SPAM: Stochastic Proximal Point Method with Momentum Variance Reduction for Non-convex Cross-Device Federated Learning

Anonymous Authors[1]

Abstract

Cross-device training is a crucial subfield of federated learning, where the number of clients can reach the billions. Standard approaches and local methods are prone to client drift and insensitivity to data similarities. We propose a novel algorithm (SPAM) for cross-device federated learning with non-convex and non-smooth losses. We provide a sharp analysis under second-order (Hessian) similarity, a condition satisfied by various machine learning problems in practice. Additionally, we extend our results to the partial participation setting, where a cohort of selected clients communicate with the server at each communication round. We then conduct a complexity analysis of our convergence results, showing the improvement of our methods upon prior work. Finally, we back up our results with experiments.

1. Introduction

Federated learning (FL) is a general learning mechanism where multiple entities, known as *clients*, work together to solve a machine learning problem under the guidance of a *central server* (Kairouz et al., 2021; Konečný et al., 2016; McMahan et al., 2017). Each client's raw data stays on their local devices and is not shared or transferred; local updates are aggregated on the central server (Kairouz et al., 2021).

This paper focuses on *cross-device* training, where the clients are mobile or IoT devices (Karimireddy et al., 2021). To model such a large number of clients, we study the following stochastic optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x), \quad \text{where} \quad f(x) := \mathrm{E}_{\xi\sim\mathcal{D}}\left[f_\xi(x)\right], \qquad (1)$$

where f_ξ may be non-convex. Here, we do not have access to the full function f, nor its gradient. This framework reflects the cross-device setting, where the number of clients is huge (e.g., billions of mobile phones), so each client participates in the training process only a few times or only even once. Therefore, we cannot expect full participation to obtain the exact gradient.

Instead, we can sample from the distribution $\mathcal{D}$ and compute $f_\xi(x)$ and $\nabla f_\xi(x)$ at each point x. We assume that the gradient and the expectation are interchangeable, meaning $\mathrm{E}_{\xi\sim\mathcal{D}}\left[\nabla f_\xi(x)\right] = \nabla f(x)$. In the context of cross-device training, f_ξ represents the loss of client ξ on its local data (Karimireddy et al., 2021).

The formulation (1) is more appropriate than the finite-sum (*cross-silo*) formulation (Wang et al., 2021):

$$\min_{x\in\mathbb{R}^d} f(x), \quad \text{where} \quad f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x).$$

The latter setting is applicable for collaborative training by organizations when n is moderately large (e.g., medical (Ogier du Terrail et al., 2022)).

Communication bottleneck. In federated learning, it is essential to broadcast or communicate information between computing nodes, such as the current gradient vector or model state. This communication often becomes the main challenge, particularly in the cross-device setting where the nodes are less powerful devices with slow network connections (Konečný et al., 2016; Caldas et al., 2018; Kairouz et al., 2021). Two main approaches to reducing communication overhead are compression and local training. Communication compression uses inexact but relevant approximations of the transferred messages at each round. These approximations often rely on (stochastic) compression operators, which can be applied to both the gradient and the model. For a more detailed discussion on compression mechanisms and algorithms, see (Xu et al., 2020; Beznosikov et al., 2020; Shulgin & Richtárik, 2022).

Local training. The second technique for reducing communication overhead is to get better client updates by performing local training. Local SGD steps have been a crucial component of practical federated training algorithms since the inception of the field, demonstrating strong empirical performance by improving communication efficiency

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

(Mangasarian & Solodov, 1993; McDonald et al., 2010; McMahan et al., 2017). However, rigorous theoretical explanations for this phenomenon were lacking until the recent introduction of the ProxSkip method by (Mishchenko et al., 2022). ScaffNew (ProxSkip specialized for the distributed setting) has been shown to provide accelerated communication complexity in the convex setting. While ScaffNew works for any level of heterogeneity, it does not benefit from the similarity between clients. In addition, methods like ScaffNew, designed to fix the client drift issue (Acar et al., 2020; Karimireddy et al., 2020), require each client to maintain state (control variate), which is incompatible with cross-device FL (Reddi et al., 2020).

Partial participation. In cross-silo federated learning, periodically, all clients may be active in a single communication round. However, an important property of cross-device learning is the impracticality of accessing all clients simultaneously. Most clients might be available only once during the entire training process. Therefore, it is crucial to design federated learning methods where only a small cohort of devices participates in each round. Modeling the problem according to (1) naturally avoids the possibility of engaging all clients at once. We refer the reader to (Reddi et al., 2020; Karimireddy et al., 2021) and (Khaled & Jin, 2022) for more details on partial participation.

Data heterogeneity. Despite recent progress in federated learning, handling data heterogeneity across clients remains a significant challenge (Kairouz et al., 2021). Empirical observations show that clients' labels for similar inputs can vary significantly (Arivazhagan et al., 2019; Silva et al., 2022). This variation arises from clients having different preferences. When local steps are used in this context, clients tend to overfit their data, a phenomenon known as client drift.

An alternative to local gradient steps is a local proximal point operator oracle, which involves solving a regularized local optimization problem on the selected client(s). This approach underlies FedProx (Li et al., 2020), which relies on a restrictive heterogeneity assumption. The algorithm was analyzed from the perspective of the Stochastic Proximal Point Method (SPPM) in (Yuan & Li, 2022). Independently, the theory of SPPM is compatible with the second-order similarity condition (Assumption 2) from an analytical perspective (Mishchenko et al., 2023). Based on these connections, various studies have explored SPPM-based federated learning algorithms, and we refer the reader to (Khaled & Jin, 2022) and (Lin et al., 2024) for more details.

1.1. Prior work

Momentum. Momentum Variance Reduction (MVR) was introduced in the context of server-only stochastic non-convex optimization (Cutkosky & Orabona, 2019). The primary motivation behind this method was to avoid computing full gradients (which is impractical in the stochastic setting) or requiring "giant batch sizes" of order $\mathcal{O}(1/\varepsilon^2)$. Such large batches are necessary for other methods like PAGE (Li et al., 2021) to find an ε-stationary point.

The authors assume bounded variance for stochastic gradients ∇f_ξ with a noise variance σ^2. Their convergence result for non-convex objectives includes σ^2 in the upper bound. To eliminate the dependence on this parameter, they propose an adaptive stepsize schedule under the additional assumption that f_ξ is Lipschitz continuous.

MIME. MIME is a flexible framework that makes existing optimization algorithms applicable in the distributed setting by combining them with local SGD updates (Karimireddy et al., 2021). The authors then study particular instances of the framework, such as MIME + ADAM (Kingma & Ba, 2014) and MIME + MVR (Cutkosky & Orabona, 2019).

However, their analysis with local steps is limited for the non-convex cross-device setting. First, they assume smoothness also in the case of one sampled client. More importantly, MIME suffers from a common issue of local methods. In Theorem 4 of (Karimireddy et al., 2021), the stepsize is taken to be of order $O(1/Lm)$, where L is the smoothness parameter of the client loss and m is the number of local steps. Thus, the stepsize is so tiny that multiple steps become equivalent to a single, smoother stochastic gradient descent step, negating the potential benefits of local SGD. Finally, their analysis requires an additional weak convexity assumption for the objective in the partial participation setting.

CE-LSGD. The Communication Efficient Local Stochastic Gradient Descent (CE-LSGD) was introduced by (Patel et al., 2022). They propose and analyze two algorithms, with the second one tailored for the cross-device setting (1). This algorithm comprises two components: the MVR update on the server and SARAH local steps on the selected client. The latter, known as the Stochastic Recursive Gradient Algorithm, is a variance-reduced version of SGD that periodically requires the full gradient of the objective function (Nguyen et al., 2017).

The analysis by (Patel et al., 2022) explicitly describes how to choose the number of local updates and the local stepsize. They also provide lower bounds for two-point first-order oracle-based federated learning algorithms. The drawback of their setting is that to have meaningful local updates; they need the smoothness of each client function f_ξ. In addition, similar to MIME, the stepsize depends on the number of local steps, which limits the benefit of doing many local steps.

SABER. The SABER algorithm by (Mishchenko et al., 2023) combines SPPM updates on the clients with PAGE

updates on the server. Their paper utilizes Hessian similarity (Assumption 2) and leverages it for the finite-sum optimization objective. However, their analysis for the partial participation setting relies on an assumption that is difficult to verify in the general non-convex regime. If the function is not weakly convex, as in the case of MIME, this assumption may not hold. Specifically, it requires that $f\left(\frac{1}{B}\sum_{i=1}^{B} w_i\right) \leq \frac{1}{B}\sum_{i=1}^{B} f(w_i)$, where w_i are arbitrary vectors in $\mathbb{R}^d$ obtained using proximal point operators.

1.2. Contributions

This paper introduces a novel method called Stochastic Proximal Point And Momentum (SPAM). Our method combines Momentum Variance Reduction (MVR) on the server side to leverage its efficiency in stochastic optimization while employing Stochastic Proximal Point Method (SPPM) updates on the clients' side. We analyze four versions of the proposed algorithm:

- SPAM - exact PPM with constant parameters,
- SPAM - exact PPM with varying parameters,
- SPAM-inexact - inexact PPM with varying parameters,
- SPAM-PP - inexact PPM with varying parameters and partial participation.

We then carry out an in-depth theoretical analysis of the proposed methods, showcasing their advantages compared to relevant competitors and addressing the limitations present in those works. Specifically, we demonstrate convergence upper bounds on the average expected gradient norm for all variants of SPAM.

We also conduct a communication complexity analysis based on our convergence results. Namely, we show that SPAM can provably benefit from similarity. In addition, we designed a varying stepsize schedule that removed the neighborhood from the stationarity boundaries. Our algorithm achieves the optimal convergence rate of $O(1/K^{1/3})$, where K denotes the number of iterations leveraging this scheme.

Our algorithms, in particular SPAM-PP, shine in the cross-device setting compared to the competitors. First, in contrast to non-SPPM-based algorithms, such as MIME and CE-LSGD, we allow greater *flexibility for the local solvers*. Thus, unlike MIME and CE-LSGD, we do not require either convexity or smoothness of the local objectives. Our algorithm is compatible with any local solver when the latter satisfies certain conditions outlined in Definition 4.1. Furthermore, compared to SABER, our partial participation setting does not require (weak) convexity of the objective. Moreover, we offer substantially simpler analysis than prior works and can be of independent interest outside of the FL context. We compare the relevant methods in Table 1.

Table 1: Comparison of the proposed algorithm with other relevant methods. The columns are: **HS** - Hessian Similarity, **PP** - Partial Participation, **NSA** - No Smoothness Assumption, **CD** - Cross Device, **SU** - Server Update, **CO** - Client Oracle.

Algorithm	HS	PP	NSA	CD	SU	CO
FedProx	✗	✔	✔	✔	–	PPM
SABER	✔	✗	✔	✗	PAGE	PPM
MIME	✔	✗	✗	✔	MVR	SGD
CE-LSGD	✔	✔	✗	✔	MVR	SARAH
SPAM	✔	✔	✔	✔	MVR	PPM

As opposed to many standard federated learning techniques (Karimireddy et al., 2020; Acar et al., 2020; Mishchenko et al., 2022), our algorithms do not need local states/control variates to be stored on each client. Not using local states is crucial for cross-device learning as each client may participate in training only once.

Finally, we validate our theoretical findings through meticulously designed experiments. Specifically, we tackle a federated ridge regression problem, where we can precisely control the second-order heterogeneity parameter δ, as well as the computation of the local proximal operator.

Paper Organization. The rest of the paper is organized as follows. Section 2 presents the mathematical notation and the theoretical assumptions we use in the analysis. The main algorithm with the exact proximal point oracle and its theoretical analysis are presented in Section 3. Sections 4 and A contain, respectively, the version of the algorithm with an inexact proximal operator and our most general method, which uses random cohorts of clients. We present experiments in Appendix B and conclude the paper in Section 5.

2. Notation and assumptions

We use ∇f for the gradient, $\|\cdot\|$ for the Euclidean norm, and $\mathrm{E}\left[\cdot\right]$ for the expectation. $\mathsf{Unif}(S)$ denotes uniform distribution over the discrete set S. The proximal point operator of a real-valued function $g : \mathbb{R}^d \to \mathbb{R}$ is defined as the solution of the following optimization

$$\mathrm{prox}_g\left(x\right) := \arg\min_{y\in\mathbb{R}^d}\left\{g(y) + \frac{1}{2}\|x-y\|^2\right\}. \quad (2)$$

We refer the reader to (Beck, 2017) for the properties of the proximal point operator. We assume there exists a lower bound for function f, and it is denoted as $f_{\inf} > -\infty$.

We use index i for a non-random client, while ξ is used for a randomly selected client. One of the main assumptions

of our analysis is that we have access to stochastic samples $\xi \sim \mathcal{D}$ and in particular, we can evaluate the gradient ∇f_ξ at any point $x \in \mathbb{R}^d$.

Assumption 1 (Bounded variance). *We assume there exists $\sigma \geq 0$ such that for any $x \in \mathbb{R}^d$*

$$\mathrm{E}\left[\|\nabla f_\xi(x) - \nabla f(x)\|^2\right] \leq \sigma^2. \tag{3}$$

We say that the function f is L-smooth, if its gradient is Lipschitz continuous $\forall x, y \in \mathbb{R}^d$:

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|x - y\|. \tag{4}$$

In many machine learning scenarios, the non-convex objective functions do not satisfy (4). Moreover, several prior works (Zhang et al., 2019; Crawshaw et al., 2022) showed that such smoothness condition does not capture the properties of popular models like LSTM, Recurrent Neural Networks, and Transformers.

Our second assumption is the second-order heterogeneity. Further in the analysis, this assumption will take the role of smoothness.

Assumption 2 (Hessian similarity). *Assume there exists $\delta \geq 0$ such that for any i and $x, y \in \mathbb{R}^d$*

$$\|\nabla f_i(x) - \nabla f(x) - \nabla f_i(y) + \nabla f(y)\| \leq \delta\|x - y\|. \tag{5}$$

When all functions f_i are twice-differentiable the above condition can also be formulated as

$$\left\|\nabla^2 f_i(x) - \nabla^2 f(x)\right\| \leq \delta, \tag{6}$$

motivating the name *second-order heterogeneity* used interchangeably with *Hessian similarity* (Khaled & Jin, 2022).

This assumption holds for a large class of machine learning problems (Mairal, 2013; Shamir et al., 2014; Mairal, 2015). Typical examples include least squares regression, classification with logistic loss (Woodworth et al., 2023), statistical learning for quadratics (Shamir et al., 2014), generalized linear models (Hendrikx et al., 2020) etc. Furthermore, a similar assumption was used to improve convergence results in centralized (Tyurin et al., 2023) and communication-constrained distributed settings (Szlendak et al., 2022). In the distributed setting, (5) is especially relevant as the parameter δ remains small, even if different clients have similar input distributions but widely varying outputs for the same input. See more details on the assumption in (Khaled & Jin, 2022, Section 9) and (Woodworth et al., 2023, Section 3) for discussion on synthetic data, private learning, etc.

