanduri, Pranay Sharma, Haibo Yang, Mingyi Hong, Jia Liu, Ketan Rajawat, and Pramod K. Varshney. STEM: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6050–6061, virtual, 2021.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5065–5076, virtual, 2021.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5809–5819, Virtual Event, 2020.

Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J. Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 661–670, New York, NY, 2014.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. FedDANE: A federated newton-type method. In *Proceedings of the 53rd Asilomar Conference on Signals, Systems, and Computers (Asilomar)*, pages 1227–1231, Pacific Grove, CA, 2019.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.*, 37(3):50–60, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys)*, Austin, TX, 2020b.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-IID data. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020c.

Xiaoyun Li and Ping Li. Analysis of error feedback in federated non-convex optimization with biased compression. *arXiv preprint arXiv:2211.14292*, 2022.

Xiaoyun Li, Belhal Karimi, and Ping Li. On distributed adaptive optimization with gradient compression. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*, Virtual Event, 2022.

Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance reduced local SGD with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282, Fort Lauderdale, FL, 2017.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

Hung T Nguyen, Vikash Sehwag, Seyyedali Hosseinalipour, Christopher G Brinton, Mung Chiang, and H Vincent Poor. Fast-convergent federated learning. *IEEE Journal on Selected Areas in Communications*, 39(1):201–218, 2020.

Hung T. Nguyen, Vikash Sehwag, Seyyedali Hosseinalipour, Christopher G. Brinton, Mung Chiang, and H. Vincent Poor. Fast-convergent federated learning. *IEEE J. Sel. Areas Commun.*, 39(1): 201–218, 2021.

Reese Pathak and Martin J. Wainwright. Fedsplit: an algorithmic framework for fast federated optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2020.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021.

Omar Rivasplata, Csaba Szepesvári, John Shawe-Taylor, Emilio Parrado-Hernández, and Shiliang Sun. PAC-Bayes bounds for stable algorithms with instance-dependent priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9234–9244, Montréal, Canada, 2018.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, 2010.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, 2019.

Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed updates. *J. Mach. Learn. Res.*, 21:1–36, 2020.

Qianqian Tong, Guannan Liang, Tan Zhu, and Jinbo Bi. Federated nonconvex sparse learning. *arXiv preprint arXiv:2101.00052*, 2020.

Jialei Wang, Weiran Wang, and Nathan Srebro. Memory and communication efficient distributed stochastic optimization with minibatch prox. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, pages 1882–1919, Amsterdam, The Netherlands, 2017.

Blake E. Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local SGD for heterogeneous distributed learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2020.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2):12:1–12:19, 2019.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 5693–5700, Honolulu, HI, 2019.

Honglin Yuan, Manzil Zaheer, and Sashank J. Reddi. Federated composite optimization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 12253–12266, Virtual Event, 2021.

Xiao-Tong Yuan and Ping Li. Boosting the confidence of generalization for l_2-stable randomized learning algorithms. *arXiv preprint arXiv:2206.03834*, 2022.

Tong Zhang. Leave-one-out bounds for kernel methods. *Neural Comput.*, 15(6):1397–1437, 2003.

Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. FedPD: A federated learning framework with optimal rates and adaptivity to non-iid data. *arXiv preprint arXiv:2005.11418*, 2020.

Weijie Zhao, Xuewu Jiao, Mingqing Hu, Xiaoyun Li, Xiangyu Zhang, and Ping Li. Communication-efficient terabyte-scale model training framework for online advertising. In *Proceedings of the IEEE International Conference on Big Data (IEEE BigData)*, Osaka, Japan, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] *In Section 1.2, we highlighted the core contributions made in this paper which are respectively expanded with details in Section 2 and 3.*
   (b) Did you describe the limitations of your work? [Yes] *We have provided a remark on the limitations of our theory at the end of the main paper.*
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] *Our contribution is theoretical in nature. As far as we are aware of, there are no foreseeable societal or ethical consequences for the present research.*
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] *The key assumptions on the structure of loss functions (such as Lipschitzness, smoothness, weak-convexity), learning rates and local oracle update sub-optimality have all been explicitly stated in details in Theorems 1, 2, 3, 4.*
   (b) Did you include complete proofs of all theoretical results? [Yes] *Due to space limit, all the full proofs of results are relegated to the appendix sections which can be found in the supplementary document. Particularly for Theorem 2 and Theorem 3, we further provide sketched proofs in the main submission to highlight the proof road maps.*
3. If you ran experiments... [N/A] *Our work is focused on providing deeper theoretical understandings of `FedProx` and its stochastic variants under milder conditions.*
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets... [N/A] *We did not use any of such assets in this work.*
5. If you used crowdsourcing or conducted research with human subjects... [N/A] *The present research was carried out with no crowdsourcing or human subjects involved in.*

The appendix is structured as follows:

- In Appendix A, we present a few preliminary lemmas to be used in our analysis.
- In Appendix B, we provide the technical proofs for the main results in Section 2.
- In Appendix C, we provide the technical proofs for the main results in Section 3.
- In Appendix D, we collect the technical proofs for some preliminary lemmas in Appendix A.
- In Appendix E, we present some experimental results on the evaluation of `FedMSPP`.

A Preliminaries

We present in this section some preliminary results on the classic algorithmic stability theory to be used in our analysis. Let us consider an algorithm $A : \mathcal{Z}^N \mapsto \mathcal{W}$ that maps a training data set $S = \{z_i\}_{i\in[N]} \in \mathcal{Z}^N$ to a model $A(S)$ in a closed subset $\mathcal{W} \subseteq \mathbb{R}^p$ such that the following population risk function (with a slight abuse of notation) evaluated at the model is as small as possible:

$$R(A(S)) := \mathbb{E}_{Z\sim\mathcal{D}}[\ell(A(S); Z)].$$

The corresponding empirical risk is defined by

$$R_S(A(S)) := \mathbb{E}_{Z\sim\texttt{Unif}(S)}[\ell(A(S); Z)] = \frac{1}{N}\sum_{i=1}^{N} \ell(A(S); z_i).$$

We denote by $S \doteq S'$ if a pair of data sets S and S' differ in a single data point. The following concept of stability that serves as a powerful tool for analyzing the generalization bounds of learning algorithms (Hardt et al., 2016; Elisseeff et al., 2005; Bassily et al., 2020).

Definition 4 (Uniform Argument Stability). *Let $A : \mathcal{Z}^N \mapsto \mathcal{W}$ be a learning algorithm that maps a data set $S \in \mathcal{Z}^N$ to a model $A(S) \in \mathcal{W}$. Then A is said to have γ-uniform stability if for every $N \geq 1$,*

$$\sup_{S\doteq S'} \|A(S) - A(S')\| \leq \gamma.$$

The following basic lemma is about the uniform argument stability of an inexact regularized empirical risk minimization (ERM) estimator. See Appendix D.1 for its proof.

Lemma 1. *Assume that the loss function ℓ is G-Lipschitz with respect to its first argument. Suppose that the regularized objective $R^r_S(w) := \frac{1}{N}\sum_{i=1}^N \ell(w; z_i) + r(w)$ is λ-strongly convex for any S. Consider the inexact estimator w_S that satisfies the following for some $\varepsilon_t \geq 0$:*

$$R^r_S(w_S) \leq \min_w R^r_S(w) + \varepsilon_t.$$

Then w_S has uniform argument stability with parameter $\frac{4G}{\lambda N} + 2\sqrt{\frac{2\varepsilon_t}{\lambda}}$.

We further need to use the following variant of Efron-Stein inequality to random vector-valued functions (see, e.g., Lemma 6, Rivasplata et al., 2018).

Lemma 2 (Efron-Stein inequality for vector-valued functions). *Let $S = \{Z_1, Z_2, ..., Z_N\}$ be a set of i.i.d. random variables valued in $\mathcal{Z}$. Suppose that the function $h : \mathcal{Z}^N \mapsto \mathcal{H}$ valued in a Hilbert space $\mathcal{H}$ is measurable and satisfies the bounded differences property, i.e., the following inequality holds for any $i \in [N]$ and any $z_1, ..., z_N, z'_i$:*

$$\|h(z_1, ..., z_{i-1}, z_i, z_{i+1}, ..., z_N) - h(z_1, ..., z_{i-1}, z'_i, z_{i+1}, ..., z_N)\| \leq \beta.$$

Then it holds that

$$\mathbb{E}_S\left[\|h(S) - \mathbb{E}_S[h(S)]\|^2\right] \leq \beta^2 N.$$

Based on the Efron-Stein inequality in Lemma 2, we can establish the following lemma which states the generalization bounds of a uniformly stable learning algorithm in terms of gradient. A proof of this result can be found in Appendix D.2.

Lemma 3. *Suppose that a learning algorithm $A : \mathcal{Z}^N \mapsto \mathcal{W}$ has γ-uniform stability. Assume that the loss function ℓ is G-Lipschitz and L-smooth with respect to its first argument. Then the following bounds hold:*

$$\begin{aligned}\|\mathbb{E}_S\left[\nabla R(A(S)) - \nabla R_S(A(S))\right]\| \leq& L\gamma,\\ \mathbb{E}_S\left[\|\nabla R(A(S)) - \mathbb{E}_S\left[\nabla R(A(S))\right]\|^2\right] \leq& L^2\gamma^2 N.\end{aligned}$$

B Proofs for Section 2

B.1 Proof of Theorem 1

Let $d_t^{(m)} = \nabla R^{(m)}_{\texttt{erm}}(w_t^{(m)})$. We define the following quantities

$$d_t := \frac{1}{|I_t|}\sum_{\xi\in I_t} d_t^{(\xi)}, \quad \bar{d}_t := \frac{1}{M}\sum_{m=1}^{M} d_t^{(m)}. \tag{9}$$

The following elementary lemma is useful in our analysis.

Lemma 4. *Assume that for each $m \in [M]$, the loss function $\ell^{(m)}$ is G-Lipschitz. Set $|I_t| \equiv I$. Then it holds that*

$$\mathbb{E}\left[d_t\right] = \bar{d}_t, \quad \mathbb{E}\left[\|d_t - \bar{d}_t\|^2\right] \le \frac{G^2}{I}.$$

Proof. By uniform sampling strategy we have

$$\mathbb{E}\left[d_{I_t}\right] = \mathbb{E}\left[\frac{1}{|I_t|}\sum_{\xi\in I_t} d_t^{(\xi)}\right] = \frac{1}{I}\sum_{\xi\in I_t}\mathbb{E}\left[d_t^{(\xi)}\right] = \frac{1}{I}\sum_{\xi\in I_t}\frac{1}{M}\sum_{m=1}^{M} d_t^{(m)} = \bar{d}_t.$$

Then it follows that

$$\begin{aligned}
\mathbb{E}\left[\|d_t - \bar{d}_t\|^2\right] =& \mathbb{E}\left[\left\|\frac{1}{|I_t|}\sum_{\xi\in I_t} d_t^{(\xi)} - \bar{d}_t\right\|^2\right] \\
=& \frac{1}{I^2}\mathbb{E}\left[\left\|\sum_{\xi\in I_t}(d_t^{(\xi)} - \bar{d}_t)\right\|^2\right] \\
=& \frac{1}{I^2}\sum_{\xi\in I_t}\mathbb{E}\left[\left\|d_t^{(\xi)} - \bar{d}_t\right\|^2\right] \le \quad \frac{1}{I}\mathbb{E}\left[(d_t^{(\xi)})^2\right] \le \frac{G^2}{I},
\end{aligned}$$

where we have used the fact $\mathbb{E}\left[d_t^{(\xi)}\right] = \bar{d}_t$, the independence among the indices in I_t and the G-Lipschitzness of losses. The desired bounds are proved. $\square$

We also need the following lemma which quantifies the impact of local update precision to the gradient norm at the inexact solution.

