SUPPLEMENTARY MATERIAL

Notations and conventions. For the sake of simplicity, with little abuse, we shall use the same notations for a probability distribution and its associated probability density function. For $n \geq 1$, we refer to the set of integers between 1 and n with the notation $[n]$. The d-multidimensional Gaussian probability distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ is denoted by $\mathrm{N}(\mu, \Sigma)$. Equations of the form (1) (resp. (S1)) refer to equations in the main paper (resp. in the supplement).

Denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$, and for $f : \mathbb{R}^d \to \mathbb{R}$ measurable, $\|f\|_\infty = \sup_{x \in \mathbb{R}^d} |f(x)|$. For μ a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and f a μ-integrable function, denote by $\mu(f)$ the integral of f w.r.t. μ. For $f : \mathbb{R}^d \to \mathbb{R}$ measurable, the V-norm of f is given by $\|f\|_V = \sup_{x \in \mathbb{R}^d} |f(x)|/V(x)$. Let ξ be a finite signed measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. The V-total variation distance of ξ is defined as

$$\|\xi\|_V = \sup_{\|f\|_V \leq 1} \left| \int_{\mathbb{R}^d} f(x) \mathrm{d}\xi(x) \right| .$$

If $V = 1$, then $\|\cdot\|_V$ is the total variation denoted by $\|\cdot\|_{\mathrm{TV}}$. Let U be an open set of $\mathbb{R}^d$. We denote by $\mathrm{C}^k(\mathsf{U}, \mathbb{R}^p)$ the set of $\mathbb{R}^p$-valued k-differentiable functions, respectively the set of compactly supported $\mathbb{R}^p$-valued and k-differentiable functions. Let $f : \mathsf{U} \to \mathbb{R}$, we denote by ∇f, the gradient of f if it exists. f is said to me $\mathtt{m}$-convex with $m \geq 0$ if for all $x, y \in \mathbb{R}^d$ and $t \in [0, 1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \mathtt{m}t(1-t) \|x - y\|^2 / 2 .$$

For any $a \in \mathbb{R}^d$ and $R > 0$, denote $\mathrm{B}(a, R)$ the open ball centered at a with radius R. Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces. A Markov kernel P is a mapping $\mathrm{K} : \mathsf{X} \times \mathcal{Y} \to [0, 1]$ such that for any $x \in \mathsf{X}$, $\mathrm{P}(x, \cdot)$ is a probability measure and for any $\mathsf{A} \in \mathcal{Y}$, $\mathrm{P}(\cdot, \mathsf{A})$ is measurable. For any probability measure μ on $(\mathsf{X}, \mathcal{X})$ and measurable function $f : \mathsf{Y} \to \mathbb{R}_+$ we denote $\mu\mathrm{P} = \int_{\mathsf{X}} \mathrm{P}(x, \cdot) \mathrm{d}\mu(x)$ and $\mathrm{P}f = \int_{\mathsf{Y}} f(y) \mathrm{P}(\cdot, \mathrm{d}y)$. In what follows the Dirac mass at $x \in \mathbb{R}^d$ by δ_x.

Contents

S1 Theoretical analysis of FedSOUK

This section aims at recasting the proposed methodology into a stochastic approximation framework and at stating the main assumptions required to show our theoretical results regarding FedSOUK, which uses a general unadjusted Markov kernel. Then, we will use these general results to show non-asymptotic convergence guarantees for FedSOUL, which considers an unadjusted Markov kernel associated to overdamped Langevin dynamics.

S1.1 Preliminaries

We first show that FedSOUK (see Algorithm 1 in the main paper) can be cast into a general *stochastic approximation* (SA) framework which corresponds to a federated variant of the *stochastic optimization via unadjusted kernel* (SOUK) approach proposed in De Bortoli et al. [11]. Then, the convergence guarantees for FedSOUK will follow by generalizing the proof techniques used to analyze SOUK.

Recall that $\theta = (\phi, \beta) \in \Theta$ corresponds to the parameter we are seeking to optimize where $\Theta = \Phi \times \mathsf{B} \subset \mathbb{R}^{d_\Theta}$. Define $f : \Theta \to \mathbb{R}$ of the form

$$f(\theta) = b^{-1} \sum_{i=1}^{b} f_i(\theta) \,, \tag{S1}$$

where for any $i \in [b]$ and $\theta \in \Theta$,

$$f_i(\theta) = -\log p(\theta) - b \log p\left(\mathrm{D}_i \mid \phi, \beta\right) \,, \tag{S2}$$

where $p(\theta) = p(\phi, \beta) = p(\phi)p(\beta)$ and for any $i \in [b]$, $p(\mathrm{D}_i \mid \phi, \beta)$ is defined in (1). Then, under these notations, (2) can be written as

$$\theta^\star = \arg\min_{\theta \in \Theta} f(\theta) \,. \tag{S3}$$

In addition, based on (4) and (5), the gradient of f_i defined in (S2) admits the form for $i \in [b]$,

$$\nabla f_i : \begin{cases} \mathbb{R}^{d_\Phi + d_\mathsf{B}} \to \mathbb{R}^{d_\Theta} \\ \theta \mapsto \int_{\mathbb{R}^d} H_\theta^{(i)}\left(z^{(i)}\right) \pi_\theta^{(i)}\left(\mathrm{d}z^{(i)}\right) \,, \end{cases} \tag{S4}$$

where, for any $i \in [b]$ and $\theta \in \Theta$, $\pi_\theta^{(i)} : z^{(i)} \mapsto p(z^{(i)} \mid \mathrm{D}_i, \theta)$ and for any $\theta \in \Theta$, $H_\theta^{(i)} : z^{(i)} \mapsto -\nabla_\theta \log p(\theta) - b \nabla_\theta \log p(\mathrm{D}_i, z^{(i)} \mid \theta)$.

S1.2 Main Assumptions

We make the following assumption on Θ and the family of functions $\{f_i : i \in [b]\}$.

A1. *Θ is a convex, closed subset of $\mathbb{R}^{d_\Theta}$ and $\Theta \subset \mathrm{B}(0, R_\Theta)$ for $R_\Theta > 0$.*

A2. *For any $i \in [b]$, the following conditions hold.*

(i) The function f_i defined in (S1) is convex.
(ii) There exist an open set $\mathsf{U} \in \mathbb{R}^{d_\Theta}$ and $L_f > 0$ such that $\Theta \subset \mathsf{U}$, $f_i \in \mathrm{C}^1(\mathsf{U}, \mathbb{R})$ and for any $\theta_1, \theta_2 \in \Theta$,

$$\|\nabla f_i(\theta_2) - \nabla f_i(\theta_1)\| \leq L_f \|\theta_2 - \theta_1\| \,.$$

Note that **A**2-(ii) implies that the objective function f defined in (S1) is gradient-Lipschitz with Lipschitz constant L_f.

We now consider assumptions on the family of *compression* and *partial participation* operators $\{\mathscr{C}_i, \mathscr{S}_i\}_{i\in[b]}$.

A3. *There exists a probability measure ν_1 on a measurable space $(\mathsf{X}_1, \mathcal{X}_1)$ and a family of measurable functions $\{\mathscr{C}_i : \mathbb{R}^{d_\Phi} \times \mathsf{X}_1 \to \mathbb{R}^{d_\Phi}\}_{i\in[b]}$ such that the following conditions hold.*

(i) For any $v \in \mathbb{R}^{d_\Phi}$ and any $i \in [b]$, $\int_{\mathsf{X}_1} \mathscr{C}_i(v, x^{(1)})\, \nu_1(\mathrm{d}x^{(1)}) = v$.

(ii) There exist $\{\omega_i \in \mathbb{R}_+\}_{i\in[b]}$, such that for any $v \in \mathbb{R}^{d_\Phi}$ and any $i \in [b]$,

$$\int_{\mathsf{X}_1} \left\| \mathscr{C}_i(v, x^{(1)}) - v \right\|^2 \nu_1(\mathrm{d}x^{(1)}) \le \omega_i \|v\|^2 .$$

In addition, recall that we consider the partial device participation context where at each communication round $k \ge 1$, each client has a probability $p_i \in (0, 1]$ of participating, independently from other clients.