In the following sections, we present our main algorithms as well as the corresponding convergence theorems. We focus on the non-convex optimization problem (1), where the goal is to find an ε-approximate stationary point $x \in \mathbb{R}^d$ such that $\mathrm{E}\left[\|\nabla f(x)\|^2\right] \leq \varepsilon$.

3. SPAM

In this section, we describe our main algorithm in its simpler form, that is, SPAM with one sampled client and exact proximal point computations. We then provide theoretical convergence guarantees and a complexity analysis of the proposed methods.

The algorithm proceeds as follows. We first choose a step-size sequence γ_k and a momentum sequence p_k. The server samples a client. The selected client then computes the new gradient estimator g_k and assigns the new iterate as the proximal point operator with a shifted gradient term:

$$x_{k+1} = \mathrm{prox}_{\gamma_k f_{\xi_k}}\left(x_k + \gamma_k(\nabla f_{\xi_k}(x_k) - g_k)\right).$$

Then $x_{k+1} = \arg\min_y \phi_k(y)$ with

$$\phi_k(y) := f_{\xi_k}(y) + \langle g_k - \nabla f_{\xi_k}(x_k), y - x_k\rangle + \frac{\|y - x_k\|^2}{2\gamma_k}. \tag{7}$$

The new iterate is then sent to the server, and the process repeats itself. For the algorithm's pseudocode, please refer to Algorithm 1.

Algorithm 1 SPAM, SPAM-inexact

Input: Starting point $x_0 = x_{-1} \in \mathbb{R}^d$, initialize $g_0 = g_{-1}$, choose $\gamma_k > 0$ and $p_k > 0$;
for $k = 0, 1, 2, \ldots$ **do**
  The server: samples $\xi_k \sim \mathcal{D}$;
  The selected client: sets $g_k = \nabla f_{\xi_k}(x_k) + (1 - p_k)\left(g_{k-1} - \nabla f_{\xi_k}(x_{k-1})\right)$;
  The selected client: sets x_{k+1} as
  $\begin{cases} \mathrm{prox}_{\gamma_k f_{\xi_k}}\left(x_k + \gamma_k(\nabla f_{\xi_k}(x_k) - g_k)\right); & \text{(SPAM)} \\ \text{a-prox}_{\epsilon}\left(x_k, g_k, \gamma_k, \xi_k\right); & \text{(SPAM-inexact)} \end{cases}$
  The selected client: sends x_{k+1}, g_k to the server.
end for

The following proposition is the cornerstone of our analysis. It provides a recurrent bound for a certain sequence V_k, which serves as a Lyapunov function:

$$V_k = f(x_k) - f_{\inf} + \frac{15\gamma_k}{16(2p_k - p_k^2)}\|g_k - \nabla f(x_k)\|^2. \tag{8}$$

Proposition 3.1. *Let x_k be the iterates of SPAM for an objective function f, which satisfies Assumptions 1 and 2. If $\gamma_k^2 \leq \min\left\{\frac{1}{16\delta^2}, \frac{p_k}{96\delta^2(1-p_k)}\right\}$, then for every $k \geq 1$*

$$\mathrm{E}\left[V_{k+1}\right] \quad \leq \quad \mathrm{E}\left[V_k\right] - \frac{\gamma_k}{32}\mathrm{E}\left[\|\nabla f(x_{k+1})\|^2\right] + 2\gamma_k p_k \sigma^2,$$

where V_k is defined in (8).

The proof can be found in Appendix C.1. This leads us to a convergence result for SPAM with fixed parameters.

Theorem 3.2 (SPAM with constant parameters)**.** *Suppose Assumptions 1, 2 are satisfied. Then,*

$$\frac{1}{K}\sum_{k=1}^{K}\mathrm{E}\left[\left\|\nabla f(x_k)\right\|^2\right] \leq \frac{32(f(x_0)-f_{\inf})}{\gamma K} + \frac{32\left\|g_0-\nabla f(x_0)\right\|^2}{(2p-p^2)K} + 64p\sigma^2, \tag{9}$$

where $\gamma^2 \leq \min\left\{\frac{1}{16\delta^2}, \frac{p}{96\delta^2(1-p)}\right\}$.

The proof of the theorem can be found in Appendix C.2.

Corollary 3.3. *The result can also be written as*

$$\mathrm{E}\left[\left\|\nabla f(\tilde{x}_{K+1})\right\|^2\right] \leq \frac{32(f(x_0)-f_{\inf})}{\gamma K} + \frac{32\left\|g_0-\nabla f(x_0)\right\|^2}{(2p-p^2)K} + 64p\sigma^2,$$

where $\tilde{x}_{K+1}$ *is taken uniformly randomly from the iterates of the algorithm* $\{x_1, x_2, \ldots, x_{K+1}\}$.

Our primary focus is communication complexity, which is typically the main bottleneck in cross-device federated settings (Kairouz et al., 2021). Below, we present the communication complexity of SPAM with fixed parameters.

Corollary 3.4. *Define* $F := f(x_0) - f_{\inf}$. *Let* $\gamma_k = \gamma = \min\left(\frac{1}{\delta}, \left(\frac{F}{2\delta^2\sigma^2 K}\right)^{1/3}\right)$, *and* $p_k = p = \max(\gamma^2\delta^2, 1/K)$. *Then, the communication complexity of SPAM, to obtain* ε *error is of order* $\mathcal{O}\left(\frac{\delta F+\sigma^2}{\varepsilon} + \frac{\delta\sigma F}{\varepsilon^{3/2}}\right)$.

The proof is deferred to Appendix E.1. Our result indicates that higher similarity (smaller δ) leads to fewer communication rounds to solve (1).

Suppose now that we can initialize $g_0 = \nabla f(x_0)$. Then, the second term in the convergence upper bound (9) vanishes. Repeating the exact steps as in the proof of Corollary 3.4, we obtain the convergence rate: $\mathcal{O}\left(\frac{\delta F}{K} + \left(\frac{\delta\sigma F}{K}\right)^{2/3}\right)$, which leads to a communication complexity of $\mathcal{O}\left(\frac{\delta F}{\varepsilon} + \frac{\delta\sigma F}{\varepsilon^{3/2}}\right)$. Thus, our result shows that in the homogeneous case (i.e., $\delta = 0$), communication is not needed at all, as each client can solve the problem locally.

It is important to highlight that the stepsize γ in SPAM differs from the stepsize used in local methods such as MIME and CE-LSGD. In these methods, the stepsize is intended to run the algorithms locally on a selected client. However, SPAM only requires an oracle for proximal points, allowing the oracle to use any optimization method suitable for the problem at hand. Additionally, the stepsize for local SGD-based methods depends on the smoothness parameter, which is not required in our theorem. Thus, our approach allows much more flexibility for choosing local solvers that are adaptive to the curvature of the loss (Malitsky & Mishchenko, 2020; Mishkin et al., 2024). See Table 1 for a detailed comparison of the methods.

In (9), we notice that the last term, which is due to the stochastic nature of our problem, does not vanish when K is large. To remove the stationarity neighborhood, let us now consider varying stepsizes for SPAM, with decaying momentum parameters p_k.

Theorem 3.5 (SPAM)**.** *Consider SPAM for an objective function* f *that satisfies Assumptions 1 and 2. Let* γ_k *be a sequence of varying stepsizes satisfying* $\gamma_k^2 \leq \frac{1}{16\delta^2}$ *and choose* $p_k = \frac{96\delta^2\gamma_k^2}{96\delta^2\gamma_k^2+1}$. *Denote* $\Gamma_K = \sum_{k=1}^{K}\gamma_k$, *then*

$$\frac{1}{\Gamma_K}\sum_{k=1}^{K}\gamma_k\mathrm{E}\left[\left\|\nabla f(x_k)\right\|^2\right] \leq \frac{2}{\Gamma_K}\sum_{k=1}^{K}\frac{96\delta^2\gamma_k^3}{96\delta^2\gamma_k^2+1}\sigma^2 + \frac{32V_0}{\Gamma_K}, \tag{10}$$

The proof of Theorem 3.5 can be found in Appendix C.3.

Remark 3.1. *Similar to Theorem 3.2, we can represent the left-hand side of* (10) *with a single expectation:* $\mathrm{E}\left[\left\|\nabla f(\tilde{x}_K)\right\|^2\right]$, *where* $\tilde{x}_K = x_i$, *for* $i = 1, \ldots, K$ *with probability* γ_i/Γ_K.

To ensure that the right-hand side converges to zero as $K \to \infty$, we need to choose a sequence $\gamma_K \to 0$ such that $\Gamma_K \to +\infty$. This suggests using a stepsize schedule of order $\gamma_k = O(k^{\beta-1})$, implying $\Gamma_K = O(K^\beta)$ for some $\beta \in (0,1)$. Consequently, the right-hand side of (10) is of order $O(K^{-\beta} + K^{2\beta-2})$. By optimizing over β, we deduce that $\gamma_k = O(k^{-1/3})$ results in a stationarity bound of order $O(K^{-2/3})$.

Corollary 3.6 (Optimal stepsize schedule)**.** *If* $\gamma_k = \frac{1}{4\delta k^{1/3}}$ *and* $p_k = \frac{96\delta^2\gamma_k^2}{96\delta^2\gamma_k^2+1}$, *then to obtain* ε*-stationarity for SPAM we need* $K = O(\varepsilon^{-3/2})$ *iterations under assumptions 1, 2.*

4. Inexact proximal operator

In the previous theorems, we assume that each sampled client ξ_k can exactly compute the proximal operator to obtain the new iterate x_{k+1}. The latter means that this client can exactly solve a (potentially) non-convex minimization problem, which might be problematic in practice. However, in the proofs of these theorems, we do not use that the new iterate x_{k+1} is the exact solution of the proximal operator (see Appendix G.1). Instead, we use two properties of the proximal point operator:

- decrease in function value: $\phi_k(x_{k+1}) \leq \phi_k(x_k)$;
- stationarity: $\nabla\phi_k(x_{k+1}) = 0$.

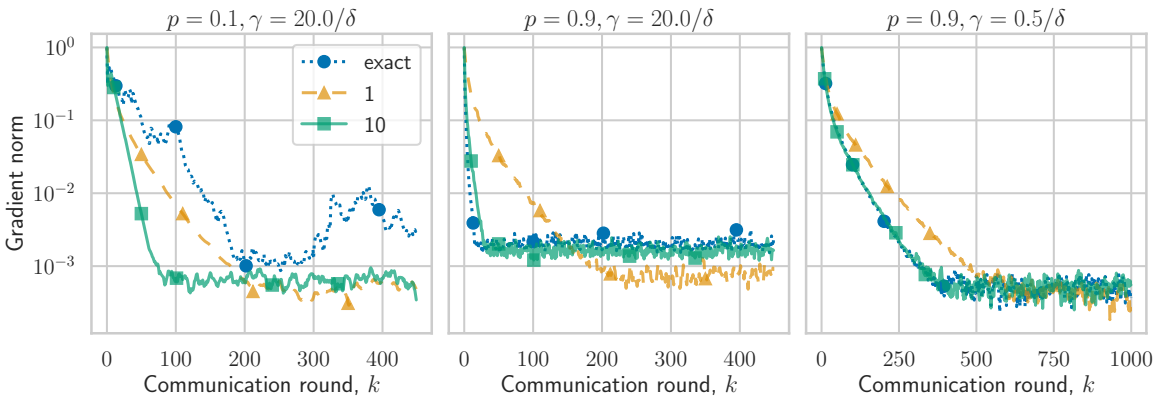

Figure 1: Convergence of SPAM-inexact on problem (27) with different p and γ.

Thus, we can replace the computation of the exact proximal point in Algorithm 1 with finding a point that satisfies the above two conditions. Furthermore, we will relax the second condition by taking an approximate stationary point. These arguments are summarized in the below assumption.

Definition 4.1 (a-prox)**.** For a given client k, a gradient estimator g_k, a current state x_k, a stepsize γ_k and a precision level ϵ, the approximate proximal point $\text{a-prox}_\epsilon(x_k, g_k, \gamma_k, k)$ is the set of vectors y_{ap}, which satisfy

- decrease in function value: $\mathrm{E}\left[\phi_k(y_{\mathsf{ap}})\right] \leq \phi_k(x^k)$,
- approximate stationarity: $\mathrm{E}\left[\left\|\nabla \phi_k(y_{\mathsf{ap}})\right\|^2\right] \leq \epsilon^2$,

where ϕ_k is defined in (7). The following theorem describes the convergence result of SPAM-inexact, whose pseudocode is described in Algorithm 1.

Theorem 4.1 (SPAM-inexact)**.** *Consider* SPAM-inexact *for an objective function f that satisfies Assumptions 1 and 2. Let γ_k be a sequence of varying stepsizes satisfying $\gamma_k^2 \leq \frac{1}{16\delta^2}$ and choose $p_k = \frac{96\delta^2\gamma_k^2}{96\delta^2\gamma_k^2+1}$. Denote $\Gamma_K = \sum_{k=1}^K \gamma_k$, then*

$$\sum_{k=1}^K \frac{\gamma_k \mathrm{E}\left[\left\|\nabla f(x_{k+1})\right\|^2\right]}{\Gamma_K} \leq \frac{40V_0}{\Gamma_K} + \frac{\epsilon^2}{8} + \sum_{k=1}^K \frac{2p_k\gamma_k^2\sigma^2}{\Gamma_K}.$$

The proof is postponed to Appendix C.4. We observe that the level of inexactness ϵ^2 appears explicitly in the theorem. In case when $\epsilon = 0$, we recover the result in Theorem 3.5 up to constants. SPAM-inexact allows to avoid solving the local minimization problem required for finding the inexact proximal point operator. This is a significant improvement over SPAM, as the latter requires minimizing (potentially) non-convex objectives at each iteration.

To compute a-prox, we can use gradient descent-type methods. This is feasible because ϕ_k is differentiable and convex when f_ξ is L-smooth and $\gamma_k \leq 1/L$. These properties allow us to apply efficient optimization techniques that achieve better convergence rates than the standard $\mathcal{O}(1/T)$, where T represents the number of iterations for the inner method. Instead we can achieve $\mathcal{O}(1/T^4)$ using more elaborate techniques (Nesterov et al., 2021; Sadiev et al., 2022). In this paper, however, we are more concerned with communication complexity rather than oracle complexity, and thus, we will not discuss this further in the paper.