Lemma 5. *Assume that for each $m \in [M]$, the loss function $\ell^{(m)}$ is L-smooth with respect to its first argument. Suppose that the local update oracle of `FedProx` is ε_t-inexactly solved and $\eta_t < \frac{1}{L}$. Then it holds that*

$$\left\|w_t^{(m)} - w_{t-1} + \eta_t d_t^{(m)}\right\| \le \eta_t\sqrt{2(L+\eta_t^{-1})\varepsilon_t}.$$

Proof. Recall $Q^{(m)}_{\texttt{erm}}(w; w_{t-1}) = R^{(m)}_{\texttt{erm}}(w) + \frac{1}{2\eta_t}\|w - w_{t-1}\|^2$. Since the loss functions are L-smooth and $\eta_t < \frac{1}{L}$, $Q^{(m)}_{\texttt{erm}}(w; w_{t-1})$ is strongly convex and thus admits a global minimizer. Then we have

$$\begin{aligned}
&\left\|\nabla R^{(m)}_{\texttt{erm}}(w_t^{(m)}) + \frac{1}{\eta_t}(w_t^{(m)} - w_{t-1})\right\|^2 \\
=&\left\|\nabla Q^{(m)}_{\texttt{erm}}(w_t^{(m)}; w_{t-1})\right\|^2 \\
\le& 2(L+\eta_t^{-1})\left(Q^{(m)}_{\texttt{erm}}(w_t^{(m)}; w_{t-1}) - \min_w Q^{(m)}_{\texttt{erm}}(w; w_{t-1})\right) \le 2(L+\eta_t^{-1})\varepsilon_t,
\end{aligned}$$

where in the last inequality is due to Definition 2. This implies the desired bound. $\square$

Lemma 6. *Assume that for each $m \in [M]$, the loss function $\ell^{(m)}$ is G-Lipschitz and L-smooth with respect to its first argument. Suppose that the local update oracle of* `FedProx` *is ε_t-inexactly solved and $\eta_t < \frac{1}{L}$. Then the following holds almost surely:*

$$\|\nabla \bar{R}_{\mathtt{erm}}(w_{t-1}) - \bar{d}_t\|^2 \leq L^2 \left(G + \sqrt{2(L+\eta_t^{-1})\varepsilon_t} \right)^2 \eta_t^2.$$

Proof. By Lemma 5 we know that

$$\|w_t^{(m)} - w_{t-1}\| \leq \eta_t \|d_t^{(m)}\| + \eta_t \sqrt{2(L+\eta_t^{-1})\varepsilon_t} \leq \left(G + \sqrt{2(L+\eta_t^{-1})\varepsilon_t} \right) \eta_t, \tag{10}$$

where we have used the G-Lipschitz assumption of loss. By definition we can see that

$$\begin{aligned}
\|\nabla \bar{R}_{\mathtt{erm}}(w_{t-1}) - \bar{d}_t\|^2 =& \left\| \frac{1}{M} \sum_{m=1}^{M} \left(\nabla R_{\mathtt{erm}}^{(m)}(w_{t-1}) - \nabla R_{\mathtt{erm}}^{(m)}(w_t^{(m)}) \right) \right\|^2 \\
\leq & \frac{1}{M} \sum_{m=1}^{M} \left\| \nabla R_{\mathtt{erm}}^{(m)}(w_{t-1}) - \nabla R_{\mathtt{erm}}^{(m)}(w_t^{(m)}) \right\|^2 \\
\overset{\zeta_1}{\leq} & \frac{L^2}{M} \sum_{m=1}^{M} \left\| w_{t-1} - w_t^{(m)} \right\|^2 \\
\overset{\zeta_2}{\leq} & L^2 \left(G + \sqrt{2(L+\eta_t^{-1})\varepsilon_t} \right)^2 \eta_t^2,
\end{aligned}$$

where in "ζ_1" we have used the L-smoothness of loss, in "ζ_2" we have used (10). This proves the desired bound. □

With all the above lemmas in place, we can prove the main result in Theorem 1. Let $\{\mathcal{F}_t\}_{t\geq 1}$ be the filtration generated by the random iterates $\{w_t\}_{t\geq 1}$ as $\mathcal{F}_t = \sigma(w_1, w_2, ..., w_t)$, where the randomness comes from the sampling of devices for partial participation.

Proof of Theorem 1. Let us denote $\delta_t^{(m)} := \eta_t^{-1}(w_t^{(m)} - w^{(t-1)}) + d_t^{(m)}$, $\delta_t := \frac{1}{|I_t|} \sum_{\xi \in I_t} \delta_t^{(\xi)}$ and $\bar{\delta}_t := \frac{1}{M} \sum_{m=1}^{M} \delta_t^{(m)}$. Then we have $\mathbb{E}[\delta_t] = \bar{\delta}_t$ and

$$w_t = w_{t-1} - \eta_t(d_t - \delta_t).$$

It can be verified based on Lemma 5 and triangle inequality that the following holds almost surely:

$$\max\left\{ \|\bar{\delta}_t\|, \|\delta_t\| \right\} \leq \sqrt{2(L+\eta_t^{-1})\varepsilon_t}. \tag{11}$$

Since the loss is L-smooth, we can show that

$$
\begin{aligned}
&\mathbb{E}[\bar{R}_{\texttt{erm}}(w_t) \mid \mathcal{F}_{t-1}] \\
\leq & \mathbb{E}\left[\bar{R}_{\texttt{erm}}(w_{t-1}) + \left\langle \nabla \bar{R}_{\texttt{erm}}(w_{t-1}), w_t - w_{t-1}\right\rangle + \frac{L}{2}\|w_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
= & \mathbb{E}\left[\bar{R}_{\texttt{erm}}(w_{t-1}) - \eta_t \left\langle \nabla \bar{R}_{\texttt{erm}}(w_{t-1}), d_t - \delta_t\right\rangle + \frac{L\eta_t^2}{2}\left\|d_t - \delta_t\right\|^2 \mid \mathcal{F}_{t-1}\right] \\
= & \bar{R}_{\texttt{erm}}(w_{t-1}) + \mathbb{E}\left[-\eta_t \left\langle \nabla \bar{R}_{\texttt{erm}}(w_{t-1}), \bar{d}_t - \bar{\delta}_t\right\rangle + \frac{L\eta_t^2}{2}\left\|d_t - \delta_t\right\|^2 \mid \mathcal{F}_{t-1}\right] \\
\overset{\zeta_1}{\leq} & \bar{R}_{\texttt{erm}}(w_{t-1}) \\
& + \mathbb{E}\left[-\eta_t \left\langle \nabla \bar{R}_{\texttt{erm}}(w_{t-1}), \bar{d}_t\right\rangle + \eta_t G\|\bar{\delta}_t\| + \frac{3L\eta_t^2}{2}\|\bar{d}_t\|^2 + \frac{3L\eta_t^2}{2}\|d_t - \bar{d}_t\|^2 + \frac{3L\eta_t^2}{2}\|\delta_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
\overset{\zeta_2}{\leq} & \bar{R}_{\texttt{erm}}(w_{t-1}) + \mathbb{E}\left[-\frac{\eta_t}{2}\|\nabla \bar{R}_{\texttt{erm}}(w_{t-1})\|^2 - \frac{\eta_t}{2}\|\bar{d}_t\|^2 + \frac{\eta_t}{2}\|\nabla \bar{R}_{\texttt{erm}}(w_{t-1}) - \bar{d}_t\|^2 + G\eta_t\sqrt{2(L+\eta_t^{-1})\varepsilon_t}\right. \\
& \left. + \frac{3L\eta_t^2}{2}\left\|\bar{d}_t\right\|^2 + \frac{3LG^2\eta_t^2}{2I} + 3L(L+\eta_t^{-1})\eta_t^2\varepsilon_t \mid \mathcal{F}_{t-1}\right] \\
\overset{\zeta_3}{\leq} & \bar{R}_{\texttt{erm}}(w_{t-1}) - \frac{\eta_t}{2}\|\nabla \bar{R}_{\texttt{erm}}(w_{t-1})\|^2 + \mathbb{E}\left[\frac{\eta_t}{2}\|\nabla \bar{R}_{\texttt{erm}}(w_{t-1}) - \bar{d}_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
& + \frac{3LG^2\eta_t^2}{2I} + G\eta_t\sqrt{2(L+\eta_t^{-1})\varepsilon_t} + 3L(L+\eta_t^{-1})\eta_t^2\varepsilon_t \\
\overset{\zeta_4}{\leq} & \bar{R}_{\texttt{erm}}(w_{t-1}) - \frac{\eta_t}{2}\|\nabla \bar{R}_{\texttt{erm}}(w_{t-1})\|^2 + \frac{L^2\left(G + \sqrt{2(L+\eta_t^{-1})\varepsilon_t}\right)^2\eta_t^3}{2} + \frac{3LG^2\eta_t^2}{2I} \\
& + G\eta_t\sqrt{2(L+\eta_t^{-1})\varepsilon_t} + 3L(L+\eta_t^{-1})\eta_t^2\varepsilon_t \\
\leq & \bar{R}_{\texttt{erm}}(w_{t-1}) - \frac{\eta_t}{2}\|\nabla \bar{R}_{\texttt{erm}}(w_{t-1})\|^2 + 2L^2G^2\eta_t^3 + \frac{5LG^2\eta_t^2}{I},
\end{aligned}
$$

where in "ζ_1" we have used the G-Lipschitz of loss and triangle inequality, in "ζ_2" we have used Lemma 4 and (11), in "ζ_3" we have used $\eta_t \leq \frac{1}{3L}$, in "ζ_4" we have used the first bound of Lemma 6, and in the last inequality we have used the condition of $\varepsilon_t \leq \min\left\{\frac{2L^2G^2\eta_t^3}{I^2(L\eta_t+1)}, \frac{G^2\eta_t}{2I(L\eta_t+1)}\right\}$. Rearranging the terms and taking expectation over $\mathcal{F}_{t-1}$ in the above yields

$$
\mathbb{E}\left[\|\nabla \bar{R}_{\texttt{erm}}(w_{t-1})\|^2\right\} \leq \frac{2}{\eta_t}\mathbb{E}\left[\bar{R}_{\texttt{erm}}(w_{t-1}) - \bar{R}_{\texttt{erm}}(w_t)\right] + 4L^2G^2\eta_t^2 + \frac{10LG^2\eta_t}{I}.
$$

Averaging the above from over $t = 1, 2, ..., T$ with $\eta_t \equiv \eta$ yields

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla \bar{R}_{\texttt{erm}}(w_t)\|^2\right] \leq & \frac{2}{\eta T}\mathbb{E}\left[\bar{R}_{\texttt{erm}}(w_0) - \bar{R}_{\texttt{erm}}(w_T)\right] + 4L^2G^2\eta^2 + \frac{10LG^2\eta}{I} \\
\leq & \frac{2}{\eta T}\bar{\Delta}_{\texttt{erm}}^{(0)} + 4L^2G^2\eta^2 + \frac{10LG^2\eta}{I}.
\end{aligned}
$$

If $T < I^3$, setting $\eta = \frac{1}{3LT^{1/3}}$ yields

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla \bar{R}_{\texttt{erm}}(w_t)\|^2\right] \lesssim \frac{L\bar{\Delta}_{\texttt{erm}}^{(0)} + G^2}{T^{2/3}} + \frac{G^2}{T^{1/3}I} \lesssim \frac{L\bar{\Delta}_{\texttt{erm}}^{(0)} + G^2}{T^{2/3}}.
$$

If $T \geq I^3$, setting $\eta = \frac{1}{3L}\sqrt{\frac{I}{T}}$ yields

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla \bar{R}_{\texttt{erm}}(w_t)\|^2\right] \lesssim \frac{L\bar{\Delta}_{\texttt{erm}}^{(0)} + G^2}{\sqrt{TI}} + \frac{G^2 I}{T} \lesssim \frac{L\bar{\Delta}_{\texttt{erm}}^{(0)} + G^2}{\sqrt{TI}}.
$$

Combining the preceding two inequalities and appealing to the definition of w_{t^*} yields the desired bound. $\square$

B.2 Proof of Theorem 2

We first present the following elementary lemma which will be used in the proof. It can be viewed as an inexact extension of the well-known three-point lemma to weakly convex functions.

