FedPop: A Bayesian Approach for Personalised Federated Learning

Nikita Kotelevskii, Maxime Vono, Eric Moulines, Alain Durmus

FedPop: A Bayesian Approach for Personalised Federated Learning

Nikita Kotelevskii[*]
Skolkovo Institute of Science and Technology
Moscow, Russia
Nikita.Kotelevskii@skoltech.ru

Maxime Vono[*]
Criteo AI Lab
Paris, France
m.vono@criteo.com

Eric Moulines
Ecole Polytechnique
eric.moulines@polytechnique.edu

Alain Durmus
ENS Paris-Saclay
alain.durmus@ens-paris-saclay.fr

Abstract

Personalised federated learning (FL) aims at collaboratively learning a machine learning model taylored for each client. Albeit promising advances have been made in this direction, most of existing approaches works do not allow for uncertainty quantification which is crucial in many applications. In addition, personalisation in the cross-device setting still involves important issues, especially for new clients or those having small number of observations. This paper aims at filling these gaps. To this end, we propose a novel methodology coined `FedPop` by recasting personalised FL into the population modeling paradigm where clients' models involve *fixed* common population parameters and *random* effects, aiming at explaining data heterogeneity. To derive convergence guarantees for our scheme, we introduce a new class of federated stochastic optimisation algorithms which relies on Markov chain Monte Carlo methods. Compared to existing personalised FL methods, the proposed methodology has important benefits: it is robust to client drift, practical for inference on new clients, and above all, enables uncertainty quantification under mild computational and memory overheads. We provide non-asymptotic convergence guarantees for the proposed algorithms and illustrate their performances on various personalised federated learning tasks.

1 Introduction

Federated learning (FL) is a recent machine learning paradigm in which distributed clients holding sensitive data collaborate in solving a learning problem, usually under the coordination of a central server (Kairouz et al., 2021; Wang et al., 2021). One of the main focus of FL is on so-called *cross-device* applications where a large number of personal electronic devices such as mobile phones, wearable devices or home assistants collect and store data at the edges of a decentralised network (McMahan et al., 2017).

While standard FL methods (Alistarh et al., 2017; Horváth et al., 2019; Karimireddy et al., 2020; Li et al., 2020; McMahan et al., 2017) have focused on training a global model that can be applied to individual agents, the relevance of such inferences has recently been questioned due to statistical *heterogeneity* between clients. Indeed, the considered common model may not generalise well on a client with a local data distribution that differs significantly from the global data distribution, especially if that client has not participated in the training process. In fact, it might even be better for such clients to derive a local model from their own data set. To circumvent these issues, a number of

[*]Both authors contributed equally to this work.

personalised federated learning approaches have recently been proposed, that use local models to fit client-specific data distribution while capturing some shared knowledge via a federated scheme (Tan et al., 2022). Personalisation has previously been approached using multi-task learning (Smith et al., 2017), meta-learning (Jiang et al., 2019; Khodak et al., 2019), client clustering (Briggs et al., 2020), data interpolation (Mansour et al., 2020), model interpolation (Hanzely and Richtárik, 2020; Hanzely et al., 2020) or partially local models (Collins et al., 2021; Singhal et al., 2021). While these methodologies are promising, they only partially solve the personalisation problem in highly heterogeneous federated settings and have no means of quantifying uncertainty. In addition, cross-device FL also faces other important challenges such as (extreme) partial device participation, small local data sets, limited upload bandwidth and device capabilities (Kairouz et al., 2021). Addressing these problems for personalised FL requires new paradigms regarding how model knowledge is shared and personalisation is performed locally.

Proposed Approach. In this paper, we adopt a novel perspective to model the problem of personalised FL. This paradigm, called *mixed-effects modeling* (also known as multi-level or population approach) is widely used to analyse data that have a clustered or nested structure, as in medical or biological research where multiple observations per patient are available (Gelman and Hill, 2007; Lavielle, 2014; Long, 2011). Although the hierarchical structure of FL has already been noted (Grant et al., 2018; Hong et al., 2022; Plassier et al., 2021), the mixed-effects paradigm has interestingly never been considered. Leveraging this framework, we develop a new model for personalised FL called `FedPop` and propose an efficient computational solution to perform inference under this model. More precisely, we introduce a novel class of federated stochastic approximation algorithms based on parallel Markov Chain Monte Carlo (MCMC) methods. In the proposed framework, we also pay special attention to the cross-device setting by taking into account partial client participation, and by addressing the communication bottleneck with both multiple local updates and the use of lossy compression operators.