A4. *For any $i \in [b]$, the unbiased partial participation operator $\mathscr{S}_i : \mathbb{R}^{d_\Theta} \times \mathsf{X}_2 \to \mathbb{R}^{d_\Theta}$ is defined, for any $\theta \in \mathbb{R}^{d_\Theta}$ and $x^{(2)} = \{x_i^{(2)}\}_{i\in[b]} \in \mathsf{X}_2$ with $\mathsf{X}_2 = [0, 1]^b$ by*

$$\mathscr{S}_i(\theta, x^{(2)}) = \mathbf{1}\{x_i^{(2)} \le p_i\}\theta/p_i ,$$

where $p_i \in (0, 1]$.

Note that the assumption **A**4 is equivalent to **H**4 in the main paper.

Let $V : \mathbb{R}^d \to [1, \infty)$ a measurable function. We consider the following assumption on the family $\{(H_\theta^{(i)}, \pi_\theta^{(i)}) : \theta \in \Theta, i \in [b]\}$.

A5. *For any $i \in [b]$, the following conditions hold.*

(i) For any $\theta \in \Theta$, $\pi_\theta^{(i)}(\|H_\theta^{(i)}\|) < \infty$ and $(\theta, z^{(i)}) \mapsto H_\theta^{(i)}(z^{(i)})$ is measurable.

(ii) There exists $L_H \ge 0$ such that for any $z \in \mathbb{R}^d$ and $\theta_1, \theta_2 \in \Theta$,

$$\left\| H_{\theta_2}^{(i)}(z) - H_{\theta_1}^{(i)}(z) \right\| \le L_H \|\theta_2 - \theta_1\| V^{1/2}(z) .$$

S1.3 Stochastic Approximation Framework

Let $(X_k^{(i,1)})_{k\in\mathbb{N},i\in[b]}$ a sequence of independent an identically distributed (i.i.d.) random variables with distribution ν_1 independent of the sequence $(X_k^{(i,2)})_{k\in\mathbb{N},i\in[b]}$ which is i.i.d. and with uniform distribution on $[0, 1]$. We consider a family of unadjusted Markov kernels $\{Q_{\gamma,\theta}^{(i)} : \gamma \in (0, \bar\gamma], \theta \in \Theta, i \in [b]\}$. Let $(\gamma_k)_{k\in\mathbb{N}^*} \in (\mathbb{R}_+^*)^{\mathbb{N}^*}$ a sequence of step-sizes which will be used to obtain approximate samples from $\pi_\theta^{(i)}$ using $Q_{\gamma,\theta}^{(i)}$.

We now recast the proposed approach detailed in Algorithm 1 into a stochastic approximation framework.

Starting from some initialization $(Z_0^{(1,0)}, \dots, Z_0^{(b,0)}, \theta_0) \in \mathbb{R}^{bd} \times \Theta$, we define on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the sequence $((Z_k^{(1,m)}, \dots, Z_k^{(b,m)})_{m\in[M]}, \theta_k)_{k\in\mathbb{N}}$ via the recursion for $k \in \mathbb{N}$,

$$\begin{aligned} &\text{for any } i \in [b], \text{ given } \mathcal{F}_{k-1}, (Z_k^{(i,m)})_{m\in\{0,\dots,M\}} \text{ is a Markov chain with Markov kernel } Q_{\gamma_k,\theta_k}^{(i)} \\ &\text{with } Z_k^{(i,0)} = Z_{k-1}^{(i,M)} , \\ &\theta_{k+1} = \Pi_\Theta \left[\theta_k - \boldsymbol{\eta}_{k+1} \odot \boldsymbol{\Delta}_{\theta_k} \left(Z_{k+1}^{(1:M)}, X_{k+1}^{(1)}, X_{k+1}^{(2)} \right) \right] , \end{aligned} \tag{S5}$$

where $\odot$ denotes the Hadamard product and for any $k \in \mathbb{N}$, $\mathcal{F}_k = \sigma(\theta_0, \{\{Z_l^{(i,m)}\}_{m\in[M]} : l \in \{0, \dots, k\}, i \in [b]\})$ and $\mathcal{F}_{-1} = \sigma(\theta_0, \{Z_0^{(i,0)} : i \in [b]\})$. In addition, for any $k \in \mathbb{N}$, $\boldsymbol{\eta}_{k+1} =$

$(\eta_{k+1}^{(1)}, \eta_{k+1}^{(2)})^\top$, $Z_{k+1}^{(1:M)} = ([Z_{k+1}^{(1,1:M)}]^\top, \ldots, [Z_k^{(b,1:M)}]^\top)^\top$ and for any $\theta \in \Theta$, $z^{(1:M)} \in \mathbb{R}^{Md}$, $x^{(1)} \in \mathsf{X}_1$, $x^{(2)} \in \mathsf{X}_2$,

$$\begin{aligned}\boldsymbol{\Delta}_\theta\left(z^{(1:M)}, x^{(1)}, x^{(2)}\right) &= \begin{pmatrix}\Delta_\phi\left(z^{(1:M)}, x^{(1)}, x^{(2)}\right)\\ \Delta_\beta\left(z^{(1:M)}, x^{(2)}\right)\end{pmatrix}, \\ &= \begin{pmatrix}\sum_{i=1}^b \mathscr{S}_i\left[\mathscr{C}_i\left(\Delta_\phi^{(i)}(z^{(i,1:M)}), x^{(i,1)}\right), x^{(i,2)}\right]\\ \sum_{i=1}^b \mathscr{S}_i\left[\Delta_\beta^{(i)}(z^{(i,1:M)}), x^{(i,2)}\right]\end{pmatrix}, \end{aligned} \tag{S6}$$

where $\{\Delta_\beta^{(i)}, \Delta_\phi^{(i)}\}_{i\in[b]}$ defined by

$$\Delta_\beta^{(i)}(z^{(i,1:M)}) = -\frac{1}{M}\sum_{m=1}^M \left\{(1/b)\nabla_\beta p(\beta) + \nabla_\beta \log p(z^{(i,m)} \mid \beta)\right\}$$

$$\Delta_\phi^{(i)}(z^{(i,1:M)}) = -\frac{1}{M}\sum_{m=1}^M \left\{(1/b)\nabla_\phi p(\phi) + \nabla_\phi \log p(\mathrm{D}_i \mid z^{(i,m)}, \phi)\right\} .$$

S1.4 Main Result

In order to show non-asymptotic convergence guarantees for `FedSOUK` detailed in Algorithm 1, we need additional assumptions ensuring some stability of the sequence $(Z_k^{(i,m)} : m \in \{0,\ldots,M\}, i \in [b])_{k\in\mathbb{N}}$. These conditions are stated hereafter.

A6. *For any $i \in [b]$, the following conditions hold.*

(i) There exists $A_1 \geq 1$ such that for any $p, k \in \mathbb{N}$ and $m \in \{0,\ldots,M\}$,

$$\mathbb{E}\left[[Q_{\gamma_k,\theta_k}^{(i)}]^p V(Z_k^{(i,m)}) \mid Z_0^{(i,0)}\right] \leq A_1 V(Z_0^{(i,0)}), \ \mathbb{E}\left[V(Z_0^{(i,0)})\right] < \infty ,$$

where $(Z_k^{(i,m)} : m \in \{0,\ldots,M\}, i \in [b])_{k\in\mathbb{N}}$ is defined in (S5).

(ii) There exists $A_2, A_3 \geq 1$, $\rho \in [0,1)$ such that for any $\gamma \in (0,\bar\gamma]$, $\theta \in \Theta$, $z \in \mathbb{R}^d$ and $k \in \mathbb{N}$, $Q_{\gamma,\theta}^{(i)}$ admits $\pi_{\gamma,\theta}^{(i)}$ as stationary distribution and

$$\left\|\delta_z[Q_{\gamma,\theta}^{(i)}]^k - \pi_{\gamma,\theta}^{(i)}\right\|_V \leq A_2\rho^{k\gamma}V(z)$$

$$\pi_{\gamma,\theta}^{(i)}(V) \leq A_3 .$$

(iii) There exists $\boldsymbol{\Psi} : \mathbb{R}_+^* \to \mathbb{R}_+$ such that for any $\gamma \in (0,\bar\gamma]$ and $\theta \in \Theta$,

$$\left\|\pi_{\gamma,\theta}^{(i)} - \pi_\theta^{(i)}\right\|_{V^{1/2}} \leq \boldsymbol{\Psi}(\gamma) .$$

A7. *There exists a measurable function $V : \mathbb{R}^d \to [1,\infty)$, $\boldsymbol{\Gamma}_1 : (\mathbb{R}_+^*)^2 \to \mathbb{R}_+$ and $\boldsymbol{\Gamma}_2 : (\mathbb{R}_+^*)^2 \to \mathbb{R}_+$ such that for any $\gamma_1, \gamma_2 \in (0,\bar\gamma]$ with $\gamma_2 < \gamma_1$, $\theta_1, \theta_2 \in \Theta$, $z \in \mathbb{R}^d$, $a \in [1/4, 1/2]$, we have for any $i \in [b]$,*

$$\left\|\delta_z Q_{\gamma_2,\theta_2}^{(i)} - \delta_z Q_{\gamma_1,\theta_1}^{(i)}\right\|_{V^a} \leq [\boldsymbol{\Gamma}_1(\gamma_1,\gamma_2) + \boldsymbol{\Gamma}_2(\gamma_1,\gamma_2)\,\|\theta_2 - \theta_1\|]V^{2a}(z) .$$

We are now ready to show our main result. To ease the presentation, assume for any $k \in \mathbb{N}$ that $\eta_{k+1}^{(1)} = \eta_{k+1}^{(2)} = \eta_{k+1}$ and, for any $i \in [b]$, $\gamma_{k+1}^{(i)} = \gamma_{k+1}$.