Lemma 7. *Let f be a ν-weakly convex function and $\eta < \frac{1}{\nu}$. Consider*

$$w^+ = \arg\min_u \left\{ f(u) + \frac{1}{2\eta}\|u - w\|^2 \right\}.$$

Then for any u, we have

$$f(w^+) + \frac{1}{2\eta}\|w^+ - w\|^2 \leq f(u) + \frac{1}{2\eta}\|u - w\|^2 - \frac{1/\eta - \nu}{2}\|w^+ - u\|^2.$$

Proof. Since $\eta < \frac{1}{\nu}$, we must have that the regularized objective $f(u) + \frac{1}{2\eta}\|u - w\|^2$ is $(1/\eta - \nu)$-strongly convex with respect to u, which immediately implies the desired bound. $\square$

We will make use of the following lemma which shows that $w_t^{(m)}$ will be close to w_{t-1} if the learning rate η_t is small enough.

Lemma 8. *Assume that for each $m \in [M]$, the loss function $\ell^{(m)}$ is G-Lipschitz and ν-weakly convex with respect to its first argument. Suppose that the local update oracle of `FedProx` is exactly solved and $\eta_t < \frac{1}{\nu}$. Then it holds that*

$$\left\| w_t^{(m)} - w_{t-1} \right\| \leq G\eta_t.$$

Proof. Recall $Q_{\texttt{erm}}^{(m)}(w; w_{t-1}) = R_{\texttt{erm}}^{(m)}(w) + \frac{1}{2\eta_t}\|w - w_{t-1}\|^2$. Since the loss function is ν-weakly convex and $\eta_t < \frac{1}{\nu}$, $Q_{\texttt{erm}}^{(m)}(w; w_{t-1})$ is strongly convex and thus admits a global minimizer. Since the local update oracle is exactly solved, we must have

$$\left\| \nabla R_{\texttt{erm}}^{(m)}(w_t^{(m)}) + \frac{1}{\eta_t}(w_t^{(m)} - w_{t-1}) \right\| = 0,$$

which implies the desired bound due to the G-Lipschitzness. $\square$

With the above two preliminary lemmas in place, we are now in the position to prove the main result in Theorem 2.

Proof of Theorem 2. Since the losses are ν-weakly convex and $\eta_t < \frac{1}{\nu}$, in view of Lemma 7 we can show for each $m \in [M]$ that the following holds for any w,

$$R_{\texttt{erm}}^{(m)}(w_t^{(m)}) + \frac{1}{2\eta_t}\|w_t^{(m)} - w_{t-1}\|^2 \leq R_{\texttt{erm}}^{(m)}(w) + \frac{1}{2\eta_t}\|w - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t^{(m)} - w\|^2. \quad (12)$$

Let us denote

$$\bar{w}_{t-1} := \text{prox}_{\rho \bar{R}_{\texttt{erm}}}(w_{t-1}) = \arg\min_w \left\{ \bar{R}_{\texttt{erm}}(w) + \frac{1}{2\rho}\|w - w_{t-1}\|^2 \right\}.$$

Setting $w = \bar{w}_{t-1}$ in the right hand side of (12) yields

$$R_{\texttt{erm}}^{(m)}(w_t^{(m)}) + \frac{1}{2\eta_t}\|w_t^{(m)} - w_{t-1}\|^2 \leq R_{\texttt{erm}}^{(m)}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t^{(m)} - \bar{w}_{t-1}\|^2.$$

In view of the above inequality we can show that for any $\xi \in I_t$,

$$
\begin{aligned}
& R^{(\xi)}_{\texttt{erm}}(w_{t-1}) + \frac{1}{2\eta_t}\|w^{(\xi)}_t - w_{t-1}\|^2 \\
=& R^{(\xi)}_{\texttt{erm}}(w^{(\xi)}_t) + \frac{1}{2\eta_t}\|w^{(\xi)}_t - w_{t-1}\|^2 + R^{(\xi)}_{\texttt{erm}}(w_{t-1}) - R^{(\xi)}_{\texttt{erm}}(w^{(\xi)}_t) \\
\le& R^{(\xi)}_{\texttt{erm}}(w^{(\xi)}_t) + \frac{1}{2\eta_t}\|w^{(\xi)}_t - w_{t-1}\|^2 + G\|w_{t-1} - w^{(\xi)}_t\| \\
\le& R^{(\xi)}_{\texttt{erm}}(w^{(\xi)}_t) + \frac{1}{2\eta_t}\|w^{(\xi)}_t - w_{t-1}\|^2 + G^2\eta_t \\
\le& R^{(\xi)}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w^{(\xi)}_t - \bar{w}_{t-1}\|^2 + G^2\eta_t,
\end{aligned}
\tag{13}
$$

where in the last but one inequality we have applied Lemma 8. Now recall that $w_t = \frac{1}{I}\sum_{\xi\in I_t} w^{(\xi)}_t$. Then based on triangle inequality we can see that

$$
\begin{aligned}
& \frac{1}{I}\sum_{\xi\in I_t} R^{(\xi)}_{\texttt{erm}}(w_{t-1}) + \frac{1}{2\eta_t}\|w_t - w_{t-1}\|^2 \\
=& \frac{1}{I}\sum_{\xi\in I_t} R^{(\xi)}_{\texttt{erm}}(w_{t-1}) + \frac{1}{2\eta_t}\left\|\frac{1}{I}\sum_{\xi\in I_t} w^{(\xi)}_t - w_{t-1}\right\|^2 \\
\le& \frac{1}{I}\sum_{\xi\in I_t}\left\{R^{(\xi)}_{\texttt{erm}}(w_{t-1}) + \frac{1}{2\eta_t}\left\|w^{(\xi)}_t - w_{t-1}\right\|^2\right\} \\
\overset{(13)}{\le}& \frac{1}{I}\sum_{\xi\in I_t}\left\{R^{(\xi)}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w^{(\xi)}_t - \bar{w}_{t-1}\|^2 + G^2\eta_t\right\} \\
\le& \frac{1}{I}\sum_{\xi\in I_t} R^{(\xi)}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\left\|\frac{1}{I}\sum_{\xi\in I_t} w^{(\xi)}_t - \bar{w}_{t-1}\right\|^2 + G^2\eta_t \\
=& \frac{1}{I}\sum_{\xi\in I_t} R^{(\xi)}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t - \bar{w}_{t-1}\|^2 + G^2\eta_t.
\end{aligned}
$$

Conditioned on $\mathcal{F}_{t-1}$, taking expectation over both sides of the above inequality leads to the following inequality:

$$
\begin{aligned}
& \mathbb{E}\left[\bar{R}_{\texttt{erm}}(w_{t-1}) + \frac{1}{2\eta_t}\|w_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
=& \mathbb{E}\left[\frac{1}{I}\sum_{\xi\in I_t} R^{(\xi)}_{\texttt{erm}}(w_{t-1}) + \frac{1}{2\eta_t}\|w_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
\le& \mathbb{E}\left[\frac{1}{I}\sum_{\xi\in I_t} R^{(\xi)}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t - \bar{w}_{t-1}\|^2 + G^2\eta_t \mid \mathcal{F}_{t-1}\right] \\
=& \mathbb{E}\left[\bar{R}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t - \bar{w}_{t-1}\|^2 + G^2\eta_t \mid \mathcal{F}_{t-1}\right].
\end{aligned}
$$

Based the above inequality and by applying Lemma 8 again we can show that

$$\begin{aligned}
&\mathbb{E}\left[\bar{R}_{\texttt{erm}}(w_t) + \frac{1}{2\eta_t}\|w_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\bar{R}_{\texttt{erm}}(w_{t-1}) + \frac{1}{2\eta_t}\|w_t - w_{t-1}\|^2 + \bar{R}_{\texttt{erm}}(w_t) - \bar{R}_{\texttt{erm}}(w_{t-1}) \mid \mathcal{F}_{t-1}\right] \\
\leq&\mathbb{E}\left[\bar{R}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t - \bar{w}_{t-1}\|^2 + G\|w_t - w_{t-1}\| \mid \mathcal{F}_{t-1}\right] \\
\leq&\mathbb{E}\left[\bar{R}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t - \bar{w}_{t-1}\|^2 + 2G^2\eta_t \mid \mathcal{F}_{t-1}\right],
\end{aligned} \tag{14}$$

where in the last inequality we have used $\|w_t - w_{t-1}\| \leq \frac{1}{I}\sum_{\xi \in I_t}\|w_t^{(\xi)} - w_{t-1}\| \leq G\eta_t$ due to triangle inequality and Lemma 8.

Since $\bar{R}_{\texttt{erm}}$ is also ν-weakly convex, invoking Lemma 7 to $\bar{w}_{t-1} = \text{prox}_{\rho\bar{R}_{\texttt{erm}}}(w_{t-1})$ yields

$$\bar{R}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\rho}\|\bar{w}_{t-1} - w_{t-1}\|^2 \leq \bar{R}_{\texttt{erm}}(w_t) + \frac{1}{2\rho}\|w_t - w_{t-1}\|^2 - \frac{1/\rho - \nu}{2}\|\bar{w}_{t-1} - w_t\|^2,$$

which immediately leads to the following conditioned expectation bound:

$$\begin{aligned}
&\mathbb{E}\left[\bar{R}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\rho}\|\bar{w}_{t-1} - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
\leq&\mathbb{E}\left[\bar{R}_{\texttt{erm}}(w_t) + \frac{1}{2\rho}\|w_t - w_{t-1}\|^2 - \frac{1/\rho - \nu}{2}\|\bar{w}_{t-1} - w_t\|^2 \mid \mathcal{F}_{t-1}\right].
\end{aligned} \tag{15}$$

By summing up (14) and (15) we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1/\eta_t - 1/\rho}{2}\|w_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
\leq&\mathbb{E}\left[\frac{1/\eta_t - 1/\rho}{2}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t + 1/\rho - 2\nu}{2}\|\bar{w}_{t-1} - w_t\|^2 + 2G^2\eta_t \mid \mathcal{F}_{t-1}\right].
\end{aligned}$$

Since by assumption $\eta_t \leq \rho$, rearranging the terms in the above yields

$$\begin{aligned}
&\mathbb{E}\left[\|w_t - \bar{w}_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
\leq&\frac{1/\eta_t - 1/\rho}{1/\eta_t + 1/\rho - 2\nu}\|\bar{w}_{t-1} - w_{t-1}\|^2 + \frac{4G^2\eta_t}{1/\eta_t + 1/\rho - 2\nu} \\
\leq&\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{2(1/\rho - \nu)}{1/\eta_t + 1/\rho - 2\nu}\|\bar{w}_{t-1} - w_{t-1}\|^2 + \frac{4G^2\eta_t}{1/\eta_t + 1/\rho - 2\nu}.
\end{aligned}$$

Then based on the above and the definition of Moreau envelope we can show that

$$\begin{aligned}
&\mathbb{E}\left[\bar{R}_{\texttt{erm},\rho}(w_t) \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\bar{R}_{\texttt{erm}}(\bar{w}_t) + \frac{1}{2\rho}\|\bar{w}_t - w_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
\leq&\mathbb{E}\left[\bar{R}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\rho}\|\bar{w}_{t-1} - w_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
=&\bar{R}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\rho}\mathbb{E}\left[\|\bar{w}_{t-1} - w_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
\leq&\bar{R}_{\texttt{erm}}(\bar{w}_{t-1}) + \frac{1}{2\rho}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{(1/\rho - \nu)/\rho}{1/\eta_t + 1/\rho - 2\nu}\|\bar{w}_{t-1} - w_{t-1}\|^2 + \frac{2G^2\eta_t/\rho}{1/\eta_t + 1/\rho - 2\nu} \\
=&\bar{R}_{\texttt{erm},\rho}(w_{t-1}) - \frac{(1/\rho - \nu)/\rho}{1/\eta_t + 1/\rho - 2\nu}\|\bar{w}_{t-1} - w_{t-1}\|^2 + \frac{2G^2\eta_t/\rho}{1/\eta_t + 1/\rho - 2\nu} \\
=&\bar{R}_{\texttt{erm},\rho}(w_{t-1}) - \frac{1 - \rho\nu}{1/\eta_t + 1/\rho - 2\nu}\left\|\nabla\bar{R}_{\texttt{erm},\rho}(w_{t-1})\right\|^2 + \frac{2G^2\eta_t/\rho}{1/\eta_t + 1/\rho - 2\nu},
\end{aligned}$$

where in the last equality we have used the identity $\|\bar{w}_{t-1} - w_{t-1}\|^2 = \rho^2 \left\|\nabla \bar{R}_{\mathtt{erm},\rho}(w_{t-1})\right\|^2$ (see, e.g., Davis and Drusvyatskiy, 2019). By rearranging the terms in the above and taking expectation over $\mathcal{F}_{t-1}$ we obtain that

$$\frac{1-\rho\nu}{1/\eta_t + 1/\rho - 2\nu}\mathbb{E}\left[\left\|\nabla \bar{R}_{\mathtt{erm},\rho}(w_{t-1})\right\|^2\right] \le \mathbb{E}\left[\bar{R}_{\mathtt{erm},\rho}(w_{t-1})\right] - \mathbb{E}\left[\bar{R}_{\mathtt{erm},\rho}(w_t)\right] + \frac{2G^2\eta_t/\rho}{1/\eta_t + 1/\rho - 2\nu}.$$