5. Conclusion

We introduced SPAM, an algorithm tailored for cross-device federated learning, which combines momentum variance reduction with the stochastic proximal point method. Operating under second-order heterogeneity and bounded variance conditions, SPAM does not necessitate smoothness of the objective function. In its most general form, SPAM achieves faster communication complexity than its competitors. Furthermore, it does not prescribe a specific local method for analysis, providing practitioners with flexibility and responsibility in selecting suitable local solver.

Limitations and future work. The paper is of theoretical nature and focuses on improving the understanding of stochastic non-convex optimization under Hessian similarity in the context of cross-device federated learning. We believe that separate experiments should be conducted to evaluate the experimental performance in a setting close to real life.

In standard optimization, the stepsize usually depends on the smoothness parameter. Adaptive methods allow iterative adjustment of the stepsize without additional information. In our case, the smoothness parameter is replaced by the second-order heterogeneity parameter δ, on which the stepsize and momentum sequences of SPAM depend. Removing this dependence using adaptive techniques under general assumptions remains an open problem even for the server-only MVR, which serves as the basis for our algorithm.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Acar, D. A. E., Zhao, Y., Matas, R., Mattina, M., Whatmough, P., and Saligrama, V. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2020. (Cited on pages 2 and 3)

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019. (Cited on page 2)

Beck, A. *First-order methods in optimization*. SIAM, 2017. (Cited on page 3)

Beznosikov, A., Horváth, S., Richtárik, P., and Safaryan, M. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020. (Cited on page 1)

Caldas, S., Konečny, J., McMahan, H. B., and Talwalkar, A. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*, 2018. (Cited on page 1)

Crawshaw, M., Liu, M., Orabona, F., Zhang, W., and Zhuang, Z. Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35:9955–9968, 2022. (Cited on page 4)

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. *Advances in Neural Information Processing Systems*, 32, 2019. (Cited on page 2)

Hendrikx, H., Xiao, L., Bubeck, S., Bach, F., and Massoulie, L. Statistically preconditioned accelerated gradient method for distributed optimization. In *International conference on machine learning*, pp. 4203–4227. PMLR, 2020. (Cited on page 4)

Horváth, S., Sanjabi, M., Xiao, L., Richtárik, P., and Rabbat, M. Fedshuffle: Recipes for better use of local work in federated learning. *Transactions on Machine Learning Research*, 2022. (Cited on page 13)

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K. A., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2):1–210, 2021. doi: 10.1561/2200000083. URL https://doi.org/10.1561/2200000083. (Cited on pages 1, 2, and 5)

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020. (Cited on pages 2 and 3)

Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S., Stich, S. U., and Suresh, A. T. Breaking the centralized barrier for cross-device federated learning. *Advances in Neural Information Processing Systems*, 34:28663–28676, 2021. (Cited on pages 1, 2, 13, and 17)

Khaled, A. and Jin, C. Faster federated optimization under second-order similarity. In *The Eleventh International Conference on Learning Representations*, 2022. (Cited on pages 2, 4, and 12)

Khaled, A., Mishchenko, K., and Richtárik, P. Tighter theory for local SGD on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4519–4529. PMLR, 2020. (Cited on page 12)

Kim, D. and Fessler, J. A. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *Journal of optimization theory and applications*, 188(1):192–219, 2021. (Cited on page 27)

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (Cited on page 2)

Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *NIPS Private Multi-Party Machine Learning Workshop*, 2016. (Cited on page 1)

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020. (Cited on page 2)

Li, Z., Bao, H., Zhang, X., and Richtárik, P. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pp. 6286–6295. PMLR, 2021. (Cited on page 2)

Lin, D., Han, Y., Ye, H., and Zhang, Z. Stochastic distributed optimization under average second-order similarity: Algorithms and analysis. *Advances in Neural Information Processing Systems*, 36, 2024. (Cited on pages 2 and 27)

Mairal, J. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pp. 783–791. PMLR, 2013. (Cited on page 4)

Mairal, J. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015. (Cited on page 4)

Malitsky, Y. and Mishchenko, K. Adaptive gradient descent without descent. In *International Conference on Machine Learning*, pp. 6702–6712. PMLR, 2020. (Cited on page 5)

Mangasarian, O. L. and Solodov, M. V. Backpropagation convergence via deterministic nonmonotone perturbed minimization. *Advances in Neural Information Processing Systems*, 6, 1993. (Cited on page 2)

McDonald, R., Hall, K., and Mann, G. Distributed training strategies for the structured perceptron. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 456–464, 2010. (Cited on page 2)

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017. (Cited on pages 1 and 2)

Mishchenko, K., Malinovsky, G., Stich, S., and Richtárik, P. ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally! In *International Conference on Machine Learning*, pp. 15750–15769. PMLR, 2022. (Cited on pages 2 and 3)

Mishchenko, K., Li, R., Fan, H., and Venieris, S. Federated learning under second-order data heterogeneity. *Openreview,* https://openreview.net/forum?id=jkhVrIllKg, 2023. (Cited on pages 2 and 14)

Mishkin, A., Khaled, A., Wang, Y., Defazio, A., and Gower, R. M. Directional smoothness and gradient methods: Convergence and adaptivity. *arXiv preprint arXiv:2403.04081*, 2024. (Cited on page 5)

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013. (Cited on page 27)

Nesterov, Y., Gasnikov, A., Guminov, S., and Dvurechensky, P. Primal–dual accelerated gradient methods with small-dimensional relaxation oracle. *Optimization Methods and Software*, 36(4):773–810, 2021. (Cited on page 6)

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pp. 2613–2621. PMLR, 2017. (Cited on page 2)

Ogier du Terrail, J., Ayed, S.-S., Cyffers, E., Grimberg, F., He, C., Loeb, R., Mangold, P., Marchand, T., Marfoq, O., Mushtaq, E., et al. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *Advances in Neural Information Processing Systems*, 35:5315–5334, 2022. (Cited on page 1)

Patel, K. K., Wang, L., Woodworth, B. E., Bullins, B., and Srebro, N. Towards optimal communication complexity in distributed non-convex optimization. *Advances in Neural Information Processing Systems*, 35:13316–13328, 2022. (Cited on pages 2, 11, 12, 13, and 27)

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations*, 2020. (Cited on page 2)

Sadiev, A., Kovalev, D., and Richtárik, P. Communication acceleration of local gradient methods via an accelerated primal-dual algorithm with an inexact prox. *Advances in Neural Information Processing Systems*, 35:21777–21791, 2022. (Cited on page 6)

Shamir, O., Srebro, N., and Zhang, T. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pp. 1000–1008. PMLR, 2014. (Cited on page 4)

Shulgin, E. and Richtárik, P. Shifted compression framework: Generalizations and improvements. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. (Cited on page 1)

Silva, A., Metcalf, K., Apostoloff, N., and Theobald, B.-J. Fedembed: Personalized private federated learning. *arXiv preprint arXiv:2202.09472*, 2022. (Cited on page 2)

Szlendak, R., Tyurin, A., and Richtárik, P. Permutation compressors for provably faster distributed nonconvex optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=GugZ5DzzAu. (Cited on page 4)

Tyurin, A., Sun, L., Burlachenko, K., and Richtárik, P. Sharper rates and flexible framework for nonconvex SGD with client and data sampling. *Transactions on Machine Learning Research*, 2023. (Cited on page 4)

Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021. (Cited on page 1)

Woodworth, B., Mishchenko, K., and Bach, F. Two losses are better than one: Faster optimization using a cheaper proxy. In *International Conference on Machine Learning*, pp. 37273–37292. PMLR, 2023. (Cited on page 4)

Xu, H., Ho, C.-Y., Abdelmoniem, A. M., Dutta, A., Bergou, E. H., Karatsenidis, K., Canini, M., and Kalnis, P. Compressed communication for distributed deep learning: Survey and quantitative evaluation. *Technical report,* `http://hdl.handle.net/10754/662495`, 2020. URL `http://hdl.handle.net/10754/662495`. (Cited on page 1)

Yuan, X. and Li, P. On convergence of FedProx: Local dissimilarity invariant bounds, non-smoothness and beyond. *Advances in Neural Information Processing Systems*, 35: 10752–10765, 2022. (Cited on page 2)

Zhang, J., He, T., Sra, S., and Jadbabaie, A. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations (ICLR)*, 2019. (Cited on page 4)

Contents

A. Partial participation

Algorithm 2 SPAM-PP

Input: learning rate $\gamma > 0$, cohort size B, starting point $x_0 \in \mathbb{R}^d$; proximal precision level ϵ; initialize $g_0 = g_{-1}$;
for $k = 0, 1, 2, \ldots$ **do**
  The server: samples a subset of clients S_k, with size $|S_k| = B$;
  The server: broadcasts x_k to the clients from S_k.
  for $i \in S_k$ in parallel **do**
    The selected clients: set $g_k^i = \nabla f_i(x_k) + (1 - p_k)\left(g_{k-1} - \nabla f_i(x_{k-1})\right)$;
    The selected clients: send g_k^i to the server;
  end for
  The server: aggregates $g_k = \frac{1}{B}\sum_{i \in S_k} g_k^i$;
  The server: samples $\xi^{k+1} \sim \mathcal{D}$;
  The selected client: computes $x_{k+1} \in \text{a-prox}_{\epsilon}\left(x_k, g_k, \gamma_k, \xi^{k+1}\right)$;
end for

In this section, we present the most general form of our algorithm, which works with the approximate proximal operator and samples multiple clients (cohort) at each round. Specifically, it uses the random cohort S_k to construct a better gradient estimator g_k. This gradient estimator is then broadcasted to a single random client $\xi_k \sim \mathcal{D}$, who locally computes the approximate proximal point. The pseudocode can be found in Algorithm 2.

Theorem A.1 (SPAM-PP)**.** *Suppose Assumptions 1 and 2 are satisfied. If $\xi_k \sim \mathsf{Unif}(S_k)$ at every iteration, then the iterates of SPAM-PP with $\gamma_k \leq \frac{1}{4\delta}$ and $p_k = \frac{96\delta^2\gamma_k^2}{96\delta^2\gamma_k^2 + B^2}$ satisfy*

$$\frac{1}{\Gamma_K}\sum_{k=0}^{K-1}\gamma_k \mathrm{E}\left[\|\nabla f(x_{k+1})\|^2\right] \leq \frac{40}{\Gamma_K}\left(V_0 - \mathrm{E}\left[V_K\right]\right) + \frac{240}{\Gamma_K}\sum_{k=0}^{K-1} p_k\gamma_k\frac{\sigma^2}{B} + 7.5\epsilon^2.$$

The proof of the theorem is postponed to Appendix D. When the client cohort size B increases, the neighborhood shrinks. This is intuitive as when $B \to \infty$, we can access the exact objective f, and the neighborhood will vanish.

Corollary A.2. *For a properly chosen constant $\gamma_k = \gamma$ and a momentum parameter $p_k = p$, the convergence rate of exact SPAM-PP is*

$$\mathcal{O}\left(\frac{\delta F}{K} + \frac{\sigma^2}{BK} + \frac{1}{B}\left(\frac{\delta F\sigma}{K}\right)^{2/3}\right), \tag{11}$$

corresponding to communication complexity of order $\mathcal{O}\left(\frac{\delta F}{\varepsilon} + \frac{\sigma^2}{B\varepsilon} + \frac{\delta F\sigma}{(B\varepsilon)^{3/2}}\right)$.

Our result improves upon the prior best rate

$$\mathcal{O}\left(\frac{(\delta + L)F}{K} + \frac{\sigma^2}{BK} + \frac{1}{B^{2/3}}\left(\frac{\delta F\sigma}{K}\right)^{2/3}\right)$$

for cross-device FL with partial participation obtained by Patel et al. (2022) for the CE-LGD method. Firstly, our rate does not have the term LF/K, where L is the smoothness parameter, which is not a requirement for SPAM-PP. Secondly, the third term in bound (11) has a better dependence on the cohort size B. Nevertheless, (11) requires exact proximal point computation.

In Appendix F, we present another version of SPAM-PP, called SPAM-PPA, which uses the sampled cohort of clients to compute local proximal points. These points are then communicated to the server, and the new iterate is their average. Hence, the name SPAM-PP with Averaging.

B. Experiments

To empirically validate our theoretical framework and its implications, we focus on a carefully controlled experimental setting similar to (Khaled & Jin, 2022). Specifically, we consider a distributed ridge regression problem formulated in (27), which allows us to calculate and control the Hessian similarity δ. An essential advantage of this optimization problem is that the proximal operator has an explicit (closed-form) representation and can be computed precisely (up to machine accuracy). This allows us to isolate the effect of varying parameters on the method's performance. Appendix H provides a larger discussion on experiments.

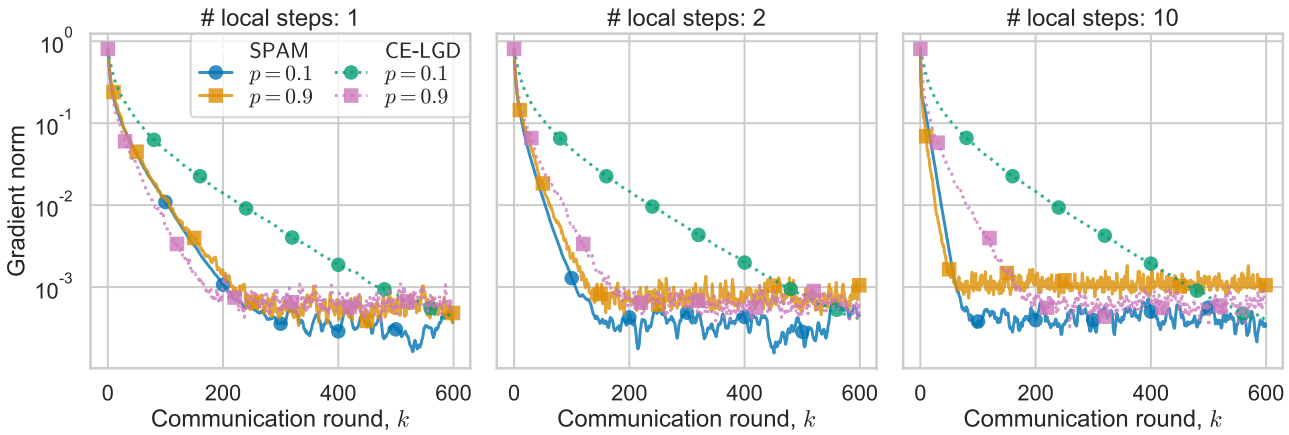

Figure 2: Comparison of SPAM-inexact ($\gamma = 5/\delta$) and CE-LGD on problem (27) with different p and number of local steps.