Theorem S2. *Assume A1, A2, A3, A4, A5, A6 and A7 and let for any $k \in [K]$, $\eta_k \in (0, 1/L_f]$. In addition, for any $\theta \in \Theta$, $z \in \mathbb{R}^d$ and $i \in [b]$, assume that $\|H_\theta^{(i)}(z)\| \leq V^{1/4}(z)$. Then, for any $K \in \mathbb{N}^*$, we have*

$$\mathbb{E}\left[\frac{\sum_{k=1}^K \eta_k\{f(\theta_k) - f(\theta^\star)\}}{\sum_{k=1}^K \eta_k}\right] \leq \frac{E_K}{\sum_{k=1}^K \eta_k} ,$$

where, for any $K \in \mathbb{N}^*$,

$$E_K = 2R_\Theta^2 + 2A_1 \sup_{i\in[b], m\in[M]}\left\{\mathbb{E}\left[V^{1/2}(Z_0^{(i,m)})\right]\right\}\sum_{k=1}^K \eta_k^2\left(8bL_f^2R_\Theta^2 + \sum_{i=1}^b \frac{(\omega_i + 1 + p_i)}{p_i}\right)$$

$$+ b \sup_{i\in[b],m\in[M]} \left\{C_3^{(i,m)}\right\} \left[\sum_{k=1}^{K} |\eta_k - \eta_{k-1}|\gamma_{k-1}^{-1} + \sum_{k=1}^{K} \eta_k^2 \gamma_{k-1}^{-1} + \eta_K/\gamma_K - \eta_1/\gamma_1\right]$$

$$+ b.A_1 C_{c,2} \sup_{i\in[b],m\in[M]} \left\{\mathbb{E}\left[V(Z_0^{(i,m)})\right]\right\} \sum_{k=1}^{K} \eta_k \gamma_k^{-1} \left[\gamma_k^{-1} \left\{\mathbf{\Lambda}_1(\gamma_{k-1},\gamma_k) + \mathbf{\Lambda}_2(\gamma_{k-1},\gamma_k)\eta_k\right\} + \eta_k\right]$$

$$+ b \sum_{k=1}^{K} \eta_k \mathbf{\Psi}(\gamma_{k-1}) \,,$$

with $\{C_3^{(i,m)}\}_{i\in[b],m\in[M]}$ *defined in Lemma S5 and* $C_{c,2}$ *defined in Lemma S6.*

Proof. The proof follows by using the fact that (S23) is a $(\mathcal{F}_{k-1})_{k\in\mathbb{N}^*}$-martingale increment and by combining Lemma S1-S7. □

S1.5 Supporting Lemmata

For convenience, we define the following quantities that will naturally appear in our derivations. For any $k \in \mathbb{N}^*$, let

$$\boldsymbol{\epsilon}_k = \mathbf{\Delta}_{\theta_{k-1}}\left(Z_k^{(1:M)}, X_k^{(1)}, X_k^{(2)}\right) - \nabla f(\theta_{k-1}) \,, \tag{S7}$$

where $\mathbf{\Delta}_\theta$ is defined in (S6).

The following lemma first provides a non-asymptotic upper bound on $\sum_{k=1}^{K} \eta_k \{f(\theta_k) - f(\theta^\star)\}$ involving key quantities to control such as the Monte Carlo approximation error term (S7).

Lemma S1. *Assume* **A***1 and* **A***2, and let for any* $k \in [K]$, $\eta_k \in (0, 1/L_f]$. *Then, for any* $K \in \mathbb{N}^*$, *we have*

$$\sum_{k=1}^{K} \eta_k \{f(\theta_k) - f(\theta^\star)\} \leq 2R_\Theta^2 + \sum_{k=1}^{K} \eta_k^2 \left\|\boldsymbol{\epsilon}_k\right\|^2 - \sum_{k=1}^{K} \eta_k \left\langle \Pi_\Theta\left(\theta_{k-1} - \eta_k \nabla f(\theta_{k-1})\right) - \theta^\star, \boldsymbol{\epsilon}_k \right\rangle \,,$$

where $\{\boldsymbol{\epsilon}_k\}_{k=1}^K$ *is defined in* (S7).

Proof. Let $k \in \mathbb{N}$. Since Θ is closed and convex by **A**1, the indicator function ι_Θ, defined for any $u \in \mathbb{R}^{d_\Phi + d_\mathsf{B}}$ by $\iota_\Theta(u) = 0$ if $u \in \Theta$ and $\iota_\Theta(u) = \infty$ otherwise, is lower semi-continuous and convex. Therefore by Atchadé et al. [3, Lemma 7] we have

$$\iota_\mathsf{B}(\beta_{k+1}) - \iota_\mathsf{B}(\beta_\star) \leq -\frac{1}{\eta_{k+1}} \left\langle \beta_{k+1} - \beta_\star, \beta_{k+1} - \beta_k + \eta_{k+1} \Delta_{\beta_k}\left(Z_{k+1}^{(1:M)}, X_{k+1}^{(2)}\right) \right\rangle \,, \tag{S8}$$

$$\iota_\Phi(\phi_{k+1}) - \iota_\Phi(\phi_\star) \leq -\frac{1}{\eta_{k+1}} \left\langle \phi_{k+1} - \phi_\star, \phi_{k+1} - \phi_k + \eta_{k+1} \Delta_{\phi_k}\left(Z_{k+1}^{(1:M)}, X_{k+1}^{(1)}, X_{k+1}^{(2)}\right) \right\rangle \,, \tag{S9}$$

where $\theta^\star = (\phi_\star, \beta_\star)$ is defined in (S3). In addition by **A**2-(ii), we have for any $i \in [b]$,

$$f_i(\theta_{k+1}) - f_i(\theta_k) \leq \langle \nabla f_i(\theta_k), \theta_{k+1} - \theta_k \rangle + \frac{L_f}{2} \left\|\theta_{k+1} - \theta_k\right\|^2 \,. \tag{S10}$$

Using (S10) and the fact that for any $k \in \mathbb{N}$, $\eta_{k+1} \leq 1/L_f$, we have

$$\begin{aligned} f(\theta_{k+1}) - f(\theta_k) &\leq \langle \nabla_\beta f(\theta_k), \beta_{k+1} - \beta_k \rangle + \frac{L_f}{2} \left\|\beta_{k+1} - \beta_k\right\|^2 \\ &\quad + \langle \nabla_\phi f(\theta_k), \phi_{k+1} - \phi_k \rangle + \frac{L_f}{2} \left\|\phi_{k+1} - \phi_k\right\|^2 \\ &\leq \langle \nabla_\beta f(\theta_k), \beta_{k+1} - \beta_k \rangle + \frac{1}{2\eta_{k+1}} \left\|\beta_{k+1} - \beta_k\right\|^2 \\ &\quad + \langle \nabla_\phi f(\theta_k), \phi_{k+1} - \phi_k \rangle + \frac{1}{2\eta_{k+1}} \left\|\phi_{k+1} - \phi_k\right\|^2 \,. \end{aligned} \tag{S11}$$

Finally, A2-(i) implies for any $i \in [b]$,

$$f_i(\theta_k) - f_i(\theta^\star) \leq -\langle \nabla f_i(\theta_k), \theta_\star - \theta_k \rangle \,. \tag{S12}$$