Averaging the above over $t = 1, ..., T$ yields

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla \bar{R}_{\mathtt{erm},\rho}(w_t)\right\|^2\right] \le & \frac{1/\eta_t + 1/\rho - 2\nu}{T(1-\rho\nu)}\mathbb{E}\left[\bar{R}_{\mathtt{erm},\rho}(w_0) - \bar{R}_{\mathtt{erm},\rho}(w_T)\right] + \frac{2G^2\eta_t}{\rho(1-\rho\nu)} \\
\le & \frac{1/\eta_t + 1/\rho - 2\nu}{T(1-\rho\nu)}\bar{\Delta}^{(0)}_{\mathtt{erm},\rho} + \frac{2G^2\eta_t}{\rho(1-\rho\nu)} \\
= & \frac{(1-2\rho\nu)\bar{\Delta}^{(0)}_{\mathtt{erm},\rho}}{T\rho(1-\rho\nu)} + \frac{\bar{\Delta}^{(0)}_{\mathtt{erm},\rho}}{\eta_t T(1-\rho\nu)} + \frac{2G^2\eta_t}{\rho(1-\rho\nu)} \\
\le & \frac{\bar{\Delta}^{(0)}_{\mathtt{erm},\rho}}{T\rho} + \frac{2\bar{\Delta}^{(0)}_{\mathtt{erm},\rho}}{\eta_t T} + \frac{4G^2\eta_t}{\rho} \\
= & \frac{\bar{\Delta}^{(0)}_{\mathtt{erm},\rho}}{T\rho} + \frac{2\bar{\Delta}^{(0)}_{\mathtt{erm},\rho} + 4G^2\rho}{\rho\sqrt{T}},
\end{aligned}$$

where in the last but one inequality we have used $\rho < \frac{1}{2\nu}$, and in the last inequality we have used the choice of $\eta_t \equiv \frac{\rho}{\sqrt{T}}$. The desired bound follows by preserving the dominant terms in the above bound and appealing to the definition of t^*. □

C Proofs for Section 3

C.1 Proof of Theorem 3

For each time instance t, let us overload the notation $d_t^{(m)}$ as

$$d_t^{(m)} = \nabla R^{(m)}_{B_t^{(m)}}(w_t^{(m)}) = \frac{1}{b}\sum_{i=1}^{b}\nabla \ell^{(m)}(w_t^{(m)}; z_{i,t}^{(m)}).$$

We then accordingly overload the quantities d_t and $\bar{d}_t$ as defined in (9). We then have the following lemma analogous to Lemma 5.

Lemma 9. *Assume that for each $m \in [M]$, the loss function $\ell^{(m)}$ is L-smooth with respect to its first argument. Suppose that the local update oracle of* `FedMSPP` *is ε_t-inexactly solved and $\eta_t < \frac{1}{L}$. Then it holds that*

$$\left\|w_t^{(m)} - w_{t-1} + \eta_t d_t^{(m)}\right\| \le \eta_t\sqrt{2(L+\eta_t^{-1})\varepsilon_t}.$$

Proof. Consider $Q^{(m)}_{B_t^{(m)}}(w; w_{t-1}) = R^{(m)}_{B_t^{(m)}}(w) + \frac{1}{2\eta_t}\|w - w_{t-1}\|^2$. Since the loss functions are L-smooth and $\eta_t < \frac{1}{L}$, $Q^{(m)}_{B_t^{(m)}}(w; w_{t-1})$ must be strongly convex and thus admits a global minimizer. Then we have

$$\begin{aligned}
&\left\|\nabla R^{(m)}_{B_t^{(m)}}(w_t^{(m)}) + \frac{1}{\eta_t}(w_t^{(m)} - w_{t-1})\right\|^2 \\
=&\left\|\nabla Q^{(m)}_{B_t^{(m)}}(w_t^{(m)}; w_{t-1})\right\|^2 \\
\le& 2(L+\eta_t^{-1})\left(Q^{(m)}_{B_t^{(m)}}(w_t^{(m)}; w_{t-1}) - \min_w Q^{(m)}_{B_t^{(m)}}(w; w_{t-1})\right) \le 2(L+\eta_t^{-1})\varepsilon_t,
\end{aligned}$$

where in the last inequality is due to Definition 3. This implies the desired bound. □

Let $\{\mathcal{F}_t\}_{t\geq 1}$ be the filtration generated by the random iterates $\{w_t\}_{t\geq 1}$ as $\mathcal{F}_t = \sigma(w_1, w_2, ..., w_t)$, where the randomness jointly comes from the sampling of devices for partial participation and sampling of minibatch for local update on each chosen device.

Lemma 10. *Assume that for each $m \in [M]$, the loss function $\ell^{(m)}$ is G-Lipschitz and L-smooth with respect to its first argument. Suppose that $\eta_t < \frac{1}{L}$ and the local update oracle of* `FedMSPP` *is ε_t-inexactly solved with $\varepsilon_t \leq \frac{G^2\eta_t}{8b^2}$. Then it holds for every $m \in [M]$ that*

$$\left\|\mathbb{E}\left[\nabla R^{(m)}(w_t^{(m)}) - d_t^{(m)} \mid \mathcal{F}_{t-1}\right]\right\| \leq \frac{5LG\eta_t}{(1-\eta_t L)b},$$
$$\mathbb{E}\left[\left\|\nabla R^{(m)}(w_t^{(m)}) - \mathbb{E}[\nabla R^{(m)}(w_t^{(m)}) \mid \mathcal{F}_{t-1}]\right\|^2 \mid \mathcal{F}_{t-1}\right] \leq \frac{25L^2G^2\eta_t}{(1-\eta_t L)^2 b}.$$

Proof. Let us recall Definition 3 where the inexact solution $w_t^{(m)}$ is given by

$$Q^{(m)}_{B_t^{(m)}}(w_t^{(m)}; w_{t-1}) \leq \min_w Q^{(m)}_{B_t^{(m)}}(w; w_{t-1}) + \varepsilon_t.$$

Since the loss functions are L-smooth and $\frac{1}{\eta_t} > L$, it is easy to verify that the regularized objective $Q^{(m)}_{B_t^{(m)}}(w; w_{t-1})$ is $(\frac{1}{\eta_t} - L)$-strongly convex. Then invoking Lemma 1 yields that $w_t^{(m)}$ uniformly stable with parameter $\frac{4G}{(1/\eta_t - L)b} + 2\sqrt{\frac{2\varepsilon_t}{1/\eta_t - L}} \leq \frac{5G}{(1/\eta_t - L)b}$, which is due to the condition on ε_t. Conditioned on the sigma-field $\mathcal{F}_{t-1}$, the desired bounds follows immediately from Lemma 3. □

The next lemma, which can be proved based on the previous lemmas, is key to our analysis.

Lemma 11. *Assume that for each $m \in [M]$, the loss function $\ell^{(m)}$ is G-Lipschitz and L-smooth with respect to its first argument. Suppose that $\eta_t < \frac{1}{L}$ and the local update oracle of* `FedMSPP` *is ε_t-inexactly solved with $\varepsilon_t \leq \min\left\{\frac{G^2\eta_t}{2(L\eta_t+1)}, \frac{G^2\eta_t}{8b^2}\right\}$. Then we have*

$$\mathbb{E}\left[\left\|\nabla \bar{R}(w_{t-1}) - d_t\right\|^2 \mid \mathcal{F}_{t-1}\right] \leq 8L^2G^2\eta_t^2 + \frac{2G^2}{b|I_t|},$$

and

$$\left\|\nabla \bar{R}(w_{t-1}) - \mathbb{E}[\bar{d}_t \mid \mathcal{F}_{t-1}]\right\|^2 \leq 12L^2G^2\eta_t^2 + \frac{75L^2G^2\eta_t^2}{(1-\eta_t L)^2 b^2} + \frac{75L^2G^2\eta_t^2}{(1-\eta_t L)^2 b}.$$

Proof. By Lemma 9 we know that for each $m \in [M]$,

$$\|w_t^{(m)} - w_{t-1}\| \leq \eta_t \|d_t^{(m)}\| + \eta_t\sqrt{2(L+\eta_t^{-1})\varepsilon_t} \leq \left(G + \sqrt{2(L+\eta_t^{-1})\varepsilon_t}\right)\eta_t \leq 2G\eta_t, \quad (16)$$

where we have used the G-Lipschitz assumption of loss and $\varepsilon_t \leq \frac{G^2\eta_t}{2(L\eta_t+1)}$. By definition we can see that

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla\bar{R}(w_{t-1}) - d_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\left\|\nabla\bar{R}(w_{t-1}) - \frac{1}{|I_t|}\sum_{\xi\in I_t}\nabla R^{(\xi)}_{B_t^{(\xi)}}(w_t^{(\xi)})\right\|^2 \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\left\|\nabla\bar{R}(w_{t-1}) - \frac{1}{|I_t|}\sum_{\xi\in I_t}\nabla R^{(\xi)}_{B_t^{(\xi)}}(w_{t-1}) + \frac{1}{|I_t|}\sum_{\xi\in I_t}\nabla R^{(\xi)}_{B_t^{(\xi)}}(w_{t-1}) - \frac{1}{|I_t|}\sum_{\xi\in I_t}\nabla R^{(\xi)}_{B_t^{(\xi)}}(w_t^{(\xi)})\right\|^2 \mid \mathcal{F}_{t-1}\right] \\
\leq&\mathbb{E}\left[2\left\|\nabla\bar{R}(w_{t-1}) - \frac{1}{b|I_t|}\sum_{\xi\in I_t}\sum_{i\in[b]}\nabla\ell^{(\xi)}(w_{t-1};z_{i,t}^{(\xi)})\right\|^2 + \frac{2}{|I_t|}\sum_{\xi\in I_t}\left\|\nabla R^{(\xi)}_{B_t^{(\xi)}}(w_{t-1}) - \nabla R^{(\xi)}_{B_t^{(\xi)}}(w_t^{(\xi)})\right\|^2 \mid \mathcal{F}_{t-1}\right] \\
\overset{\zeta_1}{\leq}&\frac{2}{b^2|I_t|^2}\sum_{\xi\in I_t}\sum_{i\in[b]}\mathbb{E}\left[\left\|\nabla\bar{R}(w_{t-1}) - \nabla\ell^{(\xi)}(w_{t-1};z_{i,t}^{(\xi)})\right\|^2 \mid \mathcal{F}_{t-1}\right] + \frac{2L^2}{|I_t|}\sum_{\xi\in I_t}\mathbb{E}\left[\left\|w_{t-1} - w_t^{(\xi)}\right\|^2 \mid \mathcal{F}_{t-1}\right] \\
\overset{\zeta_2}{\leq}&\frac{2}{b^2|I_t|^2}\sum_{\xi\in I_t}\sum_{i\in[b]}\mathbb{E}\left[\left\|\nabla\ell^{(\xi)}(w_{t-1};z_{i,t}^{(\xi)})\right\|^2 \mid \mathcal{F}_{t-1}\right] + 8L^2G^2\eta_t^2 \\
\leq&\frac{2G^2}{b|I_t|} + 8L^2G^2\eta_t^2,
\end{aligned}
$$

where in "ζ_1" we have used the independent sampling of data and devices and the L-smoothness of loss, in "ζ_2" we have used the fact $\mathbb{E}[\|Z-\mathbb{E}[Z]\|^2] \leq \mathbb{E}[\|Z\|^2]$ and (16), and in the last inequality we have used the G-Lipschitzness of loss . This proves the first desired bound.