SPAM-inexact study. In Figure 1, we display convergence of Algorithm 1 with constant parameters p and γ. The legend is shared, and labels refer to proximal operator computations: "exact" means using closed-form solution, "1" and "10" correspond to the number of local gradient descent steps. We evaluate the logarithm of a relative gradient norm $\log(\|\nabla f(x_k)\|/\|\nabla f(x_0)\|)$ in the vertical axis.

All the plots indicate convergence of the method to the neighborhood of the stationary point, followed by subsequent oscillations around the error floor. The first (left) plot shows that for small momentum $p = 0.1$ and γ exceeding the theoretical bound $1/\delta$, the algorithm can be very unstable with exact proximal point computations. Interestingly, approximate computation (1 or 10 local steps) results in more robust convergence. The second (middle) plot demonstrates that a greater $p = 0.9$ results in steady convergence even for misspecified (too large) γ. In addition, one can observe that in this case, more accurate proximal point evaluation results in significantly faster convergence but to a larger neighborhood than for one local step. This agrees well with observations for local gradient descent methods (Khaled et al., 2020). The last (right) figure shows that a properly chosen, smaller $\gamma = 0.5/\delta$ slows down convergence (twice as many communication rounds are shown). However, the method reaches a significantly lower error floor (as the vertical axis is shared across plots), which does not depend much on the accuracy of proximal point operator calculation. Moreover, 10 local steps are enough for basically the same fast convergence as with exact proximal point computation.

SPAM and CE-LGD comparison. As a reminder, CE-LGD (Patel et al., 2022) leverages MVR on the server (similarly to SPAM) but applies SARAH locally on the selected client. Note that the theory by Patel et al. (2022) requires decreasing the step size as the number of local steps (denoted by τ) increases. In contrast, our Algorithm 1 computes approximate proximal point operator by running a simple gradient descent method for the same number of τ local steps.

Figure 3 illustrates the convergence behavior of the methods towards a neighborhood of the stationary point. The vertical and horizontal axes are shared across all plots. We vary the momentum parameter $p \in \{0.1, 0.9\}$ (within each subplot), the number of local steps: $\{1, 2, 10\}$ (across columns). Appendix I also presents a study of the algorithm for varied parameter γ.

The convergence speed of CE-LGD is more sensitive to the choice of momentum parameter p: a small $p = 0.1$ significantly slows convergence, while $p = 0.9$ accelerates it. In contrast, SPAM's speed shows minimal sensitivity to variations in p. The size of the convergence neighborhood for both methods is primarily influenced by the value of p, which is especially evident in the final plot for SPAM, where a larger p results in larger gradient norm oscillations.

Varying the number of local steps has little impact on the performance of CE-LGD, as the theory of Patel et al. (2022) mandates that the local step size be inversely proportional to τ. Meanwhile, SPAM benefits from more precise local subproblem solutions: increasing the number of local steps leads to faster convergence, which aligns with standard practices in federated learning.

We want to note that momentum-based variance reduction has already shown empirical success (Karimireddy et al., 2021; Horváth et al., 2022) in practical federated learning scenarios. That is why our experiments focus on simpler but insightful settings to study the properties of the proposed algorithm carefully.

C. Convergence analysis for SPAM

C.1. Proof of Proposition 3.1

Recall that

$$V_k = f(x_k) - f_{\inf} + \frac{3\gamma_k}{2p_k - p_k^2}\|g_k - \nabla f(x_k)\|^2.$$

We bound each term separately. We formulate three technical lemmas that are inspired by (Mishchenko et al., 2023). Their proofs can be found in Appendix G. We start by bounding the first term of the Lyapunov function, which is the function values.

Lemma C.1. *Under the conditions of Proposition 3.1, the following recurrent inequality takes place*

$$f(x_{k+1}) - f_{\inf} \leq f(x_k) - f_{\inf} - \frac{1}{4\gamma_k}\|x_{k+1} - x_k\|^2 + 2\gamma_k\|\nabla f(x_k) - g_k\|^2 \tag{12}$$

Then, we bound the second term of V_k.

Lemma C.2. *Under the conditions of Proposition 3.1, the following recurrent inequality takes place*

$$\mathrm{E}\left[\|g_{k+1} - \nabla f(x_{k+1})\|^2 \middle| \mathcal{F}_k\right] \leq (1-p_k)^2\|g_k - \nabla f(x_k)\|^2 + 2(1-p_k)^2\delta^2\|x_{k+1} - x_k\|^2 + 2p_k^2\sigma^2. \tag{13}$$

We observe that the term $\|x_{k+1} - x_k\|^2$ is in both upper bounds. The following lemma provides a lower bound for this expression.

Lemma C.3. *Under the conditions of Proposition 3.1, the following recurrent inequality is true*

$$\mathrm{E}\left[\|x_{k+1} - x_k\|^2\right] \geq \frac{\gamma_k^2}{4}\mathrm{E}\left[\|\nabla f(x_{k+1})\|^2\right] - \gamma_k^2\mathrm{E}\left[\|g_k - \nabla f(x_k)\|^2\right]. \tag{14}$$

We now combine the results of the lemmas to bound V_{K+1}:

$$\begin{aligned}
\mathrm{E}\left[V_{k+1}\right] &\overset{(12)+(13)}{\leq} \alpha(1-p_k)^2\|g_k - \nabla f(x_k)\|^2 + 2\alpha\delta^2(1-p_k)^2\|x_{k+1} - x_k\|^2 + 2\alpha p_k^2\sigma^2 \\
&\quad + \mathrm{E}\left[f(x_k) - f_{\inf}\right] - \frac{1}{4\gamma_k}\mathrm{E}\left[\|x_{k+1} - x_k\|^2\right] + 2\gamma_k\mathrm{E}\left[\|\nabla f(x_k) - g_k\|^2\right] \\
&= \mathrm{E}\left[V_k\right] + \left(2\alpha\delta^2(1-p_k)^2 - \frac{1}{4\gamma_k}\right)\mathrm{E}\left[\|x_{k+1} - x_k\|^2\right] + 2\alpha p_k^2\sigma^2 \\
&\quad + (2\gamma_k - \alpha(2p_k - p_k^2))\mathrm{E}\left[\|\nabla f(x_k) - g_k\|^2\right].
\end{aligned}$$

The last inequality is true for every positive value of α. Let us now choose $\alpha = \frac{3\gamma_k}{2p_k - p_k^2}$. Then,

$$2\alpha\delta^2(1-p_k)^2 - \frac{1}{4\gamma_k} = \frac{6\gamma_k\delta^2(1-p_k)^2}{2p_k - p_k^2} - \frac{1}{4\gamma_k} \leq -\frac{1}{8\gamma_k},$$

where the latter is due to $4\delta\gamma_k \leq \sqrt{p_k/6(1-p_k)}$. Therefore, we deduce

$$\begin{aligned}
\mathrm{E}\left[V_{k+1}\right] &\leq \mathrm{E}\left[V_k\right] - \frac{1}{8\gamma_k}\mathrm{E}\left[\|x_{k+1} - x_k\|^2\right] - \gamma_k\mathrm{E}\left[\|\nabla f(x_k) - g_k\|^2\right] + 2\alpha p_k^2\sigma^2 \\
&\overset{(14)}{\leq} \mathrm{E}\left[V_k\right] - \frac{\gamma_k}{32}\mathrm{E}\left[\|\nabla f(x_{k+1})\|^2\right] + \frac{\gamma_k}{8}\mathrm{E}\left[\|\nabla f(x_k) - g_k\|^2\right] \\
&\quad - \gamma_k\mathrm{E}\left[\|\nabla f(x_k) - g_k\|^2\right] + \frac{6\gamma_k p_k}{2 - p_k}\sigma^2 \\
&\leq \mathrm{E}\left[V_k\right] - \frac{\gamma_k}{32}\mathrm{E}\left[\|\nabla f(x_{k+1})\|^2\right] + 6\gamma_k p_k\sigma^2.
\end{aligned}$$

This concludes the proof of the proposition.

C.2. Proof of Theorem 3.2

Let us apply Proposition 3.1 for the fixed stepsize $\gamma_k = \gamma$ and a fixed momentum coefficient $p_k = p$.

$$\mathrm{E}\left[V_{k+1}\right] \quad \leq \quad \mathrm{E}\left[V_k\right] - \frac{\gamma}{32}\mathrm{E}\left[\left\|\nabla f(x_{k+1})\right\|^2\right] + 6\gamma p\sigma^2.$$

Summing up these inequalities for $k = 0, \ldots, K-1$ leads to

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathrm{E}\left[\left\|\nabla f(x_k)\right\|^2\right] &\leq \frac{32}{\gamma K}\left(V_0 - \mathrm{E}\left[V_K\right]\right) + 192p\sigma^2 \\ &\leq \frac{32(f(x_0) - f_{\inf})}{\gamma K} + \frac{30\left\|g_0 - \nabla f(x_0)\right\|^2}{(2p-p^2)K} + 192p\sigma^2.\end{aligned}$$

where $\gamma^2 \leq \min\left\{\frac{1}{16\delta^2}, \frac{4p}{3\delta^2(1-p)}\right\}$. This concludes the proof of the theorem.

C.3. Proof of Theorem 3.5

From Proposition 3.1 we have

$$-\frac{\gamma_k}{32}\mathrm{E}\left[\left\|\nabla f(x_{k+1})\right\|^2\right] \quad \leq \quad \mathrm{E}\left[V_k\right] - \mathrm{E}\left[V_{k+1}\right] + 6\gamma_k p_k\sigma^2.$$

Let us sum up these inequalities for $k = 0, 1, \ldots, K-1$. We have a telescoping sum on the right-hand side. Then, dividing both sides on $\Gamma_K = \sum_{i=1}^{K}\gamma_i$, we deduce the following bound:

$$\frac{1}{\Gamma_K}\sum_{k=1}^{K}\gamma_k\mathrm{E}\left[\left\|\nabla f(x_k)\right\|^2\right] \leq \frac{32V_0}{\Gamma_K} + \frac{2}{\Gamma_K}\sum_{k=1}^{K}\frac{15\delta^2\gamma_k^3}{15\delta^2\gamma_k^2+4}\sigma^2.$$

This concludes the proof.

C.4. Proof of Theorem 4.1

We start by repeating the steps of the proof for Proposition 3.1. Notice that the proposition statement assumes that the iterate is exactly equal to the proximal point operator. However, as stated in Section 4, in the proofs of lemmas C.1 and C.2 we only use the property that $\phi_k(x_{k+1}) \leq \phi_k(x_k)$ (see (23)). Thus, both (12) and (13) are true for SPAM-inexact. Therefore,

$$\begin{aligned}\mathrm{E}\left[V_{k+1}\right] \quad \leq \quad \mathrm{E}\left[V_k\right] \quad &-\frac{1}{8\gamma_k}\mathrm{E}\left[\left\|x_{k+1}-x_k\right\|^2\right] - (2\gamma_k - \alpha(2p_k - p_k^2))\mathrm{E}\left[\left\|\nabla f(x_k) - g_k\right\|^2\right] \\ &+2\alpha p_k^2\sigma^2.\end{aligned}$$

Below, reformulate the adaptation of Lemma C.3 for the inexact case to lower bound the second term on the right-hand side.

Lemma C.4. *Under the conditions of Proposition 3.1, we have the following bound*

$$\mathrm{E}\left[\left\|x_{k+1}-x_k\right\|^2\right] \geq \frac{\gamma_k^2}{5}\mathrm{E}\left[\left\|\nabla f(x_{k+1})\right\|^2\right] - \gamma_k^2\mathrm{E}\left[\left\|g_k - \nabla f(x_k)\right\|^2\right] - \gamma_k^2\epsilon^2. \tag{15}$$

The proof can be found in Appendix G.4. Thus,

$$\begin{aligned}\mathrm{E}\left[V_{k+1}\right] &\overset{(15)}{\leq} \mathrm{E}\left[V_k\right] - \frac{\gamma_k}{40}\mathrm{E}\left[\left\|\nabla f(x_{k+1})\right\|^2\right] + \frac{15\gamma_k p_k}{8(2-p_k)}\sigma^2 + \frac{\gamma_k\epsilon^2}{8} \\ &\quad + \frac{\gamma_k}{8}\mathrm{E}\left[\left\|\nabla f(x_k) - g_k\right\|^2\right] - (2\gamma_k - \alpha(2p_k - p_k^2))\mathrm{E}\left[\left\|\nabla f(x_k) - g_k\right\|^2\right] \\ &\leq \mathrm{E}\left[V_k\right] - \frac{\gamma_k}{40}\mathrm{E}\left[\left\|\nabla f(x_{k+1})\right\|^2\right] + 2\gamma_k p_k\sigma^2 + \frac{\gamma_k\epsilon^2}{8}.\end{aligned}$$

Repeating this step for $k = 0, \ldots, K-1$, we deduce

$$\frac{1}{\Gamma_K}\sum_{k=0}^{K-1}\gamma_k\mathrm{E}\left[\left\|\nabla f(x_{k+1})\right\|^2\right] \quad \leq \quad \frac{40V_0}{\Gamma_K} + \frac{2}{\Gamma_K}\sum_{k=0}^{K-1}\frac{15\sigma^2\gamma_k^3}{15\sigma^2\gamma_k^2+4}\sigma^2 + \frac{\epsilon^2}{8}.$$

D. Proof of Theorem A.1

The proof follows the logic of Proposition 3.1. Recall that

$$V_k = f(x_k) - f_{\inf} + \frac{3\gamma_k}{2p_k - p_k^2}\|g_k - \nabla f(x_k)\|^2.$$

Recall that Lemma C.1 is true for any gradient estimator g_k. Thus, (12) is also valid for SPAM-PP. Next, we estimate the second term of the Lyapunov function. Recall that

$$\begin{aligned} g_{k+1} &= \frac{1}{S_k}\sum_{i\in S_k}\left\{\nabla f_i(x_{k+1}) + (1-p_k)\left(g_k - \nabla f_i(x_k)\right)\right\} \\ &= \nabla \tilde{f}_k(x_{k+1}) + (1-p_k)\left(g_k - \nabla \tilde{f}_k(x_k)\right), \end{aligned}$$

where $\tilde{f}_k(x) := \frac{1}{S_k}\sum_{i\in S_k}\nabla f_i(x)$. Notice also that $\mathrm{E}\left[\tilde{f}_k(x)\right] = f(x)$, for every fixed $x \in \mathbb{R}^d$. Furthermore, combining the convexity of the Euclidean norm and Hessian similarity (5) we deduce that the estimator $\tilde{f}_k$ satisfies the Hessian similarity condition

$$\begin{aligned} \left\|\nabla \tilde{f}_k(x) - \nabla f(x) - \nabla \tilde{f}_k(y) + \nabla f(y)\right\| &\leq \frac{1}{B}\sum_{i\in S_k}\|\nabla f_i(x) - \nabla f(x) - \nabla f_i(y) + \nabla f(y)\| \\ &\leq \frac{\delta}{B}\|x - y\|. \end{aligned}$$