For any $i \in [b]$, let $F_i = f_i + \iota_\Theta$ and let $F = (1/b)\sum_{i=1}^b F_i$. Using this notation and combining (S8), (S9), (S11) and (S12), we have

$$\begin{aligned}
&F(\theta_{k+1}) - F(\theta^\star) \\
&= f(\theta_{k+1}) - f(\theta_k) + f(\theta_k) - f(\theta^\star) + \iota_\Phi(\phi_{k+1}) - \iota_\Phi(\phi_\star) + \iota_{\mathsf{B}}(\beta_{k+1}) - \iota_{\mathsf{B}}(\beta_\star) \\
&\leq -\left\langle \beta_{k+1} - \beta_\star, \Delta_{\beta_k}\left(Z_{k+1}^{(i,1:M)}, X_{k+1}^{(2)}\right) - \nabla_\beta f(\theta_k) \right\rangle - \langle \beta_{k+1} - \beta_\star, \beta_{k+1} - \beta_k \rangle \\
&\quad - \left\langle \phi_{k+1} - \phi_\star, \Delta_{\phi_k}\left(Z_{k+1}^{(1:M)}, X_{k+1}^{(1)}, X_{k+1}^{(2)}\right) - \nabla_\phi f(\theta_k) \right\rangle - \langle \phi_{k+1} - \phi_\star, \phi_{k+1} - \phi_k \rangle \\
&\quad + \frac{1}{2\eta_{k+1}} \|\beta_{k+1} - \beta_k\|^2 + \frac{1}{2\eta_{k+1}} \|\phi_{k+1} - \phi_k\|^2 \\
&= -\left\langle \theta_{k+1} - \theta_\star, \boldsymbol{\Delta}_{\theta_k}\left(Z_{k+1}^{(1:M)}, X_{k+1}^{(1)}, X_{k+1}^{(2)}\right) - \nabla f(\theta_k) \right\rangle \\
&\quad + \frac{1}{2\eta_{k+1}} \left[\|\phi_k - \phi_\star\|^2 - \|\phi_{k+1} - \phi_\star\|^2\right] + \frac{1}{2\eta_{k+1}} \left[\|\beta_k - \beta_\star\|^2 - \|\beta_{k+1} - \beta_\star\|^2\right] \,.
\end{aligned} \tag{S13}$$

From (S13), it follows for any $K \in \mathbb{N}^*$ that

$$\begin{aligned}
&\sum_{k=1}^K \eta_k \{F(\theta_k) - F(\theta^\star)\} \\
&\leq -\sum_{k=1}^K \eta_k \left\langle \theta_k - \theta_\star, \boldsymbol{\Delta}_{\theta_{k-1}}\left(Z_k^{(1:M)}, X_k^{(1)}, X_k^{(2)}\right) - \nabla f(\theta_{k-1}) \right\rangle \\
&+ \frac{1}{2}\|\phi_0 - \phi_\star\|^2 - \frac{1}{2}\|\phi_K - \phi_\star\|^2 + \frac{1}{2}\|\beta_0 - \beta_\star\|^2 - \frac{1}{2}\|\beta_K - \beta_\star\|^2 \\
&\leq -\sum_{k=1}^K \eta_k \left\langle \theta_k - \theta_\star, \boldsymbol{\Delta}_{\theta_{k-1}}\left(Z_k^{(1:M)}, X_k^{(1)}, X_k^{(2)}\right) - \nabla f(\theta_{k-1}) \right\rangle + \frac{1}{2}\|\theta_0 - \theta^\star\|^2 \\
&= -\sum_{k=1}^K \eta_k \left\langle \theta_k - \Pi_\Theta\left(\theta_{k-1} - \eta_k \nabla f(\theta_{k-1})\right), \boldsymbol{\Delta}_{\theta_{k-1}}\left(Z_k^{(1:M)}, X_k^{(1)}, X_k^{(2)}\right) - \nabla f(\theta_{k-1}) \right\rangle \\
&- \sum_{k=1}^K \eta_k \left\langle \Pi_\Theta\left(\theta_{k-1} - \eta_k \nabla f(\theta_{k-1})\right) - \theta^\star, \boldsymbol{\Delta}_{\theta_{k-1}}\left(Z_k^{(1:M)}, X_k^{(1)}, X_k^{(2)}\right) - \nabla f(\theta_{k-1}) \right\rangle \\
&+ \frac{1}{2}\|\theta_0 - \theta^\star\|^2 \\
&\leq \sum_{k=1}^K \eta_k^2 \left\| \boldsymbol{\Delta}_{\theta_{k-1}}\left(Z_k^{(1:M)}, X_k^{(1)}, X_k^{(2)}\right) - \nabla f(\theta_{k-1}) \right\|^2 + \frac{1}{2}\|\theta_0 - \theta^\star\|^2 \\
&- \sum_{k=1}^K \eta_k \left\langle \Pi_\Theta\left(\theta_{k-1} - \eta_k \nabla f(\theta_{k-1})\right) - \theta^\star, \boldsymbol{\Delta}_{\theta_{k-1}}\left(Z_k^{(1:M)}, X_k^{(1)}, X_k^{(2)}\right) - \nabla f(\theta_{k-1}) \right\rangle \,,
\end{aligned}$$

where we used Atchadé et al. [3, Lemma 7] and the Cauchy-Schwarz inequality in the last inequality. The proof is concluded using $f \leq F$, $f(\theta^\star) = F(\theta^\star)$ since $\theta^\star \in \Theta$, and by noting that under A1 we have $\|\theta_0 - \theta^\star\| \leq 2R_\Theta$. □

Lemma S1 involves two key quantities to upper bound namely $\|\boldsymbol{\epsilon}_k\|$ and $\langle \Pi_\Theta\left(\theta_{k-1} - \eta_k \nabla f(\theta_{k-1})\right) - \theta^\star, \boldsymbol{\epsilon}_k \rangle$ for any $k \in \mathbb{N}^*$. Our next lemmata aim at controlling the expectations of these two terms. In particular, Lemma S2 and Lemma S3 show that the impacts of Monte Carlo approximation, partial participation and compression can be decoupled.

To this end, define for any $k \in \mathbb{N}^*$ and $i \in [b]$

$$\varepsilon_{\beta,k}^{(i)} = \frac{1}{M}\sum_{m=1}^M H_{\beta_{k-1}}^{(i)}\left(Z_k^{(i,m)}\right) - \nabla_\beta f_i(\theta_{k-1}) \,,$$

$$\varepsilon_{\phi,k}^{(i)} = \frac{1}{M}\sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}\left(Z_k^{(i,m)}\right) - \nabla_\phi f_i(\theta_{k-1})\,,$$
$$\varepsilon_{\theta,k}^{(i)} = \left[\varepsilon_{\beta,k}^{(i)}, \varepsilon_{\phi,k}^{(i)}\right]\,, \tag{S14}$$

where, for any $k \in \mathbb{N}^*$ and $i \in [b]$, $H_{\theta_{k-1}}^{(i)}(Z_k^{(i,m)}) = [H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), H_{\beta_{k-1}}^{(i)}(Z_k^{(i,m)})]$ is defined in (S4).

Lemma S2 shows that $\|\boldsymbol{\epsilon}_k\|$ can be upper bounded by a quantity involving the norm of $\{H_\theta^{(i)}\}_{i\in[b]}$.

Lemma S2. *Assume **A**1, **A**2, **A**3 and **A**4. Then, for any $k \in \mathbb{N}^*$, we have*

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_k\|^2\right] \le \frac{1}{M}\sum_{i=1}^{b}\frac{(\omega_i+1+p_i)}{p_i}\left\{\sum_{m=1}^{M}\mathbb{E}\left[\left\|H_{\theta_{k-1}}^{(i)}(Z_k^{(i,m)})\right\|^2\right]\right\} + 8bL_f^2R_\Theta^2\,,$$

where $\{\boldsymbol{\epsilon}_k\}_{k=1}^K$ is defined in (S7).