To prove the second bound, by definition we can see that

$$
\begin{aligned}
&\left\|\nabla\bar{R}(w_{t-1}) - \mathbb{E}[\bar{d}_t \mid \mathcal{F}_{t-1}]\right\|^2 \\
=&\left\|\frac{1}{M}\sum_{m=1}^{M}\left(\nabla R^{(m)}(w_{t-1}) - \mathbb{E}[d_t^{(m)} \mid \mathcal{F}_{t-1}]\right)\right\|^2 \\
=&\left\|\frac{1}{M}\sum_{m=1}^{M}\left(\nabla R^{(m)}(w_{t-1}) - \nabla R^{(m)}(w_t^{(m)}) + \nabla R^{(m)}(w_t^{(m)}) - \mathbb{E}[\nabla R^{(m)}(w_t^{(m)}) \mid \mathcal{F}_{t-1}]\right.\right. \\
&\qquad\left.\left. + \mathbb{E}\left[\nabla R^{(m)}(w_t^{(m)}) - d_t^{(m)} \mid \mathcal{F}_{t-1}\right]\right)\right\|^2 . \\
\leq&\underbrace{\frac{3}{M}\sum_{m=1}^{M}\left\|\nabla R^{(m)}(w_{t-1}) - \nabla R^{(m)}(w_t^{(m)})\right\|^2}_{A'} + \underbrace{\frac{3}{M}\sum_{m=1}^{M}\left\|\mathbb{E}\left[\nabla R^{(m)}(w_t^{(m)}) - d_t^{(m)} \mid \mathcal{F}_{t-1}\right]\right\|^2}_{B'} \\
&+\underbrace{\frac{3}{M}\sum_{m=1}^{M}\left\|\nabla R^{(m)}(w_t^{(m)}) - \mathbb{E}[\nabla R^{(m)}(w_t^{(m)}) \mid \mathcal{F}_{t-1}]\right\|^2}_{C'}.
\end{aligned}
$$

By smoothness and (16) we can show that the following holds almost surely:

$$A' \leq 3L^2\|w_t^{(m)} - w_{t-1}\|^2 \leq 12L^2G^2\eta_t^2.$$

For the component B', based on the first bound of Lemma 10 we can easily show that

$$B' \leq \frac{75L^2G^2}{(1/\eta_t - L)^2b^2}.$$

In terms of the component C', it can be bounded via invoking the second bound of Lemma 10 that

$$\mathbb{E}\left[C' \mid \mathcal{F}_{t-1}\right] \leq \frac{75L^2G^2}{(1/\eta_t - L)^2b}.$$

Finally, by combing the preceding three bounds we obtain that

$$\begin{aligned}\left\|\nabla\bar{R}(w_{t-1})-\mathbb{E}[\bar{d}_t\mid\mathcal{F}_{t-1}]\right\|^2=&\mathbb{E}\left[\left\|\nabla\bar{R}(w_{t-1})-\mathbb{E}[\bar{d}_t\mid\mathcal{F}_{t-1}]\right\|^2\mid\mathcal{F}_{t-1}\right]\\ \le&12L^2G^2\eta_t^2+\frac{75L^2G^2}{(1/\eta_t-L)^2b^2}+\frac{75L^2G^2}{(1/\eta_t-L)^2b}.\end{aligned}$$

This proves the second desired bound. □

With all the above preliminary results in place, we are now ready to prove Theorem 3.

Proof of Theorem 3. Let us denote $\delta_t^{(m)}:=\eta_t^{-1}(w_t^{(m)}-w^{(t-1)})+d_t^{(m)}$, $\delta_t:=\frac{1}{|I_t|}\sum_{\xi\in I_t}\delta_t^{(\xi)}$ and $\bar{\delta}_t:=\frac{1}{M}\sum_{m=1}^M\delta_t^{(m)}$. Then we have $\mathbb{E}[\delta_t]=\bar{\delta}_t$ and $w_t=w_{t-1}-\eta_t(d_t-\delta_t)$. It can be verified based on Lemma 9 and triangle inequality that the following holds almost surely:

$$\max\left\{\|\bar{\delta}_t\|,\|\delta_t\|\right\}\le\sqrt{2(L+\eta_t^{-1})\varepsilon_t}. \tag{17}$$

Since the objective is L-smooth, we can show that

$$\begin{aligned}&\mathbb{E}\left[\bar{R}(w_t)\mid\mathcal{F}_{t-1}\right]\\
\le&\mathbb{E}\left[\bar{R}(w_{t-1})+\left\langle\nabla\bar{R}(w_{t-1}),w_t-w_{t-1}\right\rangle+\frac{L}{2}\|w_t-w_{t-1}\|^2\mid\mathcal{F}_{t-1}\right]\\
=&\mathbb{E}\left[\bar{R}(w_{t-1})-\eta_t\left\langle\nabla\bar{R}(w_{t-1}),d_t-\delta_t\right\rangle+\frac{L\eta_t^2}{2}\|d_t-\delta_t\|^2\mid\mathcal{F}_{t-1}\right]\\
=&\mathbb{E}\left[\bar{R}(w_{t-1})-\eta_t\left\langle\nabla\bar{R}(w_{t-1}),\mathbb{E}[\bar{d}_t-\bar{\delta}_t\mid\mathcal{F}_{t-1}]\right\rangle+\frac{L\eta_t^2}{2}\|d_t-\delta_t\|^2\mid\mathcal{F}_{t-1}\right]\\
\le&\bar{R}(w_{t-1})-\eta_t\left\langle\nabla\bar{R}(w_{t-1}),\mathbb{E}[\bar{d}_t\mid\mathcal{F}_{t-1}]\right\rangle+\mathbb{E}\left[\eta_tG\|\bar{\delta}_t\|+L\eta_t^2\|d_t\|^2+L\eta_t^2\|\delta_t\|^2\mid\mathcal{F}_{t-1}\right]\\
=&\bar{R}(w_{t-1})-\frac{\eta_t}{2}\left\|\nabla\bar{R}(w_{t-1})\right\|^2-\frac{\eta_t}{2}\left\|\mathbb{E}[\bar{d}_t\mid\mathcal{F}_{t-1}]\right\|^2+\frac{\eta_t}{2}\left\|\nabla\bar{R}(w_{t-1})-\mathbb{E}[\bar{d}_t\mid\mathcal{F}_{t-1}]\right\|^2\\
&+\mathbb{E}\left[\eta_tG\|\bar{\delta}_t\|+L\eta_t^2\|d_t\|^2+L\eta_t^2\|\delta_t\|^2\mid\mathcal{F}_{t-1}\right]\\
\overset{\zeta_1}{\le}&\bar{R}(w_{t-1})-\frac{\eta_t}{2}\left\|\nabla\bar{R}(w_{t-1})\right\|^2+\frac{\eta_t}{2}\left\|\nabla\bar{R}(w_{t-1})-\mathbb{E}[\bar{d}_t\mid\mathcal{F}_{t-1}]\right\|^2\\
&+\mathbb{E}\left[G\eta_t\sqrt{2(L+\eta_t^{-1})\varepsilon_t}+L\eta_t^2\left\|d_t\right\|^2+2L(L+\eta_t^{-1})\eta_t^2\varepsilon_t\mid\mathcal{F}_{t-1}\right]\\
\le&\bar{R}(w_{t-1})-\frac{\eta_t}{2}\|\nabla\bar{R}(w_{t-1})\|^2+\frac{\eta_t}{2}\|\nabla\bar{R}(w_{t-1})-\mathbb{E}[\bar{d}_t\mid\mathcal{F}_{t-1}]\|^2\\
&+\mathbb{E}\left[2L\eta_t^2\left\|d_t-\nabla\bar{R}(w_{t-1})\right\|^2+2L\eta_t^2\left\|\nabla\bar{R}(w_{t-1})\right\|^2+G\eta_t\sqrt{2(L+\eta_t^{-1})\varepsilon_t}+2L(L+\eta_t^{-1})\eta_t^2\varepsilon_t\mid\mathcal{F}_{t-1}\right]\\
\overset{\zeta_2}{\le}&\bar{R}(w_{t-1})-\frac{\eta_t}{4}\left\|\nabla\bar{R}(w_{t-1})\right\|^2+\frac{\eta_t}{2}\left\|\nabla\bar{R}(w_{t-1})-\mathbb{E}[\bar{d}_t\mid\mathcal{F}_{t-1}]\right\|^2\\
&+\mathbb{E}\left[2L\eta_t^2\left\|d_t-\nabla\bar{R}(w_{t-1})\right\|^2\mid\mathcal{F}_{t-1}\right]+G\eta_t\sqrt{2(L+\eta_t^{-1})\varepsilon_t}+2L(L+\eta_t^{-1})\eta_t^2\varepsilon_t\\
\overset{\zeta_3}{\le}&\bar{R}(w_{t-1})-\frac{\eta_t}{4}\left\|\nabla\bar{R}(w_{t-1})\right\|^2+6L^2G^2\eta_t^3+\frac{38L^2G^2\eta_t^3}{(1-\eta_tL)^2b^2}+\frac{38L^2G^2\eta_t^3}{(1-\eta_tL)^2b}\\
&+16L^3G^2\eta_t^4+\frac{4LG^2\eta_t^2}{bI}+G\eta_t\sqrt{2(L+\eta_t^{-1})\varepsilon_t}+2L(L+\eta_t^{-1})\eta_t^2\varepsilon_t\\
\overset{\zeta_4}{\le}&\bar{R}(w_{t-1})-\frac{\eta_t}{4}\left\|\nabla\bar{R}(w_{t-1})\right\|^2+6L^2G^2\eta_t^3+\frac{43L^2G^2\eta_t^3}{b^2}+\frac{43L^2G^2\eta_t^3}{b}\\
&+16L^3G^2\eta_t^4+\frac{4LG^2\eta_t^2}{bI}+G\eta_t\sqrt{2(L+\eta_t^{-1})\varepsilon_t}+2L(L+\eta_t^{-1})\eta_t^2\varepsilon_t\\
\overset{\zeta_5}{\le}&\bar{R}(w_{t-1})-\frac{\eta_t}{4}\left\|\nabla\bar{R}(w_{t-1})\right\|^2+94L^2G^2\eta_t^3+\frac{4LG^2\eta_t^2}{bI}+G\eta_t\sqrt{2(L+\eta_t^{-1})\varepsilon_t}+2L(L+\eta_t^{-1})\eta_t^2\varepsilon_t\\
\le&\bar{R}(w_{t-1})-\frac{\eta_t}{4}\left\|\nabla\bar{R}(w_{t-1})\right\|^2+94L^2G^2\eta_t^3+\frac{6LG^2\eta_t^2}{bI},\end{aligned}$$

where in "ζ_1" we have used (17), in "ζ_2" we have used $\eta_t \le \frac{1}{8L}$, in "ζ_3" we have used Lemma 11, in "ζ_4" we have used $\eta_t \le \frac{1}{8L}$, in "ζ_5" we have used $M, b \ge 1$ and $\eta_t \le \frac{1}{8L}$, and in the last inequality we used $\varepsilon_t \le \frac{L^2G^2\eta_t^3}{2bI(L\eta_t+1)}$ and $\eta_t \le \frac{1}{8L}$. By taking expectation over $\mathcal{F}_{t-1}$ and rearranging the terms we obtain that

$$\mathbb{E}\left[\left\|\nabla \bar{R}(w_{t-1})\right\|^2\right] \le \frac{4}{\eta_t}\left(\mathbb{E}[\bar{R}(w_{t-1})] - \mathbb{E}[\bar{R}(w_t)]\right) + 376L^2G^2\eta_t^2 + \frac{24LG^2\eta_t}{bI}.$$

Averaging the above from over $t = 1, 2, .., T$ with $\eta_t \equiv \eta$ yields

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla \bar{R}(w_t)\right\|^2\right] \le& \frac{4}{\eta T}(\bar{R}(w_0) - \bar{R}(w_T)) + 376L^2G^2\eta^2 + \frac{24L\eta G^2}{bI}\\ \le& \frac{4\bar{\Delta}^{(0)}}{\eta T} + 376L^2G^2\eta^2 + \frac{24L\eta G^2}{bI}.\end{aligned}$$

If $T < (bI)^3$, setting $\eta = \frac{1}{8LT^{1/3}}$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla \bar{R}(w_t)\|^2\right] \lesssim \frac{L\bar{\Delta}^{(0)} + G^2}{T^{2/3}} + \frac{G^2}{T^{1/3}bI} \lesssim \frac{L\bar{\Delta}^{(0)} + G^2}{T^{2/3}}.$$

If $T \ge (bI)^3$, setting $\eta = \frac{1}{8L}\sqrt{\frac{bI}{T}}$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla \bar{R}(w_t)\|^2\right] \lesssim \frac{L\bar{\Delta}^{(0)} + G^2}{\sqrt{TbI}} + \frac{G^2 bI}{T} \lesssim \frac{L\bar{\Delta}^{(0)} + G^2}{\sqrt{TbI}}.$$

Combining the preceding two inequalities and appealing to the definition of t^* yield the desired bound. □

C.2 Proof of Theorem 4

The proof argument is almost identical to that of Theorem 2. We reproduce the proof in full details here for the sake of completeness.