Finally, Jensen's inequality implies that $\tilde{f}_k$ satisfies the bounded variance condition as well:

$$\mathrm{E}\left[\left\|\nabla \tilde{f}_k(x) - \nabla f(x)\right\|\right] \leq \sigma^2/B.$$

Repeating the analysis exactly as in the proof of Lemma C.2, we obtain

$$\begin{aligned} \mathrm{E}\left[\|g_{k+1} - \nabla f(x_{k+1})\|^2\right] &\leq (1-p_k)^2\mathrm{E}\left[\|g_k - \nabla f(x_k)\|^2\right] \\ &\quad +2(1-p_k)^2\frac{\delta^2}{B^2}\mathrm{E}\left[\|x_{k+1} - x_k\|^2\right] + \frac{2p_k^2\sigma^2}{B}. \end{aligned} \tag{16}$$

Let us now bound the Lyapunov function using (12) and (16):

$$\begin{aligned} \mathrm{E}\left[V_{k+1}\right] &\leq \mathrm{E}\left[f(x_k) - f_{\inf}\right] + 2\gamma_k\mathrm{E}\left[\|\nabla f(x_k) - g_k\|^2\right] - \frac{1}{4\gamma_k}\mathrm{E}\left[\|x_{k+1} - x_k\|^2\right] \\ &\quad +\alpha(1-p_k)^2\mathrm{E}\left[\|g_k - \nabla f(x_k)\|^2\right] + 2\alpha(1-p_k)^2\frac{\delta^2}{B^2}\mathrm{E}\left[\|x_{k+1} - x_k\|^2\right] + \frac{2\alpha p_k^2\sigma^2}{B} \\ &= \mathrm{E}\left[V_k\right] + \left(2\alpha\frac{\delta^2}{B^2}(1-p_k)^2 - \frac{1}{4\gamma_k}\right)\mathrm{E}\left[\|x_{k+1} - x_k\|^2\right] + \frac{2\alpha p_k^2\sigma^2}{B} \\ &\quad +(2\gamma_k - \alpha(2p_k - p_k^2))\mathrm{E}\left[\|\nabla f(x_k) - g_k\|^2\right]. \end{aligned}$$

The latter is true for every positive α. Let us now plug in the value of $\alpha = \frac{3\gamma_k}{2p_k - p_k^2}$. Then, using $\gamma \leq \sqrt{\frac{B^2 p_k}{96\delta^2(1-p_k)}}$, we obtain

$$2\alpha\frac{\delta^2}{B^2}(1-p_k)^2 - \frac{1}{4\gamma_k} \leq \frac{6\gamma_k\delta^2}{B^2(2p_k - p_k^2)}(1-p_k)^2 - \frac{1}{4\gamma_k} \leq -\frac{1}{8\gamma_k}. \tag{17}$$

Hence, we have the following bound

$$\begin{aligned}
\mathrm{E}\left[V_{k+1}\right] &\leq \mathrm{E}\left[V_k\right] - \frac{1}{8\gamma_k}\mathrm{E}\left[\left\|x_{k+1}-x_k\right\|^2\right] - \gamma_k \mathrm{E}\left[\left\|\nabla f(x_k) - g_k\right\|^2\right] + \frac{6p_k\gamma_k\sigma^2}{B(2-p_k)} \\
&\overset{(15)}{\leq} \mathrm{E}\left[V_k\right] - \frac{1}{8\gamma_k}\left(\frac{\gamma_k^2}{5}\mathrm{E}\left[\left\|\nabla f(x_{k+1})\right\|^2\right] - \gamma_k^2\mathrm{E}\left[\left\|g_k - \nabla f(x_k)\right\|^2\right] - \gamma_k^2\epsilon^2\right) \\
&\quad -\gamma_k\mathrm{E}\left[\left\|\nabla f(x_k) - g_k\right\|^2\right] + \frac{6p_k\gamma_k\sigma^2}{B} \\
&\leq \mathrm{E}\left[V_k\right] - \frac{\gamma_k}{40}\mathrm{E}\left[\left\|\nabla f(x_{k+1})\right\|^2\right] - \frac{7\gamma_k}{8}\mathrm{E}\left[\left\|\nabla f(x_k) - g_k\right\|^2\right] + \frac{6p_k\gamma_k\sigma^2}{B} + \frac{\gamma_k\epsilon^2}{8} \\
&\leq \mathrm{E}\left[V_k\right] - \frac{\gamma_k}{40}\mathrm{E}\left[\left\|\nabla f(x_{k+1})\right\|^2\right] + \frac{6p_k\gamma_k\sigma^2}{B} + \frac{\gamma_k\epsilon^2}{8}.
\end{aligned}$$

Thus, we have

$$\frac{1}{\Gamma_K}\sum_{k=0}^{K-1}\gamma_k\mathrm{E}\left[\left\|\nabla f(x_{k+1})\right\|^2\right] \leq \frac{40}{\Gamma_K}\left(V_0 - \mathrm{E}\left[V_K\right]\right) + \frac{240}{\Gamma_K}\sum_{k=0}^{K-1}p_k\gamma_k\frac{\sigma^2}{B} + 7.5\epsilon^2.$$

This concludes the proof of the theorem.

E. Complexity analysis of the methods

We use $\lesssim$ to ignore numerical constants in the subsequent analysis.

E.1. Proof of Corollary 3.4

We have stepsize condition $\gamma \lesssim \min\{1/\delta, \sqrt{p/(\delta^2(1-p))}\}$, which implies that $\gamma \lesssim \sqrt{p}/\delta$ or $p \gtrsim (\gamma\delta)^2$. Denote $F := f(x_0) - f_{\inf}$, then convergence rate of SPAM can be expressed as

$$\begin{aligned}
R_K := \frac{1}{K}\sum_{k=1}^{K}\mathrm{E}\left[\left\|\nabla f(x_k)\right\|^2\right] &\lesssim \frac{f(x_0)-f_{\inf}}{\gamma K} + \frac{\left\|g_0 - \nabla f(x_0)\right\|^2}{(2p-p^2)K} + p\sigma^2 \\
&\lesssim \frac{F}{\gamma K} + \frac{\left\|g_0 - \nabla f(x_0)\right\|^2}{pK} + p\sigma^2,
\end{aligned}$$

where in the last inequality, we used the condition for the stepsize and the fact that $p(2-p) \geq p$. Next, by using an argument similar to that in (Karimireddy et al., 2021), we suppose (without loss of generality) that the method is run for K iterations. For the first $K/2$ iterations, we simply sample ∇f_ξ at x_0 to set $g_0 = \frac{1}{K/2}\sum_{i=1}^{K/2}\nabla f_{\xi_i}(x_0)$. Then, according to (3), we have $\mathrm{E}\left[\left\|g_0 - \nabla f(x_0)\right\|^2\right] \leq \frac{\sigma^2}{K/2}$. Now, choose $p = \max(\gamma^2\delta^2, 1/K)$

$$R_K \lesssim \frac{F}{\gamma K} + \frac{\sigma^2}{pK^2} + p\sigma^2 \lesssim \frac{F}{\gamma K} + \frac{\sigma^2}{K} + \gamma^2\delta^2\sigma^2 + \frac{\sigma^2}{K}.$$

Next set $\gamma = \min\left(\frac{1}{\delta}, \left(\frac{F}{2\delta^2\sigma^2K}\right)^{1/3}\right)$ and the rate results in

$$\begin{aligned}
R_K &\lesssim \frac{\delta F}{K} + \frac{F}{K}\left(\frac{2\delta^2\sigma^2K}{F}\right)^{1/3} + \left(\frac{F}{2\delta^2\sigma^2K}\right)^{2/3}\delta^2\sigma^2 + \frac{\sigma^2}{K} \\
&\lesssim \frac{\delta F + \sigma^2}{K} + \left(\frac{\delta\sigma F}{K}\right)^{2/3},
\end{aligned}$$

which leads to the communication complexity of $\mathcal{O}\left(\frac{\delta F+\sigma^2}{\varepsilon} + \frac{\delta\sigma F}{\varepsilon^{3/2}}\right)$. This concludes the proof.

E.2. Proof of Corollary A.2

In this part, we analyze communication complexity similar to Appendix E.1. The focus is on constant stepsize case $\gamma_k \equiv \gamma \lesssim \min\{1/\delta, \sqrt{p}B/\delta\}$ and exact proximal computation $\epsilon = 0$.

Proof of Corollary A.2. Denote $F := f(x_0) - f_{\inf}$, then convergence rate of SPAM-PP can be expressed as

$$R_K := \frac{1}{K}\sum_{k=1}^{K} \mathrm{E}\left[\|\nabla f(x_k)\|^2\right] \quad \lesssim \quad \frac{F}{\gamma K} + \frac{\|g_0 - \nabla f(x_0)\|^2}{pK} + \frac{p\sigma^2}{B}.$$

Here, we used the fact that $p(2-p) > 1$, for $p < 1$. Take g_0 as the following averaged gradient estimator

$$g_0 = \frac{1}{BK/2}\sum_{j=1}^{K/2}\sum_{i\in S_j} \nabla f_i(x_0). \tag{18}$$

From the law of large numbers, we deduce $\mathrm{E}\left[\|g_0 - \nabla f(x_0)\|^2\right] \leq \sigma^2/(BK/2)$. Then

$$R_K \lesssim \frac{F}{\gamma K} + \frac{\sigma^2}{pBK^2} + \frac{p\sigma^2}{B}.$$

The choice of $\gamma \sim \min\{1/\delta, \sqrt{p}B/\delta\}$ results in

$$R_K \lesssim \frac{\delta F}{K} + \frac{\delta F}{\sqrt{p}BK} + \frac{\sigma^2}{pBK^2} + \frac{p\sigma^2}{B}. \tag{19}$$

In order to optimize the right hand side with respect to parameter p, let us divide the expression into $\frac{\sigma^2}{pBK^2} + \frac{p\sigma^2}{B}$ and $\frac{\delta F}{\sqrt{p}BK} + \frac{p\sigma^2}{B}$. Minimizing these sums separately, we obtain the optimum values $p_1 = \frac{1}{K}$ and $p_2 = \left(\frac{\delta F}{K\sigma^2}\right)^{2/3}$, respectively. We then choose $p_0 = \max\{p_1, p_2\}$. Therefore,

$$\begin{aligned} R_K &\lesssim \frac{\delta F}{K} + \frac{\delta F}{\sqrt{p_0}BK} + \frac{\sigma^2}{p_0 BK^2} + \frac{p_0\sigma^2}{B} \\ &\lesssim \frac{\delta F}{K} + \frac{\delta F}{\sqrt{p_2}BK} + \frac{\sigma^2}{p_1 BK^2} + \frac{(p_1+p_2)\,\sigma^2}{B} \\ &\lesssim \frac{\delta F}{K} + \frac{\sigma^2}{BK} + \frac{1}{B}\left(\frac{\delta\sigma F}{K}\right)^{2/3}. \end{aligned}$$

□

Namely, for $p = \max\left\{\frac{1}{K}, \left(\frac{\delta^2 F^2 B}{K^2\sigma^4}\right)^{1/3}\right\}$ we have

$$\begin{aligned} R_K &\lesssim \frac{\delta F}{K} + \frac{\delta F}{BK}\left(\frac{K\sigma^2}{\delta F\sqrt{B}}\right)^{1/3} + \frac{\sigma^2}{BK} + \left(\frac{\delta^2 F^2 B}{K^2\sigma^4}\right)^{1/3}\frac{\sigma^2}{B} + \frac{\sigma^2}{BK} \\ &\lesssim \frac{\delta F}{K} + \frac{\sigma^2}{BK} + \left(\frac{\delta\sigma F}{BK}\right)^{2/3}, \end{aligned}$$

leading to communication complexity $\mathcal{O}\left(\frac{\delta F}{\varepsilon} + \frac{\sigma^2}{B\varepsilon} + \frac{\delta\sigma F}{B\varepsilon^{3/2}}\right)$.

F. Partial participation with averaging

Algorithm 3 SPAM-PPA

Input: learning rate $\gamma > 0$, starting point $x_0 \in \mathbb{R}^d$;
proximal precision level ϵ; initialize $g_0 = g_{-1}$;
for $k = 0, 1, 2, \ldots$ **do**
  Sample a subset of clients S_k, with size $|S_k| = B$;
  Selected clients do local SPAM updates;
  for $i \in S_k$ **do**
    Set $g_k^i = \nabla f_i(x_k) + (1 - p_k)\left(g_{k-1} - \nabla f_i(x_{k-1})\right)$;
    Broadcast g_k^i to the server;
  end for
  $g_k = \frac{1}{B}\sum_{i \in S_k} g_k^i$;
  for $i \in S_k$ **do**
    Set $x_{k+1}^i = \text{a-prox}_\epsilon\left(x_k, g_k, \gamma_k, i\right)$;
    Broadcast x_{k+1}^i to the server;
  end for
  The server aggregates the local iterates: $x_{k+1} = \frac{1}{B}\sum_{i \in S_k} x_{k+1}^i$;
end for

Theorem F.1 (SPAM-PPA). *Suppose Assumptions 1, 2 are satisfied and the objective function f is L-smooth. If $\xi_k \sim \mathsf{Unif}(S_k)$ at every iteration, then the iterates of SPAM-PPA with $\gamma_k \leq \frac{1}{4(\delta+L)}$ and $p_k = \frac{96\delta^2\gamma_k^2}{96\delta^2\gamma_k^2 + B^2}$ satisfy*

$$\frac{1}{\Gamma_K}\sum_{k=0}^{K-1} \gamma_k \mathrm{E}\left[\left\|\nabla f(x_{k+1}^{\xi})\right\|^2\right] \quad \leq \quad \frac{40}{\Gamma_K}\left(V_0 - \mathrm{E}\left[V_K\right]\right) + \frac{240}{\Gamma_K}\sum_{k=0}^{K-1} p_k\gamma_k\frac{\sigma^2}{B} + 7.5\epsilon^2.$$

The result of the theorem is similar to the one in Theorem A.1. In fact, following the proof scheme of Corollary A.2, one can derive the complexity analysis for SPAM-PPA. However, unlike previous results, we require the objective function f to be smooth.

F.1. Proof of Theorem F.1

The proof follows the logic of Proposition 3.1. Recall that

$$V_k = f(x_k) - f_{\inf} + \frac{3\gamma_k}{2p_k - p_k^2}\|g_k - \nabla f(x_k)\|^2.$$

We start with proving a descent lemma. Recall that $\xi_k \sim \mathsf{Unif}(S_k)$, for the fixed S_k.