Proof. Let $k \in \mathbb{N}^*$. Then by using (S6), we have

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{\epsilon}_k\|^2\right] &= \mathbb{E}\left[\left\|\sum_{i=1}^{b}\mathscr{S}_i\left[\mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M}H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)}\right), X_k^{(i,2)}\right] - \nabla_\phi f(\theta_{k-1})\right\|^2\right] \\ &\quad + \mathbb{E}\left[\left\|\sum_{i=1}^{b}\mathscr{S}_i\left[\frac{1}{M}\sum_{m=1}^{M}H_{\beta_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,2)}\right] - \nabla_\beta f(\theta_{k-1})\right\|^2\right]\,.\end{aligned} \tag{S15}$$

Using **A**3 and **A**4, it follows that

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{i=1}^{b}\mathscr{S}_i\left[\mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M}H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)}\right), X_k^{(i,2)}\right] - \nabla_\phi f(\theta_{k-1})\right\|^2\right] \\ &\quad= \mathbb{E}\left[\left\|\sum_{i=1}^{b}\left\{\mathscr{S}_i\left[\mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M}H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)}\right), X_k^{(i,2)}\right]\right.\right. \\ &\qquad\left.\left. - \mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M}H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)}\right)\right\}\right\|^2\right] \\ &\qquad+ \mathbb{E}\left[\left\|\sum_{i=1}^{b}\mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M}H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)}\right) - \nabla_\phi f(\theta_{k-1})\right\|^2\right]\,.\end{aligned} \tag{S16}$$

In addition, by **A**3-(i) and **A**3-(ii), we obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{i=1}^{b}\left\{\mathscr{S}_i\left[\mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M}H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)}\right), X_k^{(i,2)}\right]\right.\right. \\ &\left.\left. - \mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M}H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)}\right)\right\}\right\|^2\right] \\ &= \sum_{i=1}^{b}\mathbb{E}\left[\left\|\mathscr{S}_i\left[\mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M}H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)}\right), X_k^{(i,2)}\right]\right.\right. \\ &\left.\left. - \mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M}H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)}\right)\right\|^2\right] \\ &\le \sum_{i=1}^{b}\left(\frac{1-p_i}{p_i}\right)\mathbb{E}\left[\left\|\mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M}H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)}\right)\right\|^2\right]\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{b} \left(\frac{1-p_i}{p_i} \right) \mathbb{E} \left[\left\| \mathscr{C}_i \left(\frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(1,i)} \right) - \frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}) \right\|^2 \right] \\
&+ \sum_{i=1}^{b} \left(\frac{1-p_i}{p_i} \right) \mathbb{E} \left[\left\| \frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}) \right\| \right]^2 \\
&\leq \sum_{i=1}^{b} \left[\left(\frac{1-p_i}{p_i} \right) (\omega_i + 1) \right] \mathbb{E} \left[\left\| \frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}) \right\|^2 \right] \\
&= \frac{1}{M^2} \sum_{i=1}^{b} \left[\left(\frac{1-p_i}{p_i} \right) (\omega_i + 1) \right] \mathbb{E} \left[\left\| \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}) \right\|^2 \right] . \qquad \text{(S17)}
\end{aligned}
$$

Similarly, by **A**3-(i) and **A**3-(ii), we have

$$
\begin{aligned}
&\mathbb{E} \left[\left\| \sum_{i=1}^{b} \mathscr{C}_i \left(\frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)} \right) - \nabla_\phi f(\theta_{k-1}) \right\|^2 \right] \\
&= \mathbb{E} \left[\left\| \sum_{i=1}^{b} \left[\mathscr{C}_i \left(\frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)} \right) - \frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}) \right] \right. \right. \\
&\left. \left. + \sum_{i=1}^{b} \frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}) - \nabla_\phi f(\theta_{k-1}) \right\|^2 \right] \\
&= \sum_{i=1}^{b} \mathbb{E} \left[\left\| \mathscr{C}_i \left(\frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)} \right) - \frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}) \right\|^2 \right] \\
&+ \mathbb{E} \left[\left\| \sum_{i=1}^{b} \frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}) - \nabla_\phi f(\theta_{k-1}) \right\|^2 \right] \\
&\leq \sum_{i=1}^{b} \omega_i \mathbb{E} \left[\left\| \frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}) \right\|^2 \right] \\
&+ \mathbb{E} \left[\left\| \sum_{i=1}^{b} \frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}) - \nabla_\phi f(\theta_{k-1}) \right\|^2 \right] \\
&= \frac{1}{M^2} \sum_{i=1}^{b} \omega_i \mathbb{E} \left[\left\| \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}) \right\|^2 \right] \\
&+ \mathbb{E} \left[\left\| \sum_{i=1}^{b} \frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}) - \nabla_\phi f(\theta_{k-1}) \right\|^2 \right] . \qquad \text{(S18)}
\end{aligned}
$$

By plugging (S17) and (S18) into (S16), we finally obtain

$$
\begin{aligned}
&\mathbb{E} \left[\left\| \sum_{i=1}^{b} \mathscr{S}_i \left[\mathscr{C}_i \left(\frac{1}{M} \sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)} \right), X_k^{(i,2)} \right] - \nabla_\phi f(\theta_{k-1}) \right\|^2 \right] \\
&\leq \frac{1}{M^2} \sum_{i=1}^{b} \frac{(\omega_i + 1 - p_i)}{p_i} \mathbb{E} \left[\left\| \sum_{m=1}^{M} H_{\theta_{k-1}}^{(i)}(Z_k^{(i,m)}) \right\|^2 \right] + \sum_{i=1}^{b} \mathbb{E} \left[\left\| \varepsilon_{\phi,k}^{(i)} \right\|^2 \right] . \qquad \text{(S19)}
\end{aligned}
$$

Finally, using the same arguments, we have under **H**4,

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{i=1}^{b}\mathscr{S}_i\left[\frac{1}{M}\sum_{m=1}^{M}H^{(i)}_{\beta_{k-1}}(Z_k^{(i,m)}),X_k^{(i,2)}\right]-\nabla_\beta f(\theta_{k-1})\right\|^2\right]\\
&\leq \frac{1}{M^2}\sum_{i=1}^{b}\left(\frac{1-p_i}{p_i}\right)\mathbb{E}\left[\left\|\sum_{m=1}^{M}H^{(i)}_{\beta_{k-1}}(Z_k^{(i,m)})\right\|^2\right]\\
&+\mathbb{E}\left[\left\|\sum_{i=1}^{b}\frac{1}{M}\sum_{m=1}^{M}H^{(i)}_{\beta_{k-1}}(Z_k^{(i,m)})-\nabla_\beta f(\theta_{k-1})\right\|^2\right]\\
&\leq \frac{1}{M^2}\sum_{i=1}^{b}\left(\frac{1-p_i}{p_i}\right)\mathbb{E}\left[\left\|\sum_{m=1}^{M}H^{(i)}_{\beta_{k-1}}(Z_k^{(i,m)})\right\|^2\right]\\
&+\sum_{i=1}^{b}\mathbb{E}\left[\left\|\varepsilon^{(i)}_{\beta,k}\right\|^2\right].
\end{aligned}
$$

Combining (S15) and (S19) and using (S14), lead to

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\epsilon}_k\|^2\right]&\leq\frac{1}{M^2}\sum_{i=1}^{b}\frac{(\omega_i+1-p_i)}{p_i}\mathbb{E}\left[\left\|\sum_{m=1}^{M}H^{(i)}_{\theta_{k-1}}(Z_k^{(i,m)})\right\|^2\right]+\sum_{i=1}^{b}\mathbb{E}\left[\left\|\varepsilon^{(i)}_{\theta,k}\right\|^2\right]\\
&\leq\frac{1}{M}\sum_{i=1}^{b}\frac{(\omega_i+1+p_i)}{p_i}\left\{\sum_{m=1}^{M}\mathbb{E}\left[\left\|H^{(i)}_{\theta_{k-1}}(Z_k^{(i,m)})\right\|^2\right]\right\}+2\sum_{i=1}^{b}\sup_{\theta\in\Theta}\|\nabla f_i(\theta)\|^2\\
&\leq\frac{1}{M}\sum_{i=1}^{b}\frac{(\omega_i+1+p_i)}{p_i}\left\{\sum_{m=1}^{M}\mathbb{E}\left[\left\|H^{(i)}_{\theta_{k-1}}(Z_k^{(i,m)})\right\|^2\right]\right\}+2L_f^2\sum_{i=1}^{b}\sup_{\theta\in\Theta}\left\|\theta-\theta^{\star,(i)}\right\|^2,
\end{aligned}
$$

where we used **A**2 for the last inequality and $\theta^{\star,(i)}$ is a minimizer of f_i. The proof is concluded using for any $i\in[b]$ that $\left\|\theta-\theta^{\star,(i)}\right\|\leq 2R_\Theta$ by **A**1. □

We now control the quantity $\langle\Pi_\Theta\left(\theta_{k-1}-\eta_k\nabla f(\theta_{k-1})\right)-\theta^\star,\boldsymbol{\epsilon}_k\rangle$ which appears in Lemma S1.