Similar to Lemma 8, we first establish the following lemma which shows that $w_t^{(m)}$ will be close to w_{t-1} if the learning rate η_t is small enough.

Lemma 12. *Assume that for each $m \in [M]$, the loss function $\ell^{(m)}$ is G-Lipschitz and ν-weakly convex with respect to its first argument. Suppose that the local update oracle of `FedMSPP` is exactly solved and $\eta_t < \frac{1}{\nu}$. Then it holds that*

$$\left\|w_t^{(m)} - w_{t-1}\right\| \le G\eta_t.$$

Proof. Recall $Q^{(m)}_{B_t^{(m)}}(w; w_{t-1}) = R^{(m)}_{B_t^{(m)}}(w) + \frac{1}{2\eta_t}\|w - w_{t-1}\|^2$. Since the loss function is ν-weakly convex and $\eta_t < \frac{1}{\nu}$, $Q^{(m)}_{B_t^{(m)}}(w; w_{t-1})$ is strongly convex with respect to w and thus admits a global minimizer. Since the local update oracle is exactly solved, we must have

$$\left\|\nabla R^{(m)}_{B_t^{(m)}}(w_t^{(m)}) + \frac{1}{\eta_t}(w_t^{(m)} - w_{t-1})\right\| = 0,$$

which implies the desired bound due to the G-Lipschitz-loss assumption. □

We are now ready to prove the main result in Theorem 4.

Proof of Theorem 4. Since the losses are ν-weakly convex and $\eta_t < \frac{1}{\nu}$, in view of Lemma 7 we can show for each $m \in [M]$ that the following holds for any w,

$$R^{(m)}_{B_t^{(m)}}(w_t^{(m)}) + \frac{1}{2\eta_t}\|w_t^{(m)} - w_{t-1}\|^2 \le R^{(m)}_{B_t^{(m)}}(w) + \frac{1}{2\eta_t}\|w - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t^{(m)} - w\|^2. \tag{18}$$

Let us denote for any $t \geq 1$

$$\bar{w}_{t-1} := \operatorname{prox}_{\rho \bar{R}}(w_{t-1}) = \arg\min_{w} \left\{ \bar{R}(w) + \frac{1}{2\rho}\|w - w_{t-1}\|^2 \right\}.$$

Setting $w = \bar{w}_{t-1}$ in the right hand side of (18) yields

$$R^{(m)}_{B_t^{(m)}}(w_t^{(m)}) + \frac{1}{2\eta_t}\|w_t^{(m)} - w_{t-1}\|^2 \leq R^{(m)}_{B_t^{(m)}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t^{(m)} - \bar{w}_{t-1}\|^2.$$

In view of the above inequality we can show that for any $\xi \in I_t$,

$$\begin{aligned}
&R^{(\xi)}_{B_t^{(\xi)}}(w_{t-1}) + \frac{1}{2\eta_t}\|w_t^{(\xi)} - w_{t-1}\|^2 \\
=&R^{(\xi)}_{B_t^{(\xi)}}(w_t^{(\xi)}) + \frac{1}{2\eta_t}\|w_t^{(\xi)} - w_{t-1}\|^2 + R^{(\xi)}_{B_t^{(\xi)}}(w_{t-1}) - R^{(\xi)}_{B_t^{(\xi)}}(w_t^{(\xi)}) \\
\leq&R^{(\xi)}_{B_t^{(\xi)}}(w_t^{(\xi)}) + \frac{1}{2\eta_t}\|w_t^{(\xi)} - w_{t-1}\|^2 + G\|w_{t-1} - w_t^{(\xi)}\| \\
\leq&R^{(\xi)}_{B_t^{(\xi)}}(w_t^{(\xi)}) + \frac{1}{2\eta_t}\|w_t^{(\xi)} - w_{t-1}\|^2 + G^2\eta_t \\
\leq&R^{(\xi)}_{B_t^{(\xi)}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t^{(\xi)} - \bar{w}_{t-1}\|^2 + G^2\eta_t,
\end{aligned} \tag{19}$$

where in the last but one inequality we have applied Lemma 12. Now recall that $w_t = \frac{1}{I}\sum_{\xi \in I_t} w_t^{(\xi)}$. Then based on triangle inequality we can see that

$$\begin{aligned}
&\frac{1}{I}\sum_{\xi \in I_t} R^{(\xi)}_{B_t^{(\xi)}}(w_{t-1}) + \frac{1}{2\eta_t}\|w_t - w_{t-1}\|^2 \\
=&\frac{1}{I}\sum_{\xi \in I_t} R^{(\xi)}_{B_t^{(\xi)}}(w_{t-1}) + \frac{1}{2\eta_t}\left\|\frac{1}{I}\sum_{\xi \in I_t} w_t^{(\xi)} - w_{t-1}\right\|^2 \\
\leq&\frac{1}{I}\sum_{\xi \in I_t}\left\{R^{(\xi)}_{B_t^{(\xi)}}(w_{t-1}) + \frac{1}{2\eta_t}\left\|w_t^{(\xi)} - w_{t-1}\right\|^2\right\} \\
\overset{(19)}{\leq}&\frac{1}{I}\sum_{\xi \in I_t}\left\{R^{(\xi)}_{B_t^{(\xi)}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t^{(\xi)} - \bar{w}_{t-1}\|^2 + G^2\eta_t\right\} \\
\leq&\frac{1}{I}\sum_{\xi \in I_t} R^{(\xi)}_{B_t^{(\xi)}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\left\|\frac{1}{I}\sum_{\xi \in I_t} w_t^{(\xi)} - \bar{w}_{t-1}\right\|^2 + G^2\eta_t \\
=&\frac{1}{I}\sum_{\xi \in I_t} R^{(\xi)}_{B_t^{(\xi)}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t - \bar{w}_{t-1}\|^2 + G^2\eta_t,
\end{aligned}$$

Conditioned on $\mathcal{F}_{t-1}$, taking expectation (w.r.t. both the randomness of device sampling and data sampling introduced associated with the iteration step t) over both sides of the above inequality leads to the following:

$$
\begin{aligned}
&\mathbb{E}\left[\bar{R}(w_{t-1}) + \frac{1}{2\eta_t}\|w_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\frac{1}{I}\sum_{\xi\in I_t} R^{(\xi)}(w_{t-1}) + \frac{1}{2\eta_t}\|w_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\frac{1}{I}\sum_{\xi\in I_t} R^{(\xi)}_{B_t^{(\xi)}}(w_{t-1}) + \frac{1}{2\eta_t}\|w_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
\le&\mathbb{E}\left[\frac{1}{I}\sum_{\xi\in I_t} R^{(\xi)}_{B_t^{(\xi)}}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t - \bar{w}_{t-1}\|^2 + G^2\eta_t \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\bar{R}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t - \bar{w}_{t-1}\|^2 + G^2\eta_t \mid \mathcal{F}_{t-1}\right].
\end{aligned}
$$

Based the above inequality and by applying Lemma 12 again we can show that

$$
\begin{aligned}
&\mathbb{E}\left[\bar{R}(w_t) + \frac{1}{2\eta_t}\|w_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\bar{R}(w_{t-1}) + \frac{1}{2\eta_t}\|w_t - w_{t-1}\|^2 + \bar{R}(w_t) - \bar{R}(w_{t-1}) \mid \mathcal{F}_{t-1}\right] \\
\le&\mathbb{E}\left[\bar{R}(w_{t-1}) + \frac{1}{2\eta_t}\|w_t - w_{t-1}\|^2 + G\|w_t - w_{t-1}\| \mid \mathcal{F}_{t-1}\right] \\
\le&\mathbb{E}\left[\bar{R}(\bar{w}_{t-1}) + \frac{1}{2\eta_t}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t - \nu}{2}\|w_t - \bar{w}_{t-1}\|^2 + 2G^2\eta_t \mid \mathcal{F}_{t-1}\right],
\end{aligned}
\tag{20}
$$

where in the last inequality we have used $\|w_t - w_{t-1}\| \le \frac{1}{I}\sum_{\xi\in I_t}\|w_t^{(\xi)} - w_{t-1}\| \le G\eta_t$ due to triangle inequality and Lemma 12.

Since $\bar{R}$ is also ν-weakly convex, invoking Lemma 7 to $\bar{w}_{t-1} = \text{prox}_{\rho\bar{R}_{\text{erm}}}(w_{t-1})$ yields

$$
\bar{R}(\bar{w}_{t-1}) + \frac{1}{2\rho}\|\bar{w}_{t-1} - w_{t-1}\|^2 \le \bar{R}(w_t) + \frac{1}{2\rho}\|w_t - w_{t-1}\|^2 - \frac{1/\rho - \nu}{2}\|\bar{w}_{t-1} - w_t\|^2,
$$

which immediately gives the following conditioned expectation bound:

$$
\begin{aligned}
&\mathbb{E}\left[\bar{R}(\bar{w}_{t-1}) + \frac{1}{2\rho}\|\bar{w}_{t-1} - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
\le&\mathbb{E}\left[\bar{R}(w_t) + \frac{1}{2\rho}\|w_t - w_{t-1}\|^2 - \frac{1/\rho - \nu}{2}\|\bar{w}_{t-1} - w_t\|^2 \mid \mathcal{F}_{t-1}\right].
\end{aligned}
\tag{21}
$$

By summing up (20) and (21) we get

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1/\eta_t - 1/\rho}{2}\|w_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
\le&\mathbb{E}\left[\frac{1/\eta_t - 1/\rho}{2}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{1/\eta_t + 1/\rho - 2\nu}{2}\|\bar{w}_{t-1} - w_t\|^2 + 2G^2\eta_t \mid \mathcal{F}_{t-1}\right].
\end{aligned}
$$

Since by assumption $\eta_t \le \rho$, rearranging the terms in the above yields

$$
\begin{aligned}
&\mathbb{E}\left[\|w_t - \bar{w}_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
\le&\frac{1/\eta_t - 1/\rho}{1/\eta_t + 1/\rho - 2\nu}\|\bar{w}_{t-1} - w_{t-1}\|^2 + \frac{4G^2\eta_t}{1/\eta_t + 1/\rho - 2\nu} \\
\le&\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{2(1/\rho - \nu)}{1/\eta_t + 1/\rho - 2\nu}\|\bar{w}_{t-1} - w_{t-1}\|^2 + \frac{4G^2\eta_t}{1/\eta_t + 1/\rho - 2\nu}.
\end{aligned}
$$

