ON CONVERGENCE OF FEDERATED AVERAGING LANGEVIN DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a federated averaging Langevin algorithm (FA-LD) for uncertainty quantification and mean predictions with distributed clients. In particular, we generalize beyond normal posterior distributions and consider a general class of models. We develop theoretical guarantees for FA-LD for strongly log-concave distributions with non-i.i.d data and study how the injected noise and the stochastic-gradient noise, the heterogeneity of data, and the varying learning rates affect the convergence. Such an analysis sheds light on the optimal choice of local updates to minimize communication cost. Important to our approach is that the communication efficiency does not deteriorate with the injected noise in the Langevin algorithms. In addition, we examine in our FA-LD algorithm both independent and correlated noise used over different clients. We observe there is a trade-off between the pairs among communication, accuracy, and data privacy. As local devices may become inactive in federated networks, we also show convergence results based on different averaging schemes where only partial device updates are available. In such a case, we discover an additional bias that does not decay to zero.

1 INTRODUCTION

Federated learning (FL) allows multiple parties to jointly train a consensus model without sharing user data. Compared to the classical centralized learning regime, federated learning keeps training data on local clients, such as mobile devices or hospitals, where data privacy, security, and access rights are a matter of vital interest. This aggregation of various data resources heeding privacy concerns yields promising potential in areas of internet of things (Chen et al., 2020), healthcare (Li et al., 2020d; 2019b), text data (Huang et al., 2020), and fraud detection (Zheng et al., 2020).

A standard formulation of federated learning is a distributed optimization framework that tackles communication costs, client robustness, and data heterogeneity across different clients (Li et al., 2020a). Central to the formulation is the efficiency of the communication, which directly motivates the communication-efficient federated averaging (FedAvg) (McMahan et al., 2017). FedAvg introduces a global model to synchronously aggregate multi-step local updates on the available clients and yields distinctive properties in communication. However, FedAvg often stagnates at inferior local modes empirically due to the data heterogeneity across the different clients (Charles & Konečnỳ, 2020; Woodworth et al., 2020). To tackle this issue, Karimireddy et al. (2020); Pathaky & Wainwright (2020) proposed stateful clients to avoid the unstable convergence, which are, however, not scalable with respect to the number of clients in applications with mobile devices (Al-Shedivat et al., 2021). In addition, the optimization framework often fails to quantify the uncertainty accurately for the parameters of interest, which are crucial for building estimators, hypothesis tests, and credible intervals. Such a problem leads to unreliable statistical inference and casts doubts on the credibility of the prediction tasks or diagnoses in medical applications.

To unify optimization and uncertainty quantification in federated learning, we resort to a *Bayesian treatment by sampling from a global posterior distribution*, where the latter is aggregated by infrequent communications from local posterior distributions. We adopt a popular approach for inferring posterior distributions for large datasets, the stochastic gradient Markov chain Monte Carlo (SG-MCMC) method (Welling & Teh, 2011; Vollmer et al., 2016; Teh et al., 2016; Chen et al., 2014; Ma et al., 2015), which enjoys theoretical guarantees beyond convex scenarios (Raginsky et al., 2017; Zhang et al., 2017; Mangoubi & Vishnoi, 2018; Ma et al., 2019). In particular, we examine

in the federated learning setting the efficacy of the stochastic gradient Langevin dynamics (SGLD) algorithm, which differs from stochastic gradient descent (SGD) in an additionally injected noise for exploring the posterior. The close resemblance naturally inspires us to adapt the optimization-based FedAvg to a distributed sampling framework. Similar ideas have been proposed in federated posterior averaging (Al-Shedivat et al., 2021), where empirical study and analyses on Gaussian posteriors have shown promising potential of this approach. Compared to the appealing theoretical guarantees of optimization-based algorithms in federated learning (Pathaky & Wainwright, 2020; Al-Shedivat et al., 2021), the convergence properties of approximate sampling algorithms in federated learning is far less understood. To fill this gap, we proceed by asking the following question:

Can we build a unified algorithm with convergence guarantees for sampling in FL?

In this paper, we make a first step in answering this question in the affirmative. We propose the federated averaging Langevin dynamics for posterior inference beyond the Gaussian distribution. We list our contributions as follows:

- We present a novel non-asymptotic convergence analysis for FA-LD from simulating strongly log-concave distributions on non-i.i.d data when the learning rate is fixed. The frequently used bounded gradient assumption of ℓ_2 norm in FedAvg optimization is not required.
- The convergence analysis indicates that injected noise, data heterogeneity, and stochastic-gradient noise are all driving factors that affect the convergence. Such an analysis provides a concrete guidance on the optimal number of local updates to minimize communications.
- We can activate partial device updates to avoid straggler's effects in practical applications and tune the correlation of injected noises to protect privacy.
- We also provide differential privacy guarantees, which shed light on the trade-off between data privacy and accuracy given a limited budget.

For **related works** of other federated learning approaches, we refer interested readers to section H.

2 Preliminaries

2.1 An optimization perspective on federated averaging

Federated averaging (FedAvg) is a standard algorithm in federated learning and is typically formulated into a distributed optimization framework as follows

$$\min_{\theta} \ell(\theta) := \frac{\sum_{c=1}^{N} \ell^c(\theta)}{\sum_{c=1}^{N} n_c}, \quad \ell^c(\theta) := \sum_{i=1}^{n_c} l(\theta; x_{c,i}), \tag{1}$$

where $\theta \in \mathbb{R}^d$, $l(\theta; x_{c,j})$ is a certain loss function based on θ and the data point $x_{c,j}$.

FedAvg algorithm requires the following three iterate:

- *Broadcast*: The center server *broadcasts* the latest model, θ_k, to all local clients.
- *Local updates*: For any $c \in [N]$, the c-th client first sets the auxiliary variable $\beta_k^c = \theta_k$ and then conducts $K \geq 1$ local steps: $\beta_{k+1}^c = \beta_k^c - \frac{\eta}{n_c} \nabla \widetilde{\ell}^c(\beta_k^c)$, where η is the learning rate and $\nabla \widetilde{\ell}^c$ is the unbiased estimate of the exact gradient $\nabla \ell^c$.
- *Synchronization*: The local models are sent to the center server and then aggregated into a unique model $\theta_{k+K} := \sum_{c=1}^{N} p_c \beta_{k+K}^c$, where p_c as the weight of the c-th client such that $p_c = \frac{n_c}{\sum_{i=1}^{N} n_i} \in (0, 1)$ and $n_c > 0$ is the number of data points in the c-th client.

From the optimization perspective, Li et al. (2020c) proved the convergence of the FedAvg algorithm on non-i.i.d data such that a larger number of local steps K and a higher order of data heterogeneity slows down the convergence. Notably, Eq.(1) can be interpreted as maximizing the likelihood function, which is a special case of maximum a posteriori estimation (MAP) given a uniform prior.

2.2 Stochastic gradient Langevin dynamics

Posterior inference offers the exact uncertainty quantification ability of the predictions. A popular method for posterior inference with large dataset is the stochastic gradient Langevin dynamics

(SGLD) (Welling & Teh, 2011), which injects additional noise into the stochastic gradient such that $\theta_{k+1} = \theta_k - \eta \nabla \widetilde{f}(\theta_k) + \sqrt{2\tau\eta}\xi_k$, where τ is the temperature and ξ_k is a standard Gaussian vector. $f(\theta) := \sum_{c=1}^N \ell^c(\theta)$ is an energy function. $\nabla \widetilde{f}(\theta)$ is an unbiased estimate of $\nabla f(\theta)$. In the longtime limit, θ_k converges weakly to the distribution $\pi(\theta) \propto \exp(-f(\theta)/\tau)$ (Teh et al., 2016) as $\eta \to 0$.

3 Posterior inference via federated averaging Langevin dynamics

The increasing concern for uncertainty estimation in federated learning motivates us to consider the simulation of the distribution $\pi(\theta) \propto \exp(-f(\theta)/\tau)$ with distributed clients.

Problem formulation We propose the federated averaging Langevin dynamics (FA-LD) based on the FedAvg framework in section 2.1. We follow the same *broadcast* step and *synchronization* step but propose to inject random noises for *local updates*. In particular, we consider the following scheme: for any $c \in [N]$, the c-th client first sets $\theta_k^c = \theta_k$ and then conducts $K \geq 1$ local steps:

$$\textit{Local updates: } \beta_{k+1}^c = \theta_k^c - \eta \nabla \widetilde{f}^c(\theta_k^c) + \sqrt{2\eta\tau}\Xi_k^c \tag{2}$$

$$\textit{Synchronization: } \theta_{k+1}^c = \begin{cases} \beta_{k+1}^c & \text{if } k+1 \bmod K \neq 0 \\ \sum_{c=1}^N p_c \beta_{k+1}^c & \text{if } k+1 \bmod K = 0. \end{cases} \tag{3}$$

where $\nabla f^c(\theta) = \frac{1}{p_c}\nabla \ell^c(\theta)$; $\nabla \widetilde{f}^c(\theta)$ is the unbiased estimate of $\nabla f^c(\theta)$; Ξ_k^c is some Gaussian vector in Eq.(6). Summing Eq.(2) from clients $c = 1$ to N and setting $\boldsymbol{\theta}_k = (\theta_k^1, \cdots, \theta_k^N)$, we have

$$\beta_{k+1} = \theta_k - \eta \boldsymbol{\nabla} \widetilde{f}(\boldsymbol{\theta}_k) + \sqrt{2\eta\tau}\xi_k,$$

$$\text{where} \quad \beta_k = \sum_{c=1}^N p_c \beta_k^c, \quad \theta_k = \sum_{c=1}^N p_c \theta_k^c, \quad \boldsymbol{\nabla} \widetilde{f}(\boldsymbol{\theta}_k) = \sum_{c=1}^N p_c \nabla \widetilde{f}^c(\theta_k^c), \quad \xi_k = \sum_{c=1}^N p_c \Xi_k^c. \tag{4}$$

By the nature of *synchronization*, we always have $\beta_k = \theta_k$ for any $k \geq 0$ and the process follows

$$\theta_{k+1} = \theta_k - \eta \boldsymbol{\nabla} \widetilde{f}(\boldsymbol{\theta}_k) + \sqrt{2\eta\tau}\xi_k, \tag{5}$$

which resembles SGLD except that we have a different "gradient" operator $\boldsymbol{\nabla}$ and θ_k is *not accessible when* k *mod* $K \neq 0$. Since our target is to simulate from $\pi(\theta) \propto \exp(-f(\theta)/\tau)$, we expect ξ_k is a standard Gaussian vector. By the concentration of independent Gaussian variables, it is natural to set

$$\Xi_k^c = \xi_k^c / \sqrt{p_c}, \tag{6}$$

where $\xi_k = \sum_{c=1}^N p_c \Xi_k^c = \sum_{c=1}^N \sqrt{p_c}\xi_k^c$ and ξ_k^c is a also standard Gaussian vector. Now we present the algorithm based on independent inject noise ($\rho = 0$) and the full-device update (3) in Algorithm 1, where ρ is the the correlation coefficient and will be further studied in section 4.3.3. We observe Eq.(7) maintains a temperature $\tau/p_c > \tau$ to converge to the stationary distribution π. Such a mechanism may limit the disclosure of individual data and shows a potential to protect the data privacy.

Algorithm 1 Federated averaging Langevin dynamics Algorithm (FA-LD), informal version of Algorithm 4. Denote by θ_k^c the model parameter in the c-th client at the k-th step. Denote the one-step intermediate result by β_k^c. ξ_k^c is an independent standard d-dimensional Gaussian vector at iteration k for each client $c \in [N]$; $\dot{\xi}_k$ is a d-dimensional Gaussian vector shared by all the clients; ρ denotes the correlation coefficient. $\mathcal{S}_k$ is sampled according to a device-sampling rule based on scheme I or II.

$$\beta_{k+1}^c = \theta_k^c - \eta \nabla \widetilde{f}^c(\theta_k^c) + \sqrt{2\eta\tau\rho^2}\dot{\xi}_k + \sqrt{2\eta\tau(1-\rho^2)/p_c}\xi_k^c, \tag{7}$$

$$\theta_{k+1}^c = \begin{cases} \beta_{k+1}^c & \text{if } k+1 \bmod K \neq 0 \\ \Pi_{\text{device}}(\beta_{k+1}^c) & \text{if } k+1 \bmod K = 0. \end{cases}$$

where $\Pi_{\text{device}}(\beta_{k+1}^c) = \sum_{c=1}^N p_c \beta_{k+1}^c$ for full device and $= \sum_{c \in \mathcal{S}_{k+1}} \frac{1}{S}\beta_{k+1}^c$ for partial device.

4 CONVERGENCE ANALYSIS

In this section, we show that FA-LD converges to the stationary distribution $\pi(\theta)$ in the 2-Wasserstein (W_2) distance at a rate of $O(1/\sqrt{T_\epsilon})$ for strongly log-concave and smooth density. The W_2 distance is defined between a pair of Borel probability measures μ and ν on $\mathbb{R}^d$ as follows

$$W_2(\mu,\nu) := \inf_{\gamma^2 \in \text{Couplings}(\mu,\nu)} \left(\int \|\boldsymbol{\beta}_\mu - \boldsymbol{\beta}_\nu\|_2^2 d\gamma^2(\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\nu) \right)^{\frac{1}{2}},$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm on $\mathbb{R}^d$ and the pair of random variables $(\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\nu) \in \mathbb{R}^d \times \mathbb{R}^d$ is a coupling with the marginals following $\mathcal{L}(\boldsymbol{\beta}_\mu) = \mu$ and $\mathcal{L}(\boldsymbol{\beta}_\nu) = \nu$. Note that $\mathcal{L}(\cdot)$ denotes a distribution of a random variable.

4.1 ASSUMPTIONS

We make standard assumptions on the smoothness and convexity of the functions $f^1, f^2, \cdots, f^N$, which naturally yields appealing tail properties of the stationary measure π. Thus, we no longer require a restrictive assumption on the bounded gradient in ℓ_2 norm as in Koloskova et al. (2019); Yu et al. (2019); Li et al. (2020c). In addition, to control the distance between ∇f^c and $\nabla \widetilde{f}^c$, we also assume a bounded variance of the stochastic gradient in assumption 4.3.

Assumption 4.1 (Smoothness)**.** *For each $c \in [N]$, f^c is L-smooth if for some $L > 0$*

$$f^c(y) \leq f^c(x) + \langle \nabla f^c(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^d.$$

Assumption 4.2 (Strongly convexity)**.** *For each $c \in [N]$, f^c is m-strongly convex if for some $m > 0$*

$$f^c(x) \geq f^c(y) + \langle \nabla f^c(y), x - y \rangle + \frac{m}{2}\|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^d.$$

Assumption 4.3 (Bounded variance)**.** *For each $c \in [N]$, the variance of noise in the stochastic gradient $\nabla \widetilde{f}^c(x)$ in each client is upper bounded such that*

$$\mathbb{E}[\|\nabla \widetilde{f}^c(x) - \nabla f^c(x)\|_2^2] \leq \sigma^2 d, \quad \forall x \in \mathbb{R}^d.$$

Quality of non-i.i.d data Denote by θ_* the global minimum of f. Next, we quantify the degree of the non-i.i.d data by $\gamma := \max_{c \in [N]} \|\nabla f^c(\theta_*)\|_2$, which is non-negative and yields a larger scale if the data is less identically distributed.

4.2 PROOF SKETCH

The proof hinges on showing the one-step result in the W_2 distance. To facilitate the analysis, we first define an auxiliary continuous-time processes $(\bar{\theta}_t)_{t \geq 0}$ that synchronizes for any time $t \geq 0$ [1]

$$\mathrm{d}\bar{\theta}_t = -\nabla f(\bar{\theta}_t)\mathrm{d}t + \sqrt{2\tau}\mathrm{d}\overline{W}_t, \tag{8}$$

where $\bar{\theta}_t = \sum_{c=1}^N p_c \bar{\theta}_t^c$, $\nabla f(\bar{\theta}_t) = \sum_{c=1}^N p_c \nabla f^c(\bar{\theta}_t^c)$, $\bar{\theta}_t^c$ is the continuous-time variable at client c, and $\overline{W}$ is a d-dimensional Brownian motion. The continuous-time algorithm is known to converge to the stationary distribution $\pi(\theta) \propto e^{-\frac{f(\theta)}{\tau}}$, where $f(\theta) = \sum_{c=1}^N p_c f^c(\theta)$. Assume that $\bar{\theta}_0$ simulates from the stationary distribution π, then it follows that $\bar{\theta}_t \sim \pi$ for any $t \geq 0$.

4.2.1 DOMINATED CONTRACTION IN FEDERATED LEARNING

The first target is to show a certain contraction property of $\|\beta - \theta - \eta(\nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta}))\|_2^2$ based on distributed clients with infrequent communications. Consider a standard decomposition

$$\|\beta - \theta - \eta(\nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta}))\|_2^2 = \|\beta - \theta\|_2^2 - 2\eta \underbrace{\langle \beta - \theta, \nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta}) \rangle}_{\mathcal{I}} + \eta^2 \|\nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta})\|_2^2.$$

Using Eq.(4), we decompose $\mathcal{I}$ and apply Jensen's inequality to obtain a lower bound of $\mathcal{I}$. In what follows, we have the following lemma.

[1] As such, $\bar{\theta}_t^c = \sum_{c=1}^N p_c \bar{\beta}_t^c = \bar{\theta}_t$ for $\forall t \geq 0$ and $\forall c \in [N]$. It also indicates $\nabla f(\bar{\theta}_t) = \sum_{c=1}^N p_c \nabla f^c(\bar{\theta}_t^c) = \sum_{c=1}^N p_c \nabla f^c(\bar{\theta}_t)$; hence the process recovers the standard Langevin diffusion and naturally converges to π.

Lemma 4.4 (Dominated contraction property, informal version of Lemma B.1)**.** *Assume assumptions 4.1 and 4.2 hold. For any learning rate $\eta \in (0, \frac{1}{L+m}]$, any $\{\theta^c\}_{c=1}^N, \{\beta^c\}_{c=1}^N \in \mathbb{R}^d$, we have*

$$\|\beta - \theta - \eta(\nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta}))\|_2^2 \leq (1 - \eta m) \cdot \|\beta - \theta\|_2^2 + 4\eta L \sum_{c=1}^N p_c \cdot \underbrace{(\|\beta^c - \beta\|_2^2 + \|\theta^c - \theta\|_2^2)}_{\textit{divergence term}},$$

where $\beta = \sum_{c=1}^N p_c \beta^c$, $\theta = \sum_{c=1}^N p_c \theta^c$, $\nabla f(\beta) = \sum_{c=1}^N p_c \nabla f^c(\beta^c)$, and $\boldsymbol{\nabla} f(\boldsymbol{\theta}) = \sum_{c=1}^N p_c \nabla f^c(\theta^c)$. It implies that as long as the local parameters θ^c, β^c and global θ, β don't differ each other too much, we can guarantee the desired convergence. In a special case when the communication is conducted at every iteration, the divergence term disappears and recovers the standard contraction (Dalalyan & Karagulyan, 2019).

4.2.2 Bounding divergence

The following result shows that given a finite number of local steps K, the divergence between θ^c in local client and θ in the center is bounded in ℓ_2 norm. Notably, since the Brownian motion leads to a lower order term $O(\eta)$ instead of $O(\eta^2)$, a naïve proof framework such as Li et al. (2020c) may lead to a crude upper bound for the final convergence.

Lemma 4.5 (Bounded divergence, informal version of Lemma B.3)**.** *Assume assumptions 4.1, 4.2, and 4.3 hold. For any $\eta \in (0, 2/m)$ and $\|\theta_0^c - \theta_*\|_2^2 \leq d\mathcal{D}^2$ for any $c \in [N]$ and some constant $\mathcal{D}$, we have the ℓ_2 upper bound of the divergence between local clients and the center*

$$\sum_{c=1}^N p_c \mathbb{E}\|\theta_k^c - \theta_k\|_2^2 \leq O((K-1)^2 \eta^2 d) + O((K-1)\eta d).$$

The result relies on a uniform upper bound in ℓ_2 norm, which avoids bounded gradient assumptions.

4.2.3 Coupling to the stationary process

Note that $\bar{\theta}_t$ is initialized from the stationary distribution π. The solution to the continuous-time process Eq.(8) follows:

$$\bar{\theta}_t = \bar{\theta}_0 - \int_0^t \nabla f(\bar{\theta}_s) \mathrm{d}s + \sqrt{2\tau} \cdot \overline{W}_t, \qquad \forall t \geq 0. \tag{9}$$

Set $t \to (k+1)\eta$ and $\bar{\theta}_0 \to \bar{\theta}_{k\eta}$ for Eq.(9) and consider a *synchronous coupling* such that $W_{(k+1)\eta} - W_{k\eta} := \sqrt{\eta}\xi_k$ is used to cancel the noise terms, we have

$$\bar{\theta}_{(k+1)\eta} = \bar{\theta}_{k\eta} - \int_{k\eta}^{(k+1)\eta} \nabla f(\bar{\theta}_s) \mathrm{d}s + \sqrt{2\tau\eta}\xi_k. \tag{10}$$

Subtracting Eq.(5) from Eq.(10) and taking square and expectation on both sides yield that

$$\mathbb{E}\|\bar{\theta}_{(k+1)\eta} - \theta_{k+1}\|_2^2 \leq (1 - \eta m/2) \cdot \mathbb{E}\|\bar{\theta}_{k\eta} - \theta_k\|_2^2 + \text{divergence term} + \text{time error}.$$

Eventually, we arrive at the one-step error bound for establishing the convergence results.