Lemma S3. *Assume **A**1, **A**3 and **A**4. Then, for any $k\in\mathbb{N}^*$, we have*

$$
\mathbb{E}\left[\langle\Pi_\Theta\left(\theta_{k-1}-\eta_k\nabla f(\theta_{k-1})\right)-\theta^\star,\boldsymbol{\epsilon}_k\rangle\right]\leq\sum_{i=1}^{b}\mathbb{E}\left[\left\langle\Pi_\Theta\left(\theta_{k-1}-\eta_k\nabla f(\theta_{k-1})\right)-\theta^\star,\varepsilon^{(i)}_{\theta,k}\right\rangle\right],
$$

where $\{\boldsymbol{\epsilon}_k\}_{k=1}^K$ is defined in (S7).

Proof. Let $a_k=\Pi_\Theta\left(\theta_{k-1}-\eta_k\nabla f(\theta_{k-1})\right)-\theta^\star$, $a_k^{(\phi)}=\Pi_\Phi\left(\phi_{k-1}-\eta_k\nabla_\phi f(\theta_{k-1})\right)-\phi_\star$ and $a_k^{(\beta)}=\Pi_{\mathsf{B}}\left(\beta_{k-1}-\eta_k\nabla_\beta f(\theta_{k-1})\right)-\beta_\star$. We have

$$
\begin{aligned}
\langle a_k,\boldsymbol{\epsilon}_k\rangle&=\left\langle a_k^{(\phi)},\sum_{i=1}^{b}\left\{\mathscr{S}_i\left[\mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M}H^{(i)}_{\phi_{k-1}}(Z_k^{(i,m)}),X_k^{(i,1)}\right),X_k^{(i,2)}\right]-\frac{1}{M}\sum_{m=1}^{M}H^{(i)}_{\phi_{k-1}}(Z_k^{(i,m)})\right\}\right\rangle\\
&+\sum_{i=1}^{b}\left\langle a_k^{(\phi)},\frac{1}{M}\sum_{m=1}^{M}H^{(i)}_{\phi_{k-1}}(Z_k^{(i,m)})-\nabla_\phi f_i(\theta_{k-1})\right\rangle\\
&+\sum_{i=1}^{b}\left\langle a_k^{(\beta)},\frac{1}{M}\sum_{m=1}^{M}H^{(i)}_{\beta_{k-1}}(Z_k^{(i,m)})-\nabla_\beta f_i(\theta_{k-1})\right\rangle\\
&=\left\langle a_k^{(\phi)},\sum_{i=1}^{b}\mathscr{S}_i\left[\mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M}H^{(i)}_{\phi_{k-1}}(Z_k^{(i,m)}),X_k^{(i,1)}\right),X_k^{(i,2)}\right]-\frac{1}{M}\sum_{m=1}^{M}H^{(i)}_{\phi_{k-1}}(Z_k^{(i,m)})\right\rangle
\end{aligned}
$$

$$+ \sum_{i=1}^{b} \left\langle a_k, \varepsilon_{\theta,k}^{(i)} \right\rangle , \tag{S20}$$

where the last line follows from (S14). Using **A**3 and **H**4, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\langle a_k^{(\phi)}, \sum_{i=1}^{b} \mathscr{S}_i \left[\mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)}\right), X_k^{(i,2)}\right] - \frac{1}{M}\sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)})\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle a_k^{(\phi)}, \sum_{i=1}^{b} \mathbb{E}^{\mathcal{F}_{k-1}}\left[\mathscr{S}_i\left\{\mathscr{C}_i\left(\frac{1}{M}\sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)}), X_k^{(i,1)}\right), X_k^{(i,2)}\right\} - \frac{1}{M}\sum_{m=1}^{M} H_{\phi_{k-1}}^{(i)}(Z_k^{(i,m)})\right]\right\rangle\right] \\
&= 0 .
\end{aligned}$$

The proof is concluded by taking the expectation in (S20) and using the previous result. □

Similar to De Bortoli et al. [11, Appendix C.3], we now decompose the Monte Carlo error terms $\{\varepsilon_{\theta,k}^{(i)}\}_{i\in[b],k\in[K]}$ in order to end up with an upper bound on $\sum_{k=1}^{K}\eta_k\{f(\theta_k) - f(\theta^\star)\}/(\sum_{k=1}^{K}\eta_k)$ which vanishes when $\lim_{k\to\infty}\eta_k = 0_+$ and $\lim_{k\to\infty}\gamma_k = 0_+$.

For any $\theta \in \Theta$ and $\gamma \in (0, \bar{\gamma}]$, let for any $i \in [b]$, a function $\hat{H}_{\gamma,\theta}^{(i)} : \mathbb{R}^d \to \mathbb{R}^{d_\Theta}$ defined for any $z \in \mathbb{R}^d$ by

$$\hat{H}_{\gamma,\theta}^{(i)}(z) = \sum_{j\in\mathbb{N}} \left\{\left[R_{\gamma,\theta}^{(i)}\right]^j H_\theta^{(i)}(z) - \pi_{\gamma,\theta}^{(i)}(H_\theta^{(i)})\right\} ,$$

where $R_{\gamma,\theta}^{(i)}$ is the Markov kernel associated with the discretized overdamped Langevin dynamics targetting $\pi_\theta^{(i)}$, and where $\pi_{\gamma,\theta}^{(i)}$ denotes the invariant distribution of $R_{\gamma,\theta}^{(i)}$. By **A**5 and **A**6-(i)-(ii), for any $\theta \in \Theta$, $\gamma \in (0, \bar{\gamma}]$ and $i \in [b]$, $\hat{H}_{\gamma,\theta}^{(i)}$ is solution of the *Poisson* equation defined by

$$(\mathrm{Id} - R_{\gamma,\theta}^{(i)})\hat{H}_{\gamma,\theta}^{(i)} = H_\theta - \pi_{\gamma,\theta}^{(i)}(H_\theta) . \tag{S21}$$

In addition, note that using **A**6-(i) and De Bortoli et al. [11, Lemma 10], it follows for any $\theta \in \Theta$, $i \in [b]$ and $z \in \mathbb{R}^d$ that

$$\left\|\hat{H}_{\gamma,\theta}^{(i)}(z)\right\| \le C_{\hat{H}}\gamma^{-1}V^{1/4}(z) , \tag{S22}$$

where $C_{\hat{H}} = 8A_2\log^{-1}(1/\rho)\rho^{-\bar{\gamma}/4}$.

Using (S21), we can decompose the Monte Carlo error terms, for any $i \in [b], k \in [K]$ as $\varepsilon_{\theta,k}^{(i)} = (1/M)\sum_{m=1}^{M}\{\varepsilon_{\theta,k,m}^{(i,a)} + \varepsilon_{\theta,k,m}^{(i,b)} + \varepsilon_{\theta,k,m}^{(i,c)} + \varepsilon_{\theta,k,m}^{(i,d)}\}$ with, for any $m \in [M]$,

$$\begin{aligned}
\varepsilon_{\theta,k,m}^{(i,a)} &= \hat{H}_{\gamma_{k-1},\theta_{k-1}}^{(i)}(Z_k^{(i,m)}) - R_{\gamma_{k-1},\theta_{k-1}}^{(i)}\hat{H}_{\gamma_{k-1},\theta_{k-1}}^{(i)}(Z_{k-1}^{(i,m)}) \\
\varepsilon_{\theta,k,m}^{(i,b)} &= R_{\gamma_{k-1},\theta_{k-1}}^{(i)}\hat{H}_{\gamma_{k-1},\theta_{k-1}}^{(i)}(Z_{k-1}^{(i,m)}) - R_{\gamma_k,\theta_k}^{(i)}\hat{H}_{\gamma_k,\theta_k}^{(i)}(Z_k^{(i,m)}) \\
\varepsilon_{\theta,k,m}^{(i,c)} &= R_{\gamma_k,\theta_k}^{(i)}\hat{H}_{\gamma_k,\theta_k}^{(i)}(Z_k^{(i,m)}) - R_{\gamma_{k-1},\theta_{k-1}}^{(i)}\hat{H}_{\gamma_{k-1},\theta_{k-1}}^{(i)}(Z_k^{(i,m)}) \\
\varepsilon_{\theta,k,m}^{(i,d)} &= \pi_{\gamma_{k-1},\theta_{k-1}}^{(i)}(H_{\theta_{k-1}}^{(i)}) - \pi_{\theta_{k-1}}^{(i)}(H_{\theta_{k-1}}^{(i)}) .
\end{aligned} \tag{S23}$$

The following lemmata aim at upper bounding these four error terms.