Then based on the above and the definition of Moreau envelope we can show that

$$
\begin{aligned}
&\mathbb{E}\left[\bar{R}_\rho(w_t) \mid \mathcal{F}_{t-1}\right] \\
=&\mathbb{E}\left[\bar{R}(\bar{w}_t) + \frac{1}{2\rho}\|\bar{w}_t - w_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
\le&\mathbb{E}\left[\bar{R}(\bar{w}_{t-1}) + \frac{1}{2\rho}\|\bar{w}_{t-1} - w_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
=&\bar{R}(\bar{w}_{t-1}) + \frac{1}{2\rho}\mathbb{E}\left[\|\bar{w}_{t-1} - w_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
\le&\bar{R}(\bar{w}_{t-1}) + \frac{1}{2\rho}\|\bar{w}_{t-1} - w_{t-1}\|^2 - \frac{(1/\rho-\nu)/\rho}{1/\eta_t + 1/\rho - 2\nu}\|\bar{w}_{t-1} - w_{t-1}\|^2 + \frac{2G^2\eta_t/\rho}{1/\eta_t + 1/\rho - 2\nu} \\
=&\bar{R}_\rho(w_{t-1}) - \frac{(1/\rho-\nu)/\rho}{1/\eta_t + 1/\rho - 2\nu}\|\bar{w}_{t-1} - w_{t-1}\|^2 + \frac{2G^2\eta_t/\rho}{1/\eta_t + 1/\rho - 2\nu} \\
=&\bar{R}_\rho(w_{t-1}) - \frac{1-\rho\nu}{1/\eta_t + 1/\rho - 2\nu}\left\|\nabla \bar{R}_\rho(w_{t-1})\right\|^2 + \frac{2G^2\eta_t/\rho}{1/\eta_t + 1/\rho - 2\nu},
\end{aligned}
$$

where in the last equality we have used the identity $\|\bar{w}_{t-1} - w_{t-1}\|^2 = \rho^2\left\|\nabla \bar{R}_\rho(w_{t-1})\right\|^2$ (see, e.g., Davis and Drusvyatskiy, 2019). By rearranging the terms in the above and taking expectation over $\mathcal{F}_{t-1}$ we obtain that

$$
\frac{1-\rho\nu}{1/\eta_t + 1/\rho - 2\nu}\mathbb{E}\left[\left\|\nabla \bar{R}_\rho(w_{t-1})\right\|^2\right] \le \mathbb{E}\left[\bar{R}_\rho(w_{t-1})\right] - \mathbb{E}\left[\bar{R}_\rho(w_t)\right] + \frac{2G^2\eta_t/\rho}{1/\eta_t + 1/\rho - 2\nu}.
$$

Averaging the above over $t = 1, ..., T$ yields

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla \bar{R}_\rho(w_t)\right\|^2\right] \le& \frac{1/\eta_t + 1/\rho - 2\nu}{T(1-\rho\nu)}\mathbb{E}\left[\bar{R}_\rho(w_0) - \bar{R}_\rho(w_T)\right] + \frac{2G^2\eta_t}{\rho(1-\rho\nu)} \\
\le& \frac{1/\eta_t + 1/\rho - 2\nu}{T(1-\rho\nu)}\bar{\Delta}_\rho^{(0)} + \frac{2G^2\eta_t}{\rho(1-\rho\nu)} \\
=& \frac{(1-2\rho\nu)\bar{\Delta}_\rho^{(0)}}{T\rho(1-\rho\nu)} + \frac{\bar{\Delta}_\rho^{(0)}}{\eta_t T(1-\rho\nu)} + \frac{2G^2\eta_t}{\rho(1-\rho\nu)} \\
\le& \frac{\bar{\Delta}_\rho^{(0)}}{T\rho} + \frac{2\bar{\Delta}_\rho^{(0)}}{\eta_t T} + \frac{4G^2\eta_t}{\rho} \\
=& \frac{\bar{\Delta}_\rho^{(0)}}{T\rho} + \frac{2\bar{\Delta}_\rho^{(0)} + 4G^2\rho}{\rho\sqrt{T}},
\end{aligned}
$$

where in the last but one inequality we have used $\rho < \frac{1}{2\nu}$, and in the last inequality we have used the choice of $\eta_t \equiv \frac{\rho}{\sqrt{T}}$. The desired bound follows by preserving the dominant terms in the above bound and appealing to the definition of t^*. □

D Proofs of Preliminary Lemmas

Here we provide the proofs of some auxiliary lemmas introduced in Appendix A.

D.1 Proof of Lemma 1

Proof. Let $w_S^* = \arg\min_{w\in\mathbb{R}^p} R_S^r(w)$. Based on the strong convexity of $R_S^r(w_S)$ we can see that

$$
\frac{\lambda}{2}\|w_S - w_S^*\|^2 \le R_S^r(w_S) - R_S^r(w_S^*) \le \varepsilon_t,
$$

which directly implies $\|w_S - w_S^*\| \le \sqrt{\frac{2\varepsilon_t}{\lambda}}$. Let us consider a sample set $S^{(i)}$ which is identical to S except that one of the z_i is replaced by another random sample z_i'. Denote

$w^*_{S^{(i)}} = \arg\min_{w\in\mathbb{R}^p} R^r_{S^{(i)}}(w)$. Then we can show that

$$
\begin{aligned}
& R^r_S(w^*_{S^{(i)}}) - R^r_S(w^*_S) \\
=& \frac{1}{N}\sum_{j\neq i}\left(\ell(w^*_{S^{(i)}}; z_j) - \ell(w^*_S; z_j)\right) + \frac{1}{N}\left(\ell(w^*_{S^{(i)}}; z_i) - \ell(w^*_S; z_i)\right) + r(w^*_{S^{(i)}}) - r(w^*_S) \\
=& R^r_{S^{(i)}}(w^*_{S^{(i)}}) - R^r_{S^{(i)}}(w^*_S) + \frac{1}{N}\left(\ell(w^*_{S^{(i)}}; z_i) - \ell(w^*_S; z_i)\right) - \frac{1}{N}\left(\ell(w^*_{S^{(i)}}; z'_i) - \ell(w^*_S; z'_i)\right) \\
\overset{\zeta_1}{\leq}& \frac{1}{N}\left(\ell(w^*_{S^{(i)}}; z_i) - \ell(w^*_S; z_i)\right) - \frac{1}{N}\left(\ell(w^*_{S^{(i)}}; z'_i) - \ell(w^*_S; z'_i)\right) \\
\overset{\zeta_2}{\leq}& \frac{2G}{N}\|w^*_{S^{(i)}} - w^*_S\|,
\end{aligned}
$$

where "ζ_1" follows from the optimality of $w_{S^{(i)}}$ and "ζ_2" is due to the Lipschitz continuity of loss. The strong convexity of R^r_S implies

$$
R^r_S(w^*_{S^{(i)}}) - R^r_S(w^*_S) \geq \frac{\lambda}{2}\|w^*_{S^{(i)}} - w^*_S\|^2.
$$

Combining the preceding two inequalities yields

$$
\|w^*_{S^{(i)}} - w^*_S\| \leq \frac{4G}{\lambda N}.
$$

Therefore by triangle inequality and the above bounds we get

$$
\begin{aligned}
\|w_{S^{(i)}} - w_S\| =& \|w_{S^{(i)}} - w^*_{S^{(i)}} + w^*_{S^{(i)}} - w^*_S + w^*_S - w_S\| \\
\leq& \|w_{S^{(i)}} - w^*_{S^{(i)}}\| + \|w^*_{S^{(i)}} - w^*_S\| + \|w^*_S - w_S\| \\
\leq& \frac{4G}{\lambda N} + 2\sqrt{\frac{2\varepsilon_t}{\lambda}}
\end{aligned}
$$

which implies the desired uniform stability as the above holds for any pair of $S^{(i)}$ and S. □

D.2 Proof of Lemma 3

Proof. Let us consider a sample set $S^{(i)}$ which is identical to S except that one of the Z_i is replaced by another random sample Z'_i. Since S and $S^{(i)}$ are both i.i.d. samples of the data distribution. It follows that

$$
\mathbb{E}_S\left[\nabla R(A(S))\right] = \mathbb{E}_{S^{(i)}}\left[\nabla R(A(S^{(i)}))\right] = \mathbb{E}_{S^{(i)}\cup\{Z_i\}}\left[\nabla \ell(A(S^{(i)})); Z_i)\right] = \mathbb{E}_{S\cup\{Z'_i\}}\left[\nabla \ell(A(S^{(i)}); Z_i)\right].
$$

Since the above holds for all $i = 1, ..., N$, by averaging the above equality over the training data we obtain that

$$
\mathbb{E}_S\left[\nabla R(A(S))\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{S\cup\{Z'_i\}}\left[\nabla \ell(A(S^{(i)}); Z_i)\right]. \tag{22}
$$

Regarding the empirical case, by definition we have

$$
\mathbb{E}_S\left[\nabla R_S(A(S))\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_S\left[\nabla \ell(A(S); Z_i)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{S\cup\{Z'_i\}}\left[\nabla \ell(A(S); Z_i)\right].
$$

Combining the preceding two equalities gives that

$$
\begin{aligned}
\|\mathbb{E}_S\left[\nabla R(A(S)) - \nabla R_S(A(S))\right]\| =& \left\|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{S\cup\{Z'_i\}}\left[\nabla \ell(A(S^{(i)}); Z_i) - \nabla \ell(A(S); Z_i)\right]\right\| \\
\leq& L\left\|A(S^{(i)}) - A(S)\right\| \leq L\gamma,
\end{aligned}
$$

where we have used the uniform stability of A.

To prove the second inequality, again by smoothness of the loss function we have

$$
\left\|\nabla R(A(S)) - \nabla R(A(S^{(i)}))\right\| \leq L\left\|A(S) - A(S^{(i)})\right\| \leq L\gamma.
$$

Then it follows from Lemma 2 that

$$
\mathbb{E}_S\left[\|\nabla R(A(S)) - \mathbb{E}_S\left[\nabla R_S(A(S))\right]\|^2\right] \leq L^2\gamma^2 N.
$$

The proof is completed. □

E Preliminary Experimental Results

In this section, we carry out a preliminary experimental study to demonstrate the speed-up behavior of `FedMSPP` under varying minibatch sizes for achieving comparable test performances to `FedProx`. We also conventionally use `FedAvg` as a baseline algorithm for comparison.

E.1 Data and Models

We compare the considered algorithms over the following three benchmark data sets popularly used for evaluating heterogenous FL approaches:

- The MNIST (LeCun et al., 1998) dataset of handwritten digits 0-9 is used for digit image classification with a two layer convolutional neural network (CNN). The model takes as input the images of size 28×28, and first performs a 2-layer ($\{1, 32$, max-pooling$\}$, $\{32, 64$, max-pooling$\}$) convolution followed by a fully connected (FC) layer. We use 63,000 images in which 90% are for training and the rest for test. The data are distributed over 100 devices such that each device has samples of only 2 digits.
- The FEMNIST (Li et al., 2020b) dataset is a subset of the 62-class EMNIST (Cohen et al., 2017) database constructed by sub-sampling 10 lower case characters ('a'-'j'). We study the performances of the considered algorithms for character image classification using the same two layer CNN as used for MNIST, which takes as input the images of size 28×28. We use 55,050 images in which 90% are for training and the rest for test. The data are distributed over 50 devices, each of which has samples of 3 characters.
- The Sent140 (Go et al., 2009) dataset of text sentiment analysis on tweets is used for evaluating the considered algorithms for sentiment classification. The model we use is a two layer LSTM binary classifier containing 256 hidden units followed by a densely-connected layer. The input is a sequence of 25 characters represented by a 300-dimensional GloVe embedding (Pennington et al., 2014) and the output is one character per training sample. We use for our experiment a total number of $21,546$ tweets from 261 twitter accounts, each of which corresponds to a device. The training/test sample split is 80% versus 20%.

The statistics of the data and models in use are summarized in Table 2.

Dataset	Model	# Devices	# Samples (Training)
MNIST	2-layer CNN	100	$63,000$ (56700)
FEMNIST	2-layer CNN	50	55050 (49545)
Sent140	2-layer LSTM	261	21546 (17237)

Table 2: Statistics of data and models used in the experiments.

E.2 Implementation Details and Performance Metrics

We generally follow the instructions of Li et al. (2020b) for implementing `FedProx`, `FedMSPP` and `FedAvg`. More specifically, we use SGD as the local solver for `FedProx`, `FedMSPP` and `FedAvg`. For `FedMSPP`, we implement with three varying minibatch sizes on each data set as shortly reported in the next subsection about results. The hyper-parameters used in our implementation, such as number of communication rounds and number of local SGD epochs, are listed in Table 3.