Benefits. Up to the authors' knowledge, `FedPop` is the first *personalised FL* approach that allows *cheap uncertainty quantification* with a theoretically-grounded methodology. The proposed framework also comes with other interesting properties. First, in contrast to most of personalised FL methods that only consider "fixed-effects" models (Collins et al., 2021; Hanzely et al., 2021; Smith et al., 2017), `FedPop` provides credibility information (via credibility regions) and allows more accurate inference for clients with small and heterogeneous local data via *partial pooling* (Gelman and Hill, 2007). In addition, inference for new clients which did not participate in the training phase can be easily performed by sampling from the prior over the local random effects. Second, contrary to existing Bayesian FL approaches that aim to provide credibility information by sampling from a target posterior distribution (El Mekkaoui et al., 2021; Hong et al., 2022; Vono et al., 2022; Yoon et al., 2018), `FedPop` allows to perform personalisation and cheaper on-device uncertainty quantification taking an empirical Bayes prediction approach. Finally, an important benefit of `FedPop` is its ability to allow for multiple local updates without suffering from the client-drift phenomenon (Karimireddy et al., 2020).

Outline and Contributions. Our contributions are fourfold. First, in Section 2, we propose a novel probabilistic methodology, which we call `FedPop`, to address personalisation under the cross-device FL paradigm. To perform efficient inference under this model, we introduce a general class of stochastic approximation algorithms based on MCMC. Second, we provide in Section 3 non-asymptotic convergence guarantees for the proposed methodology. Then, we perform in Section 4 a comparison between the proposed approach and exisiting works. Finally, we illustrate in Section 5 the benefits of our methodology on several federated learning benchmarks involving both synthetic and real data.

2 Proposed Approach

In this section, we present the statistical estimation problem we are considering and the proposed methodology called `FedPop` to address it.

Problem Formulation. We are interested in the *cross-device* FL setting involving a large number $b \in \mathbb{N}^*$ of clients, potentially unreliable *i.e.* not necessarily available at each communication round. These clients are assumed to own sensitive local data sets $\{\mathrm{D}_i\}_{i\in[b]}$. In this framework, we aim to make both uncertainty quantification and personalised statistical inference by learning a local model

taylored to each client. To this end, and inspired by the population approach used in the biological and physical sciences (Lavielle, 2014), we consider mixed-effects modeling for each client leading to the local marginal likelihood function defined, for any $i \in [b]$, by

$$p(\mathrm{D}_i \mid \phi, \beta) = \int_{\mathbb{R}^d} p(\mathrm{D}_i \mid \phi\,, z^{(i)}) p(z^{(i)} \mid \beta) \mathrm{d}z^{(i)}\,, \tag{1}$$

where $\phi \in \Phi \subseteq \mathbb{R}^{d_\Phi}$ stands for a *fixed effect* and $\{z^{(i)}\}_{i\in[b]} \in \mathsf{Z}$, $\mathsf{Z} = \prod_{i=1}^b \mathbb{R}^d$, represent *random effects* aimed at explaining statistical heterogeneity between local data sets $\{\mathrm{D}_i\}_{i\in[b]}$. The objective of the fixed (*i.e.* constant across all clients) part is to capture a common representation (*e.g.* same features across different classes of images) while the random part, which is typically low-dimensional, performs personalisation and is assumed to be drawn from a *population* prior whose variance aims at modeling data heterogeneity.

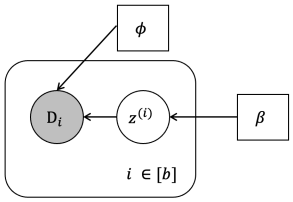

Figure 1: DAG for `FedPop`.

Figure 1 illustrates this statistical framework, referred to as `FedPop`, by showing its directed acyclic graph (DAG) where grey-filled shapes indicate observed variables, white-filled shapes unknown variables and squared shapes variables to be estimated.

When the size of the local data set D_i is small, this common prior leverages information from other clients to limit the risk of overfitting and is often called *partial pooling* in the multi-level statistical literature (Gelman and Hill, 2007, Section 12). Examples of model architectures involving ϕ and $\{z^{(i)}\}_{i\in[b]}$ include for instance *composition*-based architectures $p(\mathrm{D}_i \mid \phi, z^{(i)}) = p(\mathrm{D}_i \mid h_\phi \circ h_{z^{(i)}})$ where h_ϕ and $h_{z^{(i)}}$ are two neural networks (Arivazhagan et al., 2019; Collins et al., 2021). For the sake of generality, we propose to adopt a flexible energy-based prior distribution of the form for each $i \in [b]$,

$$p(z^{(i)} \mid \beta) = \frac{1}{Z(\beta)} \exp\left\{-E(z^{(i)}; \beta)\right\}\,, \text{ where } Z(\beta) = \int_{\mathbb{R}^d} \exp\left\{-E(z^{(i)}; \beta)\right\} \mathrm{d}z^{(i)}\,.$$