Lemma S4. *Assume **A**1, **A**2, **A**5 and **A**6, and for any $\theta \in \Theta$, $z \in \mathbb{R}^d$ and $i \in [b]$, assume that $\|H_\theta^{(i)}(z)\| \le V^{1/4}(z)$. Then, for any $i \in [b]$, $m \in [M]$, $k \in \mathbb{N}^*$, we have*

$$\mathbb{E}\left[\left\|\varepsilon_{\theta,k,m}^{(i,a)}\right\|^2\right] \le A_1 C_{\hat{H}}^2\gamma_{k-1}^{-2}\mathbb{E}\left[V^{1/2}\left(Z_0^{(i,m)}\right)\right] ,$$

where $C_{\hat{H}}$ is defined in (S22).

Proof. The proof follows from De Bortoli et al. [11, Lemma 14]. □

Lemma S5. *Assume A1, A2, A6 and for any $\theta \in \Theta$, $z \in \mathbb{R}^d$ and $i \in [b]$, assume that $\|H_\theta^{(i)}(z)\| \leq V^{1/4}(z)$. Then, for any $i \in [b]$, $m \in [M]$, $k \in \mathbb{N}^*$, we have*

$$\mathbb{E}\left[\left\|\sum_{k=1}^{K} \eta_k \langle \Pi_\Theta\left(\theta_{k-1} - \eta_k \nabla f(\theta_{k-1})\right) - \theta^\star, \varepsilon_{\theta,k,m}^{(i,b)} \rangle\right\|\right]$$
$$\leq C_3^{(i,m)} \left[\sum_{k=1}^{K} |\eta_k - \eta_{k-1}| \gamma_{k-1}^{-1} + \sum_{k=1}^{K} \eta_k^2 \gamma_{k-1}^{-1} + \eta_K/\gamma_K - \eta_1/\gamma_1\right],$$

where, for any $i \in [b]$ and $m \in [M]$,

$$C_3^{(i,m)} = A_1 C_{\hat{H}}(2R_\Theta(2 + L_f) + 1 + \eta_1 L_f)\mathbb{E}\left[V^{1/4}(Z_0^{(i,m)})\right].$$

Proof. The proof follows from De Bortoli et al. [11, Lemma 15]. □

Lemma S6. *Assume A1, A2, A5, A6 and A7. In addition, for any $\theta \in \Theta$, $z \in \mathbb{R}^d$ and $i \in [b]$, assume that $\|H_\theta^{(i)}(z)\| \leq V^{1/4}(z)$. Then, for any $i \in [b]$, $m \in [M]$, $k \in \mathbb{N}^*$, we have*

$$\mathbb{E}\left[\left\|\varepsilon_{\theta,k,m}^{(i,c)}\right\|\right] \leq A_1 \mathbb{E}\left[V(Z_0^{(i,m)})\right] C_{c,2} \gamma_k^{-1}\left[\gamma_k^{-1}\left\{\mathbf{\Gamma}_1(\gamma_{k-1}, \gamma_k) + \mathbf{\Gamma}_2(\gamma_{k-1}, \gamma_k)\eta_k\right\} + \eta_k\right],$$

where

$$C_{c,2} = 4A_2 \log^{-1}(1/\rho)\rho^{-\bar{\gamma}/2} \max\{L_H C_{c,1} + 2A_2 \log^{-1}(1/\rho)\rho^{-\bar{\gamma}/2}\},$$
$$C_{c,1} = 4A_1 A_2 \log^{-1}(1/\rho)\rho^{-\bar{\gamma}/2}\mathbb{E}\left[V(Z_0^{(i,m)})\right].$$

Proof. The proof follows from De Bortoli et al. [11, Lemma 16]. □

Lemma S7. *Assume A1, A2, A6 and for any $\theta \in \Theta$, $z \in \mathbb{R}^d$ and $i \in [b]$, assume that $\|H_\theta^{(i)}(z)\| \leq V^{1/4}(z)$. Then, for any $i \in [b]$, $m \in [M]$, $k \in \mathbb{N}^*$, we have*

$$\mathbb{E}\left[\left\|\varepsilon_{\theta,k,m}^{(i,d)}\right\|\right] \leq \mathbf{\Psi}(\gamma_{k-1}).$$

Proof. The proof follows from De Bortoli et al. [11, Lemma 17].

□

S2 Application to `FedSOUL`

We now apply Theorem S2 to `FedSOUL` where for any $i \in [b]$, $\gamma \in (0, \bar{\gamma}]$ and $\theta \in \Theta$, the Markov kernel $Q_{\gamma,\theta}^{(i)}$ is associated with a Gaussian probability density function $q_{\gamma,\theta}^{(i)}(z^{(i)}, \cdot)$ with mean $z^{(i)} - \gamma \nabla_z \log p(z^{(i)} \mid \mathrm{D}_i, \theta)$ and variance $2\gamma \mathrm{I}_d$. To this end, we show explicit conditions on the family of posterior distributions $\{\pi_\theta^{(i)}\}_{i \in [b]}$ such that **A**6 and **A**7 are satisfied.

S2.1 Assumptions

For any $i \in [b]$, let $U_\theta^{(i)} : \mathbb{R}^d \to \mathbb{R}$ such that for any $z^{(i)} \in \mathbb{R}^d$, $\pi_\theta^{(i)}(z^{(i)}) \propto \exp\{-U_\theta^{(i)}(z^{(i)})\}$. In our case, this boils down to set $U_\theta^{(i)}(z^{(i)}) = -\log p(z^{(i)} \mid \mathrm{D}_i, \phi, \beta)$ for any $z^{(i)} \in \mathbb{R}^d$.

A8. *For any $i \in [b]$, the following conditions hold.*

(i) Assume that $(\theta, z^{(i)}) \mapsto U_\theta(z^{(i)})$ is continuous, $z^{(i)} \mapsto U_\theta^{(i)}(z^{(i)})$ is differentiable for any $\theta_1, \theta_2 \in \Theta$ and there exists $L \geq 0$ such that for any $z_1, z_2 \in \mathbb{R}^d$,

$$\sup_{\theta \in \Theta} \left\|\nabla_z U_\theta^{(i)}(z_2) - \nabla_z U_\theta^{(i)}(z_1)\right\| \leq L \|\theta_2 - \theta_1\|,$$

and $\{\nabla_z U_\theta^{(i)}(0) : \theta \in \Theta\}$ is bounded.

(ii) There exist $\mathtt{m}_1, \mathtt{m}_2 > 0$ *and* $\mathtt{c}, R \geq 0$ *such that for any* $\theta \in \Theta$ *and* $z \in \mathbb{R}^d$,

$$\langle \nabla_z U_\theta^{(i)}(z), z \rangle \geq \mathtt{m}_1 \|z\| \mathbf{1}_{\mathrm{B}(0,R)^c}(z) + \mathtt{m}_2 \left\| \nabla_z U_\theta^{(i)}(z) \right\|^2 - \mathtt{c} \,.$$

(iii) There exists $L_U \geq 0$ *such that* $z \in \mathbb{R}^d$ *and* $\theta_1, \theta_2 \in \Theta$,

$$\left\| \nabla_z U_{\theta_2}^{(i)}(z) - \nabla_z U_{\theta_1}^{(i)}(z) \right\| \leq L_U \|\theta_2 - \theta_1\| V(z)^{1/2} \,,$$

where $V : \mathbb{R}^d \to \mathbb{R}$ *is defined under A8-(ii), for any* $z \in \mathbb{R}^d$, *as*

$$V(z) = \exp\left\{ \mathtt{m}_1 \sqrt{1 + \|z\|^2}/4 \right\} \,. \tag{S24}$$

S2.2 Verification of A6 and A7

Lemma S8. *Assume A8. Then, A6 and A7 are satisfied with* V *defined in* (S24) *and*