Lemma 4.6 (One step update, informal version of Lemma B.5)**.** *Assume assumptions 4.1, 4.2, and 4.3 hold. Consider Algorithm 1 with any $\eta \in (0, \frac{1}{2L})$ and $\|\theta_0^c - \theta_*\|_2^2 \leq d\mathcal{D}^2$, $\rho = 0$, and full device participation for any $c \in [N]$, where θ_* is the global minimum for the function f. Then*

$$W_2^2(\mu_{k+1}, \pi) \leq (1 - \eta m/2) \cdot W_2^2(\mu_k, \pi) + O(\eta^2 d((K-1)^2 + \kappa)),$$

where μ_k denotes the probability measure of θ_k and $\kappa = L/m$ is the condition number.

Given small enough η, the above Lemma indicates that the algorithm will eventually converge

4.3 FULL DEVICE PARTICIPATION

4.3.1 CONVERGENCE BASED ON INDEPENDENT NOISE

When the synchronization step is conducted at every iteration k, the FA-LD algorithm is essentially the standard SGLD algorithm (Welling & Teh, 2011). Theoretical analysis based on the 2-Wasserstein distance has been established in Durmus & Moulines (2016); Dalalyan (2017); Dalalyan & Karagulyan (2019). However, in scenarios of $K > 1$ with distributed clients, a divergence between the global variable θ_k and local variable θ_k^c appears and unavoidably affects the performance. The upper bound on the sampling error is presented as follows.

Theorem 4.7 (Main result, informal version of Theorem B.6)**.** *Assume assumptions 4.1, 4.2, and 4.3 hold. Given Algorithm 1 with $\eta \in (0, \frac{1}{2L}]$, $\rho = 0$, full device, and well initialized $\{\theta_0^c\}_{c=1}^N$, we have*

$$W_2(\mu_k, \pi) \leq (1 - \eta m/4)^k \cdot \left(\sqrt{2d}\big(\mathcal{D} + \sqrt{\tau/m}\big)\right) + 30\kappa\sqrt{\eta m d} \cdot \sqrt{((K-1)^2 + \kappa)H_0}.$$

where μ_k denotes the density of θ_k at iteration k, K is the local updates, $\kappa := L/m$, $\gamma := \max_{c\in[N]} \|\nabla f^c(\theta_*)\|_2$, and $H_0 := \mathcal{D}^2 + \max_{c\in[N]} \frac{\tau}{mp_c} + \frac{\gamma^2}{m^2 d} + \frac{\sigma^2}{m^2}$.

We observe that the initialization, the injected noise, the data heterogeneity of the data, and the noise in the stochastic gradient all affect the convergence. Similar to the result of Li et al. (2020c), FA-LD with K-local steps resembles the behaviour of one-step SGLD with a large learning rate. Although bias-reduction techniques (Karimireddy et al., 2021) can be further adopted to alleviate the data heterogeneity issue, it is relatively minor compared to the large error caused by multiple local steps.

Optimal choice of K. To achieve the precision ϵ based on the learning rate η, we can set

$$30\kappa\sqrt{\eta m d} \cdot \sqrt{(K^2 + \kappa)H_0} \leq \epsilon/2, \quad \exp\big(-\frac{\eta m}{4}T_\epsilon\big) \cdot \sqrt{2d}(\mathcal{D} + \sqrt{\tau/m}) \leq \epsilon/2.$$

This readily leads to $\eta m \leq O\left(\frac{\epsilon^2}{d\kappa^2(K^2+\kappa)H_0}\right), \quad T_\epsilon \geq \Omega\left(\frac{\log(d/\epsilon)}{m\eta}\right)$.

Denote by T_ϵ the number of iterations to achieve the target accuracy ϵ. Plugging into the upper bound of ηm, it suffices to set $T_\epsilon = \Omega(\epsilon^{-2} d\kappa^2(K^2 + \kappa)H_0 \cdot \log(d/\epsilon))$. Note that $H_0 = \Omega(\mathcal{D}^2)$, thus the number of communication rounds is of the order $\frac{T_\epsilon}{K} = \Omega\left(K + \frac{\kappa}{K}\right)$, where the value of $\frac{T_\epsilon}{K}$ first decreases and then increases w.r.t. K, which indicates setting K either too large or too small leads to high communication costs. Ideally, K should be selected in the scale of $\Omega(\sqrt{\kappa})$. Combining the definition of T_ϵ, this shows that the optimal K for FA-LD is in the order of $O(\sqrt{T_\epsilon})$, which matches the optimization-based results (Stich, 2019; Li et al., 2020c).

4.3.2 CONVERGENCE GUARANTEES VIA VARYING LEARNING RATES

Theorem 4.8 (Informal version of Theorem B.7)**.** *Assume assumptions 4.1, 4.2, and 4.3 hold. Consider Algorithm 1 with $\rho = 0$, full device, an initialization satisfying $\|\theta_0^c - \theta_*\|_2^2 \leq d\mathcal{D}^2$ for any $c \in [N]$, and the varying learning rate following $\eta_k = \frac{1}{2L + (1/12)mk}$. Then for any $k \geq 0$, we have $W_2(\mu_k, \pi) \leq 45\kappa\sqrt{((K-1)^2 + \kappa)H_0} \cdot \big(\eta_k m d\big)^{1/2}, \ \forall k \geq 0$.*

To achieve the precision ϵ, we need $W_2(\mu_k, \pi) \leq \epsilon$, i.e. $W_2(\mu_k, \pi) \leq 45\kappa\sqrt{(K^2 + \kappa)H_0} \cdot \left(\frac{md}{2L + (1/12)mk}\right)^{1/2}$. We therefore require $\Omega(\epsilon^{-2} d)$ iterations to achieve the precision ϵ, which improves the $\Omega(\epsilon^{-2} d \log(d/\epsilon))$ rate for FA-LD with a fixed learning rate by a $O(\log(d/\epsilon))$ factor.

4.3.3 PRIVACY-ACCURACY TRADE-OFF VIA CORRELATED NOISES

The local updates in Eq.(2) with $\Xi_k^c = \xi_k^c/\sqrt{p_c}$ requires all the local clients to generate the independent noise ξ_k^c. Such a mechanism enjoys the implementation convenience and yields a potential to protect the data privacy and alleviates the security issue. However, the large scale noise inevitably

slows down the convergence. To handle this issue, the independent noise can be generalized to correlated noise based on a correlation coefficient ρ between different clients. Replacing Eq.(7) with

$$\beta_{k+1}^c = \theta_k^c - \eta \nabla \widetilde{f}^c(\theta_k^c) + \sqrt{2\eta\tau\rho^2}\dot{\xi}_k + \sqrt{2\eta(1-\rho^2)\tau/p_c}\xi_k^c, \tag{11}$$

where $\dot{\xi}_k$ is a d-dimensional standard Gaussian vector shared by all the clients at iteration k and $\dot{\xi}_k$ is independent with ξ_k^c for any $c \in [N]$. Following the synchronization step based on Eq.(3), we have

$$\theta_{k+1} = \theta_k - \eta \nabla \widetilde{f}(\theta_k) + \sqrt{2\eta\tau}\xi_k, \tag{12}$$

where $\xi_k = \rho\dot{\xi}_k + \sqrt{1-\rho^2}\sum_{c=1}^N \sqrt{p_c}\xi_k^c$. Since the variance of i.i.d variables is additive, it is clear that ξ_k follows the standard d-dimensional Gaussian distribution. The correlated noise implicitly reduces the temperature and naturally yields a trade-off between federation and accuracy.

Since the inclusion of correlated noise doesn't affect the iterate of Eq.(12), the algorithm property maintains the same except the scale of the temperature τ and efficacy of federation are changed. Based on a target correlation coefficient $\rho \geq 0$, Eq.(11) is equivalent to applying a temperature $T_{c,\rho} = \tau(\rho^2 + (1-\rho^2)/p_c)$. In particular, setting $\rho = 0$ leads to $T_{c,0} = \tau/p_c$, which exactly recovers Algorithm 1; however, setting $\rho = 1$ leads to $T_{c,1} = \tau$, where the injected noise in local clients is reduced by $1/p_c$ times. Now we adjust the analysis as follows

Theorem 4.9 (Informal version of Theorem B.8)**.** *Assume assumptions 4.1, 4.2, and 4.3 hold. Consider Algorithm 1 with $\rho \in [0,1]$, $\eta \in (0, \frac{1}{2L}]$ and $\|\theta_0^c - \theta_*\|_2^2 \leq d\mathcal{D}^2$ for any $c \in [N]$, we have*

$$W_2(\mu_k, \pi) \leq (1-\eta m/4)^k \cdot \left(\sqrt{2d}\big(\mathcal{D} + \sqrt{\tau/m}\big)\right) + 30\kappa\sqrt{\eta m d} \cdot \sqrt{((K-1)^2+\kappa)H_\rho},$$

where μ_k denotes the probability measure of θ_k, $H_\rho := \mathcal{D}^2 + \frac{1}{m}\max_{c\in[N]} T_{c,\rho} + \frac{\gamma^2}{m^2 d} + \frac{\sigma^2}{m^2}$.

Such a mechanism leads to a trade-off between data privacy and accuracy and may motivate us to exploit the optimal ρ under differential privacy theories (Wang et al., 2015).

4.4 Partial device participation

Full device participation enjoys appealing convergence properties. However, it suffers from the straggler's effect in real-world applications, where the communication is limited by the slowest device. Partial device participation handles this issue by only allowing a small portion of devices in each communication and greatly increased the communication efficiency in a federated network.

The first device-sampling scheme I (Li et al., 2020b) selects a total of S devices, where the c-th device is selected with a probability p_c. The first theoretical justification for convex optimization has been proposed by Li et al. (2020c). The second device-sampling scheme II is to uniformly select S devices without replacement. We follow Li et al. (2020c) and assume S indices are selected uniformly without replacement. In addition, the convergence also requires an additional assumption on balanced data (Li et al., 2020c). Both schemes are formally defined in section C.1.

Theorem 4.10 (Informal version of Theorem C.3)**.** *Under mild assumptions, we run Algorithm 1 with $\rho \in [0,1]$, a fixed $\eta \in (0, \frac{1}{2L}]$ and $\|\theta_0^c - \theta_*\|_2^2 \leq d\mathcal{D}^2$ for any $c \in [N]$, we have*

$$\begin{aligned} W_2(\mu_k, \pi) \leq (1-\eta m/4)^k \cdot &\left(\sqrt{2d}\big(\mathcal{D} + \sqrt{\tau/m}\big)\right) \\ &+ 30\kappa\sqrt{\eta m d} \cdot \sqrt{H_\rho(K^2+\kappa)} + O\left(\sqrt{\frac{d}{S}(\rho^2 + N(1-\rho^2))C_S}\right), \end{aligned}$$

where $C_S = 1$ for Scheme I and $C_S = \frac{N-S}{N-1}$ for Scheme II.

Partial device participation leads to an extra bias regardless of the scale of η. To reduce it, we suggest to consider highly correlated injected noise, such as $\rho = 1$, to reduce the impact of the injected noise. Further setting $O(\sqrt{d/S}) \leq \epsilon/3$ and following a similar η as in section 4.3.1, we can achieve the precision ϵ within $\Omega(\epsilon^{-2} d \log(d/\epsilon))$ iterations given enough devices satisfying $S = \Omega(\epsilon^{-2} d)$.

The device-sampling scheme I provides a viable solution to handle the straggler's effect, which is rather robust to the data heterogeneity and doesn't require the data to be balanced. In more practical cases where a system can only operate based on the first S messages for the local updates, Scheme II can achieve a reasonable approximation given more balanced data with uniformly sampled device. If $S = 1$, our Scheme II matches the result in the Scheme I; If $S = N$, then our Scheme II recovers the result in the full device setting; If $S = N - o(N)$, our Scheme II bound is better than scheme I.

5 Differential Privacy Guarantees

We consider the (ϵ, δ)-differential privacy with respect to the substitute-one relation $\simeq_s$ (Balle et al., 2018). Two datasets $\mathcal{S} \simeq_s \mathcal{S}'$ if they have the same size and differ by exactly one data point. For $\epsilon \geq 0$ and $\delta \in [0, 1]$, a mechanism $\mathcal{M}$ is (ϵ, δ)-differentially private w.r.t. $\simeq_s$ if for any pair of input datasets $\mathcal{S} \simeq_s \mathcal{S}'$, and every measurable subset $E \subset \text{Range}(\mathcal{M})$, we have

$$\mathbb{P}[\mathcal{M}(\mathcal{S}) \in E] \leq e^{\epsilon}\mathbb{P}[\mathcal{M}(\mathcal{S}') \in E] + \delta. \tag{13}$$

Since partial device participation is more general, we focus on analyzing the differential privacy guarantee based on updates with partial devices. Here, we present the result under scheme II. For the result under scheme I, please refer to Theorem G.3 in the appendix.

Theorem 5.1 (Partial version of Theorem G.3)**.** *Assume assumptions G.1 and G.2 holds. For any* $\delta_0 \in (0, 1)$*, if* $\eta \in \left(0, \frac{\tau(1-\rho^2)\gamma^2 \min_{c\in[N]} p_c}{\Delta_l^2 \log(1.25/\delta_0)}\right]$*, then Algorithm 1 under scheme II is* $(\epsilon_{K,T}^{(3)}, \delta_{K,T}^{(3)})$*-differentially private w.r.t.* $\simeq_s$ *after* T *(*$T = EK$ *with* $E \in \mathbb{N}, E \geq 1$*) iterations where*

$$\epsilon_{K,T}^{(3)} = \widetilde{\epsilon}_K \min\left\{\sqrt{\frac{2T}{K}\log\left(\frac{1}{\delta_2}\right)} + \frac{TS(e^{\epsilon_K}-1)}{KN}, \frac{T}{K}\right\}, \quad \delta_{K,T}^{(3)} = \frac{S}{N}\gamma T\delta_0 + \frac{TS}{KN}\delta_1 + \delta_2,$$

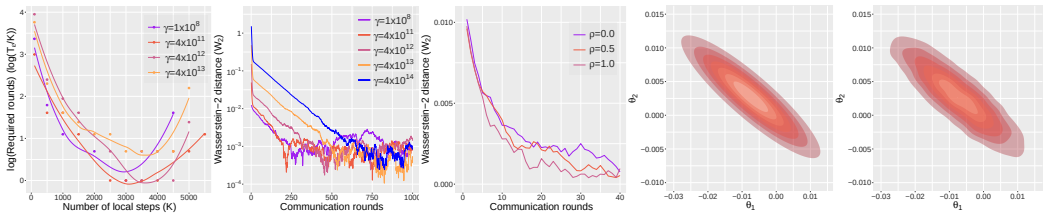
with $\widetilde{\epsilon}_K = \log\left(1 + \frac{S}{N}\left(e^{\epsilon_K} - 1\right)\right)$, $\epsilon_K = \epsilon_1 \min\left\{\sqrt{2K\log(1/\delta_1)} + K(e^{\epsilon_1} - 1),\ K\right\}$, $\epsilon_1 = 2\Delta_l\sqrt{\frac{\eta\log(1.25/\delta_0)}{\tau(1-\rho^2)\min_{c\in[N]} p_c}}$*, and* $\delta_1, \delta_2 \in [0, 1)$.

According to Theorem 5.1 and section G, Algorithm 1 is at least $(\frac{T}{K}\log\left(1 + \frac{S}{N}(e^{K\epsilon_1} - 1)\right), \frac{S}{N}\gamma T\delta_0)$-differentially private. Moreover, if $\eta = O\left(\frac{\tau(1-\rho^2)N^2 \min_{c\in[N]} p_c \log(1/\delta_2)}{\Delta_l^2 S^2 T\log(1/\delta_0)\log(1/\delta_1)}\right)$, then $\epsilon_{K,T}^{(3)} = O\left(\frac{S\Delta_l}{N}\sqrt{\frac{\eta T\log(1/\delta_0)\log(1/\delta_1)\log(1/\delta_2)}{\tau(1-\rho^2)\min_{c\in[N]} p_c}}\right)$.

There is a trade-off between privacy and utility. By Theorem 5.1, $\epsilon_{K,T}^{(3)}$ is an increasing function of $\frac{\eta}{\tau(1-\rho^2)}$, $\frac{S}{N}$, and T. $\delta_{K,T}^{(3)}$ is an increasing function of $\frac{S}{N}$, γ, and T. However, by Theorem 4.10, the upper-bound of $W_2(\mu_T, \mu)$ is a decreasing function of ρ, T, S and is an increasing function of τ and N. There is an optimal η to minimize $W_2(\mu_T, \mu)$ for fixed T while we can make $\epsilon_{K,T}^{(3)}$ arbitrarily small by decreasing η for any fixed T. In practice, users can tune hyper-parameters based on DP and accuracy budget. For example, under some DP budget (ϵ_*, δ_*), we can select the largest $\rho \in [0, 1]$ and $S \in [N]$ such that $\epsilon_{K,T}^{(3)} \leq \epsilon_*$ and $\delta_{K,T}^{(3)} \leq \delta_*$ to achieve the target error $W_2(\mu_T, \mu)$.

6 Experiments

Simulations For each $c \in [N]$, where $N = 50$, we sample θ_c from a 2d Gaussian distribution $N(0, \alpha I_2)$ and sample n_c points from $N(\theta_c, \Sigma)$, where $\Sigma = \begin{bmatrix} 5 & -2 \\ -2 & 1 \end{bmatrix}$. Thus, $l(\theta; x_{c,i}) = \frac{1}{2}(\theta - x_{c,i})^\top\Sigma^{-1}(\theta - x_{c,i}) + \log(2\pi|\Sigma|^{\frac{1}{2}})$, $\ell^c(\theta) = \sum_{i=1}^{n_c} l(\theta; x_{c,i})$. The temperature is $\tau = 1$. The target density is $N(u, \frac{1}{n}\Sigma)$ with $u = \frac{1}{n}\sum_{c=1}^{N}\sum_{i=1}^{n_c} x_{c,i}$. We choose a Gaussian posterior to facilitate the calculation of the W2 distance to verify theoretical properties. Other details are detailed in section I.



(a) Study of K (b) Study of γ (c) Study of ρ (d) True density (e) Empirical density

Figure 1: Convergence of FA-LD based on full devices. In Figure 1(a), points may coincide.

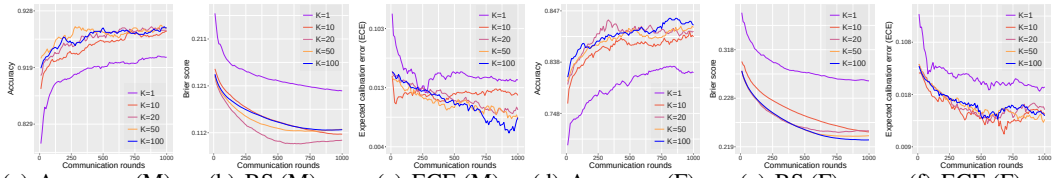
Optimal local steps: We study the choices of local step K for Algorithm 1 based on $\rho = 0$, full device, and different α's, which corresponds to different levels of data heterogeneity modelled by γ. We choose $\alpha = 0, 1, 10, 100, 1000$ and the corresponding γ is around $1\times10^8, 4\times10^{11}, 4\times10^{12}, 4\times10^{13}$, and 4×10^{14}, respectively. We fix $\eta = 10^{-7}$. We evaluate the (log) number of communication rounds to achieve the accuracy $\epsilon = 10^{-3}$ and denote it by T_ϵ. As shown in Figure 1(a), a small K leads to an excessive amount of communication costs; by contrast, a large K results in large biases, which in turn requires high communications. The optimal K that minimizes communication is around 3000 and *the communication savings can be as large as 30 times*.

Data heterogeneity and correlated noise: We study the impact of γ on the convergence based on $\rho = 0$, full device, different γ from $\{1 \times 10^8, 4 \times 10^{11}, 4 \times 10^{12}, 4 \times 10^{13}$, and $4 \times 10^{14}\}$. We set $K = 10$. As shown in Figure 1(b), the W_2 distances under different γ all converge to some levels around 10^{-3} after sufficient computations. Nevertheless, a larger γ does slow down the convergence, which suggests adopting more balanced data to facilitate the computations. In Figure 1(c), we study the impact of ρ on the convergence of the algorithm. We choose $K = 100$ and $\gamma = 10^8$ and observe that a larger correlation slightly accelerates the computation, although it risks in privacy concerns.

Approximate samples: In Figure 1(e), we plot the empirical density according to the samples from Algorithm 1 with $\rho = 0$, full device, $K = 10$ and $\gamma = 10^8$, $\eta = 10^{-7}$. For comparison, we show the true density plot of the target distribution in Figure 1(d). The empirical density approximates the true density very well, which indicates that the potential of FA-LD in federated learning.