Lemma F.2. *For an L-smooth objective f satisfying assumptions 1,2 and parameters $\gamma_k^2 \leq \min\left\{\frac{1}{16(L+\delta)^2}, \frac{4p_k}{15\delta^2(1-p_k)}\right\}$, the iterates of the SPAM-PPA algorithm satisfy*

$$\mathrm{E}\left[f(x_{k+1}) - f_{\inf}\right] \leq \mathrm{E}\left[f(x_k) - f_{\inf}\right] + 2\gamma_k\mathrm{E}\left[\left\|\nabla f(x_k) - g_k\right\|^2\right] - \frac{1}{4\gamma_k}\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2\right]. \tag{20}$$

The proof of the lemma is deferred to Appendix G.5. Next, we estimate the second term of the Lyapunov function. Recall that

$$\begin{aligned} g_{k+1} &= \frac{1}{S_k}\sum_{i \in S_k}\left\{\nabla f_i(x_{k+1}) + (1-p_k)\left(g_k - \nabla f_i(x_k)\right)\right\} \\ &= \nabla \tilde{f}_k(x_{k+1}) + (1-p_k)\left(g_k - \nabla \tilde{f}_k(x_k)\right), \end{aligned}$$

where $\tilde{f}_k(x) := \frac{1}{S_k}\sum_{i \in S_k}\nabla f_i(x)$. Notice that $\mathrm{E}\left[\tilde{f}_k(x)\right] = f(x)$, for every fixed $x \in \mathbb{R}^d$. Furthermore, combining the convexity of the Euclidean norm and Hessian similarity (5) we deduce that the estimator $\tilde{f}_k$ satisfies the Hessian similarity

condition

$$\left\|\nabla \tilde{f}_k(x) - \nabla f(x) - \nabla \tilde{f}_k(y) + \nabla f(y)\right\| \leq \frac{1}{B}\sum_{i\in S_k}\|\nabla f_i(x) - \nabla f(x) - \nabla f_i(y) + \nabla f(y)\|$$
$$\leq \frac{\delta}{B}\|x-y\|.$$

Furthermore, Jensen's inequality implies that $\tilde{f}_k$ satisfies the bounded variance condition as well:

$$\mathrm{E}\left[\left\|\nabla \tilde{f}_k(x) - \nabla f(x)\right\|\right] \leq \sigma^2/B.$$

Repeating the analysis exactly as in the proof of Lemma C.2, we obtain

$$\begin{aligned}\mathrm{E}\left[\|g_{k+1} - \nabla f(x_{k+1})\|^2\right] &\leq (1-p_k)^2\mathrm{E}\left[\|g_k - \nabla f(x_k)\|^2\right] \\ &\quad +2(1-p_k)^2\frac{\delta^2}{B^2}\mathrm{E}\left[\|x_{k+1}-x_k\|^2\right] + \frac{2p_k^2\sigma^2}{B}.\end{aligned}$$

Assume now that $\xi_k \sim \mathsf{Unif}(S_k)$, for a fixed S_k. The latter means $x_{k+1} = \mathrm{E}\left[x_{k+1}^{\xi_k}\middle|\mathcal{G}_k\right]$, and subsequently, Jensen's inequality yields

$$\begin{aligned}\mathrm{E}\left[\|g_{k+1} - \nabla f(x_{k+1})\|^2\right] \leq& (1-p_k)^2\mathrm{E}\left[\|g_k - \nabla f(x_k)\|^2\right] \\ &+ 2(1-p_k)^2\frac{\delta^2}{B^2}\mathrm{E}\left[\left\|\mathrm{E}\left[x_{k+1}^{\xi_k}\middle|\mathcal{G}_k\right] - x_k\right\|^2\right] + \frac{2p_k^2\sigma^2}{B} \\ \leq& (1-p_k)^2\mathrm{E}\left[\|g_k - \nabla f(x_k)\|^2\right] \\ &+ 2(1-p_k)^2\frac{\delta^2}{B^2}\mathrm{E}\left[\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2\middle|\mathcal{G}_k\right]\right] + \frac{2p_k^2\sigma^2}{B} \\ =& (1-p_k)^2\mathrm{E}\left[\|g_k - \nabla f(x_k)\|^2\right] \\ &+ 2(1-p_k)^2\frac{\delta^2}{B^2}\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2\right] + \frac{2p_k^2\sigma^2}{B}.\end{aligned}$$

Now, we need to bound $\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2\right]$ from below.

Lemma F.3. *Under assumptions 1 and 2, we have the following lower bound for the iterates of* SPAM-PPA *algorithm*

$$\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2\right] \geq \frac{\gamma_k^2}{5}\mathrm{E}\left[\left\|\nabla f(x_{k+1}^{\xi_k})\right\|^2\right] - \gamma_k^2\mathrm{E}\left[\|g_k - \nabla f(x_k)\|^2\right] - \gamma_k\epsilon^2. \tag{21}$$

The proof of the lemma can be found in Appendix G.6. Let us now bound the Lyapunov function using (20) and (21):

$$\begin{aligned}\mathrm{E}\left[V_{k+1}\right] &\leq \mathrm{E}\left[f(x_k) - f_{\inf}\right] + 2\gamma_k\mathrm{E}\left[\|\nabla f(x_k) - g_k\|^2\right] - \frac{1}{4\gamma_k}\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2\right] \\ &\quad +\alpha(1-p_k)^2\mathrm{E}\left[\|g_k - \nabla f(x_k)\|^2\right] + 2\alpha(1-p_k)^2\frac{\delta^2}{B^2}\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2\right] + \frac{2\alpha p_k^2\sigma^2}{B} \\ &= \mathrm{E}\left[V_k\right] + \left(2\alpha\frac{\delta^2}{B^2}(1-p_k)^2 - \frac{1}{4\gamma_k}\right)\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2\right] + \frac{2\alpha p_k^2\sigma^2}{B} \\ &\quad +(2\gamma_k - \alpha(2p_k - p_k^2))\mathrm{E}\left[\|\nabla f(x_k) - g_k\|^2\right].\end{aligned}$$

The latter is true for every positive α. Let us now plug in the value of $\alpha = \frac{3\gamma_k}{2p_k - p_k^2}$. Then, using $\gamma \leq \sqrt{\frac{B^2 p_k}{96\delta^2(1-p_k)}}$, we obtain

$$2\alpha\frac{\delta^2}{B^2}(1-p_k)^2 - \frac{1}{4\gamma_k} \leq \frac{6\gamma_k\delta^2}{B^2(2p_k - p_k^2)}(1-p_k)^2 - \frac{1}{4\gamma_k} \leq -\frac{1}{8\gamma_k}. \tag{22}$$

Hence, we have the following bound

$$\begin{aligned}
\mathrm{E}\left[V_{k+1}\right] &\leq \mathrm{E}\left[V_k\right] - \frac{1}{8\gamma_k}\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2\right] - \gamma_k \mathrm{E}\left[\left\|\nabla f(x_k) - g_k\right\|^2\right] + \frac{6p_k\gamma_k\sigma^2}{B(2-p_k)} \\
&\overset{(21)}{\leq} \mathrm{E}\left[V_k\right] - \frac{1}{8\gamma_k}\left(\frac{\gamma_k^2}{5}\mathrm{E}\left[\left\|\nabla f(x_{k+1}^{\xi_k})\right\|^2\right] - \gamma_k^2\mathrm{E}\left[\left\|g_k - \nabla f(x_k)\right\|^2\right] - \gamma_k^2\epsilon^2\right) \\
&\quad -\gamma_k \mathrm{E}\left[\left\|\nabla f(x_k) - g_k\right\|^2\right] + \frac{6p_k\gamma_k\sigma^2}{B} \\
&\leq \mathrm{E}\left[V_k\right] - \frac{\gamma_k}{40}\mathrm{E}\left[\left\|\nabla f(x_{k+1}^{\xi_k})\right\|^2\right] - \frac{7\gamma_k}{8}\mathrm{E}\left[\left\|\nabla f(x_k) - g_k\right\|^2\right] + \frac{6p_k\gamma_k\sigma^2}{B} + \frac{\gamma_k\epsilon^2}{8} \\
&\leq \mathrm{E}\left[V_k\right] - \frac{\gamma_k}{40}\mathrm{E}\left[\left\|\nabla f(x_{k+1}^{\xi_k})\right\|^2\right] + \frac{6p_k\gamma_k\sigma^2}{B} + \frac{\gamma_k\epsilon^2}{8}.
\end{aligned}$$

Thus, we have

$$\frac{1}{\Gamma_K}\sum_{k=0}^{K-1}\gamma_k\mathrm{E}\left[\left\|\nabla f(x_{k+1}^{\xi_k})\right\|^2\right] \leq \frac{40}{\Gamma_K}\left(V_0 - \mathrm{E}\left[V_K\right]\right) + \frac{240}{\Gamma_K}\sum_{k=0}^{K-1}p_k\gamma_k\frac{\sigma^2}{B} + 7.5\epsilon^2.$$

This concludes the proof of the theorem.

G. Proofs of the technical lemmas

G.1. Proof of Lemma C.1

By the main theorem of Calculus, we have

$$\begin{aligned}
f(x_{k+1}) - f(x_k) &= \int_0^1 \left\langle \nabla f(\underbrace{x_k + \tau(x_{k+1} - x_k)}_{:=x(\tau)}), x_{k+1} - x_k \right\rangle d\tau, \\
f_{\xi_k}(x_{k+1}) - f_{\xi_k}(x_k) &= \int_0^1 \left\langle \nabla f_{\xi_k}(\underbrace{x_k + \tau(x_{k+1} - x_k)}_{:=x(\tau)}), x_{k+1} - x_k \right\rangle d\tau
\end{aligned}$$

Therefore the difference in function value can be bounded as follows:

$$\begin{aligned}
f(x_{k+1}) - f(x_k) &= f_{\xi_k}(x_{k+1}) - f_{\xi_k}(x_k) \\
&\quad + \int_0^1 \langle \nabla f(x(\tau)) - \nabla f_{\xi_k}(x(\tau)), x_{k+1} - x_k\rangle \, d\tau \\
&= f_{\xi_k}(x_{k+1}) - f_{\xi_k}(x_k) + \langle g_k - \nabla f_{\xi_k}(x_k), x_{k+1} - x_k\rangle \\
&\quad + \int_0^1 \langle \nabla f(x(\tau)) - \nabla f_{\xi_k}(x(\tau)) - g_k + \nabla f_{\xi_k}(x_k), x_{k+1} - x_k\rangle \, d\tau \\
&\leq -\frac{1}{2\gamma_k}\|x_{k+1} - x_k\|^2 + \langle \nabla f(x_k) - g_k, x_{k+1} - x_k\rangle \\
&\quad + \int_0^1 \langle \nabla f(x(\tau)) - \nabla f_{\xi_k}(x(\tau)) - \nabla f(x_k) + \nabla f_{\xi_k}(x_k), x_{k+1} - x_k\rangle \, d\tau.
\end{aligned}$$

The last inequality is due to

$$f_{\xi_k}(x_{k+1}) + \langle g_k - \nabla f_{\xi_k}(x_k), x_{k+1} - x_k\rangle + \frac{1}{2\gamma_k}\|x_{k+1} - x_k\|^2 \leq f_{\xi_k}(x_k), \tag{23}$$

which is a direct consequence of $x_{k+1} = \arg\min_x \left\{ f_{\xi_k}(x) + \langle g_k - \nabla f_{\xi_k}(x_k), x - x_k\rangle + \frac{1}{2\gamma_k}\|x - x_k\|^2 \right\}$. Let us now

apply Cauchy-Schwartz inequality to bound both scalar products:

$$\begin{aligned} f(x_{k+1}) - f(x_k) \quad &\leq \quad -\frac{1}{2\gamma_k}\|x_{k+1}-x_k\|^2 + 2\gamma_k\|\nabla f(x_k) - g_k\|^2 + \frac{1}{8\gamma_k}\|x_{k+1}-x_k\|^2 \\ &\quad + \int_0^1 \|\nabla f(x(\tau)) - \nabla f_{\xi_k}(x(\tau)) - \nabla f(x_k) + \nabla f_{\xi_k}(x_k)\|\|x_{k+1}-x_k\|d\tau \\ &\overset{(5)}{\leq} \quad -\frac{1}{2\gamma_k}\|x_{k+1}-x_k\|^2 + 2\gamma_k\|\nabla f(x_k) - g_k\|^2 + \frac{1}{8\gamma_k}\|x_{k+1}-x_k\|^2 \\ &\quad + \delta\int_0^1 \|x(\tau)-x_k\|\|x_{k+1}-x_k\|d\tau \\ &= \quad -\frac{1}{2\gamma_k}\|x_{k+1}-x_k\|^2 + 2\gamma_k\|\nabla f(x_k) - g_k\|^2 + \frac{1}{8\gamma_k}\|x_{k+1}-x_k\|^2 \\ &\quad + \frac{\delta}{2}\|x_{k+1}-x_k\|^2 \\ &\overset{\gamma_k \leq \frac{1}{4\delta}}{\leq} \quad -\frac{1}{4\gamma_k}\|x_{k+1}-x_k\|^2 + 2\gamma_k\|\nabla f(x_k) - g_k\|^2. \end{aligned}$$

Thus, we have

$$f(x_{k+1}) - f_{\inf} \leq f(x_k) - f_{\inf} - \frac{1}{4\gamma_k}\|x_{k+1}-x_k\|^2 + 2\gamma_k\|\nabla f(x_k) - g_k\|^2. \tag{24}$$

This concludes the proof of the lemma.