$\bar{\gamma} < \min\{1, 2\mathtt{m}_2\}$,
$\tilde{\mathtt{m}}_1 = \mathtt{m}_1/4$,
$b = \tilde{\mathtt{m}}_1(d + \mathtt{c} + \sqrt{2}\tilde{\mathtt{m}}_1) \exp(\tilde{\mathtt{m}}_1^2\{(d + \mathtt{c} + \tilde{\mathtt{m}}_1\bar{\gamma} + \sqrt{1 + \mathtt{r}^2}\})$,
$\lambda = \exp(-\tilde{\mathtt{m}}_1^2[\sqrt{2} - 1])$,
$\mathtt{r} = \max\{1, 2(d + \mathtt{c})/\mathtt{m}_1, R\}$,
$\boldsymbol{\Gamma}_1 : (\gamma_1, \gamma_2) \mapsto \gamma_1/\gamma_2 - 1$,
$\boldsymbol{\Gamma}_2 : (\gamma_1, \gamma_2) \mapsto \gamma_2^{1/2}$,

$$\boldsymbol{\Psi} : \gamma \mapsto 2C(1-\xi)^{-1}\gamma^{1/2}\tilde{D}_1^{1/2}(1+\bar{\gamma})^{1/2} \left\{ d + 2\bar{\gamma} \left(\mathtt{L}^2 M_V + \sup_{\theta \in \Theta, i \in [b]} \left\| \nabla_z U_\theta^{(i)}(0) \right\|^2 \right) \tilde{D}_1 \right\}^{1/2} \mathtt{L} \,,$$

$$\tilde{D}_1 = \frac{\sqrt{2}\tilde{\mathtt{m}}_1 \exp(\tilde{\mathtt{m}}_1 \sqrt{1 + \max\{1, R\}^2})(1 + \tilde{\mathtt{m}}_1 + \mathtt{c} + d)}{3\tilde{\mathtt{m}}_1^2} + b\lambda^{-\bar{\gamma}} \log^{-1}(1/\lambda) \,,$$

with $M_V = \sup_{z \in \mathbb{R}^d}\{(1 + \|z\|)^2/V(z)\}$, $C \geq 0$, $\xi \in (0, 1)$.

Proof. The proof follows from De Bortoli et al. [11, Theorem 5]. □

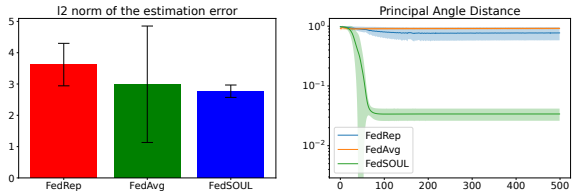

Figure S1: Small data sets - synthetic data. $b = 50$ clients.

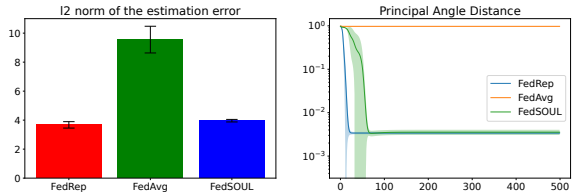

Figure S2: Small data sets - synthetic data. $b = 200$ clients.

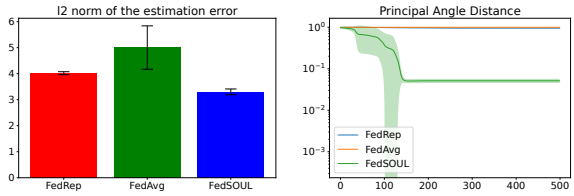

Figure S3: Small data sets - synthetic data. Raw data dimensionality is $k = 50$.

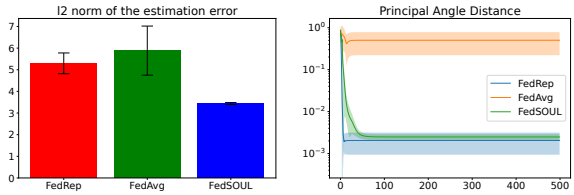

Figure S4: Small data sets - synthetic data. Raw data dimensionality is $k = 5$.

S3 Additional Experiments

S3.1 Synthetic datasets

In this section, following the experiments from the main paper, we will show additional configurations of the toy example. We still use the same model (see Section 5 and Singhal et al. [47], Collins et al. [8]), but we choose different values of (d, k, b). First, let us test, how the total number of clients b impacts the performances of the different approaches. Figure S1 and Figure S2 depict our results for $b \in \{50, 200\}$, with the size of the minimal dataset being 5 and the share of clients with the minimal dataset 90%. We can see that in both cases, `FedSOUL` outperforms its competitors.

Second, we test, how the dimensionality of raw data impacts the result. Figure S3 and Figure S4 show our results with $k \in \{5, 50\}$. All others parameters are the same as before.

One more experiment we conducted is the dependence on latent dimensionality d. We test two options $d = 2$ (as in original experiments) and $d = 5$ in Figure S5 and Figure S6. Again, the more

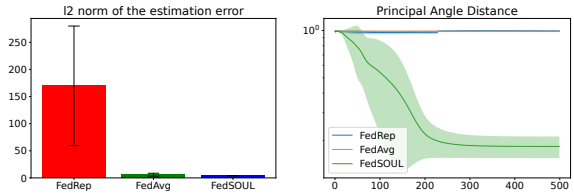

Figure S5: Small data sets - synthetic data. Latent space dimensionality is $d = 5$.

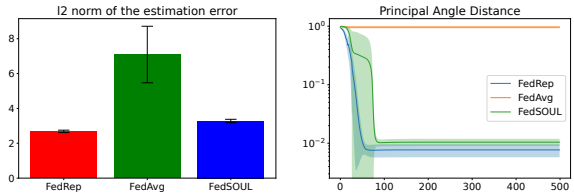

Figure S6: Small data sets - synthetic data. Latent space dimensionality is $d = 2$.

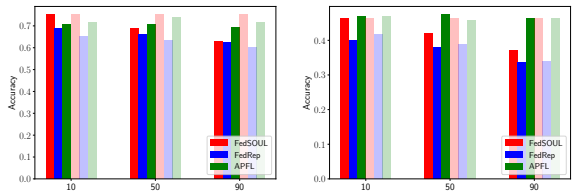

Figure S7: Small image datasets. The minimal local dataset size is 2 (top) or 5 (bottom).

parameters we have to learn (given the same small data budget), the better Bayesian methods (*i.e.* `FedSOUL`) are better.

S3.2 Image datasets classification

In this section, we provide an additional baseline for the experiments with personalization, in case we have only a few heterogeneous data. Specifically, we consider `APFL` [13] which is another personalized federated learning approach. We consider the CIFAR-10 dataset with 100 clients. Among these clients, there are 10, 50, or 90 which have a local dataset of either 5 (one setup) or 10 (another setup). Else of size 25.

We see in Figure S7 that `FedSOUL` typically performs better than `FedRep`, but on par with `APFL`. It is surprising, that `APFL` is a very good baseline in this type of problem, which it was not specially designed for.

S3.3 Image datasets uncertainty quantification

In this section, we provide additional experiments on image uncertainty with CIFAR-10 (in distribution) and SVHN (out of distribution) datasets. As a measure of uncertainty, we will use predictive entropy. On Figure S8, we present 4 different models among 100. In the left part of the figure, we see the distribution of entropy, assigned to the in-distribution objects (validation split, but same domain as training data). In the right part, we see the distribution for out-of-distribution (SVHN in our case). Contrary to MNIST vs Fashion-MNIST example, here it is not that clear that `FedSOUL` captures uncertainty well.

We also provide additional plots for calibration on CIFAR-10 again for two cases, when each client had 2 classes to predict or 5.

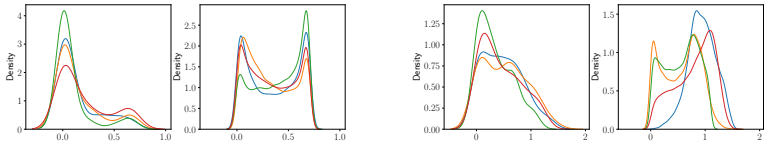

Figure S8: Out-of-distribution detection. CIFAR 10 vs SVHN. 2 classes for model (top) and 5 (bottom).

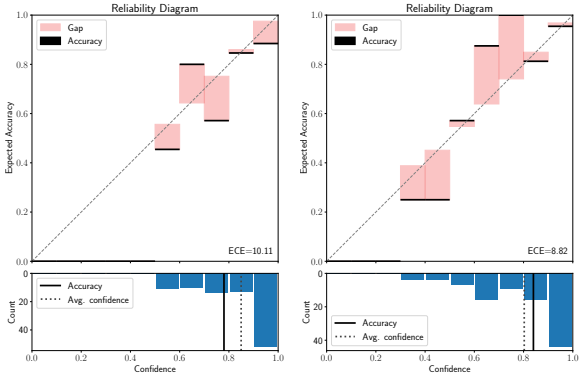

Figure S9: Reliability diagram for CIFAR10. 2 classes for model (top) and 5 (bottom).