For the convergence based on *partial device participation*, we refer interested readers to section I.

(Fashion) MNIST To evaluate the performance of FA-LD under different local steps K on real-world datsets, we apply FA-LD to train a logistic regression model with the cross entropy loss on the MNIST and Fashion-MNIST dataset. The training dataset is split uniformly at random into $N = 10$ subsets of equal size for 10 clients. In each setting, we collect one parameter sample after every 10 communication rounds and average the predicted probabilities made by all the previous collected parameter samples to calculate three test statistics, accuracy, Brier Score (BS) (Brier et al., 1950), and Expected Calibration Error (ECE) (Guo et al., 2017) on the test dataset. We tune the step sizes η for the best performance and plot the curves of those test statistics against communication rounds under different local steps $K = 1, 10, 20, 50, 100$ in Figure 2. Other details are provided in section J.



(a) Accuracy (M) (b) BS (M) (c) ECE (M) (d) Accuracy (F) (e) BS (F) (f) ECE (F)

Figure 2: Convergence of FA-LD on the MNIST (M) and Fashion-MNIST (F) dataset.

According to Figure 2, under the *same communication budget*, FA-LD with $K = 1$ (i.e. the standard SGLD algorithm) performs the worst in terms of all three test statistics, which indicates the necessity of multiple local updates in federated learning. Moreover, for different test statistics, the optimal local step K could be different; e.g., for the MNIST dataset, the optimal K in terms of accuracy is between 50 and 100 (Figure 2(a)), while the optimal K in terms of BS is around 20 (Figure 2(b)). To better visualize the convergence, we also plot the curves of accuracy, BS, and ECE when a warmup period consisting of the first 500 communication rounds is removed. Such results are detailed in section J.

7 Conclusion

We propose a novel convergence analysis for federated averaging Langevin dynamics (FA-LD) with distributed clients. Our results no longer require the bounded gradient assumption in ℓ_2 norm as in the optimization-driven literature in federated learning. The theoretical guarantees yield a concrete guidance on the selection of the optimal number of local updates to minimize communication costs. In addition, the convergence highly depends on the data heterogeneity and the injected noises, where the latter also inspires us to consider correlated injected noise and partial device updates to balance between differential privacy and prediction accuracy with theoretical guarantees.

REFERENCES

Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring. In *Proc. of the International Conference on Machine Learning (ICML)*, 2012.

Sungjin Ahn, Babak Shahbaba, and Max Welling. Distributed Stochastic Gradient MCMC. In *International Conference on Machine Learning (ICML)*, 2014.

Maruan Al-Shedivat, Jennifer Gillenwater, Eric Xing, and Afshin Rostamizadeh. Federated Learning via Posterior Averaging: A New Perspective and Practical Algorithms. In *ICLR*. arXiv:2010.05273, 2021.

Dmitrii Avdyukhin and Shiva Prasad Kasiviswanathan. Federated Learning under Arbitrary Communication Patterns. In *Proc. of the International Conference on Machine Learning (ICML)*, 2021.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Zachary Charles and Jakub Konečnỳ. On the Outsized Importance of Learning Rates in Local Update Methods. *arXiv:2007.00878*, 2020.

Changyou Chen, Nan Ding, and Lawrence Carin. On the Convergence of Stochastic Gradient MCMC Algorithms with High-order Integrators. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2278–2286, 2015.

Changyou Chen, Nan Ding, Chunyuan Li, Yizhe Zhang, and Lawrence Carin. Stochastic Gradient MCMC with Stale Gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Hong-You Chen and Wei-Lun Chao. FedBE: Making Bayesian Model Ensemble Applicable to Federated Learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Mingzhe Chen, Zhaohui Yang, Walid Saad, Changchuan Yin, H Vincent Poor, and Shuguang Cui. A Joint Learning and Communications Framework for Federated Learning over Wireless Networks. *IEEE Trans. on Wireless Communications*, 2020.

Tianqi Chen, Emily B. Fox, and Carlos Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. In *Proc. of the International Conference on Machine Learning (ICML)*, 2014.

Xiang Cheng, Niladri S Chatterji, Peter L Bartlett, and Michael I Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on Learning Theory (COLT)*, pp. 300–323. PMLR, 2018.

Arkabandhu Chowdhury and Chris Jermaine. Parallel and Distributed MCMC via Shepherding Distributions. In *AISTAT*, 2018.

Arnak S. Dalalyan. Further and Stronger Analogy Between Sampling and Optimization: Langevin Monte Carlo and Gradient Descent. In *Conference on Learning Theory (COLT)*, June 2017.

Arnak S Dalalyan and Avetik Karagulyan. User-friendly Guarantees for the Langevin Monte Carlo with Inaccurate Gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large Scale Distributed Deep Networks. In *Advances in neural information processing systems (NeurIPS)*, pp. 1223–1231, 2012.

Wei Deng, Qi Feng, Liyao Gao, Faming Liang, and Guang Lin. Non-Convex Learning via Replica Exchange Stochastic Gradient MCMC. In *Proc. of the International Conference on Machine Learning (ICML)*, 2020.

Alain Durmus and Éric Moulines. Sampling from a Strongly Log-concave Distribution with the Unadjusted Langevin Algorithm. *arXiv:1605.01559*, 2016.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60. IEEE, 2010.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Mert Gürbüzbalaban, Xuefeng Gao, Yuanhan Hu, and Lingjiong Zhu. Decentralized Stochastic Gradient Langevin Dynamics and Hamiltonian Monte Carlo. *arXiv:2007.00590v3*, 2021.

Farzin Haddadpour and Mehrdad Mahdavi. On the Convergence of Local Descent Methods in Federated Learning. 2019.

Baihe Huang, Xiaoxiao Li, Zhao Song, and Xin Yang. FL-NTK: A Neural Tangent Kernel-based Framework for Federated Learning Convergence Analysis. In *International Conference on Machine Learning (ICML)*, 2021.

Yangsibo Huang, Zhao Song, Danqi Chen, Kai Li, and Sanjeev Arora. TextHide: Tackling Data Privacy in Language Understanding Tasks. In *EMNLP*, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in neural information processing systems (NeurIPS)*, pp. 8571–8580, 2018.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic Controlled Averaging for Federated Learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *International Conference on Machine Learning (ICML)*, 2021.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First Analysis of Local GD on Heterogeneous Data. *arXiv:1909.04715*, 2019.

Anastasia Koloskova, Sebastian U.Stich, and Martin Jaggi. Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication. In *Proc. of the International Conference on Machine Learning (ICML)*, 2019.

Nikita Kotelevskii, Maxime Vono, Eric Moulines, and Alain Durmus. FedPop: A Bayesian Approach for Personalised Federated Learning. In *Advances in neural information processing systems (NeurIPS)*, 2022.

Chunyuan Li, Changyou Chen, Yunchen Pu, Ricardo Henao, and Lawrence Carin. Communication-Efficient Stochastic Gradient MCMC for Neural Networks. In *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2019a.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smithy. Federated Optimization in Heterogeneous Networks. In *Proceedings of the 3rd MLSys Conference*, 2020b.

Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M Jorge Cardoso, et al. Privacy-preserving Federated Brain Tumour Segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pp. 133–141. Springer, 2019b.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the Convergence of FedAvg on Non-IID Data. In *Proc. of the International Conference on Learning Representation (ICLR)*, 2020c.

Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence Staib, Pamela Ventola, and James S Duncan. Multi-site fMRI Analysis using Privacy-preserving Federated Learning and Domain Adaptation: ABIDE Results. *Medical Image Analysis*, 2020d.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on Non-IID Features via Local Batch Normalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Xuechen Li, Denny Wu, Lester Mackey, and Murat A. Erdogdu. Stochastic Runge-Kutta Accelerates Langevin Monte Carlo and Beyond. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7746–7758, 2019c.

Y.-A Ma, E. B. Fox, T. Chen, and L. Wu. Irreversible samplers from jump and continuous Markov processes. *Stat. Comput.*, pp. 1–26, 2018.

Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan. Is there an analog of Nesterov acceleration for MCMC? *Bernoulli*, 27:1942–1992, 2021.

Yi-An Ma, Tianqi Chen, and Emily B. Fox. A Complete Recipe for Stochastic Gradient MCMC. In *Neural Information Processing Systems (NeurIPS)*, 2015.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. Sampling Can Be Faster Than Optimization. *PNAS*, 2019.

Oren Mangoubi and Nisheeth K. Vishnoi. Convex Optimization with Unbounded Nonconvex Oracles using Simulated Annealing. In *Proc. of Conference on Learning Theory (COLT)*, 2018.

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

H. McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated Learning of Deep Networks using Model Averaging. 2016.

Khaoula el Mekkaoui, Diego Mesquita, Paul Blomstedt, and Samuel Kaski. Federated Stochastic Gradient Langevin Dynamics. In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021.

S. Minsker, S. Srivastava, L. Lin, and D. B. Dunson. Scalable and Robust Bayesian Inference via the Median Posterior. In *International Conference on Machine Learning (ICML)*, 2014.

Wenlong Mou, Yi-An Ma, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. High-Order Langevin Diffusion Yields an Accelerated MCMC Algorithm. *Journal of Machine Learning Research (JMLR)*, 22:1–41, 2021.

W. Neiswanger, C. Wang, and E. Xing. Asymptotically Exact, Embarrassingly Parallel MCMC. *arXiv:1311.4780*, 2013.

Y. Nesterov. *Introductory Lectures on Convex Optimization, in: Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

R. Nishihara, I. Murray, and R. P. Adams. Parallel MCMC with Generalized Elliptical Slice Sampling. *Journal of Machine Learning Research*, 15(1):2087–2112, 2014.

Reese Pathaky and Martin J. Wainwright. Fedsplit: An Algorithmic Framework for Fast Federated Optimization. *arXiv:2005.05238*, 2020.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex Learning via Stochastic Gradient Langevin Dynamics: a Nonasymptotic Analysis. In *Conference on Learning Theory*, June 2017.

R. Shen and Y. T. Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pp. 2098–2109, 2019.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving Deep Learning. In *SIGSAC conference on computer and communications security (CCS)*. ACM, 2015.

Sebastian U. Stich. Local SGD Converges Fast and Communicates Little. *arXiv:1805.09767v3*, 2019.

Yee Whye Teh, Alexandre Thiery, and Sebastian Vollmer. Consistency and Fluctuations for Stochastic Gradient Langevin Dynamics. *Journal of Machine Learning Research*, 17:1–33, 2016.

Sebastian J. Vollmer, Konstantinos C. Zygalakis, and Yee Whye Teh. Exploration of the (Non-) Asymptotic Bias and Variance of Stochastic Gradient Langevin Dynamics. *Journal of Machine Learning Research*, 17(159):1–48, 2016.

Maxime Vono, Vincent Plassier, Alain Durmus, Aymeric Dieuleveut, and Éric Moulines. QLSD: Quantised Langevin Stochastic Dynamics for Bayesian Federated Learning. *arXiv:2106.00797v2*, 2021.

Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive Federated Learning in Resource Constrained Edge Computing Systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019.

Xiangyu Wang and David B. Dunson. Parallelizing MCMC via Weierstrass Sampler.

Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for Free: Posterior Sampling and Stochastic Gradient Monte Carlo. In *ICML*, pp. 2493–2502, 2015.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *International Conference on Machine Learning*, 2011.

Blake Woodworth, Kumar Kshitij Patel, and Nathan Srebro. Minibatch vs Local SGD for Heterogeneous Distributed Learning. *arXiv:2006.04735*, 2020.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel Restarted SGD with Faster Convergence and Less Communication: Demystifying Why Model Averaging Works for Deep Learning. In *In Proc. of Conference on Artificial Intelligence (AAAI)*, 2019.

Xu Zhang, Yinchuan Li, Wenpeng Li, Kaiyang Guo, and Yunfeng Shao. Personalized Federated Learning via Variational Bayesian Inference. In *International Conference on Machine Learning (ICML)*, 2022.

Yuchen Zhang, Percy Liang, and Moses Charikar. A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics. In *Proc. of Conference on Learning Theory (COLT)*, pp. 1980–2022, 2017.

Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. Federated meta-learning for fraudulent credit card detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

APPENDIX

Roadmap. In Section A, we layout the formulation of the algorithm, basic notations, and definitions. In Section B, we present the main convergence analysis for full device participation. We discuss the optimal number of local updates based on a fixed learning rate, the acceleration achieved by varying learning rates, and the privacy-accuracy trade-off through correlated noises. In Section C, we analyze the convergence of partial device participation through two device-sampling schemes. In Section D, we provide lemmas to upper bound the contraction, discretization and divergence for proving the main convergence results. In Section E, we include supporting lemmas to prove results in the previous section. In Section F, we establish the initial condition. In Section G, we prove differential privacy guarantees. In Section H, we discuss the related work and literature. In Section I, we include additional experiments.

A PRELIMINARIES

A.1 BASIC NOTATIONS AND BACKGROUNDS

Let N denote the number of clients. Let T_ϵ denote the number of global steps to achieve the precision ϵ. Let K denote the number of local steps. For each $c \in [N] := \{1, 2, \cdots, N\}$, we use f^c and ∇f^c denote the loss function and gradient of the function f^c in client c. Notably, ∇f is not a standard gradient operator acting on f when multiple local steps are adopted ($K > 1$). For the stochastic gradient oracle, we denote by $\nabla \widetilde{f}^c(\cdot)$ the *unbiased* estimate of the exact gradient ∇f^c of client c. In addition, we denote p_c as the weight of the c-th client such that $p_c \geq 0$ and $\sum_{c=1}^N p_c = 1$. ξ_k^c is an independent standard d-dimensional Gaussian vector at iteration k for each client $c \in [N]$ and $\dot{\xi}_k$ is a unique Gaussian vector shared by all the clients.

Algorithm 2 Federated averaging Langevin dynamics algorithm (FA-LD). Denote by θ_k^c the model parameter in the c-th client at the k-th step. Denote the immediate result of one step SGLD update from θ_k^c by β_k^c. ξ_k^c is an independent standard d-dimensional Gaussian vector at iteration k for each client $c \in [N]$. A global synchronization is conducted every K steps. This is a transformed version of Algorithm 1 with $\rho = 0$ and full device participation for ease of analysis.

$$\beta_{k+1}^c = \theta_k^c - \eta \nabla \widetilde{f}^c(\theta_k^c) + \sqrt{2\eta\tau/p_c}\xi_k^c, \tag{14}$$

$$\theta_{k+1}^c = \begin{cases} \beta_{k+1}^c & \text{if } k+1 \bmod K \neq 0 \\ \sum_{c=1}^N p_c \beta_{k+1}^c & \text{if } k+1 \bmod K = 0. \end{cases} \tag{15}$$

Inspired by Li et al. (2020c), we define two virtual sequences

$$\beta_k = \sum_{c=1}^N p_c \beta_k^c, \qquad \theta_k = \sum_{c=1}^N p_c \theta_k^c, \tag{16}$$

which are *both inaccessible when* k *mod* $K \neq 0$. For the gradients and injected noise, we also define

$$\nabla f(\theta_k) = \sum_{c=1}^N p_c \nabla f^c(\theta_k^c), \quad \boldsymbol{\nabla} \widetilde{f}(\boldsymbol{\theta}_k) = \sum_{c=1}^N p_c \nabla \widetilde{f}^c(\theta_k^c), \quad \boldsymbol{\theta}_k = (\theta_k^1, \cdots, \theta_k^N), \quad \xi_k = \sum_{c=1}^N \sqrt{p_c}\xi_k^c. \tag{17}$$

In what follows, it is clear that $\mathbb{E}\boldsymbol{\nabla}\widetilde{f}(\boldsymbol{\theta}) = \sum_{c=1}^N p_c \mathbb{E}\nabla \widetilde{f}^c(\theta^c) = \boldsymbol{\nabla} f(\boldsymbol{\theta})$ for any $\theta^c \in \mathbb{R}^d$ and any $c \in [N]$. Summing Eq.(14) from clients $c = 1$ to N and combining Eq.(16) and Eq.(17), we have

$$\beta_{k+1} = \theta_k - \eta \boldsymbol{\nabla} \widetilde{f}(\boldsymbol{\theta}_k) + \sqrt{2\eta\tau}\xi_k. \tag{18}$$

Moreover, we always have $\beta_k = \theta_k$ whether $k + 1 \bmod K = 0$ or not by Eq.(15) and Eq.(16). In what follows, we can write

$$\theta_{k+1} = \theta_k - \eta \boldsymbol{\nabla} \widetilde{f}(\boldsymbol{\theta}_k) + \sqrt{2\eta\tau}\xi_k, \tag{19}$$

which resembles the SGLD algorithm (Welling & Teh, 2011) except that the construction of stochastic gradients is different and θ_k is *not accessible when* k *mod* $K \neq 0$. To facilitate the analysis, we also define an auxiliary continuous-time processes $(\bar{\theta}_t)_{t\geq 0}$

$$\mathrm{d}\bar{\theta}_t = -\nabla f(\bar{\theta}_t)\cdot \mathrm{d}t + \sqrt{2\tau}\cdot \mathrm{d}\overline{W}_t, \tag{20}$$

where $\bar{\theta}_t = \sum_{c=1}^N p_c \bar{\theta}_t^c$, $\nabla f(\bar{\theta}_t) = \sum_{c=1}^N p_c \nabla f^c(\bar{\theta}_t^c)$, $\bar{\theta}_t^c$ is the continuous-time variable at client c, and $\overline{W}$ is a d-dimensional Brownian motion. The continuous-time algorithm is referred to as Federated Averaging Langevin diffusion and is described as

$$\mathrm{d}\bar{\beta}_t^c = -\nabla f^c(\bar{\theta}_t^c)\cdot \mathrm{d}t + \sqrt{2\tau/p_c}\cdot \mathrm{d}\overline{W}_t^c$$
$$\bar{\theta}_t^c = \sum_{c=1}^N p_c \bar{\beta}_t^c.$$

Since the synchronization step is conducted at every time step t, the Federated Averaging Langevin diffusion performs the same as the standard Langevin diffusion with the temperature τ and convergences to the stationary distribution $\pi(\theta) \propto \exp(-f(\theta)/\tau)$, where $f(\theta) = \sum_{c=1}^N p_c f^c(\theta)$. Assume that $\bar{\theta}_0$ simulates from the stationary distribution π, then it follows that $\bar{\theta}_t \sim \pi$ for any $t \geq 0$.

A.2 Assumptions and definitions

Assumption A.1 (Smoothness, restatement of Assumption 4.1)**.** *For each $c \in [N]$, we say f^c is L-smooth if for some $L > 0$*

$$f^c(y) \leq f^c(x) + \langle \nabla f^c(x), y - x\rangle + \frac{L}{2}\|y-x\|_2^2 \quad \forall x, y \in \mathbb{R}^d.$$

Note that the above assumption is equivalent to saying that

$$\|\nabla f^c(y) - \nabla f^c(x)\|_2 \leq L\|y-x\|_2, \quad \forall x, y \in \mathbb{R}^d.$$

Assumption A.2 (Strong convexity, restatement of Assumption 4.2)**.** *For each $c \in [N]$, f^c is m-strongly convex if for some $m > 0$*

$$f^c(x) \geq f^c(y) + \langle \nabla f^c(y), x - y\rangle + \frac{m}{2}\|y-x\|_2^2 \quad \forall x, y \in \mathbb{R}^d.$$

An alternative formulation for strong convexity is that

$$\langle \nabla f^c(x) - \nabla f^c(y), x - y\rangle \geq m\,\|x-y\|_2^2 \quad \forall x, y \in \mathbb{R}^d.$$

Assumption A.3 (Bounded variance, restatement of Assumption 4.3)**.** *For each $c \in [N]$, the variance of noise in the stochastic gradient $\nabla \widetilde{f}^c(x)$ in each client is upper bounded such that*

$$\mathbb{E}[\|\nabla \widetilde{f}^c(x) - \nabla f^c(x)\|_2^2] \leq \sigma^2 d, \quad \forall x \in \mathbb{R}^d.$$

The bounded variance in the stochastic gradient is a rather standard assumption and has been widely used in Cheng et al. (2018); Dalalyan & Karagulyan (2019); Li et al. (2020c). Extension of bounded variance to unbounded cases such as $\mathbb{E}[\|\nabla \widetilde{f}^c(x) - \nabla f^c(x)\|_2^2] \leq \delta(L^2x^2 + B^2)$ for some M and $\delta \in [0, 1)$ is quite straightforward and has been adopted in assumption A.4 stated in Raginsky et al. (2017). The proof framework remains the same.

Quality of non-i.i.d data Denote by θ_* the global minimum of f. Next, we quantify the degree of the non-i.i.d data by $\gamma := \max_{c\in[N]} \|\nabla f^c(\theta_*)\|_2$, which is a non-negative constant and yields a smaller scale if the data is more evenly distributed.