G.2. Proof of Lemma C.2

Recall that $g_{k+1} = \nabla f_{\xi_{k+1}}(x_{k+1}) + (1-p_k)\left(g_k - \nabla f_{\xi_{k+1}}(x_k)\right)$. We define $\mathcal{F}_k := \{x_{k+1}, x_k, g_k\}$. Then,

$$\begin{aligned} &\mathrm{E}\left[\|g_{k+1} - \nabla f(x_{k+1})\|^2 \middle| \mathcal{F}_k\right] \\ = \ &\mathrm{E}\left[\left\|\nabla f_{\xi_{k+1}}(x_{k+1}) + (1-p_k)\left(g_k - \nabla f_{\xi_{k+1}}(x_k)\right) - \nabla f(x_{k+1})\right\|^2 \middle| \mathcal{F}_k\right] \\ = \ &\mathrm{E}\left[\left\|\nabla f_{\xi_{k+1}}(x_{k+1}) - \nabla f(x_{k+1}) + (1-p_k)\left(\nabla f(x_k) - \nabla f_{\xi_{k+1}}(x_k)\right)\right\|^2 \middle| \mathcal{F}_k\right] \\ &+(1-p_k)^2\|g_k - \nabla f(x_k)\|^2. \end{aligned}$$

The last equality is due to the bias-variance formula and the fact that ξ_{k+1} is independent of $\mathcal{F}_k$ and that the stochastic gradients are unbiased. Using the Cauchy-Schwartz inequality, we deduce the following bound for the first term on the right-hand side, where $\alpha > 0$ is an arbitrary constant:

$$\begin{aligned} &\mathrm{E}\left[\left\|\nabla f_{\xi_{k+1}}(x_{k+1}) - \nabla f(x_{k+1}) + (1-p_k)\left(\nabla f(x_k) - \nabla f_{\xi_{k+1}}(x_k)\right)\right\|^2 \mid \mathcal{F}_k\right] \\ &\quad = \mathrm{E}\left[\|p_k\left(\nabla f_{\xi_{k+1}}(x_{k+1}) - \nabla f(x_{k+1})\right)\right. \qquad (25) \\ &\quad \left. + (1-p_k)\left(\nabla f_{\xi_{k+1}}(x_{k+1}) - \nabla f(x_{k+1}) + \nabla f(x_k) - \nabla f_{\xi_{k+1}}(x_k)\right)\|^2 \mid \mathcal{F}_k\right] \\ &\quad \leq (1+\alpha)p_k^2\mathrm{E}\left[\left\|\nabla f_{\xi_{k+1}}(x_{k+1}) - \nabla f(x_{k+1})\right\|^2 \middle| \mathcal{F}_k\right] \qquad (26) \\ &\qquad + (1+\alpha^{-1})(1-p_k)^2\mathrm{E}\left[\left\|\nabla f_{\xi_{k+1}}(x_{k+1}) - \nabla f(x_{k+1}) + \nabla f(x_k) - \nabla f_{\xi_{k+1}}(x_k)\right\|^2 \middle| \mathcal{F}_k\right]. \end{aligned}$$

We apply (3) and (5) to bound, respectively, the first term and the second term on the right-hand side of (25):

$$\begin{aligned} &\mathrm{E}\left[\left\|\nabla f_{\xi_{k+1}}(x_{k+1}) - \nabla f(x_{k+1}) + (1-p_k)\left(\nabla f(x_k) - \nabla f_{\xi_{k+1}}(x_k)\right)\right\|^2 \middle| \mathcal{F}_k\right] \\ \leq \ &(1+\alpha)p_k^2\sigma^2 + (1+\alpha^{-1})(1-p_k)^2\delta^2\|x_{k+1}-x_k\|^2. \end{aligned}$$

Taking $\alpha = 1$, we obtain the following

$$\mathrm{E}\left[\|g_{k+1} - \nabla f(x_{k+1})\|^2 \middle| \mathcal{F}_k\right] \leq (1-p_k)^2\|g_k - \nabla f(x_k)\|^2 + 2(1-p_k)^2\delta^2\|x_{k+1}-x_k\|^2 + 2p_k^2\sigma^2.$$

This concludes the proof of the lemma.

G.3. Proof of Lemma C.3

By the definition of x_{k+1}, we have

$$\begin{aligned}\|x_{k+1}-x_k\|^2 &= \gamma_k^2\|\nabla f_{\xi_k}(x_{k+1})+g_k-\nabla f_{\xi_k}(x_k)\|^2\\ &= \gamma_k^2\|\nabla f(x_{k+1})+g_k-\nabla f(x_k)+\nabla f_{\xi_k}(x_{k+1})-\nabla f(x_{k+1})-\nabla f_{\xi_k}(x_k)+\nabla f(x_k)\|^2\\ &\geq \frac{\gamma_k^2}{3}\|\nabla f(x_{k+1})\|^2-\gamma_k^2\|g_k-\nabla f(x_k)\|^2\\ &\quad -\gamma_k^2\|\nabla f_{\xi_k}(x_{k+1})-\nabla f(x_{k+1})-\nabla f_{\xi_k}(x_k)+\nabla f(x_k)\|^2\\ &\geq \frac{\gamma_k^2}{3}\|\nabla f(x_{k+1})\|^2-\gamma_k^2\|g_k-\nabla f(x_k)\|^2-\gamma_k^2\delta^2\|x_{k+1}-x_k\|^2,\end{aligned}$$

where we used a variant of Jensen's inequality $3(a^2+b^2+c^2)\geq(a+b+c)^2$, for $a,b,c>0$. Therefore, we have

$$\begin{aligned}\|x_{k+1}-x_k\|^2 &\geq \frac{1}{1+\gamma_k^2\delta^2}\left(\frac{\gamma_k^2}{3}\|\nabla f(x_{k+1})\|^2-\gamma_k^2\|g_k-\nabla f(x_k)\|^2\right)\\ &\geq \frac{16}{17}\left(\frac{\gamma_k^2}{3}\|\nabla f(x_{k+1})\|^2-\gamma_k^2\|g_k-\nabla f(x_k)\|^2\right)\\ &\geq \frac{\gamma_k^2}{4}\|\nabla f(x_{k+1})\|^2-\gamma_k^2\|g_k-\nabla f(x_k)\|^2.\end{aligned}$$

Thus, we have

$$\mathrm{E}\left[\|x_{k+1}-x_k\|^2\right]\geq\frac{\gamma_k^2}{4}\mathrm{E}\left[\|\nabla f(x_{k+1})\|^2\right]-\gamma_k^2\mathrm{E}\left[\|g_k-\nabla f(x_k)\|^2\right].$$

This concludes the proof of the lemma.

G.4. Proof of Lemma C.4

Let $x_{k+1}=\text{a-prox}_\epsilon\left(x_k,g_k,\gamma_k,\xi_k\right)$. Then, from the definition of the function ϕ_k (7), we have

$$\begin{aligned}\epsilon^2 &\geq \|\nabla\phi_k(x_{k+1})\|^2\\ &= \left\|\nabla f_{\xi_k}(x_{k+1})+g_k-\nabla f_{\xi_k}(x_k)+\frac{1}{\gamma_k}\left(x_{k+1}-x_k\right)\right\|^2\\ &= \left\|\nabla f(x_{k+1})+\nabla f_{\xi_k}(x_{k+1})-\nabla f(x_{k+1})+\nabla f(x_k)-\nabla f_{\xi_k}(x_k)+g_k-\nabla f(x_k)+\frac{1}{\gamma_k}\left(x_{k+1}-x_k\right)\right\|^2\\ &\geq \frac{1}{4}\|\nabla f(x_{k+1})\|^2-\|g_k-\nabla f(x_k)\|^2-\delta^2\|x_{k+1}-x_k\|^2-\frac{1}{\gamma_k^2}\|x_{k+1}-x_k\|^2.\end{aligned}$$

Where in the last inequality we used $\|a_1+a_2+a_3+a_4\|^2\leq 4\left(\|a_1\|^2+\|a_2\|^2+\|a_3\|^2+\|a_4\|^2\right)$ for any vectors $a_i\in\mathbb{R}^d$. Thus, we deduce

$$\begin{aligned}\|x_{k+1}-x_k\|^2 &\geq \frac{\gamma_k^2}{1+\gamma_k^2\delta^2}\left(\frac{1}{4}\|\nabla f(x_{k+1})\|^2-\|g_k-\nabla f(x_k)\|^2-\epsilon^2\right)\\ &\geq \frac{1}{1+\gamma_k^2\delta^2}\left(\frac{\gamma_k^2}{4}\|\nabla f(x_{k+1})\|^2-\gamma_k^2\|g_k-\nabla f(x_k)\|^2-\gamma_k^2\epsilon^2\right)\\ &\geq \frac{16}{17}\left(\frac{\gamma_k^2}{4}\|\nabla f(x_{k+1})\|^2-\gamma_k^2\|g_k-\nabla f(x_k)\|^2-\gamma_k^2\epsilon^2\right)\\ &\geq \frac{\gamma_k^2}{5}\|\nabla f(x_{k+1})\|^2-\gamma_k^2\|g_k-\nabla f(x_k)\|^2-\gamma_k^2\epsilon^2.\end{aligned}$$

Taking expectations on both sides leads to

$$\mathrm{E}\left[\|x_{k+1}-x_k\|^2\right]\geq\frac{\gamma_k^2}{5}\mathrm{E}\left[\|\nabla f(x_{k+1})\|^2\right]-\gamma_k^2\mathrm{E}\left[\|g_k-\nabla f(x_k)\|^2\right]-\gamma_k^2\epsilon^2.$$

This concludes the proof of the lemma.

G.5. Proof of Lemma F.2

Recalling that $x_{k+1}^{\xi_k} = \text{a-prox}_\epsilon\left(x_k, g_k, \gamma_k, \xi_k\right)$, we have

$$f_{\xi_k}(x_{k+1}^{\xi_k}) + \left\langle g_k - \nabla f_{\xi_k}(x_k), x_{k+1}^{\xi_k} - x_k \right\rangle + \frac{1}{2\gamma_k}\|x_{k+1}^{\xi_k} - x_k\|^2 \leq f_{\xi_k}(x_k).$$

Similar to the proof of Proposition 3.1 we start with

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= \int_0^1 \left\langle \nabla f(\underbrace{x_k + \tau(x_{k+1} - x_k)}_{:=x(\tau)}), x_{k+1} - x_k \right\rangle d\tau, \\ f_{\xi_k}(x_{k+1}^{\xi_k}) - f_{\xi_k}(x_k) &= \int_0^1 \left\langle \nabla f_{\xi_k}(\underbrace{x_k + \tau(x_{k+1}^{\xi_k} - x_k)}_{:=x^{\xi_k}(\tau)}), x_{k+1}^{\xi_k} - x_k \right\rangle d\tau. \end{aligned}$$

Thus, we have

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= f_{\xi_k}(x_{k+1}^{\xi_k}) - f_{\xi_k}(x_k) \\ &\quad + \int_0^1 \langle \nabla f(x(\tau)), x_{k+1} - x_k \rangle d\tau \\ &\quad + \int_0^1 \left\langle -\nabla f_{\xi_k}(x^{\xi_k}(\tau)), x_{k+1}^{\xi_k} - x_k \right\rangle d\tau \\ &= f_{\xi_k}(x_{k+1}^{\xi_k}) - f_{\xi_k}(x_k) + \left\langle g_k - \nabla f_{\xi_k}(x_k), x_{k+1}^{\xi_k} - x_k \right\rangle \\ &\quad + \int_0^1 \langle \nabla f(x(\tau)), x_{k+1} - x_k \rangle d\tau \\ &\quad + \int_0^1 \left\langle -\nabla f_{\xi_k}(x^{\xi_k}(\tau)) - g_k + \nabla f_{\xi_k}(x_k), x_{k+1}^{\xi_k} - x_k \right\rangle d\tau. \end{aligned}$$

Applying the descent property of a-prox (see Definition 4.1), we deduce the following:

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq -\frac{1}{2\gamma_k}\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 + \left\langle \nabla f(x_k) - g_k, x_{k+1}^{\xi_k} - x_k \right\rangle \\ &\quad + \int_0^1 \langle \nabla f(x(\tau)), x_{k+1} - x_k \rangle d\tau \\ &\quad + \int_0^1 \left\langle -\nabla f_{\xi_k}(x^{\xi_k}(\tau)) - \nabla f(x_k) + \nabla f_{\xi_k}(x_k), x_{k+1}^{\xi_k} - x_k \right\rangle d\tau. \end{aligned}$$

Let us take expectation from both sides conditioned to $\mathcal{G}_k = \{x_k, x_{k+1}, S_k, g_k\}$. In other words, we take expectation with respect to the random index ξ_k chosen uniformly from the already chosen S_k:

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \mathrm{E}\left[-\frac{1}{2\gamma_k}\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 + \left\langle \nabla f(x_k) - g_k, x_{k+1}^{\xi_k} - x_k \right\rangle \mid \mathcal{G}_k\right] \\ &\quad + \mathrm{E}\left[\int_0^1 \langle \nabla f(x(\tau)), x_{k+1} - x_k \rangle d\tau \mid \mathcal{G}_k\right] \\ &\quad + \mathrm{E}\left[\int_0^1 \left\langle -\nabla f_{\xi_k}(x^{\xi_k}(\tau)) - \nabla f(x_k) + \nabla f_{\xi_k}(x_k), x_{k+1}^{\xi_k} - x_k \right\rangle d\tau \middle| \mathcal{G}_k\right] \\ &= \mathrm{E}\left[-\frac{1}{2\gamma_k}\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right] + \langle \nabla f(x_k) - g_k, x_{k+1} - x_k \rangle \\ &\quad + \mathrm{E}\left[\int_0^1 \left\langle \nabla f(x(\tau)) - \nabla f(x_k) - \nabla f_{\xi_k}(x^{\xi_k}(\tau)) + \nabla f_{\xi_k}(x_k), x_{k+1}^{\xi_k} - x_k \right\rangle d\tau \mid \mathcal{G}_k\right]. \end{aligned}$$

Here the last equality is due to the fact that ξ_k is independent of $\mathcal{G}_k$ and $x_{k+1} = \mathrm{E}\left[x_{k+1}^{\xi_k} \mid \mathcal{G}_k\right]$. Therefore, applying the Cauchy-Schwartz inequality

$$
\begin{aligned}
f(x_{k+1}) - f(x_k) &\leq \mathrm{E}\left[-\frac{1}{2\gamma_k}\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right] + \langle \nabla f(x_k) - g_k, x_{k+1} - x_k\rangle \\
&\quad + \mathrm{E}\left[\int_0^1 \left\langle \nabla f(x(\tau)) - \nabla f(x^{\xi_k}(\tau)), x_{k+1}^{\xi_k} - x_k\right\rangle d\tau \mid \mathcal{G}_k\right] \\
&\quad + \mathrm{E}\left[\int_0^1 \left\langle \nabla f(x^{\xi_k}(\tau)) - \nabla f(x_k) - \nabla f_{\xi_k}(x^{\xi_k}(\tau)) + \nabla f_{\xi_k}(x_k), x_{k+1}^{\xi_k} - x_k\right\rangle d\tau \mid \mathcal{G}_k\right] \\
&\leq \mathrm{E}\left[-\frac{1}{2\gamma_k}\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right] + \langle \nabla f(x_k) - g_k, x_{k+1} - x_k\rangle \\
&\quad + \mathrm{E}\left[\int_0^1 \left\|\nabla f(x(\tau)) - \nabla f(x^{\xi_k}(\tau))\right\|\left\|x_{k+1}^{\xi_k} - x_k\right\| d\tau \mid \mathcal{G}_k\right] \\
&\quad + \mathrm{E}\left[\int_0^1 \left\|\nabla f(x^{\xi_k}(\tau)) - \nabla f(x_k) - \nabla f_{\xi_k}(x^{\xi_k}(\tau)) + \nabla f_{\xi_k}(x_k)\right\|\left\|x_{k+1}^{\xi_k} - x_k\right\| d\tau \mid \mathcal{G}_k\right].
\end{aligned}
$$