Hyper-parameter	MNIST	FEMNIST	Sent140
#Communication rounds	200	300	300
#Local SGD epochs	2	5	10
Local SGD minibatch size	567	512	100
Local SGD learning rate	0.25	0.06	0.1
Strength of regularization μ_t	0.1	0.1	0.001

Table 3: Hyper-parameter settings.

Since the chief goal of this empirical study is to illustrate the benefit of `FedMSPP` for speeding up the convergence of `FedProx`, we use the numbers of data points and communication rounds needed for reaching the desired solution accuracy as performance metrics. The desired test accuracies are $\{80\%, 90\%, 95\%\}$ on MNIST, $\{80\%, 85\%, 91\%\}$ on FEMNIST, and $\{68\%, 70\%, 73\%\}$ on Sent140.

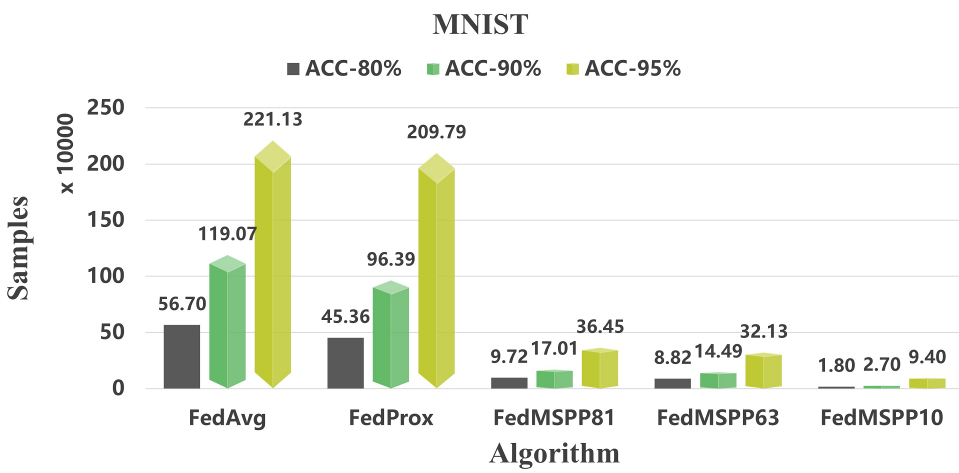
E.3 Results

In Figure 1, we show the numbers of data samples accessed by the considered algorithms to reach comparable test accuracies. For `FedMSPP`, we test with minibatch sizes $\{81, 63, 10\}$ on MNIST, $\{128, 64, 16\}$ on FEMNIST, and $\{75, 50, 20\}$ on Sent140. From this set of results we can observe that:

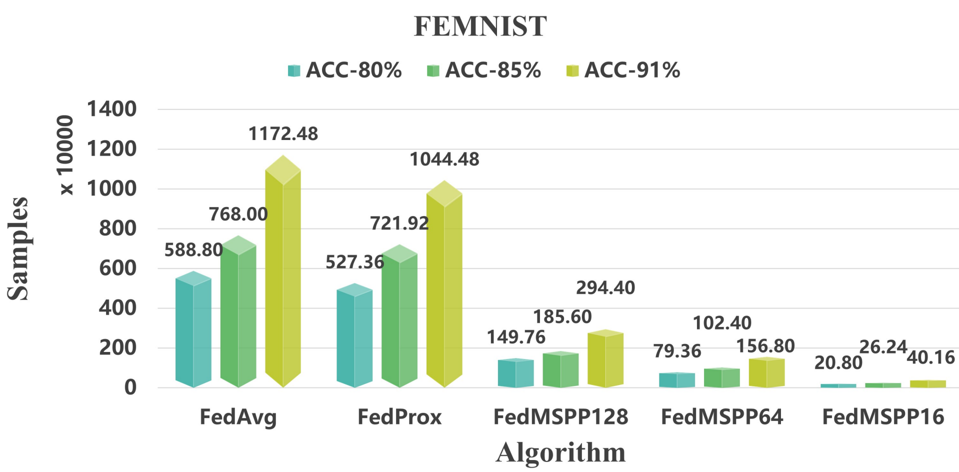
- On all the three data sets in use, `FedMSPP` with varying minibatch sizes consistently needs significantly fewer samples than `FedProx` and `FedAvg` to reach the desired test accuracies.
- `FedMSPP` with smaller minibatch size tends to have better sample efficiency.

Figure 2 shows the corresponding rounds of communication needed to reach comparable test accuracies. From this group results we can see that in most cases, `FedMSPP` just needs slightly increased rounds of communication than `FedProx` and `FedAvg` to reach comparable generalization accuracy.

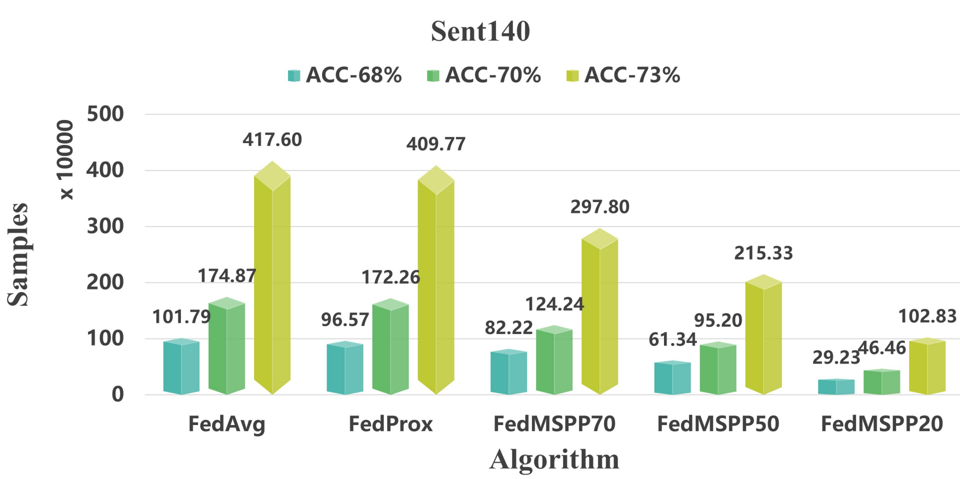
Overall, our numerical results confirm that `FedMSPP` can be served as a safe and computationally more efficient replacement to `FedProx` on the considered heterogenous FL tasks.



(a) MNIST: Numbers of data points needed to reach 80%, 90% and 95% test accuracies. For `FedMSPP`, we test with different minibatch sizes 81, 63, 10.

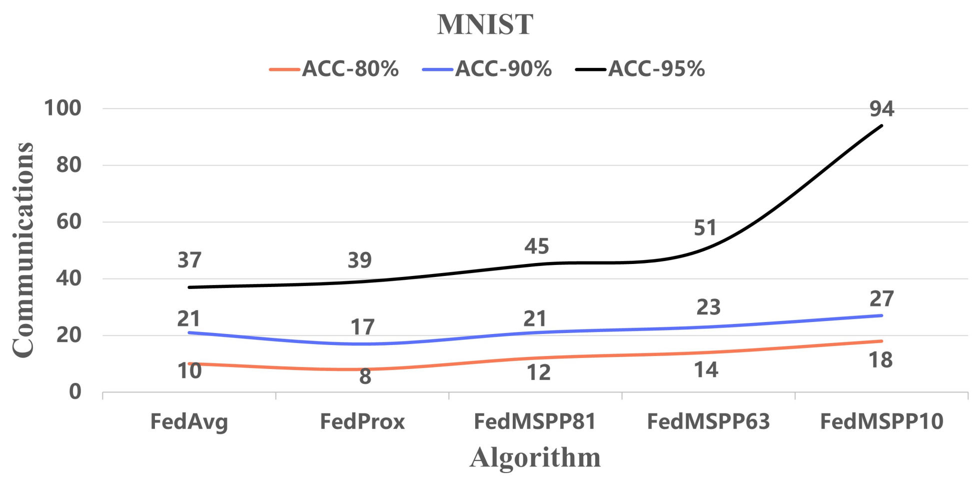


(b) FEMNIST: Numbers of data points needed to reach 80%, 85% and 91% test accuracies. For `FedMSPP`, we test with different minibatch sizes 128, 64, 16.

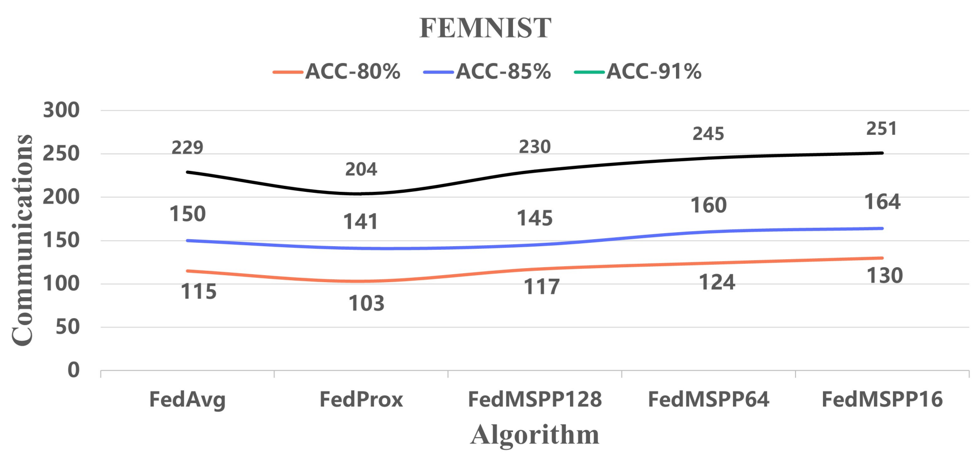


(c) Sent140: Numbers of data points needed to reach 68%, 70% and 73% test accuracies. For `FedMSPP`, we test with different minibatch sizes 70, 50, 20.

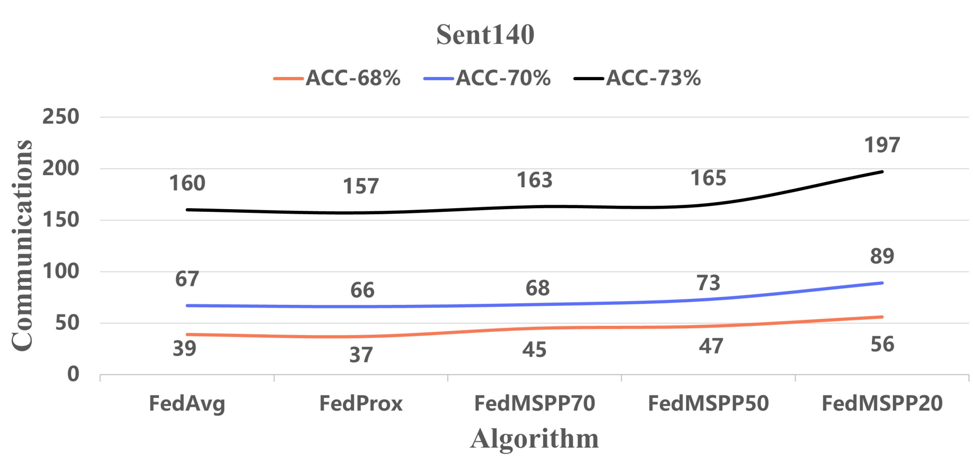
Figure 1: Comparison of numbers of data points accessed by the considered algorithms to reach varying desired test accuracies.



(a) MNIST: Rounds of communication needed to reach 80%, 90% and 95% test accuracies. For `FedMSPP`, we test with different minibatch sizes 81, 63, 10.



(b) FEMNIST: Rounds of communication needed to reach 80%, 85% and 91% test accuracies. For `FedMSPP`, we test with different minibatch sizes 128, 64, 16.



(c) Sent140: Rounds of communication needed to reach 68%, 70% and 73% test accuracies. For `FedMSPP`, we test with different minibatch sizes 70, 50, 20.

Figure 2: Comparison of rounds of communication needed by the considered algorithms to reach varying desired test accuracies.