Definition A.4. *We define parameter $T_{c,\rho}$ H_ρ, κ and γ^2*

$$T_{c,\rho} := \tau(\rho^2 + (1-\rho^2)/p_c),$$
$$H_\rho := \underbrace{\mathcal{D}^2}_{\text{initialization}} + \underbrace{\frac{1}{m}\max_{c\in[N]} T_{c,\rho}}_{\text{injected noise}} + \underbrace{\frac{\gamma^2}{m^2 d}}_{\text{data heterogeneity}} + \underbrace{\frac{\sigma^2}{m^2}}_{\text{stochastic noise}},$$
$$\kappa := L/m,$$
$$\gamma^2 := \max_{c\in[N]} \|\nabla f^c(\theta_*)\|_2^2.$$

B FULL DEVICE PARTICIPATION

B.1 ONE-STEP UPDATE

Wasserstein distance We define the 2-Wasserstein distance between a pair of Borel probability measures μ and ν on $\mathbb{R}^d$ as follows

$$W_2(\mu,\nu) := \inf_{\gamma^2 \in \text{Couplings}(\mu,\nu)} \left(\int \|\boldsymbol{\beta}_\mu - \boldsymbol{\beta}_\nu\|_2^2 d\gamma^2(\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\nu) \right)^{\frac{1}{2}},$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm on $\mathbb{R}^d$ and the pair of random variables $(\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\nu) \in \mathbb{R}^d \times \mathbb{R}^d$ is a coupling with the marginals following $\mathcal{L}(\boldsymbol{\beta}_\mu) = \mu$ and $\mathcal{L}(\boldsymbol{\beta}_\nu) = \nu$. $\mathcal{L}(\cdot)$ denotes a distribution of a random variable.

The following result provides a crucial contraction property based on distributed clients with infrequent synchronizations.

Lemma B.1 (Dominated contraction property, restatement of Lemma 4.4)**.** *Assume assumptions A.1 and A.2 hold. For any learning rate $\eta \in (0, \frac{1}{L+m}]$, any $\{\theta^c\}_{c=1}^N, \{\beta^c\}_{c=1}^N \in \mathbb{R}^d$, we have*

$$\|\beta - \theta - \eta(\nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta}))\|_2^2 \le (1-\eta m)\cdot \|\beta-\theta\|_2^2 + 4\eta L \sum_{c=1}^N p_c \cdot (\|\beta^c - \beta\|_2^2 + \|\theta^c - \theta\|_2^2).$$

where $\beta = \sum_{c=1}^N p_c \beta^c$, $\theta = \sum_{c=1}^N p_c \theta^c$, $\boldsymbol{\nabla} f(\boldsymbol{\theta}) = \sum_{c=1}^N p_c \nabla f^c(\theta^c)$, and $\nabla f(\beta) = \sum_{c=1}^N p_c \nabla f^c(\beta^c)$. We postpone the proof into Section D.1. The above result implies that as long as the local parameters θ^c, β^c and global θ, β don't differ each other too much, we can guarantee the desired convergence.

The following result ensures a bounded gap between $\bar{\theta}_s^c$ and $\bar{\theta}^c_{\eta\lfloor \frac{s}{\eta} \rfloor}$ in ℓ_2 norm for any $s \ge 0$ and $c \in [N]$. We postpone the proof of Lemma B.2 into Section D.2.

Lemma B.2 (Discretization error)**.** *Assume assumptions A.1, A.2, and A.3 hold. For any $s \ge 0$, any learning rate $\eta \in (0, 2/m)$ and $\|\theta_0^c - \theta_*\|_2^2 \le d\mathcal{D}^2$ for any $c \in [N]$, the iterates of $(\bar{\theta}_s)$ based on the continuous dynamics of Eq.(20) satisfy the following estimate*

$$\mathbb{E}\|\bar{\theta}_s^c - \bar{\theta}^c_{\eta\lfloor \frac{s}{\eta} \rfloor}\|_2^2 \le 8\eta^2 d\kappa \left(\frac{\kappa\gamma^2}{d} + L\tau \right) + 16\eta d\tau / p_c.$$

The following result shows that given a finite number of local steps K, the divergence between θ^c in local client and θ in the center is bounded in ℓ_2 norm. Notably, since the non-differentiable Brownian motion leads to a lower order term $O(\eta)$ instead of $O(\eta^2)$ in ℓ^2 norm, a naïve proof may lead to a crude upper bound. We delay the proof of Lemma B.3 into Section D.3.

Lemma B.3 (Bounded divergence, restatement of Lemma 4.5)**.** *Assume assumptions A.1, A.2, and A.3 hold. For any learning rate $\eta \in (0, 2/m)$ and $\|\theta_0^c - \theta_*\|_2^2 \le d\mathcal{D}^2$ for any $c \in [N]$, we have the ℓ_2 upper bound of the divergence between local clients and the center as follows*

$$\sum_{c=1}^N p_c \mathbb{E}\|\theta_k^c - \theta_k\|_2^2 \le 112(K-1)^2\eta^2 d L^2 H_\rho + 8(K-1)\eta d\tau(\rho^2 + N(1-\rho^2)),$$

where H_ρ, κ and γ^2 are defined as Definition A.4.

The following presents a standard result for bounding the gap between $\boldsymbol{\nabla} f(\boldsymbol{\theta})$ and $\boldsymbol{\nabla} \widetilde{f}(\boldsymbol{\theta})$. We delay the proof of Lemma B.4 into Setion D.

Lemma B.4 (Bounded variance)**.** *Given assumption A.3, we have*

$$\mathbb{E}\|\boldsymbol{\nabla} f(\boldsymbol{\theta}) - \boldsymbol{\nabla} \widetilde{f}(\boldsymbol{\theta})\|_2^2 \le d\cdot\sigma^2, \qquad \forall\, \theta \in \mathbb{R}^d.$$

Having all the preliminary results ready, now we present a crucial lemma for proving the convergence of all the algorithms.

Lemma B.5 (One step update, restatement of Lemma 4.6)**.** *Assume assumptions A.1, A.2, and A.3 hold. Consider Algorithm 2 with independently injected noise $\rho = 0$, any learning rate $\eta \in (0, \frac{1}{2L})$ and $\|\theta_0^c - \theta_*\|_2^2 \leq d\mathcal{D}^2$ for any $c \in [N]$, where θ_* is the global minimum for the function f. Then*

$$W_2^2(\mu_{k+1}, \pi) \leq \left(1 - \frac{\eta m}{2}\right) \cdot W_2^2(\mu_k, \pi) + 400\eta^2 dL^2 H_0((K-1)^2 + \kappa),$$

where μ_k denotes the probability measure of θ_k, H_0, κ and γ^2 are defined as Definition A.4.

Proof. The solution of the continuous-time process Eq.(20) follows that

$$\bar{\theta}_t = \bar{\theta}_0 - \int_0^t \nabla f(\bar{\theta}_s)\mathrm{d}s + \sqrt{2\tau} \cdot \overline{W}_t, \qquad \forall t \geq 0. \tag{21}$$

Set $t \to (k+1)\eta$ and $\bar{\theta}_0 \to \bar{\theta}_{k\eta}$ for Eq.(21) and consider a synchronous coupling such that $W_{(k+1)\eta} - W_{k\eta} := \sqrt{\eta}\xi_k$

$$\begin{aligned}
\bar{\theta}_{(k+1)\eta} &= \bar{\theta}_{k\eta} - \int_{k\eta}^{(k+1)\eta} \nabla f(\bar{\theta}_s)\mathrm{d}s + \sqrt{2\tau}(W_{(k+1)\eta} - W_{k\eta}) \\
&= \bar{\theta}_{k\eta} - \int_{k\eta}^{(k+1)\eta} \nabla f(\bar{\theta}_s)\mathrm{d}s + \sqrt{2\tau\eta}\xi_k.
\end{aligned} \tag{22}$$

We first denote $\zeta_k := \boldsymbol{\nabla}\widetilde{f}(\boldsymbol{\theta}_k) - \boldsymbol{\nabla} f(\boldsymbol{\theta}_k)$. Subtracting Eq.(19) from Eq.(22) yields that

$$\begin{aligned}
&\bar{\theta}_{(k+1)\eta} - \theta_{k+1} \\
&= \bar{\theta}_{k\eta} - \theta_k + \eta\boldsymbol{\nabla}\widetilde{f}(\boldsymbol{\theta}_k) - \int_{k\eta}^{(k+1)\eta} \nabla f(\bar{\theta}_s)\mathrm{d}s \\
&= \bar{\theta}_{k\eta} - \theta_k - \eta\Big(\nabla f(\theta_k + \bar{\theta}_{k\eta} - \theta_k) - \boldsymbol{\nabla}\widetilde{f}(\boldsymbol{\theta}_k)\Big) - \int_{k\eta}^{(k+1)\eta} \Big(\nabla f(\bar{\theta}_s) - \nabla f(\bar{\theta}_{k\eta})\Big)\mathrm{d}s \qquad (23) \\
&= \bar{\theta}_{k\eta} - \theta_k - \eta\Big(\underbrace{\nabla f(\theta_k + \bar{\theta}_{k\eta} - \theta_k) - \boldsymbol{\nabla} f(\boldsymbol{\theta}_k)}_{:=X_k}\Big) - \underbrace{\int_{k\eta}^{(k+1)\eta} \Big(\nabla f(\bar{\theta}_s) - \nabla f(\bar{\theta}_{k\eta})\Big)\mathrm{d}s}_{:=Y_k} + \eta\zeta_k.
\end{aligned}$$

Taking square and expectation on both sides, we have

$$\begin{aligned}
&\mathbb{E}\|\bar{\theta}_{(k+1)\eta} - \theta_{k+1}\|_2^2 \\
&= \mathbb{E}\|\bar{\theta}_{k\eta} - \theta_k - \eta X_k - Y_k\|_2^2 + \mathbb{E}\|\eta\zeta_k\|_2^2 + 2\eta \underbrace{\mathbb{E}\langle \bar{\theta}_{k\eta} - \theta_k - \eta X_k - Y_k, \zeta_k\rangle}_{\mathbb{E}\zeta_k = 0 \text{ and mutual independence}} \\
&\leq (1+q) \cdot \mathbb{E}\|\bar{\theta}_{k\eta} - \theta_k - \eta X_k\|_2^2 + (1 + 1/q) \cdot \mathbb{E}\|Y_k\|_2^2 + \mathbb{E}\|\eta\zeta_k\|_2^2 \\
&\leq (1+q) \cdot \Big((1 - \eta m) \cdot \mathbb{E}\|\bar{\theta}_{k\eta} - \theta_k\|_2^2 + 4\eta L \sum_{c=1}^N p_c \cdot \big(\mathbb{E}\|\bar{\theta}_{k\eta}^c - \bar{\theta}_{k\eta}\|_2^2 + \mathbb{E}\|\theta_k^c - \theta_k\|_2^2\big)\Big) \\
&\qquad + (1 + 1/q) \cdot \mathbb{E}\|Y_k\|_2^2 + \eta^2\sigma^2 d \\
&\leq (1+q) \cdot \Bigg(\underbrace{(1 - \eta m)}_{\phi} \mathbb{E}\|\bar{\theta}_{k\eta} - \theta_k\|_2^2 + 448\eta^3 d(K-1)^2 L^3 H_0 + 32(K-1)\eta^2 dL\tau N\Bigg) \\
&\qquad + (1 + 1/q) \cdot \mathbb{E}\|Y_k\|_2^2 + \eta^2\sigma^2 d,
\end{aligned} \tag{24}$$

where the first inequality follows by the AM-GM inequality for any $q > 0$, the second inequality follows by Lemma B.1 and Assumption A.3. The third inequality follows by Lemma B.3 with $\rho = 0$; moreover, the continuous-time process conducts synchronization at any time step, hence $\bar{\theta}_{k\eta}^c = \bar{\theta}_{k\eta}$. Since the learning rate follows $\frac{1}{2L} \leq \frac{1}{m+L} \leq \frac{2}{m}$, the requirement of the learning rate for Lemma B.1 and Lemma B.3 is clearly satisfied.

Recall that $\phi = 1 - \eta m$, we get $\frac{1+\phi}{2} = 1 - \frac{1}{2}\eta m$. Choose $q = \frac{1+\phi}{2\phi} - 1$ so that $(1+q)\phi = \frac{(1+\phi)}{2} = 1 - \frac{1}{2}\eta m$. In addition, we have $1 + \frac{1}{q} = \frac{1+q}{q} = \frac{1+\phi}{1-\phi} \leq \frac{2}{\eta m}$. It follows that

$$(1+q)\cdot(1-\eta m) \leq 1 - \frac{1}{2}\eta m, \quad 1+q \leq \frac{1-\frac{1}{2}\eta m}{1-\eta m} \leq 1.5, \quad (1+1/q) \leq \frac{2}{m\eta}, \tag{25}$$

where the second inequality holds because $\eta \in (0, \frac{1}{2L}] \leq \frac{1}{2m}$.

For the term $\mathbb{E}\|Y_k\|_2^2$ in Eq.(24), we have the following estimate

$$\begin{aligned}
\mathbb{E}\|Y_k\|_2^2 &= \mathbb{E}\left\|\int_{k\eta}^{(k+1)\eta}\Big(\nabla f(\bar{\theta}_s) - \nabla f(\bar{\theta}_{k\eta})\Big)\mathrm{d}s\right\|_2^2 \\
&\leq \eta\int_{k\eta}^{(k+1)\eta}\mathbb{E}\big\|\nabla f(\bar{\theta}_s) - \nabla f(\bar{\theta}_{k\eta})\big\|_2^2\mathrm{d}s \\
&= \eta\int_{k\eta}^{(k+1)\eta}\mathbb{E}\left\|\sum_{c=1}^{N}p_c\Big(\nabla f^c(\bar{\theta}_s^c) - \nabla f^c(\bar{\theta}_{k\eta}^c)\Big)\right\|_2^2\mathrm{d}s \\
&\leq \eta\int_{k\eta}^{(k+1)\eta}\sum_{c=1}^{N}p_c\cdot\mathbb{E}\big\|\nabla f^c(\bar{\theta}_s^c) - \nabla f^c(\bar{\theta}_{k\eta}^c)\big\|_2^2\mathrm{d}s \\
&\leq \eta L^2\int_{k\eta}^{(k+1)\eta}\sum_{c=1}^{N}p_c\cdot\mathbb{E}\big\|\bar{\theta}_s^c - \bar{\theta}_{k\eta}^c\big\|_2^2\mathrm{d}s \\
&\leq \eta L^2\int_{k\eta}^{(k+1)\eta}\left(8\eta^2 d\kappa\left(\frac{\kappa\gamma^2}{d} + L\tau\right) + 16\eta d\tau N\right)\mathrm{d}s \\
&= 8\eta^4 dL^4 H_0 + 16\eta^3 L^2 d\tau N,
\end{aligned} \tag{26}$$

where the first inequality follows by Hölder's inequality, the second inequality follows by Jensen's inequality, the third inequality follows by Assumption A.1, and the last inequality follows by Lemma B.2. The last equality holds since $\frac{\kappa}{d}\gamma^2 + L\tau \leq LmH_0$ and $\kappa = L/m$.

Plugging Eq.(25) and Eq.(26) into Eq.(24), we have

$$\begin{aligned}
\mathbb{E}\|\bar{\theta}_{(k+1)\eta} - \theta_{k+1}\|_2^2 \leq (1 - \frac{\eta m}{2})\cdot\mathbb{E}\|\bar{\theta}_{k\eta} - \theta_k\|_2^2 \\
+ 672\eta^3 d(K-1)^2 L^3 H_0 + 48\eta^2 d(K-1)L\tau N \\
+ 16\eta^3 dL^3\kappa H_0 + 32\eta^2 d\frac{L^2}{m}\tau N + \eta^2\sigma^2 d.
\end{aligned}$$

Choose the specific Langevin diffusion $\bar{\theta}$ in stationary regime, we have $W_2^2(\mu_k, \pi) = \mathbb{E}\|\bar{\theta}_{k\eta} - \theta_k\|_2^2$ and $W_2^2(\mu_{k+1}, \pi) \leq \mathbb{E}\|\bar{\theta}_{(k+1)\eta} - \theta_{k+1}\|_2^2$. Arranging the terms, we have

$$W_2^2(\mu_{k+1}, \pi) \leq (1 - \frac{\eta m}{2})\cdot W_2^2(\mu_k, \pi) + 400\eta^2 dL^2 H_0((K-1)^2 + \kappa),$$

where $\eta \leq \frac{1}{2L}$, $\kappa \geq 1$, $m\tau \leq L\tau \leq L\tau N \leq L\max_{c\in[N]} T_{c,0} \leq LmH_0$, and $\sigma^2 \leq L^2 H_0$ are applied to the result.

□

B.2 Convergence via independent noises

Theorem B.6 (Restatement of Theorem 4.7)**.** *Assume assumptions A.1, A.2, and A.3 hold. Consider Algorithm 2 with a fixed learning rate $\eta \in (0, \frac{1}{2L}]$ and $\|\theta_0^c - \theta_*\|_2^2 \leq d\mathcal{D}^2$ for any $c \in [N]$, we have*

$$W_2(\mu_k, \pi) \leq \left(1 - \frac{\eta m}{4}\right)^k\cdot\left(\sqrt{2d}\big(\mathcal{D} + \sqrt{\tau/m}\big)\right) + 30\kappa\sqrt{\eta m d}\cdot\sqrt{((K-1)^2 + \kappa)H_0}.$$

where μ_k denotes the probability measure of θ_k, H_0, κ and γ^2 are defined as Definition A.4.

Proof. Iteratively applying Theorem B.5 and arranging terms, we have that

$$\begin{aligned} W_2^2(\mu_k,\pi) &\le \left(1-\frac{\eta m}{2}\right)^k W_2^2(\mu_0,\pi) + \frac{1-(1-\frac{\eta m}{2})^k}{1-(1-\frac{\eta m}{2})}\left(400\eta^2 dL^2 H_0((K-1)^2+\kappa)\right) \\ &\le \left(1-\frac{\eta m}{2}\right)^k W_2^2(\mu_0,\pi) + \frac{2}{\eta m}\left(400\eta^2 dL^2 H_0((K-1)^2+\kappa)\right) \\ &\le \left(1-\frac{\eta m}{2}\right)^k W_2^2(\mu_0,\pi) + 800\kappa^2\eta m d((K-1)^2+\kappa)H_0, \end{aligned} \tag{27}$$

where $\kappa = \frac{L}{m}$. By Lemma F.1 and the initialization condition $\|\theta_0^c - \theta_*\|_2^2 \le d\mathcal{D}^2$ for any $c \in [N]$, we have that

$$W_2(\mu_0,\pi) \le \sqrt{2d}(\mathcal{D} + \sqrt{\tau/m}).$$

Applying the inequality $(1-\frac{\eta m}{2}) \le (1-\frac{\eta m}{4})^2$ completes the proof. □

Discussions

Optimal choice of K. To ensure the algorithm to achieve the ϵ precision based on the total number of steps T_ϵ and the learning rate η, we can set

$$\begin{aligned} &30\kappa\sqrt{\eta m d}\cdot\left(\sqrt{((K-1)^2+\kappa)H_0}\right) \le \frac{\epsilon}{2} \\ &e^{-\frac{\eta m}{4}T_\epsilon}\cdot\left(\sqrt{2d}(\mathcal{D}+\sqrt{\tau/m})\right) \le \frac{\epsilon}{2}. \end{aligned}$$

This directly leads to

$$\eta m \le \min\left\{\frac{m}{2L}, O\left(\frac{\epsilon^2}{d\kappa^2((K-1)^2+\kappa)H_0}\right)\right\}, \quad T_\epsilon \ge \Omega\left(\frac{\log\left(\frac{d}{\epsilon}\right)}{m\eta}\right).$$

Plugging into the upper bound of η, it implies that to reach the precision level ϵ, it suffices to set

$$T_\epsilon = \Omega\left(\frac{d\kappa^2((K-1)^2+\kappa)H_0}{\epsilon^2}\cdot\log\left(\frac{d}{\epsilon}\right)\right). \tag{28}$$

Since $H_0 = \Omega(\mathcal{D}^2 + \frac{\tau}{m})$, we observe that the number of communication rounds is around the order

$$\frac{T_\epsilon}{K} = \Omega\left(K + \frac{\kappa}{K}\right),$$

where the value of $\frac{T_\epsilon}{K}$ first decreases and then increases with respect to K, indicating that setting K either too large or too small may lead to high communication costs and hurt the performance. Ideally, K should be selected in the scale of $\Omega(\sqrt{\kappa})$. Combining the definition of T_ϵ in Eq.(28), this suggests an interesting result that the optimal K should be in the order of $O(\sqrt{T_\epsilon})$. Similar results have been achieved by Stich (2019); Li et al. (2020c).