Here, $Z(\beta)$ is a normalising constant and $E(\cdot; \beta)$ represents an *energy* function, typically a neural network, parameterised by a set of parameters $\beta \in \mathsf{B} \subseteq \mathbb{R}^{d_\mathsf{B}}$ (LeCun et al., 2006). This framework is particularly interesting in the cross-device setting where the number of clients b is large as it allows for efficient enrichment of the model. However, in the case where b is small, the inference of the parameter β is difficult. In this situation, a more pragmatic solution is to consider a common prior for the local random effects $\{z^{(i)}\}_{i\in[b]}$ which is held fixed, *i.e.* $p(z^{(i)} \mid \beta) \propto \exp\{-E(z^{(i)})\}$ for any $\beta \in \mathsf{B}$. Finally, for completeness, we allow the use of a prior model $p(\phi, \beta) = p(\phi)p(\beta)$ for the hyperparameters $\{\phi, \beta\}$. Using Bayes' rule (Robert, 2001) and by denoting $\mathrm{D} = \sqcup_{i=1}^b \mathrm{D}_i$ the global data set, the posterior distribution associated with these hyperparameters admits a probability density function which can be written as

$$p\left(\phi, \beta \mid \mathrm{D}\right) = p(\phi)p(\beta) \prod_{i=1}^b \left[\int_{\mathbb{R}^d} p(\mathrm{D}_i \mid \phi, z^{(i)}) p(z^{(i)} \mid \beta) \mathrm{d}z^{(i)}\right]\,.$$

Set $\theta = \{\phi, \beta\} \in \Theta$ with $\Theta = \Phi \times \mathsf{B}$. In the sequel, we will be interested in solving the maximum a posteriori problem given by

$$\theta^\star \in \underset{\theta\in\Theta}{\arg\max}\ \log p(\phi, \beta \mid \mathrm{D})\,, \tag{2}$$

$$\log p(\phi, \beta \mid \mathrm{D}) = \log p(\phi) + \log p(\beta) + \sum_{i=1}^b \left[\log \int_{\mathbb{R}^d} p(\mathrm{D}_i \mid \phi, z^{(i)}) p(z^{(i)} \mid \beta) \mathrm{d}z^{(i)}\right]\,. \tag{3}$$

Once we have estimated $\theta^\star$, using an empirical Bayesian approach, we can perform "for free" on-device uncertainty quantification for each client $i \in [b]$ by sampling from the local posterior distribution $p(z^{(i)} \mid \mathrm{D}_i, \phi^\star, \beta^\star)$, which is typically designed to be low-dimensional.

Algorithm. To solve the optimisation problem (2), we can either use an *alternating maximisation* algorithm or perform global maximisation over Θ. Since the former approach requires more upload

bandwidth, in this work we consider the second alternative which is more suitable for FL. The gradient of the objective function (3) being intractable, we propose to resort to the stochastic approximation framework (Robbins and Monro, 1951) which iteratively defines $(\phi_k, \beta_k)_{k\in\mathbb{N}}$, starting from any $(\phi_0, \beta_0) \in \Theta$, via the recursions for any $k \in \mathbb{N}$,

$$\beta_{k+1} = \Pi_{\mathsf{B}}\left(\beta_k + \eta_{k+1}^{(1)}\left[\nabla_\beta \log p(\beta) + \sum_{i=1}^{b} g_k^{(i)}(\phi_k, \beta_k)\right]\right),$$
$$\phi_{k+1} = \Pi_{\Phi}\left(\phi_k + \eta_{k+1}^{(2)}\left[\nabla_\phi \log p(\phi) + \sum_{i=1}^{b} h_k^{(i)}(\phi_k, \beta_k)\right]\right),$$

where Π_{C} denotes the projection onto $\mathsf{C} \in \{\Phi, \mathsf{B}\}$, $(\eta_k^{(1)}, \eta_k^{(2)})_{k\in\mathbb{N}^*}$ are sequences of step-sizes, and $\{g_k^{(i)} : i \in [b], k \in \mathbb{N}^*\}$ and $\{h_k^{(i)} : i \in [b], k \in \mathbb{N}^*\}$ are estimators of the intractable gradients $(\phi, \beta) \mapsto \nabla_\beta \log p(\mathrm{D}_i \mid \phi, \beta)$ and $(\phi, \beta) \mapsto \nabla_\phi \log p(\mathrm{D}_i \mid \phi, \beta)$ at (ϕ_k, β_k), where $p(\mathrm{D}_i \mid \phi, \beta)$ is defined in (1) for any $i \in [b]$.