Applying Cauchy-Schwartz inequality once again, we deduce

$$
\begin{aligned}
f(x_{k+1}) - f(x_k) &\leq \mathrm{E}\left[-\frac{1}{2\gamma_k}\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right] + \frac{C}{2}\|\nabla f(x_k) - g_k\|^2 + \frac{1}{2C}\|x_{k+1} - x_k\|^2 \\
&\quad + \mathrm{E}\left[\int_0^1 L\left\|x(\tau) - x^{\xi_k}(\tau)\right\|\left\|x_{k+1}^{\xi_k} - x_k\right\| d\tau \mid \mathcal{G}_k\right] \\
&\quad + \mathrm{E}\left[\int_0^1 \delta\left\|x^{\xi_k}(\tau) - x_k\right\|\left\|x_{k+1}^{\xi_k} - x_k\right\| d\tau \mid \mathcal{G}_k\right] \\
&\leq \mathrm{E}\left[-\frac{1}{2\gamma_k}\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right] + \frac{C}{2}\|\nabla f(x_k) - g_k\|^2 + \frac{1}{2C}\|x_{k+1} - x_k\|^2 \\
&\quad + \mathrm{E}\left[\int_0^1 L\tau\left\|x_{k+1} - x_{k+1}^{\xi_k}\right\|\left\|x_{k+1}^{\xi_k} - x_k\right\| d\tau \mid \mathcal{G}_k\right] \\
&\quad + \mathrm{E}\left[\int_0^1 \delta\tau\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 d\tau \mid \mathcal{G}_k\right].
\end{aligned}
$$

Computing the integral with respect to τ, we obtain

$$
\begin{aligned}
f(x_{k+1}) - f(x_k) &\leq \mathrm{E}\left[-\frac{1}{2\gamma_k}\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right] + \frac{C}{2}\|\nabla f(x_k) - g_k\|^2 + \frac{1}{2C}\|x_{k+1} - x_k\|^2 \\
&\quad + \frac{L}{2}\mathrm{E}\left[\left\|x_{k+1} - x_{k+1}^{\xi_k}\right\|\left\|x_{k+1}^{\xi_k} - x_k\right\| \mid \mathcal{G}_k\right] + \frac{\delta}{2}\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right] \\
&\leq \mathrm{E}\left[-\frac{1}{2\gamma_k}\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right] + \frac{C}{2}\|\nabla f(x_k) - g_k\|^2 + \frac{1}{2C}\|x_{k+1} - x_k\|^2 \\
&\quad + \frac{L}{4}\mathrm{E}\left[\left\|x_{k+1} - x_{k+1}^{\xi_k}\right\|^2 \mid \mathcal{G}_k\right] + \frac{2\delta + L}{4}\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right].
\end{aligned}
$$

Recall again that $x_{k+1} = \mathrm{E}\left[x_{k+1}^{\xi_k} \mid \mathcal{G}_k\right]$, thus $x_{k+1} = \arg\min_{a \in \mathbb{R}^d} \mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - a\right\|^2 \mid \mathcal{G}_k\right]$. Therefore,

$$
\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_{k+1}\right\|^2 \mid \mathcal{G}_k\right] \leq \mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right].
$$

Furthermore,

$$
\|x_{k+1} - x_k\|^2 = \left\|\mathrm{E}\left[x_{k+1}^{\xi_k} \mid \mathcal{G}_k\right] - x_k\right\|^2 \leq \mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right].
$$

Combining these two bounds, we deduce

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \frac{C}{2}\|\nabla f(x_k) - g_k\|^2 \\ &\quad + \left(\frac{1}{2C} + \frac{\delta + L}{2} - \frac{1}{2\gamma_k}\right) \mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right]. \end{aligned}$$

The previous bound is true for every positive value of C. Thus, it is true also for $C = 4\gamma_k$. Taking into account that $\gamma_k < \frac{1}{4(L+\delta)}$, we get

$$\frac{1}{2C} + \frac{\delta + L}{2} - \frac{1}{2\gamma_k} \leq \frac{1}{8\gamma_k} + \frac{1}{8\gamma_k} - \frac{1}{2\gamma_k} = -\frac{1}{4\gamma_k}.$$

Therefore,

$$f(x_{k+1}) - f(x_k) \leq 2\gamma_k\|\nabla f(x_k) - g_k\|^2 - \frac{1}{4\gamma_k}\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 \mid \mathcal{G}_k\right].$$

Thus, taking full expectation on both sides, we have

$$\mathrm{E}\left[f(x_{k+1}) - f_{\inf}\right] \leq \mathrm{E}\left[f(x_k) - f_{\inf}\right] + 2\gamma_k\mathrm{E}\left[\|\nabla f(x_k) - g_k\|^2\right] - \frac{1}{4\gamma_k}\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2\right].$$

This concludes the proof.

G.6. Proof of Lemma F.3

By the definition of $x_{k+1}^{\xi_k}$, for every $\xi \in S_k$ we have

$$\begin{aligned} \left\|x_{k+1}^{\xi_k} - x_k\right\|^2 &= \gamma_k^2\left\|\nabla f_{\xi_k}(x_{k+1}^{\xi_k}) + g_k - \nabla f_{\xi_k}(x_k) - \nabla\phi_k(x_{k+1})\right\|^2 \\ &= \gamma_k^2\Big\|\nabla f(x_{k+1}^{\xi_k}) + g_k - \nabla f(x_k) + \nabla f_{\xi_k}(x_{k+1}^{\xi_k}) \\ &\quad -\nabla f(x_{k+1}^{\xi_k}) - \nabla f_{\xi_k}(x_k) + \nabla f(x_k) - \nabla\phi_k(x_{k+1})\Big\|^2 \\ &\geq \frac{\gamma_k^2}{4}\left\|\nabla f(x_{k+1}^{\xi_k})\right\|^2 - \gamma_k^2\|g_k - \nabla f(x_k)\|^2 \\ &\quad -\gamma_k^2\left\|\nabla f_{\xi_k}(x_{k+1}^{\xi_k}) - \nabla f(x_{k+1}^{\xi_k}) - \nabla f_{\xi_k}(x_k) + \nabla f(x_k)\right\|^2 - \gamma_k^2\epsilon^2 \\ &\geq \frac{\gamma_k^2}{4}\left\|\nabla f(x_{k+1}^{\xi_k})\right\|^2 - \gamma_k^2\|g_k - \nabla f(x_k)\|^2 - \gamma_k^2\delta^2\left\|x_{k+1}^{\xi_k} - x_k\right\|^2 - \gamma_k^2\epsilon^2. \end{aligned}$$

The third inequality is due to Cauchy-Schwartz and the second property of the approximate proximal operator (See Definition 4.1). Therefore, we have

$$\begin{aligned} \left\|x_{k+1}^{\xi_k} - x_k\right\|^2 &\geq \frac{1}{1 + \gamma_k^2\delta^2}\left(\frac{\gamma_k^2}{4}\left\|\nabla f(x_{k+1}^{\xi_k})\right\|^2 - \gamma_k^2\|g_k - \nabla f(x_k)\|^2 - \gamma_k^2\epsilon^2\right) \\ &\geq \frac{16}{17}\left(\frac{\gamma_k^2}{4}\left\|\nabla f(x_{k+1}^{\xi_k})\right\|^2 - \gamma_k^2\|g_k - \nabla f(x_k)\|^2 - \gamma_k^2\epsilon^2\right) \\ &\geq \frac{\gamma_k^2}{5}\left\|\nabla f(x_{k+1}^{\xi_k})\right\|^2 - \gamma_k^2\|g_k - \nabla f(x_k)\|^2 - \gamma_k^2\epsilon^2. \end{aligned}$$

We deduce

$$\mathrm{E}\left[\left\|x_{k+1}^{\xi_k} - x_k\right\|^2\right] \geq \frac{\gamma_k^2}{5}\mathrm{E}\left[\left\|\nabla f(x_{k+1}^{\xi_k})\right\|^2\right] - \gamma_k^2\mathrm{E}\left[\|g_k - \nabla f(x_k)\|^2\right] - \gamma_k^2\epsilon^2.$$

This concludes the proof of the lemma.

H. Experimental details

We provide additional details on the experimental settings from Section B. Consider a distributed ridge regression problem defined as

$$f(x) = \mathrm{E}_\xi \left[\|A_\xi x - y_\xi\|^2 \right] + \frac{\lambda}{2}\|x\|^2, \tag{27}$$

where ξ is uniform random variable over $\{1, \ldots, n\}$ for $n = 10$, $\lambda = 0.1$. At every iteration, one client is sampled uniformly at random.

We follow a similar to (Lin et al., 2024) procedure for synthetic data generation, which allows us to calculate and control Hessian similarity δ. Namely, a random matrix $A_0 \in \mathbb{R}^{d\times d}$ ($d = 100$) is generated with entries from a standard Gaussian distribution $\mathcal{N}(0, 1)$. Then we obtain $A = A_0 A_0^\top$ (to ensure symmetry) and set $A'_\xi = A + B_\xi$ by adding a random symmetric matrix B_ξ (generated similarly to A). Afterwards we modify $A_\xi = A'_\xi + I\lambda_{\min}(A'_\xi)$ by adding an identity matrix I times minimum eigenvalue to guarantee $A_\xi \succeq 0$. Entries of vectors $y_\xi \in \mathbb{R}^d$, and initialization $x_0 \in \mathbb{R}^d$ are generated from a standard Gaussian distribution $\mathcal{N}(0, 1)$.

In the case of inexact proximal point computation (1/10 local steps), local subproblems (7) are solved by gradient descent with a fixed step size of $1/(2L_l)$, where L_l is the local smoothness constant. A more efficient method (e.g., (Nesterov, 2013), (Kim & Fessler, 2021)) could be used for local optimization instead.

Simulations were performed on a machine with 24 $\mathrm{Intel(R)\ Xeon(R)}$ Gold 6246 CPU @ 3.30 GHz.

I. Additional experiments

In this section, we present complementary experimental results to compare SPAM-inexact (with varying parameter γ) and CE-LGD (Patel et al., 2022). The problem setup remains consistent with Section B and Appendix H.

Figure 3 illustrates the convergence behavior of the methods towards a neighborhood of the stationary point. The vertical and horizontal axes are shared across all plots. We vary the momentum parameter $p \in \{0.1, 0.9\}$ (within each subplot), the number of local steps: $\{1, 2, 10\}$ (across columns), and the parameter $\gamma \in \{1, 2, 5\}$ (divided by δ) for SPAM (across rows). The size of the convergence neighborhood for both methods is primarily influenced by the value of p, which is especially evident in the final plot for SPAM, where a bigger p results in larger gradient norm oscillations.

Overall, we observe that CE-LGD may outperform SPAM when using a small number of local steps and a small parameter γ. However, the fastest overall convergence is achieved by SPAM when γ is sufficiently large and the number of local steps exceeds 1.

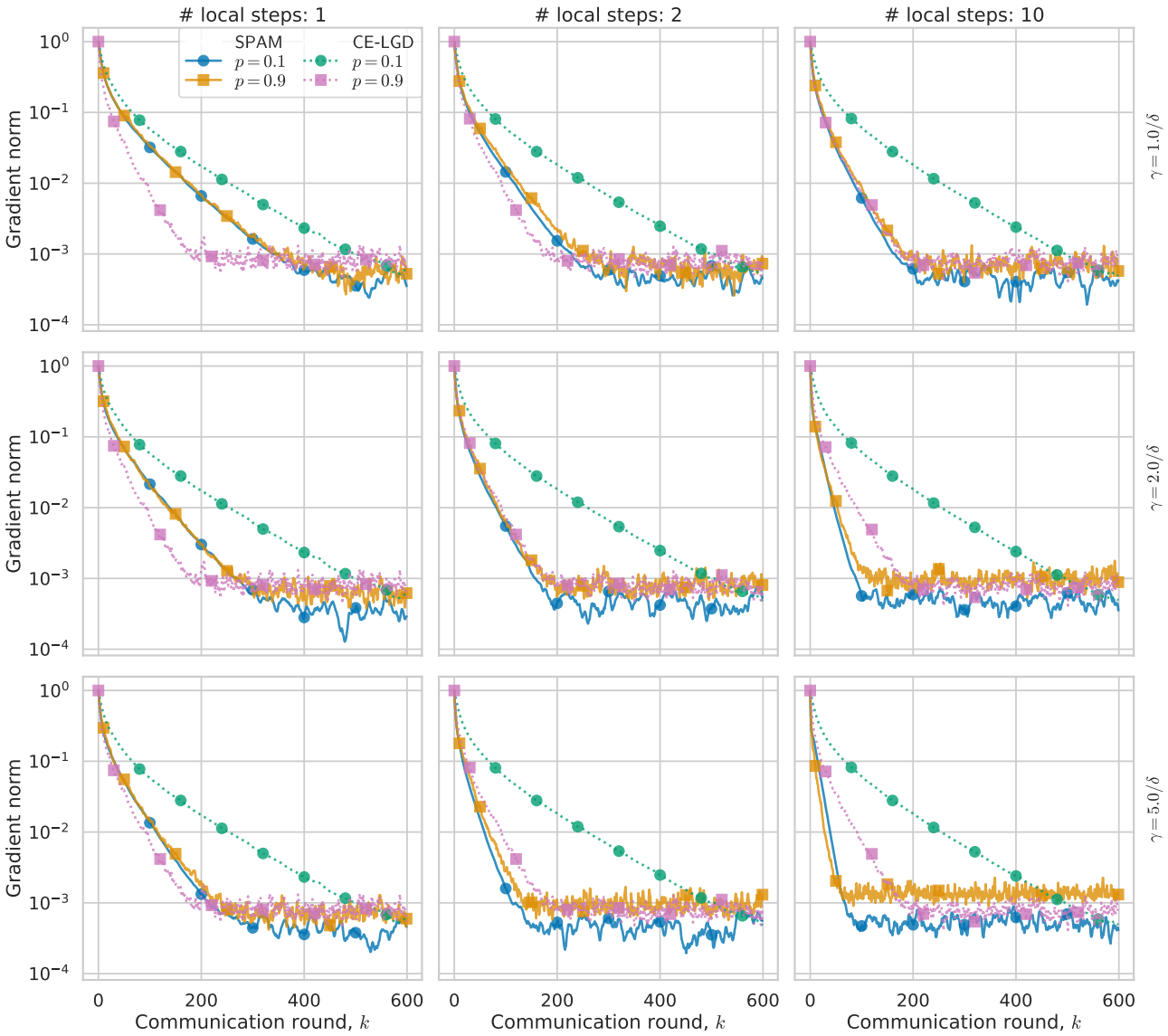

Figure 3: Comparison of SPAM-inexact and CE-LGD on problem (27) with different p, γ and number of local steps.