B.3 Convergence via varying learning rates

Theorem B.7 (Restatement of Theorem 4.8)**.** *Assume assumptions A.1, A.2, and A.3 hold. Consider Algorithm 2 with an initialization satisfying $\|\theta_0^c - \theta_*\|_2^2 \le d\mathcal{D}^2$ for any $c \in [N]$ and varying learning rate following*

$$\eta_k = \frac{1}{2L + (1/12)mk}, \qquad k = 1, 2, \cdots.$$

Then for any $k \ge 0$, we have

$$W_2(\mu_k,\pi) \le 45\kappa\sqrt{((K-1)^2+\kappa)H_0}\cdot\left(\eta_k m d\right)^{1/2}, \qquad \forall k \ge 0,$$

Proof. We first denote

$$C_\kappa = 30\kappa\sqrt{((K-1)^2+\kappa)H_0}.$$

Next, we proceed to show the following inequality by the induction method

$$W_2(\mu_k,\pi) \le 1.5C_\kappa\left(\frac{d}{2L+(1/12)mk}\right)^{1/2} = 1.5C_\kappa\big(\eta_k md\big)^{1/2}, \qquad \forall k\ge 0, \tag{29}$$

where the decreasing learning rate follows that

$$\eta_k = \frac{1}{2L+(1/12)mk}.$$

(i) For the case of $k=0$, since

$$\begin{aligned} C_\kappa &\ge 4\sqrt{\kappa}\sqrt{H_0} \ge 4\sqrt{\kappa}\sqrt{\mathcal{D}^2+\frac{1}{m}\max_{c\in[N]}T_{c,0}} \ge 4\sqrt{\kappa/d}\left(\sqrt{d\mathcal{D}^2}+\sqrt{\frac{d}{m}\max_{c\in[N]}T_{c,0}}\right)\\ &\ge 4\sqrt{\kappa/d}W_2(\mu_0,\pi), \end{aligned} \tag{30}$$

where the last inequality follows by Lemma F.1 and $\|\theta_0^c-\theta_*\|_2^2 \le d\mathcal{D}^2$ for any $c\in[N]$.

It is clear that $W_2(\mu_0,\pi)\le \frac{1}{4}C_\kappa\sqrt{\frac{md}{L}} \le 1.5C_\kappa\sqrt{\eta_0 md}$ by Eq.(30).

(ii) If now that Eq.(29) holds for some $k\ge 0$, it follows by Lemma B.5 that

$$\begin{aligned} W_2^2(\mu_{k+1},\pi) &\le \big(1-\frac{\eta_k m}{2}\big)\cdot W_2^2(\mu_k,\pi) + 400\eta_k^2 dL^2H_0((K-1)^2+\kappa)\\ &\le \big(1-\frac{\eta_k m}{2}\big)\cdot W_2^2(\mu_k,\pi) + \frac{\eta_k^2m^2}{2}C_\kappa^2 d\\ &\le \big(1-\frac{\eta_k m}{2}\big)\cdot 2.25C_\kappa^2\eta_k md + \frac{\eta_k m}{3}2.25C_\kappa^2\eta_k md\\ &\le \big(1-\frac{\eta_k m}{6}\big)\cdot 2.25C_\kappa^2\eta_k md. \end{aligned}$$

Since $\big(1-\frac{\eta_k m}{6}\big)\le\big(1-\frac{\eta_k m}{12}\big)^2$, we have

$$W_2(\mu_{k+1},\pi)\le\big(1-\frac{\eta_k m}{12}\big)\cdot 1.5C_\kappa\big(\eta_k md\big)^{1/2}.$$

To prove $W_2(\mu_{k+1},\pi)\le 1.5C_\kappa\big(\eta_{k+1}md\big)^{1/2}$, it suffices to show $\big(1-\frac{\eta_k m}{12}\big)\eta_k^{1/2}\le\eta_{k+1}$, which is detailed as follows

$$\begin{aligned} \big(1-\frac{\eta_k m}{12}\big)\eta_k^{1/2} &= \frac{\sqrt{12}(24L+mk-m)}{(24L+mk)^{3/2}}\\ &\le \frac{\sqrt{12}(24L+mk-m)^{1/2}}{24L+mk}\\ &\le \frac{\sqrt{12}}{(24L+m(k+1))^{1/2}} := \eta_{k+1}, \end{aligned}$$

where the last inequality follows since

$$(24L+mk-m)(24L+mk+m)) \le (24L+mk)^2.$$

□

The above result implies that to achieve the precision ϵ, we require

$$W_2(\mu_k,\pi)\le 1.5C_\kappa\left(\frac{md}{2L+(1/12)mk}\right)^{1/2}\le\epsilon.$$

The means that we only require $k=\Omega(\frac{d}{\epsilon^2})$ to achieve the precision ϵ. By contrast, the fixed learning rate requires $T_\epsilon=\Omega\Big(\frac{d}{\epsilon^2}\cdot\log\big(d/\epsilon\big)\Big)$, which is much slower than the algorithm with varying learning rate by $O\big(\log\big(d/\epsilon\big)\big)$ times.

B.4 PRIVACY-ACCURACY TRADE-OFF VIA CORRELATED NOISES

Note that Algorithm 2 requires all the local clients to generate the independent noise ξ_k^c. Such a mechanism enjoys the convenience of the implementation and yields a potential to protect the privacy of data and alleviates the security issue. However, the scale of noises is maximized and inevitable slows down the convergence. For extensions, it can be naturally generalized to correlated noise based on a hyperparameter, namely the correlation coefficient ρ between different clients. Replacing Eq.(14) with

$$\beta_{k+1}^c = \theta_k^c - \eta \nabla \widetilde{f}^c(\theta_k^c) + \sqrt{2\eta\tau\rho^2}\dot{\xi}_k + \sqrt{2\eta(1-\rho^2)\tau/p_c}\xi_k^c, \tag{31}$$

where $\dot{\xi}_k$ is a d-dimensional standard Gaussian vector shared by all the clients at iteration k, ξ_k^c is a unique d-dimensional Gaussian vector generated by client $c \in [N]$ only. Moreover, $\dot{\xi}_k$ is dependent with ξ_k^c for any $c \in [N]$. Following the same synchronization step based Eq.(15), we have

$$\theta_{k+1} = \theta_k - \eta \boldsymbol{\nabla} \widetilde{f}(\boldsymbol{\theta}_k) + \sqrt{2\eta\tau}\xi_k, \tag{32}$$

where $\xi_k = \rho\xi_k + \sqrt{1-\rho^2}\sum_{c=1}^N \sqrt{p_c}\xi_k^c$. Since the variance of i.i.d variables is additive, it is clear that ξ_k follows the standard d-dimensional Gaussian distribution. The inclusion of the correlated noise implicitly reduces the temperature and naturally yields a trade-off between federation and accuracy. We refer to the algorithm with correlated noise as the hybrid Federated Averaging Langevin dynamics (gFA-LD) and present it in Algorithm 3.

Since the inclusion of correlated noise doesn't affect the formulation of Eq.(32), the algorithm property maintains the same except the scale of the temperature τ and federation are changed. Based on a target correlation coefficient $\rho \geq 0$, Eq.(31) is equivalent to applying a temperature $T_{c,\rho} = \tau(\rho^2 + (1-\rho^2)/p_c)$. In particular, setting $\rho = 0$ leads to $T_{c,0} = (1-\rho^2)/p_c$, which exactly recovers Algorithm 2; however, setting $\rho = 1$ leads to $T_{c,1} = \tau$, where the injected noise in local clients is reduced by $1/p_c$ times. Now we adjust the analysis as follows

Theorem B.8 (Restatement of Theorem 4.9)**.** *Assume assumptions A.1, A.2, and A.3 hold. Consider Algorithm 3 with a correlation coefficient $\rho \in [0,1]$, a fixed learning rate $\eta \in (0, \frac{1}{2L}]$ and $\|\theta_0^c - \theta_*\|_2^2 \leq d\mathcal{D}^2$ for any $c \in [N]$, we have*

$$W_2(\mu_k, \pi) \leq \left(1 - \frac{\eta m}{4}\right)^k \cdot \left(\sqrt{2d}\big(\mathcal{D} + \sqrt{\tau/m}\big)\right) + 30\kappa\sqrt{\eta m d} \cdot \sqrt{((K-1)^2 + \kappa)H_\rho},$$

where μ_k denotes the probability measure of θ_k, H_ρ, κ and γ^2 are defined as Definition A.4.

Proof. The proof follows the same techniques as in Theorem B.6 except that H_0 is generalized to H_ρ to accommodate to the changes of the *injected noise*. The details are omitted. □

Algorithm 3 Hybrid federated averaging Langevin dynamics algorithm. Denote by θ_k^c the model parameter in the c-th client at the k-th step. Denote the immediate result of one step SGLD update from θ_k^c by β_k^c. ξ_k^c is an independent standard d-dimensional Gaussian vector at iteration k for each client $c \in [N]$ and $\dot{\xi}_k$ is a d-dimensional standard Gaussian vector shared by all the clients. ρ denotes the correlation coefficient of the injected noises. A global synchronization is conducted every K steps. This is a clean version of Algorithm 1 based on full device updates for ease of analysis.

$$\beta_{k+1}^c = \theta_k^c - \eta \nabla \widetilde{f}^c(\theta_k^c) + \sqrt{2\eta\tau\rho^2}\dot{\xi}_k + \sqrt{2\eta(1-\rho^2)\tau/p_c}\xi_k^c,$$

$$\theta_{k+1}^c = \begin{cases} \beta_{k+1}^c & \text{if } k+1 \bmod K \neq 0 \\ \sum_{c=1}^N p_c \beta_{k+1}^c & \text{if } k+1 \bmod K = 0. \end{cases}$$

C PARTIAL DEVICE PARTICIPATION

Full device participation enjoys appealing convergence properties. However, it suffers from the straggler's effect in real-world applications, where the communication is limited by the slowest device. Partial device participation handles this issue by only allowing a small portion of devices in each communication and greatly increased the communication efficiency in a federated network.

C.1 UNBIASED SAMPLING SCHEMES

The first device-sampling scheme I (Li et al., 2020b) selects a total of S devices, where the c-th device is selected with a probability p_c. The first theoretical justification for convex optimization has been proposed by Li et al. (2020c).

(Scheme I: with replacement). Assume $\mathcal{S}_k = \{n_1, n_2, \cdots, n_S\}$, where $n_j \in [N]$ is a random number that takes a value of c with a probability p_c for any $j \in \{1, 2, \cdots, S\}$. The synchronization step follows that $\theta_k = \frac{1}{S}\sum_{c\in\mathcal{S}_k}\theta_k^c$.

Another strategy is to uniformly select S devices without replacement. We follow Li et al. (2020c) and assume S indices are selected uniformly without replacement and the synchronization step is the same as before. In addition, the convergence also requires an additional assumption on balanced data (Li et al., 2020c).

(Scheme II: without replacement). Assume $\mathcal{S}_k = \{n_1, n_2, \cdots, n_S\}$, where $n_j \in [N]$ is a random number that takes a value of c with a probability $\frac{1}{S}$ for any $j \in \{1, 2, \cdots, S\}$. Assume the data is balanced such that $p_1 = \cdots = p_N = \frac{1}{N}$. The synchronization step follows that $\theta_k = \frac{N}{S}\sum_{c\in\mathcal{S}_k} p_c\theta_k^c = \frac{1}{S}\sum_{c\in\mathcal{S}_k}\theta_k^c$.

Algorithm 4 Hybrid federated averaging Langevin dynamics algorithm with partial device participation. ξ_k^c is the independent Gaussian vector proposed by each client $c \in [N]$ and $\dot{\xi}_k$ is a unique Gaussian vector shared by all the clients. ρ denotes the correlation coefficient. A global synchronization is conducted every K steps. $\mathcal{S}_k$ is a subset that contains S indices according to a device-sampling rule based on scheme I or II. This is a clean version of Algorithm 1 for ease of analysis.

$$\beta_{k+1}^c = \theta_k^c - \eta\nabla\widetilde{f}^c(\theta_k^c) + \sqrt{2\eta\tau\rho^2}\dot{\xi}_k + \sqrt{2\eta(1-\rho^2)\tau/p_c}\xi_k^c,$$

$$\theta_{k+1}^c = \begin{cases} \beta_{k+1}^c & \text{if } k+1 \bmod K \neq 0 \\ \sum_{c\in\mathcal{S}_{k+1}}\frac{1}{S}\beta_{k+1}^c & \text{if } k+1 \bmod K = 0. \end{cases}$$

Lemma C.1 (Unbiased sampling scheme)**.** *For any k mod $K = 0$ based on scheme I or II, we have*

$$\mathbb{E}\theta_k = \mathbb{E}\sum_{c\in\mathcal{S}_k}\theta_k^c = \beta_k := \sum_{c=1}^N p_c\beta_k^c.$$

Proof. According to the definition of scheme I or II, we have $\theta_k = \frac{1}{S}\sum_{c\in\mathcal{S}_k}\theta_k^c$. In what follows, $\mathbb{E}\theta_k = \frac{1}{S}\mathbb{E}\sum_{c\in\mathcal{S}_k}\theta_k^c = \frac{1}{S}\sum_{c_0\in\mathcal{S}_k}\sum_{c=1}^N p_c\beta_k^c = \sum_{c=1}^N p_c\beta_k^c$, where $p_1 = p_2 = \cdots = p_N$ for scheme II in particular. □

C.2 BOUNDED DIVERGENCE BASED ON PARTIAL DEVICE

Lemma C.2 (Bounded divergence based on partial device)**.** *Assume assumptions A.1, A.2, and A.3 hold. Consider Algorithm 4 with a correlation coefficient $\rho \in [0, 1]$, any learning rate $\eta \in (0, 2/m)$ and $\|\theta_0^c - \theta_*\|_2^2 \leq d\mathcal{D}^2$ for any $c \in [N]$, we have the following results*

For Scheme I, the divergence between θ_k and β_k is upper bounded by

$$\mathbb{E}\|\beta_k - \theta_k\|_2^2 \leq \frac{112}{S}K^2\eta^2 dL^2 H_\rho + \frac{8}{S}K\eta d\tau(\rho^2 + N(1-\rho^2)).$$

For Scheme II, assuming the data is balanced such that $p_1 = \cdots = p_N = \frac{1}{N}$, the divergence between θ_k and β_k is upper bounded by

$$\mathbb{E}\|\beta_k - \theta_k\|_2^2 \leq \frac{N-S}{S(N-1)}\Big(112K^2\eta^2 dL^2 H_\rho + 8K\eta d\tau(\rho^2 + N(1-\rho^2))\Big).$$

where H_ρ, κ and γ^2 are defined as Definition A.4.

Proof. We prove the bounded divergence for the two schemes, respectively.

For **scheme I** with replacement, $\bar{\theta}_k = \sum_{c\in\mathcal{S}_k} \frac{1}{S}\beta_k^c$ for a subset of indices $\mathcal{S}_k$. Taking expectation with respect to $\mathcal{S}_k$, we have

$$\mathbb{E}\|\theta_k - \beta_k\|_2^2 = \frac{1}{S^2}\sum_{i=1}^{S}\mathbb{E}\|\beta_k^{n_i} - \beta_k\|_2^2 = \frac{1}{S}\sum_{c=1}^{N} p_c \|\beta_k^c - \beta_k\|_2^2, \tag{33}$$

where the first equality follows by the independence and unbiasedness of $\theta_k^{n_i}$ for any $i \in [S]$.

To further upper bound Eq.(33), we follow the same technique as in Lemma B.3. Since $k \bmod K = 0$, $k_0 = k - K$ is also the communication time, which yields the same $\theta_{k_0}^c$ for any $c \in [N]$. in what follows,

$$\begin{aligned}\sum_{c=1}^{N} p_c \|\beta_k^c - \beta_k\|_2^2 &= \sum_{c=1}^{N} p_c \|\beta_k^c - \theta_{k_0} - (\beta_k - \theta_{k_0})\|_2^2 \\ &\le \sum_{c=1}^{N} p_c \|\beta_k^c - \theta_{k_0}\|_2^2,\end{aligned} \tag{34}$$

where the last inequality follows by $\beta_k = \sum_{c=1}^{N} p_c\beta_k^c$ and $\mathbb{E}\|x - \mathbb{E}x\|_2^2 \le \mathbb{E}\|x\|_2^2$. Combining Eq.(33) and Eq.(34), we have

$$\begin{aligned}\mathbb{E}\|\theta_k - \beta_k\|_2^2 &\le \frac{1}{S}\sum_{c=1}^{N} p_c \|\beta_k^c - \theta_{k_0}\|_2^2 \\ &\le \frac{1}{S}\sum_{c=1}^{N} p_c \left\|\beta_k^c - \theta_{k_0}^c\right\|_2^2 \\ &\le \frac{1}{S}\sum_{c=1}^{N} p_c \mathbb{E}\sum_{k=k_0}^{k-1} 2K\eta^2 \left\|\nabla \widetilde{f}^c(\theta_k^c)\right\|_2^2 + 4K\eta d\tau\big(\rho^2 + (1-\rho^2)/p_c\big) \\ &\le \frac{1}{S}\sum_{c=1}^{N} p_c \left(\sum_{k=k_0}^{k-1} 2K\eta^2 \mathbb{E}\left\|\nabla \widetilde{f}^c(\theta_k^c)\right\|_2^2 + 4K\eta d\tau\big(\rho^2 + (1-\rho^2)/p_c\big)\right) \\ &\le \frac{28}{S}K^2\eta^2 dL^2 H_\rho + \frac{4}{S}K\eta d\tau(\rho^2 + N(1-\rho^2))\end{aligned}$$

where the last inequality follows a similar argument as in Lemma B.3.

For **scheme II**, given $p_1 = p_2 = \cdots = p_N = \frac{1}{N}$, we have $\theta_k = \frac{1}{S}\sum_{c\in\mathcal{S}_k}\beta_k^c$, which leads to

$$\mathbb{E}\|\theta_k - \beta_k\|_2^2 = \mathbb{E}\left\|\frac{1}{S}\sum_{c\in\mathcal{S}_k}\beta_k^c - \beta_k\right\|_2^2 = \frac{1}{S^2}\mathbb{E}\left\|\sum_{c=1}^{N}\mathbb{I}_{c\in\mathcal{S}_k}(\beta_k^c - \beta_k)\right\|_2^2,$$

where $\mathbb{I}_A$ is an indicator function that equals to 1 if the event A happens.

Plugging the facts that $\mathbb{P}(c \in \mathcal{S}_k) = \frac{S}{N}$ and $\mathbb{P}(c_1, c_2 \in \mathcal{S}_k) = \frac{S(S-1)}{N(N-1)}$ for any $c_1 \neq c_2 \in [N]$ into the above equation, we have

$$\begin{aligned}&\mathbb{E}\|\theta_k - \beta_k\|_2^2 \\ &= \frac{1}{S^2}\left[\sum_{c\in[N]}\mathbb{P}(c\in\mathcal{S}_k)\|\beta_k^c - \beta_k\|_2^2 + \sum_{c_1\neq c_2}\mathbb{P}(c_1, c_2\in\mathcal{S}_k)\langle\beta_k^{c_1} - \beta_k, \beta_k^{c_2} - \beta_k\rangle\right] \\ &= \frac{1}{SN}\sum_{c=1}^{N}\|\beta_k^c - \beta_k\|_2^2 + \sum_{c_1\neq c_2}\frac{S-1}{SN(N-1)}\langle\beta_k^{c_1} - \beta_k, \beta_k^{c_2} - \beta_k\rangle \\ &= \frac{1 - \frac{S}{N}}{S(N-1)}\sum_{c=1}^{N}\|\beta_k^c - \beta_k\|_2^2,\end{aligned}$$

where the last equality holds since $\sum_{c\in[N]}\|\beta_k^c-\beta_k\|_2^2+\sum_{c_1\neq c_2}\langle\beta_k^{c_1}-\beta_k,\beta_k^{c_2}-\beta_k\rangle=\|\beta_k-\beta_k\|_2^2=0$.

Eventually, we have

$$\begin{aligned}\mathbb{E}\|\theta_k-\beta_k\|_2^2&=\frac{N-S}{S(N-1)}\mathbb{E}\frac{1}{N}\sum_{c=1}^{N}\|\beta_k^c-\beta_k\|_2^2\\&\leq\frac{N-S}{S(N-1)}\mathbb{E}\frac{1}{N}\sum_{c=1}^{N}\|\beta_k^c-\theta_{k_0}\|_2^2\\&\leq\frac{N-S}{S(N-1)}\Big(28K^2\eta^2dL^2H_\rho+4K\eta d\tau\big(\rho^2+N(1-\rho^2)\big)\Big),\end{aligned}$$

where the first inequality follows a similar argument as in Eq.(34) and the last inequality follows by Lemma B.3.

□

C.3 Convergence via partial device participation

Theorem C.3 (Restatement of Theorem 4.10)**.** *Assume assumptions A.1, A.2, and A.3 hold. Consider Algorithm 4 with a correlation coefficient $\rho\in[0,1]$, a fixed learning rate $\eta\in(0,\frac{1}{2L}]$ and $\|\theta_0^c-\theta_*\|_2^2\leq d\mathcal{D}^2$ for any $c\in[N]$, we have*

$$\begin{aligned}W_2(\mu_k,\pi)\leq&\left(1-\frac{\eta m}{4}\right)^k\cdot\left(\sqrt{2d}\big(\mathcal{D}+\sqrt{\tau/m}\big)\right)\\&+30\kappa\sqrt{\eta md}\cdot\sqrt{H_\rho((K-1)^2+\kappa)}+2\sqrt{\frac{C_Kd\tau}{Sm}(\rho^2+N(1-\rho^2))C_S},\end{aligned}$$

where $C_K=\frac{\eta mK}{1-e^{-\frac{\eta mK}{2}}}$, $C_S=1$ for Scheme I *and $C_S=\frac{N-S}{N-1}$ for* Scheme II.

Proof. Note that

$$\begin{aligned}&\mathbb{E}\big\|\bar{\theta}_{(k+1)\eta}-\theta_{k+1}\big\|_2^2\\=&\,\mathbb{E}\big\|\bar{\theta}_{(k+1)\eta}-\beta_{k+1}+\beta_{k+1}-\theta_{k+1}\big\|_2^2\\=&\,\mathbb{E}\big\|\bar{\theta}_{(k+1)\eta}-\beta_{k+1}\big\|_2^2+\mathbb{E}\|\beta_{k+1}-\theta_{k+1}\|_2^2+\mathbb{E}2\langle\bar{\theta}_{(k+1)\eta}-\beta_{k+1},\beta_{k+1}-\theta_{k+1}\rangle\\=&\,\mathbb{E}\big\|\bar{\theta}_{(k+1)\eta}-\beta_{k+1}\big\|_2^2+\mathbb{E}\|\beta_{k+1}-\theta_{k+1}\|_2^2,\end{aligned}$$

where the last equality follows by the unbiasedness of the device-sampling scheme in Lemma C.1.

If $k+1 \bmod K\neq 0$, we always have $\beta_{k+1}=\theta_{k+1}$ and $\mathbb{E}\|\beta_{k+1}-\theta_{k+1}\|_2^2=0$. Following the same argument as in Lemma B.5, both schemes lead to the one-step iterate as follows

$$W_2^2(\mu_{k+1},\pi)\leq(1-\frac{\eta m}{2})\cdot W_2^2(\mu_k,\pi)+400\eta^2dL^2H_\rho((K-1)^2+\kappa).\tag{35}$$

If $k+1 \bmod K=0$, combining Lemma C.2 and Lemma B.5, we have

$$W_2^2(\mu_{k+1},\pi)\leq(1-\frac{\eta m}{2})\cdot W_2^2(\mu_k,\pi)+450\eta^2dL^2H_\rho(K^2+\kappa)+\frac{4Kd\eta\tau}{S}(\rho^2+N(1-\rho^2))C_S,\tag{36}$$

where $C_S=1$ for *Scheme I* and $C_S=\frac{N-S}{N-1}$ for *Scheme II*.

Repeatedly applying Eq.(35) and Eq.(36) and arranging terms, we have that

$$
\begin{aligned}
W_2^2(\mu_k, \pi) &\leq \left(1 - \frac{\eta m}{2}\right)^k W_2^2(\mu_0, \pi) + \frac{2}{\eta m}\left(450\eta^2 d L^2 H_\rho(K^2 + \kappa)\right) \\
&\quad + \frac{(1 - (1 - \frac{\eta m}{2})^K)^{\lfloor k/K \rfloor}}{1 - (1 - \frac{\eta m}{2})^K}\left(\frac{4Kd\eta\tau}{S}(\rho^2 + N(1-\rho^2))C_S\right) \\
&\leq \left(1 - \frac{\eta m}{2}\right)^k W_2^2(\mu_0, \pi) + 900\eta m d\kappa^2 H_0((K-1)^2 + \kappa) \\
&\quad + \underbrace{\frac{\eta m K}{1 - e^{-\frac{\eta m K}{2}}}}_{C_K} \frac{4Kd\eta\tau}{\eta m K S}(\rho^2 + N(1-\rho^2))C_S, \\
&= \left(1 - \frac{\eta m}{2}\right)^k W_2^2(\mu_0, \pi) + 900\eta m d\kappa^2 H_0((K-1)^2 + \kappa) \\
&\quad + \frac{4C_K d\tau}{Sm}(\rho^2 + N(1-\rho^2))C_S,
\end{aligned}
$$

where the second inequality follows by $(1-r)^K \leq e^{-rK}$ for any $r \geq 0$.

□

D Bounding contraction, discretization, and divergence

D.1 Dominated contraction property

Proof of Lemma B.1 . Given a client index $c \in [N]$, applying Theorem 2.1.12 (Nesterov, 2004) leads to

$$
\langle y - x, \nabla f^c(y) - \nabla f^c(x)\rangle \geq \frac{mL}{L+m}\|y - x\|_2^2 + \frac{1}{L+m}\|\nabla f^c(y) - \nabla f^c(x)\|_2^2, \quad \forall x, y \in \mathbb{R}^d. \tag{37}
$$

In what follows, we have

$$
\begin{aligned}
&\|\beta - \theta - \eta(\nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta}))\|_2^2 \\
&= \|\beta - \theta\|_2^2 - 2\eta \underbrace{\langle \beta - \theta, \nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta})\rangle}_{\mathcal{I}} + \eta^2 \|\nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta})\|_2^2 .
\end{aligned} \tag{38}
$$

For the second item $\mathcal{I}$ in the right hand side, we have

$$
\begin{aligned}
\mathcal{I} &= \sum_{c=1}^{N} p_c \big\langle \beta - \theta, \nabla f^c(\beta^c) - \nabla f^c(\theta^c) \big\rangle \\
&= \sum_{c=1}^{N} p_c \big\langle \beta - \beta^c + \beta^c - \theta^c + \theta^c - \theta, \nabla f^c(\beta^c) - \nabla f^c(\theta^c) \big\rangle \\
&= -\sum_{c=1}^{N} p_c \left(\big\langle \beta^c - \beta, \nabla f^c(\beta^c) - \nabla f^c(\theta^c) \big\rangle + \big\langle \theta - \theta^c, \nabla f^c(\beta^c) - \nabla f^c(\theta^c) \big\rangle \right) \\
&\qquad + \sum_{c=1}^{N} p_c \big\langle \beta^c - \theta^c, \nabla f^c(\beta^c) - \nabla f^c(\theta^c) \big\rangle \\
&\geq -\sum_{c=1}^{N} p_c \cdot \big((m+L) \left\| \beta^c - \beta \right\|_2^2 + (m+L) \left\| \theta^c - \theta \right\|_2^2 + \frac{1}{2(m+L)} \left\| \nabla f^c(\beta^c) - \nabla f^c(\theta^c) \right\|_2^2 \big) \\
&\qquad + \sum_{c=1}^{N} p_c \cdot \big(\frac{mL}{L+m} \left\| \beta^c - \theta^c \right\|_2^2 + \frac{1}{L+m} \left\| \nabla f^c(\beta^c) - \nabla f^c(\theta^c) \right\|_2^2 \big) \\
&\geq -(m+L) \sum_{c=1}^{N} p_c \left(\left\| \beta^c - \beta \right\|_2^2 + \left\| \theta^c - \theta \right\|_2^2 \right) + \frac{mL}{L+m} \left\| \beta - \theta \right\|_2^2 \\
&\qquad + \frac{1}{2(L+m)} \left\| \nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta}) \right\|_2^2 ,
\end{aligned} \tag{39}
$$

where the first inequality follows by the AM-GM inequality and Eq.(37), respectively; the last inequality follows by Jensen's inequality such that

$$
\begin{aligned}
\sum_{c=1}^{N} p_c \|\beta^c - \theta^c\|_2^2 &\geq \left\| \sum_{c=1}^{N} p_c (\beta^c - \theta^c) \right\|_2^2 = \|\beta - \theta\|_2^2 \\
\sum_{c=1}^{N} p_c \left\| \nabla f^c(\beta^c) - \nabla f^c(\theta^c) \right\|_2^2 &\geq \left\| \sum_{c=1}^{N} p_c \Big(\nabla f^c(\beta^c) - \nabla f^c(\theta^c) \Big) \right\|_2^2 = \left\| \nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta}) \right\|_2^2 .
\end{aligned}
$$

Plugging Eq.(39) into Eq.(38), we have

$$
\begin{aligned}
&\left\| \beta - \theta - \eta \cdot (\nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta})) \right\|_2^2 \\
&\leq \big(1 - \frac{2\eta m L}{m+L}\big) \cdot \|\beta - \theta\|_2^2 + \eta \underbrace{\big(\eta - \frac{1}{m+L}\big)}_{\leq 0 \text{ if } \eta \leq \frac{1}{m+L}} \cdot \left\| \nabla f(\beta) - \boldsymbol{\nabla} f(\boldsymbol{\theta}) \right\|_2^2 \\
&\qquad + 2\eta(m+L) \sum_{c=1}^{N} p_c \cdot (\|\beta^c - \beta\|_2^2 + \|\theta^c - \theta\|_2^2) \\
&\leq (1 - \eta m) \|\beta - \theta\|_2^2 + 4\eta L \sum_{c=1}^{N} p_c \cdot \big(\|\beta^c - \beta\|_2^2 + \|\theta^c - \theta\|_2^2 \big),
\end{aligned}
$$

where the last inequality follows by $\frac{2L}{m+L} \geq 1$, $m \leq L$, $1 - 2a \leq (1-a)^2$ for any a, and $\eta \in (0, \frac{1}{m+L}]$.

□

D.2 DISCRETIZATION ERROR

Proof of Lemma B.2. For any $s \in [0, \infty)$, there exists a certain $k \in \mathbb{N}^+$ such that $s \in [k\eta, (k+1)\eta)$. By the continuous dynamics of Eq.(20), we have

$$\bar{\theta}_s^c = -\int_{k\eta}^{s} \nabla f^c(\bar{\theta}_t^c)dt + \sqrt{2\tau/p_c}\int_{k\eta}^{s} \mathrm{d}\overline{W}_t,$$

which suggests that

$$\sup_{s\in[k\eta,(k+1)\eta)} \left\|\bar{\theta}_s^c - \bar{\theta}_{\eta\lfloor\frac{s}{\eta}\rfloor}^c\right\|_2 \leq \left\|\int_{k\eta}^{s} \nabla f^c(\bar{\theta}_t^c)dt\right\|_2 + \sup_{s\in[k\eta,(k+1)\eta)} \left\|\int_{k\eta}^{s} \sqrt{2\tau/p_c}\mathrm{d}\overline{W}_t\right\|_2.$$

We first square the terms on both sides and take Young's inequality and expectation

$$\mathbb{E}\sup_{s\in[k\eta,(k+1)\eta)} \left\|\bar{\theta}_s^c - \bar{\theta}_{\eta\lfloor\frac{s}{\eta}\rfloor}^c\right\|_2^2 \leq 2\mathbb{E}\left\|\int_{k\eta}^{s} \nabla f^c(\bar{\theta}_t^c)dt\right\|_2^2 + 2\mathbb{E}\sup_{s\in[k\eta,(k+1)\eta)} \left\|\int_{k\eta}^{s} \sqrt{2\tau/p_c}\mathrm{d}\overline{W}_t\right\|_2^2.$$

Then, by Cauchy Schwarz inequality and the fact that $|s - k\eta| \leq \eta$, we have

$$\begin{aligned}\mathbb{E}\sup_{s\in[k\eta,(k+1)\eta)} \left\|\bar{\theta}_s^c - \bar{\theta}_{\eta\lfloor\frac{s}{\eta}\rfloor}^c\right\|_2^2 &\leq 2\eta\mathbb{E}\int_{k\eta}^{s} \left\|\nabla f^c(\bar{\theta}_t^c)dt\right\|_2^2 dt + 8\sum_{i=1}^{d}\mathbb{E}\int_{k\eta}^{s} 2\tau/p_c\mathrm{d}t \\ &\leq 2\eta^2 \sup_s \mathbb{E}\left\|\nabla f^c(\bar{\theta}_s^c)\right\|_2^2 + 16\eta d\tau/p_c, \end{aligned} \tag{40}$$

where the last inequality follows by Burkholder-Davis-Gundy inequality (51) and Itô isometry.

By Young's inequality and the smoothness assumption A.1, we have

$$\begin{aligned}\sup_s \mathbb{E}\|\nabla f^c(\bar{\theta}_s)\|_2^2 &= \sup_s \mathbb{E}\|\nabla f^c(\bar{\theta}_s^c) - \nabla f^c(\theta_*) + \nabla f^c(\theta_*)\|_2^2 \\ &\leq 2\sup_s \mathbb{E}\|\nabla f^c(\bar{\theta}_s^c) - \nabla f^c(\theta_*)\|_2^2 + 2\|\nabla f^c(\theta_*)\|_2^2 \\ &\leq 2L^2 \sup_s \mathbb{E}\|\bar{\theta}_s^c - \theta_*\|_2^2 + 2\gamma^2 \\ &\leq 2L^2\left(\frac{1}{m}\left(\frac{\gamma^2}{m} + 2d\tau\right)\right) + 2\gamma^2 \\ &\leq 4d\kappa\left(\frac{\kappa\gamma^2}{d} + 4L\tau\right), \end{aligned} \tag{41}$$

where the third inequality follows by Lemma E.2, the fourth step holds since $\kappa \geq 1$. Combining Eq.(40) and Eq.(41), we have

$$\mathbb{E}\sup_{s\in[k\eta,(k+1)\eta)} \left\|\bar{\theta}_s^c - \bar{\theta}_{\eta\lfloor\frac{s}{\eta}\rfloor}^c\right\|_2^2 \leq 8\eta^2 d\kappa\left(\frac{\kappa\gamma^2}{d} + L\tau\right) + 16\eta d\tau/p_c.$$

□

D.3 BOUNDED DIVERGENCE

Proof of Lemma B.3. For any $k \geq 0$, consider $k_0 = K\lfloor\frac{k}{K}\rfloor$ such that $k \leq k_0$ and $\theta_{k_0}^c = \theta_{k_0}$ for any $k \geq 0$. It is clear that $k - k_0 \leq K - 1$ for all $k \geq 0$. Consider the non-increasing learning rate such that $\eta_{k_0} \leq 2\eta_k$ for all $k - k_0 \leq K - 1$.

By the iterate Eq.(19), we have

$$
\begin{aligned}
&\sum_{c=1}^{N} p_c \mathbb{E}\|\theta_k^c - \theta_k\|_2^2 \\
=& \sum_{c=1}^{N} p_c \mathbb{E}\|\theta_k^c - \theta_{k_0} - (\theta_k - \theta_{k_0})\|_2^2 \\
\leq& \sum_{c=1}^{N} p_c \mathbb{E}\|\theta_k^c - \theta_{k_0}\|_2^2 \\
\leq& \sum_{c=1}^{N} p_c \mathbb{E} \sum_{k=k_0}^{k-1} 2(K-1)\eta_k^2 \left\|\nabla \widetilde{f}^c(\theta_k^c)\right\|_2^2 + 4(K-1)\eta_k d\tau(\rho^2 + (1-\rho^2)/p_c) \\
\leq& \sum_{c=1}^{N} p_c \left(\sum_{k=k_0}^{k-1} 2(K-1)\eta_{k_0}^2 \mathbb{E}\left\|\nabla \widetilde{f}^c(\theta_k^c)\right\|_2^2 + 4(K-1)\eta_{k_0} d\tau(\rho^2 + (1-\rho^2)/p_c) \right) \\
\leq& 112(K-1)^2 \eta_k^2 d L^2 H_\rho + 8(K-1)\eta_k d\tau(\rho^2 + N(1-\rho^2)),
\end{aligned}
$$

where the first inequality holds by $\mathbb{E}\|\theta - \mathbb{E}\theta\|_2^2 \leq \mathbb{E}\|\theta\|_2^2$ for a stochastic variable θ; the second inequality follows by $(\sum_{i=1}^{K-1} a_i)^2 \leq (K-1)\sum_{i=1}^{K-1} a_i^2$; the last inequality follows by Lemma E.3 and $\eta_{k_0}^2 \leq 4\eta_k^2$. H_ρ is defined in Definition A.4.

□

D.4 Bounded variance

Proof of Lemma B.4. By assumption A.3, we have

$$
\begin{aligned}
\mathbb{E}\left\|\boldsymbol{\nabla} f(\boldsymbol{\theta}) - \boldsymbol{\nabla} \widetilde{f}(\boldsymbol{\theta})\right\|_2^2 &= \mathbb{E}\left\|\sum_{c=1}^{N} p_c \left(\nabla f^c(\theta^c) - \nabla \widetilde{f}^c(\theta^c)\right)\right\|_2^2 \\
&= \sum_{c=1}^{N} p_c^2 \mathbb{E}\left\|\nabla f^c(\theta^c) - \nabla \widetilde{f}^c(\theta^c)\right\|_2^2 \\
&\leq d\sigma^2 \sum_{c=1}^{N} p_c^2 \leq d\sigma^2 \left(\sum_{c=1}^{N} p_c\right)^2 := d\sigma^2.
\end{aligned}
$$

□

E Uniform upper bound

E.1 Discrete dynamics

Lemma E.1 (Discrete dynamics)**.** *Assume assumptions A.1, A.2, and A.3 hold. We consider the generalized formulation in Algorithm 3 with the temperature*

$$T_{c,\rho} = \tau(\rho^2 + (1-\rho^2)/p_c)$$

given a correlation coefficient ρ. For any learning rate $\eta \in (0, 2/m)$ and $\|\theta_0^c - \theta_*\|_2^2 \leq d\mathcal{D}^2$ for any $c \in [N]$, we have the ℓ_2 norm upper bound as follows

$$\sup_k \mathbb{E}\|\theta_k^c - \theta_*\|_2^2 \leq d\mathcal{D}^2 + \frac{6d}{m}\left(\max_{c\in[N]} T_{c,\rho} + \frac{\sigma^2}{m} + \frac{\gamma^2}{md}\right),$$

where $\gamma := \max_{c\in[N]} \|\nabla f^c(\theta_*)\|_2$ and θ_* denotes the global minimum for the function f.

Proof. First, we consider the k-th iteration, where $k \in \{1, 2, \cdots, K-2, (K-1)_-\}$ and $(K-1)_-$ denotes the $(K-1)$-step before synchronization. Following the iterate of Eq.(14) in a local client of $c \in [N]$, we have

$$\begin{aligned}
&\mathbb{E}\left\|\theta_{k+1}^c - \theta_*\right\|_2^2 \\
&= \mathbb{E}\|\theta_k^c - \theta_* - \eta\nabla\widetilde{f}^c(\theta_k^c)\|_2^2 + \sqrt{8\eta T_{c,\rho}}\mathbb{E}\langle\theta_k^c - \theta_* - \eta\nabla\widetilde{f}^c(\theta_k^c), \xi_k\rangle + 2\eta T_{c,\rho}\mathbb{E}\|\xi_k\|_2^2 \\
&= \mathbb{E}\|\theta_k^c - \theta_* - \eta\nabla\widetilde{f}^c(\theta_k^c)\|_2^2 + 2\eta d T_{c,\rho},
\end{aligned} \tag{42}$$

where the last equality follows from $\mathbb{E}\xi_k = 0$ and the conditional independence of $\theta_k^c - \theta_* - \widetilde{f}^c(\theta_k^c)$ and ξ_k. Note that

$$\begin{aligned}
&\mathbb{E}\|\theta_k^c - \theta_* - \eta\widetilde{f}^c(\theta_k^c)\|_2^2 \\
&= \mathbb{E}\|\theta_k^c - \theta_* - \eta\nabla f^c(\theta_k^c)\|_2^2 + \eta^2\mathbb{E}\|\nabla f^c(\theta_k^c) - \nabla\widetilde{f}^c(\theta_k^c)\|_2^2 \\
&\qquad + 2\eta\mathbb{E}\langle\theta_k^c - \theta_* - \eta\nabla f^c(\theta_k^c), \nabla f^c(\theta_k^c) - \nabla\widetilde{f}^c(\theta_k^c)\rangle \\
&= \mathbb{E}\|\theta_k^c - \theta_* - \eta\nabla f^c(\theta_k^c)\|_2^2 + \eta^2\mathbb{E}\|\nabla f^c(\theta_k^c) - \nabla\widetilde{f}^c(\theta_k^c)\|_2^2 \\
&\leq \mathbb{E}\|\theta_k^c - \theta_* - \eta\nabla f^c(\theta_k^c)\|_2^2 + \eta^2 d\sigma^2,
\end{aligned} \tag{43}$$

where the first step follows from simple algebra, the second step follows from the unbiasedness of the stochastic gradient, and the last step follows from Assumption A.3. For any $q > 0$, we can upper bound the first term of Eq.(43) as follows

$$\begin{aligned}
&\mathbb{E}\|\theta_k^c - \theta_* - \eta\nabla f^c(\theta_k^c)\|_2^2 \\
&= \mathbb{E}\|\theta_k^c - \theta_* - \eta(\nabla f^c(\theta_k^c) - \nabla f^c(\theta_*)) - \eta\nabla f^c(\theta_*)\|_2^2 \\
&\leq (1+q)\mathbb{E}\|\theta_k^c - \theta_* - \eta(\nabla f^c(\theta_k^c) - \nabla f^c(\theta_*))\|_2^2 + \eta^2\left(1 + \frac{1}{q}\right)\|\nabla f^c(\theta_*)\|_2^2 \\
&\leq (1+q)\underbrace{\left(1 - \frac{\eta m}{2}\right)^2}_{\psi^2}\mathbb{E}\|\theta_k^c - \theta_*\|_2^2 + \eta^2\left(1 + \frac{1}{q}\right)\gamma^2,
\end{aligned} \tag{44}$$

where the first inequality follows by the AM-GM inequality; the second inequality is a special case of Lemma B.1 based on Assumption A.2, where no local steps is involved before the synchronization step. Similar results have been achieved in Theorem 3 (Dalalyan, 2017). In addition, $\gamma := \max_{c\in[N]}\|\nabla f^c(\theta_*)\|_2$.

Choose $q = (\frac{1+\psi}{2\psi})^2 - 1$ so that $(1+q)\psi^2 = \frac{(1+\psi)^2}{4}$. Moreover, since $\psi = 1 - \frac{\eta m}{2}$, we get $\frac{1+\psi}{2} = 1 - \frac{1}{4}\eta m$. In addition, we have $1 + \frac{1}{q} = \frac{1+q}{q} = \frac{(1+\psi)^2}{(1-\psi)(1+3\psi)} \leq \frac{2}{\eta m}$. It follows that

$$\eta^2\left(1 + \frac{1}{q}\right) \leq \frac{2\eta}{m}. \tag{45}$$

Combining Eq.(42), Eq.(43), Eq.(44), and Eq.(45), we have the following iterate

$$\mathbb{E}\|\theta_{k+1}^c - \theta_*\|_2^2 \leq \underbrace{\left(1 - \frac{\eta m}{4}\right)^2}_{:=g(\eta)}\mathbb{E}\|\theta_k^c - \theta_*\|_2^2 + 2\eta d T_{c,\rho} + \eta^2 d\sigma^2 + \frac{2\eta\gamma^2}{m}.$$

Note that $\frac{1}{1-g(\eta)} = \frac{1}{\frac{\eta m}{2}(1-\frac{\eta m}{8})} \leq \frac{3}{\eta m}$ given $\eta \in (0, \frac{2}{m})$. Recursively applying the above equation k times, where $k \in \{1, 2, \cdots, K-1, K_-\}$ and K_- denotes the K-step without synchronization, it follows that

$$\begin{aligned}
\mathbb{E}\|\theta_k^c - \theta_*\|_2^2 &\leq g(\eta)^k\|\theta_0^c - \theta_*\|_2^2 + \frac{1-g(\eta)^k}{1-g(\eta)}\cdot\left(2\eta d T_{c,\rho} + \eta^2 d\sigma^2 + \frac{2\eta\gamma^2}{m}\right) \\
&\leq \|\theta_0^c - \theta_*\|_2^2 + \frac{3}{\eta m}\cdot\left(2\eta d T_{c,\rho} + \eta^2 d\sigma^2 + \frac{2\eta\gamma^2}{m}\right) \\
&\leq d\mathcal{D}^2 + \underbrace{\frac{6d}{m}\left(\max_{c\in[N]} T_{c,\rho} + \frac{\sigma^2}{m} + \frac{\gamma^2}{md}\right)}_{:=U},
\end{aligned} \tag{46}$$

where the second inequality holds by $g(\eta) \leq 1$, the third inequality holds because $\|\theta_0^c - \theta_*\|_2^2 \leq d\mathcal{D}^2$ for any $c \in [N]$ and $\eta < \frac{2}{m}$. In particular, the K-th step before synchronization yields that

$$\mathbb{E}\|\theta_{K_-}^c - \theta_*\|_2^2 \leq d\mathcal{D}^2 + U. \tag{47}$$

Having all the results ready, for the K-local step after synchronization, applying Jensen's inequality

$$\begin{aligned}\mathbb{E}\|\theta_K^c - \theta_*\|_2^2 &= \mathbb{E}\left\|\sum_{c=1}^N p_c \theta_{K-}^c - \theta_*\right\|_2^2 \\ &\leq \sum_{c=1}^N p_c \mathbb{E}\left\|\theta_{K-}^c - \theta_*\right\|_2^2 \\ &\leq d\mathcal{D}^2 + U,\end{aligned} \tag{48}$$

Now starting from iteration K, we adapt the recursion of Eq.(46) for the k-th step, where $k \in \{K+1, \cdots, 2K-1, (2K)_-\}$ and $(2K)_-$ denotes the $2K$-step without synchronization, we have

$$\begin{aligned}&\mathbb{E}\|\theta_k^c - \theta_*\|_2^2 \\ \leq\ & g(\eta)^{k-K} \cdot \mathbb{E}\|\theta_K^c - \theta_*\|_2^2 + \frac{1-g(\eta)^{k-K}}{1-g(\eta)} \cdot \left(2\eta d \max_{c\in[N]} T_{c,\rho} + \eta^2 d\sigma^2 + \frac{2\eta\gamma^2}{m}\right) \\ \leq\ & g(\eta)^{k-K}(d\mathcal{D}^2 + U) + \frac{1-g(\eta)^{k-K}}{m\eta/3}\frac{m\eta}{3}U \\ \leq\ & d\mathcal{D}^2 + g(\eta)^{k-K}U + (1-g(\eta)^{k-K})U \\ \leq\ & d\mathcal{D}^2 + U,\end{aligned} \tag{49}$$

where the second inequality follows by Eq.(48), the fact that $1 - g(\eta) \geq \eta m/3$ and $\eta \leq \frac{2}{m}$, and the definition of U. The third one holds since $g(\eta) \leq 1$.

By repeating Eq.(48) and (49), we have that for all $k \geq 0$, we can obtain the desired uniform upper bound. □

Discussions: Since the above result is independent of the learning rate η, it can be naturally applied to the setting with decreasing learning rates. The details are omitted.

E.2 Continuous diffusion

Lemma E.2 (Continuous time)**.** *Assume assumption A.2 holds. We have the ℓ_2 norm upper bound as follows*

$$\sup_t \mathbb{E}\left\|\bar{\theta}_t^c - \theta_*\right\|_2^2 \leq \frac{1}{m}\left(\frac{\gamma^2}{m} + 2d\tau\right),$$

where $\gamma := \max_{c\in[N]} \|\nabla f^c(\theta_*)\|_2$ and θ_* denotes the global minimum for the function f.

Proof. Since the synchronization is conducted at every time t, the essential temperature applied to each client is τ. Let $q(\bar{\theta}_t^c) = \left\|\bar{\theta}_t^c - \theta_*\right\|_2^2$. For any time $t \geq 0$, applying Itô's lemma leads to

$$\begin{aligned}\mathrm{d}q(\bar{\theta}_t^c) &= -2\langle \bar{\theta}_t^c - \theta_*, \nabla f^c(\bar{\theta}_t^c)\rangle \mathrm{d}t + 2d\tau \mathrm{d}t + \sqrt{8\tau}\langle \bar{\theta}_t^c - \theta_*, \mathrm{d}\overline{W}_t\rangle \\ &= -2\langle \bar{\theta}_t^c - \theta_*, \nabla f^c(\bar{\theta}_t^c) - \nabla f^c(\theta_*) + \nabla f^c(\theta_*)\rangle \mathrm{d}t + 2d\tau \mathrm{d}t + \sqrt{8\tau}\langle \bar{\theta}_t^c - \theta_*, \mathrm{d}\overline{W}_t\rangle \\ &\leq -2m\left\|\bar{\theta}_t^c - \theta_*\right\|_2^2 \mathrm{d}t - 2\langle \bar{\theta}_t^c - \theta_*, \nabla f^c(\theta_*)\rangle \mathrm{d}t + 2d\tau \mathrm{d}t + \sqrt{8\tau}\langle \bar{\theta}_t^c - \theta_*, \mathrm{d}\overline{W}_t\rangle \\ &\leq -2m\left\|\bar{\theta}_t^c - \theta_*\right\|_2^2 \mathrm{d}t + m\left\|\bar{\theta}_t^c - \theta_*\right\|_2^2 \mathrm{d}t + \frac{\|\nabla f^c(\theta_*)\|_2^2}{m}\mathrm{d}t + 2d\tau \mathrm{d}t + \sqrt{8\tau}\langle \bar{\theta}_t^c - \theta_*, \mathrm{d}\overline{W}_t\rangle \\ &\leq -mq(\bar{\theta}_t^c)\mathrm{d}t + \left(\frac{\gamma^2}{m} + 2d\tau\right)\mathrm{d}t + \sqrt{8\tau}\langle \bar{\theta}_t^c - \theta_*, \mathrm{d}\overline{W}_t\rangle,\end{aligned}$$

where the first inequality follows by Assumption A.2; the second inequality follows by the AM-GM inequality; the third inequality follows by the definition that $\gamma^2 = \max_{c\in[N]} \|\nabla f^c(\theta_*)\|_2^2$.

In other words, we have

$$\begin{aligned}
\mathrm{d}(e^{mt}q(\bar{\theta}_t^c)) &= me^{mt}q(\bar{\theta}_t^c)\mathrm{d}t + e^{mt}\mathrm{d}q(\bar{\theta}_t^c) \\
&\leq me^{mt}q(\bar{\theta}_t^c)\mathrm{d}t + e^{mt}\left(-mq(\bar{\theta}_t^c)\mathrm{d}t + \left(\frac{\gamma^2}{m} + 2d\tau\right)\mathrm{d}t + \sqrt{8\tau}\langle\bar{\theta}_t^c - \theta_*, \mathrm{d}\overline{W}_t\rangle\right) \\
&\leq e^{mt}\left(\frac{\gamma^2}{m} + 2d\tau\right)\mathrm{d}t + \sqrt{8\tau}e^{mt}\langle\bar{\theta}_t^c - \theta_*, \mathrm{d}\overline{W}_t\rangle.
\end{aligned}$$

The solution is upper bounded by

$$e^{mt}q(\bar{\theta}_t^c) \leq e^{m\cdot 0}q(\bar{\theta}_0^c) + \int_0^t \left(e^{ms}\left(\frac{\gamma^2}{m} + 2d\tau\right)\mathrm{d}s + \sqrt{8\tau}e^{ms}\langle\bar{\theta}_s^c - \theta_*, \mathrm{d}\overline{W}_s\rangle\right).$$

By the martingale property of Itô integral, taking expectations yields

$$\begin{aligned}
\mathbb{E}q(\bar{\theta}_t^c) &\leq e^{-mt}\mathbb{E}q(\bar{\theta}_0^c) + e^{-mt}\left(\frac{\gamma^2}{m} + 2d\tau\right)\int_0^t e^{ms}\mathrm{d}s \\
&= e^{-mt}\mathbb{E}q(\bar{\theta}_0^c) + \frac{1-e^{-mt}}{m}\left(\frac{\gamma^2}{m} + 2d\tau\right) \\
&\leq e^{-mt}\mathbb{E}q(\bar{\theta}_0^c) + \frac{1-e^{-mt}}{m}\underbrace{\left(\frac{\gamma^2}{m} + 2d\tau\right)}_{:=V},
\end{aligned} \tag{50}$$

where the last inequality follows since the synchronization is conducted at any time step t. Since $\bar{\theta}_0^c$ is simulated from the stationary distribution π, by Lemma 12 (Durmus & Moulines, 2016) or Theorem 17 (Cheng et al., 2018), we have

$$\mathbb{E}q(\bar{\theta}_0^c) = \mathbb{E}\|\bar{\theta}_0^c - \theta_*\|_2^2 \leq \frac{d\tau}{m} \leq \frac{1}{m}\left(\frac{\gamma^2}{m} + 2d\tau\right) = \frac{V}{m},$$

which completes the proof.

□

E.3 Bounded gradient

Lemma E.3 (Bounded gradient in ℓ_2 norm)**.** *Given assumptions A.1, A.2, and A.3 hold, for any client c and any learning rate $\eta \in (0, 2/m)$ and $\|\theta_0^c - \theta_*\|_2^2 \leq d\mathcal{D}^2$ for any $c \in [N]$, we have the ℓ_2 norm upper bound as follows*

$$\mathbb{E}\|\nabla \widetilde{f}^c(\theta_k^c)\|_2^2 \leq 14dL^2H_\rho,$$

where $H_\rho = \mathcal{D}^2 + \frac{1}{m}\max_{c\in[N]} T_{c,\rho} + \frac{\gamma^2}{m^2 d} + \frac{\sigma^2}{m^2}$.

Proof. Decompose the ℓ_2 of the gradient as follows

$$\begin{aligned}
\mathbb{E}\left\|\nabla \widetilde{f}^c(\theta_k^c)\right\|_2^2 &= \mathbb{E}\left\|\nabla \widetilde{f}^c(\theta_k^c) - \nabla f^c(\theta_k^c) + \nabla f^c(\theta_k^c)\right\|_2^2 \\
&= \mathbb{E}\|\nabla f^c(\theta_k^c)\|_2^2 + \mathbb{E}\left\|\nabla \widetilde{f}^c(\theta_k^c) - \nabla f^c(\theta_k^c)\right\|_2^2 \\
&\quad + 2\mathbb{E}\left\langle \nabla \widetilde{f}^c(\theta_k^c) - \nabla f^c(\theta_k^c), \nabla f^c(\theta_k^c)\right\rangle \\
&\leq \mathbb{E}\|\nabla f^c(\theta_k^c)\|_2^2 + \sigma^2 d \\
&= \mathbb{E}\|\nabla f^c(\theta_k^c) - \nabla f^c(\theta_*) + \nabla f^c(\theta_*)\|_2^2 + \sigma^2 d \\
&\leq 2\mathbb{E}\|\nabla f^c(\theta_k^c) - \nabla f^c(\theta_*)\|_2^2 + 2\mathbb{E}\left\|\nabla f^c(\theta_*)\right\|_2^2 + \sigma^2 d \\
&\leq 2L^2\mathbb{E}\|\theta_k^c - \theta_*\|_2^2 + 2\gamma^2 + \sigma^2 d \\
&\leq 2dL^2\mathcal{D}^2 + \frac{12dL^2}{m}\cdot\left(\max_{c\in[N]} T_{c,\rho} + \frac{\sigma^2}{m} + \frac{\gamma^2}{md}\right) + 2\gamma^2 + \sigma^2 d \\
&\leq 14dL^2\cdot\left(\mathcal{D}^2 + \frac{1}{m}\max_{c\in[N]} T_{c,\rho} + \frac{\gamma^2}{m^2 d} + \frac{\sigma^2}{m^2}\right) := 14dL^2 H_\rho,
\end{aligned}$$

where the first inequality follows by Assumption A.3; the second inequality follows by Young's inequality; the third inequality follows by Assumption A.1 and the definition that $\gamma := \max_{c\in[N]}\|\nabla f^c(\theta_*)\|_2$; the fourth inequality follows by Lemma E.1; the last inequality follows by $\kappa := \frac{L}{m} \geq 1$. □

F Initial condition

Lemma F.1 (Initial condition)**.** *Let μ_0 denote the Dirac delta distribution at θ_0. Then, we have*

$$W_2(\mu_0, \pi) \leq \sqrt{2}(\|\theta_0 - \theta_*\|_2 + \sqrt{d\tau/m}).$$

Proof. By Cheng et al. (2018), there exists an optimal coupling between μ_0 and π such that

$$\begin{aligned}
W_2^2(\mu_0, \pi) &\leq \mathbb{E}_{\theta\sim\pi}[\|\theta_0 - \theta\|_2^2] \\
&\leq 2\mathbb{E}_{\theta\sim\pi}[\|\theta_0 - \theta_*\|_2^2] + 2\mathbb{E}_{\theta\sim\pi}[\|\theta - \theta_*\|_2^2] \\
&= 2\|\theta_0 - \theta_*\|_2^2 + 2\mathbb{E}_{\theta\sim\pi}[\|\theta - \theta_*\|_2^2] \\
&\leq 2\|\theta_0 - \theta_*\|_2^2 + 2d\tau/m,
\end{aligned}$$

where the second step follows from triangle inequality, the last step follows from Lemma 12 (Durmus & Moulines, 2016) and the temperature τ is included to adapt to the time scaling. □

Burkholder-Davis-Gundy inequality Let $\phi : [0, \infty) \to \mathbb{R}^{r\times d}$ for some positive integers r and d. In addition, we assume $\mathbb{E}\int_0^\infty |\psi(s)|^2 \mathrm{d}s < \infty$ and let $Z(t) = \int_0^t \psi(s)\mathrm{d}W_s$, where W_s is a d-dimensional Brownian motion. Then for all $t \geq 0$, we have

$$\mathbb{E}\sup_{0\leq s\leq t}|Z(s)|^2 \leq 4\mathbb{E}\int_0^t |\phi(s)|^2 \mathrm{d}s. \tag{51}$$

G More on Differential Privacy Guarantees

We make the following assumptions for the analysis of DP.

Assumption G.1 (Bounded ℓ_2-sensitivity)**.** *The gradient of loss function $l : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R},\ (\theta, x) \mapsto l(\theta; x)$ w.r.t. θ has a uniformly bounded ℓ_2-sensitivity for $\forall\theta \in \mathbb{R}^d$:*

$$\Delta_l := \sup_{\theta\in\mathbb{R}^d}\sup_{x,x'\in\mathcal{X}}\|\nabla l(\theta; x) - \nabla l(\theta; x')\|_2 < \infty \tag{52}$$

For example, when $l(\theta;\cdot)$ is M-Lipschitz for any $\theta \in \mathbb{R}^d$, $\Delta_l \leq 2M$. Following the tradition of DP guarantees, we assume that the unbiased gradient is computed as follows.

Assumption G.2 (Unbiased gradient estimates)**.** *The unbiased estimate of $\nabla f^c(\theta)$ is calculated using a subset (denoted by $\mathcal{S}^c$) of $\{x_{c,i}\}_{i=1}^{n_c}$ sampled uniformly at random from all the subsets of size γn_c of $\{x_{c,i}\}_{i=1}^{n_c}$:*

$$\nabla \widetilde{f}^c(\theta) = \frac{1}{\gamma p_c} \sum_{i \in \mathcal{S}^c} \nabla l(\theta; x_{c,i}). \tag{53}$$

Theorem G.3. *Assume assumptions G.1 and G.2 holds. For any $\delta_0 \in (0,1)$, if $\eta \in \left(0, \frac{\tau(1-\rho^2)\gamma^2 \min_{c\in[N]} p_c}{\Delta_l^2 \log(1.25/\delta_0)}\right]$, then Algorithm 1 is $(\epsilon_{K,T}^{(3)}, \delta_{K,T}^{(3)})$-differentially private w.r.t. $\simeq_s$ after executed for T ($T = EK$ with $E \in \mathbb{N}, E \geq 1$) iterations where*

$$\epsilon_{K,T}^{(3)} = \widetilde{\epsilon}_K \min\left\{\sqrt{\frac{2T}{K}\log\left(\frac{1}{\delta_2}\right)} + \frac{T(e^{\widetilde{\epsilon}_K}-1)}{K},\ \frac{T}{K}\right\},$$

$$\delta_{K,T}^{(3)} = \frac{T}{K}\widetilde{\delta}_K + \delta_2,$$

with

$$\widetilde{\epsilon}_K = \begin{cases} \log\left(1 + \left(1 - \left(1 - \frac{1}{N}\right)^S\right)(e^{\epsilon_K} - 1)\right), & \textit{under scheme I}, \\ \log\left(1 + \frac{S}{N}\left(e^{\epsilon_K} - 1\right)\right), & \textit{under scheme II}, \end{cases}$$

$$\widetilde{\delta}_K = \begin{cases} \sum_{s=1}^{S} \binom{S}{s} \left(\frac{1}{N}\right)^s \left(1 - \frac{1}{N}\right)^{S-s} \delta_{K,s}, & \textit{under scheme I}, \\ \frac{S}{N}\left(K\gamma\delta_0 + \delta_1\right), & \textit{under scheme II}, \end{cases}$$

$$\delta_{K,s} = \frac{\left(e^{\epsilon_K} - 1\right)\delta_{K,s,0}}{e^{\epsilon_K/s} - 1},$$

$$\delta_{K,s,0} = 1.25K\gamma\left(\frac{\delta_0}{1.25}\right)^{1/s^2} + \delta_1,$$

$$\epsilon_K = \epsilon_1 \min\left\{\sqrt{2K\log(1/\delta_1)} + K(e^{\epsilon_1} - 1),\ K\right\},$$

$$\epsilon_1 = 2\Delta_l\sqrt{\frac{\eta\log(1.25/\delta_0)}{\tau(1-\rho^2)\min_{c\in[N]} p_c}},$$

and $\delta_1, \delta_2 \in [0,1)$.

Proof. Let $\mathcal{D}_c := \{x_{c,i}\}_{i\in[n_c]}$ denote the dataset of the c-th client for $c \in [N]$. Let $\mathcal{D} := \cup_{c\in[N]} \mathcal{D}_c$ denote the whole dataset.

As FedAvg algorithms can be divided into the processes of local updates, synchronization, and broadcasting with risks of information leakage in synchronization (local model uploading and aggregation) and broadcasting, we consider the differential privacy guarantees in synchronization and broadcasting similar to Wei et al. (2020). Since there is no involvement of data in model aggregation and broadcasting, they are post-processing processes. Thus, it suffices to analyze the differential privacy guarantees in local model uploading.

For any two datasets $\mathcal{D} \simeq_s \mathcal{D}'$, there exists $c \in [N]$ such that $\mathcal{D}_c \simeq_s \mathcal{D}'_c$ and $\mathcal{D}_{c'} = \mathcal{D}'_{c'}$ for any $c' \in [N], c' \neq c$.

Consider the function $m_c(\mathcal{S};\theta) = \theta - \frac{\eta}{\gamma}\nabla f^c(\theta;\mathcal{S}) = \theta - \frac{\eta}{\gamma p_c}\sum_{x\in\mathcal{S}} \nabla l^c(\theta;x)$ with $|\,\mathcal{S}\,| = \gamma n_c$ ($|A|$ denotes the cardinality of set A). By Assumption G.1, for any $\theta \in \mathbb{R}^d$, the sensitivity of $m_c(\mathcal{D}_c;\theta)$ is

$$\Delta m_c := \sup_{\mathcal{S}_c \simeq_s \mathcal{S}'_c} \|m_c(\mathcal{S}_c;\theta) - m_c(\mathcal{S}'_c;\theta)\| = \frac{\eta}{\gamma p_c}\Delta_l \tag{54}$$

For the mechanism $\mathcal{M}_c(\mathcal{S};\theta) := m_c(\mathcal{S};\theta) + \sqrt{2\eta\tau\rho^2}\dot{\xi} + \sqrt{2\eta(1-\rho^2)\tau/p_c}\xi$ with $\dot{\xi}$ and ξ being two independent standard d-dimensional Gaussian vector, since $\dot{\xi}$ is broadcasted to all the clients, it

can be treated as some known constant which does not contribute to the differential privacy. Thus, the standard deviation of the added Gaussian noise is $\sqrt{2\eta\tau(1-\rho^2)/p_c}$ at each dimension. Then, according to the Gaussian mechanism Dwork et al. (2014), $M_c(\mathcal{D}_c;\theta)$ is $(\epsilon_{0,c},\delta_0)$-differentially private for any $\theta\in\mathbb{R}^d$ with

$$\epsilon_{0,c}=c(\delta_0)\frac{\Delta_l}{\gamma}\sqrt{\frac{\eta}{2p_c\tau(1-\rho^2)}},\quad c(\delta_0)=\sqrt{2\log(1.25/\delta_0)},\quad \delta_0\in(0,1). \tag{55}$$

For $\mathcal{S}^c$ sampled uniformly at random from all the subsets of size γn_c of $\mathcal{D}_c$, define $\widetilde{M}_c(\mathcal{D}_c;\theta):=M_c(\mathcal{S}^c;\theta)$. Then, according to Theorem 9 in Balle et al. (2018), $\widetilde{M}_c$ is $(\log\left(1+\gamma(e^{\epsilon_{0,c}}-1)\right),\gamma\delta_0)$-differentially private. Notice that for any $\epsilon_{0,c}\in[0,1]$ (i.e., $0\leq\eta\leq\frac{2p_c\tau(1-\rho^2)\gamma^2}{\Delta_l^2c(\delta_0)^2}$), we have $0\leq e^{\epsilon_{0,c}}-1\leq 2\epsilon_{0,c}$ and

$$\log\left(1+\gamma(e^{\epsilon_{0,c}}-1)\right)\leq\log\left(1+2\gamma\epsilon_{0,c}\right)\leq 2\gamma\epsilon_{0,c}=c(\delta_0)\Delta_l\sqrt{\frac{2\eta}{p_c\tau(1-\rho^2)}}=:\epsilon_{1,c}. \tag{56}$$

Define

$$\epsilon_1:=c(\delta_0)\Delta_l\sqrt{\frac{2\eta}{\tau(1-\rho^2)\min_{c'\in[N]}p_{c'}}}=2\Delta_l\sqrt{\frac{\eta\log(1.25/\delta_0)}{\tau(1-\rho^2)\min_{c\in[N]}p_c}}. \tag{57}$$

Then, we have $\max_{c\in[N]}\epsilon_{1,c}=\epsilon_1$ and $\max_{c\in[N]}\epsilon_{0,c}\in[0,1]$ if

$$0\leq\eta\leq\frac{2\tau(1-\rho^2)\gamma^2\min_{c'\in[N]}p_{c'}}{\Delta_l^2c(\delta_0)^2}=\frac{\tau(1-\rho^2)\gamma^2\min_{c'\in[N]}p_{c'}}{\Delta_l^2\log(1.25/\delta_0)}. \tag{58}$$

Thus, for $0\leq\eta\leq\frac{\tau(1-\rho^2)\gamma^2\min_{c'\in[N]}p_{c'}}{\Delta_l^2\log(1.25/\delta_0)}$, $\widetilde{M}_c(\mathcal{D}_c;\theta)$ is $(\epsilon_1,\gamma\delta_0)$-differentially private for any $\theta\in\mathbb{R}^d$. From now on, we assume that 58 holds.

Define $\mathcal{M}_c^K(\mathcal{D}_c;\theta)$ to be the K-fold composition of $\widetilde{\mathcal{M}}_c(\mathcal{D}_c;\theta)$. According to the composition rules of (ϵ,δ)-differential privacy (Theorem 3.1 and 3.3 in Dwork et al. (2010)), $\mathcal{M}_c^K(\mathcal{D}_c;\theta)$ is (ϵ_K,δ_K)-differentially private with

$$\epsilon_K=\min\left\{\sqrt{2K\log(1/\delta_1)}\epsilon_1+K\epsilon_1(e^{\epsilon_1}-1),\ K\epsilon_1\right\}, \tag{59}$$

$$\delta_K=K\gamma\delta_0+\delta_1. \tag{60}$$

for any $\delta_1\in[0,1)$.

By 57, if

$$0\leq\eta\leq\frac{\tau(1-\rho^2)\min_{c\in[N]}p_c}{2\Delta_l^2c(\delta_0)^2}\log^2\left(1+\sqrt{\frac{2\log(1/\delta_1)}{K}}\right), \tag{61}$$

we have $\epsilon_1\in\left[0,\log\left(1+\sqrt{\frac{2\log(1/\delta_1)}{K}}\right)\right]$ which implies that $K\epsilon_1(e^{\epsilon_1}-1)\leq\sqrt{2K\log(1/\delta_1)}\epsilon_1$ and

$$\epsilon_K\leq 2\sqrt{2K\log(1/\delta_1)}\epsilon_1. \tag{62}$$

In the synchronization process, S clients selected via device-sampling scheme I or II send their local models to the center. Thus, for scheme I (with replacement) and scheme II (without replacement), according to Theorem 10 and Theorem 9 in Balle et al. (2018) respectively, each synchronization process is $(\widetilde{\epsilon}_K,\widetilde{\delta}_K)$-differentially private with

$$\widetilde{\epsilon}_K=\begin{cases}\log\left(1+\left(1-\left(1-\frac{1}{N}\right)^S\right)(e^{\epsilon_K}-1)\right), & \text{under scheme I,}\\ \log\left(1+\frac{S}{N}\left(e^{\epsilon_K}-1\right)\right), & \text{under scheme II,}\end{cases} \tag{63}$$

$$\widetilde{\delta}_K=\begin{cases}\sum_{s=1}^S\binom{S}{s}\left(\frac{1}{N}\right)^s\left(1-\frac{1}{N}\right)^{S-s}\delta_{K,s}, & \text{under scheme I,}\\ \frac{S}{N}\delta_K=\frac{S}{N}\left(K\gamma\delta_0+\delta_1\right), & \text{under scheme II.}\end{cases} \tag{64}$$

where

$$\delta_{K,s} = \frac{(e^{\epsilon_K} - 1)\,\delta_{K,s,0}}{e^{\epsilon_K/s} - 1}, \quad \delta_{K,s,0} = 1.25K\gamma\left(\frac{\delta_0}{1.25}\right)^{1/s^2} + \delta_1.$$

The aggregation and broadcasting process is post-processing and preserves the guarantees of differential privacy (Proposition 2.1 in Dwork et al. (2014)). When executed T iterations, Algorithm 1 is the T/K-fold composition of local updates, synchronization, and broadcasting. According to the composition rules of (ϵ, δ)-differential privacy (Theorem 3.1 and 3.3 in Dwork et al. (2010)), Algorithm 1 is $(\epsilon^{(3)}_{K,T}, \delta^{(3)}_{K,T})$-differentially private after T iterations with

$$\epsilon^{(3)}_{K,T} = \min\left\{\sqrt{2\frac{T}{K}\log(1/\delta_2)}\widetilde{\epsilon}_K + \frac{T}{K}\widetilde{\epsilon}_K(e^{\widetilde{\epsilon}_K} - 1),\ \frac{T}{K}\widetilde{\epsilon}_K\right\}, \tag{65}$$

$$\delta^{(3)}_{K,T} = \frac{T}{K}\widetilde{\delta}_K + \delta_2, \tag{66}$$

for $\delta_1, \delta_2 \in [0, 1)$ and $\delta_0 \in (0, 1)$. Notice that under scheme II, $e^{\widetilde{\epsilon}_K} - 1 = \frac{S}{N}(e^{\epsilon_K} - 1)$, thus, $\epsilon^{(3)}_{K,T} = \widetilde{\epsilon}_K \min\left\{\sqrt{2\frac{T}{K}\log(1/\delta_2)} + \frac{TS}{KN}(e^{\epsilon_K} - 1),\ \frac{T}{K}\right\}$ and $\delta^{(3)}_{K,T} = \frac{S}{N}\gamma T\delta_0 + \frac{TS}{KN}\delta_1 + \delta_2$. □

Discussion on the differential privacy guarantees of Algorithm 1 under scheme II By 59, 60, 63, 64, 65, and 66, by letting $\delta_1, \delta_2 = 0$, we have that Algorithm 1 is at least $\left(\frac{T}{K}\log\left(1 + \frac{S}{N}(e^{K\epsilon_1} - 1)\right), \frac{S}{N}\gamma T\delta_0\right)$-differentially private.

If

$$\widetilde{\epsilon}_K \le \log\left(1 + \sqrt{\frac{2K\log(1/\delta_2)}{T}}\right), \tag{67}$$

we have $\frac{T}{K}\widetilde{\epsilon}_K(e^{\widetilde{\epsilon}_K} - 1) \le \sqrt{2\frac{T}{K}\log(1/\delta_2)}\widetilde{\epsilon}_K$ and therefore

$$\epsilon^{(3)}_{K,T} \le 2\sqrt{2\frac{T}{K}\log(1/\delta_2)}\widetilde{\epsilon}_K = 2\sqrt{2\frac{T}{K}\log(1/\delta_2)}\log\left(1 + \frac{S}{N}(e^{\epsilon_K} - 1)\right). \tag{68}$$

Now assume η satisfies 61, then by 62 and 68,

$$\epsilon^{(3)}_{K,T} \le 2\sqrt{2\frac{T}{K}\log(1/\delta_2)}\log\left(1 + \frac{S}{N}(e^{2\sqrt{2K\log(1/\delta_1)}\epsilon_1} - 1)\right) \tag{69}$$

Notice that if

$$0 \le \eta \le \frac{\tau(1-\rho^2)\min_{c\in[N]} p_c}{2\Delta_l^2 c(\delta_0)^2}\min\left\{\log^2\left(1 + \sqrt{\frac{2\log(1/\delta_1)}{K}}\right),\ \frac{\log^2\left(1 + \frac{N}{S}\sqrt{\frac{2K\log(1/\delta_2)}{T}}\right)}{8K\log(1/\delta_1)}\right\}, \tag{70}$$

by 57, 62, 63, and 64, we have

$$\epsilon_1 \le \frac{\log\left(1 + \frac{N}{S}\sqrt{\frac{2K\log(1/\delta_2)}{T}}\right)}{2\sqrt{2K\log(1/\delta_1)}}, \tag{71}$$

$$\epsilon_K \le 2\sqrt{2K\log(1/\delta_1)}\epsilon_1 \le \log\left(1 + \frac{N}{S}\sqrt{\frac{2K\log(1/\delta_2)}{T}}\right), \tag{72}$$

$$\widetilde{\epsilon}_K \le \log\left(1 + \sqrt{\frac{2K\log(1/\delta_2)}{T}}\right). \tag{73}$$

Thus, we have 69 holds. Plugging 57 into 69, we have

$$\epsilon^{(3)}_{K,T} \le 2\sqrt{2\frac{T}{K}\log(1/\delta_2)}\log\left(1 + \frac{S}{N}\left(\exp\left\{4c(\delta_0)\Delta_l\sqrt{\frac{\eta K\log(1/\delta_1)}{\tau(1-\rho^2)\min_{c\in[N]} p_c}}\right\} - 1\right)\right). \tag{74}$$

For any $K \geq 1$, if $\eta K \ll 1$ and $T \gg 1$, using $\log(1+x) \approx x$ and $e^x - 1 \approx x$ when $|x| \ll 1$, we can write 70 as

$$0 \leq \eta = O\left(\frac{\tau(1-\rho^2)N^2 \min_{c\in[N]} p_c \log(1/\delta_2)}{\Delta_l^2 S^2 T \log(1/\delta_0)\log(1/\delta_1)}\right), \tag{75}$$

and 74 as

$$\epsilon_{K,T}^{(3)} = O\left(\frac{S\Delta_l}{N}\sqrt{\frac{\eta T \log(1/\delta_0)\log(1/\delta_1)\log(1/\delta_2)}{\tau(1-\rho^2)\min_{c\in[N]} p_c}}\right). \tag{76}$$

H Related work

Federated Learning Current federated learning follows two paradigms. The first paradigm asks every client to learn the model using private data and communicate in model parameters. The second one uses encryption techniques to guarantee secure communication between clients. In this paper, we focus on the first paradigms Dean et al. (2012); Shokri & Shmatikov (2015); McMahan et al. (2016; 2017); Huang et al. (2021). There is a long list of works showing provable convergence algorithm for FedAvg types of algorithms in the field of optimization Li et al. (2020c; 2021); Huang et al. (2021); Khaled et al. (2019); Yu et al. (2019); Wang et al. (2019); Karimireddy et al. (2020). One line of research Li et al. (2020c); Khaled et al. (2019); Yu et al. (2019); Wang et al. (2019); Karimireddy et al. (2020) focuses on standard assumptions in optimization (such as, convex, smooth, strongly-convex, bounded gradient). The other line of work Li et al. (2021); Huang et al. (2021) proves the convergence in the regime where the model of interest is an over-parameterized neural network (also called NTK regime Jacot et al. (2018)). Extensions to general partial device participation, and arbitrary communication schemes have been well addressed in Avdyukhin & Kasiviswanathan (2021); Haddadpour & Mahdavi (2019).

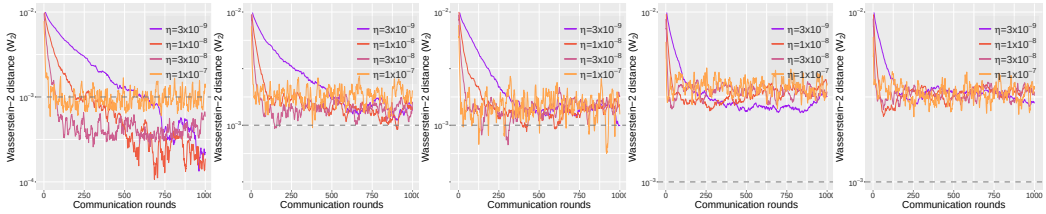
Scalable Monte Carlo methods SGLD Welling & Teh (2011) is the first stochastic gradient Monte Carlo method that tackles the scalability issue in big data problems. Ever since, variants of stochastic gradient Monte Carlo methods were proposed to accelerate the simulations by utilizing more general Markov dynamics Ma et al. (2015; 2018); Chen et al. (2014), Hessian approximation Ahn et al. (2012), parallel tempering Deng et al. (2020), as well as higher-order numerical schemes Chen et al. (2015); Li et al. (2019c); Cheng et al. (2018); Ma et al. (2021); Mou et al. (2021); Shen & Lee (2019).

Distributed Monte Carlo methods Sub-posterior aggregation was initially proposed in Neiswanger et al. (2013); Wang & Dunson; Minsker et al. (2014) to accelerate MCMC methods to cope with large datasets. Other parallel MCMC algorithms Nishihara et al. (2014); Ahn et al. (2014); Chen et al. (2016); Chowdhury & Jermaine (2018); Li et al. (2019a) propose to improve the efficiency of Monte Carlo computation in distributed or asynchronous systems. Gürbüzbalaban et al. (2021) proposed stochastic gradient Monte Carlo methods in decentralized systems. Al-Shedivat et al. (2021); Mekkaoui et al. (2021); Chen & Chao (2021) introduced empirical studies of posterior averaging in federated learning.

A Concurrent Work Parallel to our work, QLSD Vono et al. (2021) also studied the convergence of SGLD in federated settings and a compression operator was proposed to alleviate the communication overhead. By contrast, we follow the tradition in the FL community and achieve this target by solely conducting multiple local steps to balance accuracy and communication. Other interesting Bayesian federated learning algorithms can be found in Kotelevskii et al. (2022); Zhang et al. (2022).

I Simulations

Experimental details We repeat each experiment $R = 300$ times. At the k-th communication round, we obtain a set of R simulated parameters $\{\theta_{k,j}\}_{j=1}^R$, where $\theta_{k,j}$ denotes the parameter at the k-th round in the j-th independent run. The underlying distribution μ_k at round k is approximated by a Gaussian variable with the empirical mean $u_k = \frac{1}{R}\sum_{j=1}^R \theta_{k,j}$ and covariance matrix $\Sigma_k = \frac{1}{R-1}\sum_{j=1}^R (\theta_{k,j} - u_k)(\theta_{k,j} - u_k)^\top$.



(a) Full devices: $S = 50$ (b) Scheme I: $S = 40$ (c) Scheme II: $S = 40$ (d) Scheme I: $S = 30$ (e) Scheme II: $S = 30$

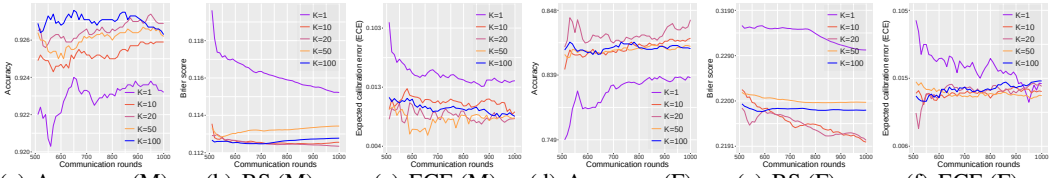
Figure 3: Convergence of FA-LD based on different device-sampling schemes.

Partial device participation We study the convergence of two popular device-sampling schemes I and II. We fix the number of local steps $K = 100$ and the total devices $N = 50$. We try to sample S devices based on different fixed learning rates η. The full device updates are also presented for a fair evaluation. As shown in Figure 3(a), larger learning rates converge faster but lead to larger biases; small learning rates, by contrast, yield diminishing biases consistently, where is in accordance with Theorem 4.7. However, in partial device scenarios, the bias becomes much less dependent on the learning rate in the long run. We observe in Figure 3(b), Figure 3(c), Figure 3(d), and Figure 3(e) that the bias caused by partial devices becomes dominant as we decrease the number of partial devices S for both schemes. Unfortunately, such a phenomenon still exists even when the algorithms converge, which suggests that the proposed partial device updates may be only appropriate for the early period of the training or simulation tasks with low accuracy demand.

J Experiments on the MNIST and Fashion-MNIST dataset

Experimental details For both the MNIST and Fashion-MNIST dataset, the temperature τ is set to be 0.05. We split the training dataset of size 60000 into 10 non-overlapping subsets uniformly at random for 10 clients. During the training, the stochastic gradient of the energy function at each step is calculated with a batch size of 200 for each client.

Warmup period To better observe the convergence under different local steps K with the existence of a warmup period, we use a warmup period of 500 communication rounds for different local steps $K = 1, 10, 20, 50, 100$. Specifically, in each setting, after the first 500 communication rounds, we collect one parameter sample every 10 communication rounds and average the predicted probabilities made by all the previous collected parameter samples to calculate three test statistics, accuracy, Brier Score (BS), and Expected Calibration Error (ECE), on the test dataset. We tune the step sizes η for the best performance and plot the curves of those test statistics against communication rounds in Figure 4.



(a) Accuracy (M) (b) BS (M) (c) ECE (M) (d) Accuracy (F) (e) BS (F) (f) ECE (F)

Figure 4: Convergence of FA-LD on the MNIST (M) and Fashion-MNIST (M) dataset with a warmup period of the first 500 communication rounds.

According to Figure 4, for the MNIST dataset, FA-LD with $K = 1$ performs the worst in terms of all three test statistics. For the Fashion-MNIST dataset, FA-LD with $K = 1$ performs the worst in terms of accuracy and BS and does not perform the best in terms of ECE. Thus, it is beneficial to have multiple local updates in federated learning under a fixed communication budget.

Among $K = 1, 10, 20, 50, 100$, according to Figure 4(a), 4(b), and 4(c), for the MNIST dataset, the optimal local step K is 20 in terms of accuracy, BS, and ECE. According to Figure 4(d), 4(e), and 4(f), for the Fashion-MNIST dataset, the optimal K in terms of accuracy is 20, the optimal K in terms of BS is 10, while the optimal K in terms of ECE is 50.

It is worth mentioning that the optimal K with a warmup period could be different from the optimal K without a warmup period (see Figure 2). For example, the optimal K in terms of BS for the Fashion-MNIST dataset is 100 without a warmup period (Figure 2(e)) but is 10 with a warmup period of 500 communication rounds (Figure 4(e)), which indicates that the budget of communication also plays a rule in determining the optimal local step K because the optimal K changes when the samples in the first 500 communication rounds are not collected.