The choices of the estimators $\{g_k^{(i)} : i \in [b], k \in \mathbb{N}^*\}$ and $\{h_k^{(i)} : i \in [b], k \in \mathbb{N}^*\}$ are motivated by the Fisher identity. More precisely, under mild regularity assumptions, and using the Lebesgue dominated convergence theorem, we have for any, $(\phi, \beta) \in \Theta$, $i \in [b]$

$$\nabla_\beta \log p(\mathrm{D}_i \mid \phi, \beta) = \int_{\mathbb{R}^d} \left[\nabla_\beta \log p(\mathrm{D}_i, z^{(i)} \mid \phi, \beta)\right] p(z^{(i)} \mid \mathrm{D}_i, \phi, \beta) \mathrm{d}z^{(i)},$$
$$\nabla_\phi \log p(\mathrm{D}_i \mid \phi, \beta) = \int_{\mathbb{R}^d} \left[\nabla_\phi \log p(\mathrm{D}_i, z^{(i)} \mid \phi, \beta)\right] p(z^{(i)} \mid \mathrm{D}_i, \phi, \beta) \mathrm{d}z^{(i)},$$

which suggests to consider

$$g_k^{(i)}(\phi, \beta) = \frac{1}{M}\sum_{m=1}^{M} \nabla_\beta \log p(Z_k^{(i,m)} \mid \beta), \tag{4}$$

$$h_k^{(i)}(\phi, \beta) = \frac{1}{M}\sum_{m=1}^{M} \nabla_\phi \log p(\mathrm{D}_i \mid Z_k^{(i,m)}, \phi), \tag{5}$$

where $M \in \mathbb{N}^*$ and $Z_k^{(i,1:M)} = (Z_k^{(i,m)})_{m\in[M]}$ are approximate samples from $p(z^{(i)} \mid \mathrm{D}_i, \phi, \beta)$. More precisely, we consider a family $\{Q_{\gamma,\theta}^{(i)} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta\}$ where for any step-size γ, $Q_{\gamma,\theta}^{(i)}$ is a Markov kernel which targets a close approximation of $p(z^{(i)} \mid \mathrm{D}_i, \theta)$ with $\theta = \{\phi, \beta\}$. As an example, we can use overdamped Langevin dynamics (Roberts and Tweedie, 1996; Welling and Teh, 2011) to generate these samples. In this case, $Q_{\gamma,\theta}^{(i)}$ 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 $2\gamma\mathrm{I}_d$. Note that the number of Monte Carlo draws per iteration k is considered constant here but we can easily generalise our scheme to the non-constant setting. In addition, our scheme can also be generalised by taking into account *stochastic* gradient estimators of (4) and (5). For the sake of simplicity, we present our approach with standard gradients.

In this framework, we present the main steps of the corresponding stochastic approximation algorithm, called `FedSOUK`, in Algorithm 1. Since we consider the *cross-device* federated setting, note that only a random subset A_{k+1} of active (*i.e.* available) clients communicates with the central server at each iteration $k \in \mathbb{N}$. In addition, due to limited upload bandwidth, the potentially high-dimensional gradient estimator (5) is compressed locally via an unbiased stochastic compression operator $\mathscr{C}_{k+1}$ before being sent to the central server (Alistarh et al., 2017; Philippenko and Dieuleveut, 2020). Finally, depending on local memory constraints, we allow for a possible warm-start strategy across communication rounds to improve the convergence properties of the proposed algorithm.

3 Theoretical Guarantees

In this section, we present non-asymptotic convergence guarantees for Algorithm 1 when the family of Markov kernels $\{Q_{\gamma,\theta}^{(i)} : \gamma \in (0, \bar{\gamma}], \theta \in \Theta, i \in [b]\}$ is associated to unadjusted, *i.e.* without

Algorithm 1 FL via Stochastic Optimisation using Unadjusted Kernel (`FedSOUK`)

Input: nb. outer iterations K, nb. local iterations M, Markov kernels $\{Q_{\gamma,\theta}^{(i)}\}_{\gamma,\theta,i}$, step-sizes $\{\eta_k^{(1)}, \eta_k^{(2)}\}_{k\in[K],i\in[b]}$ and initial points $Z_0^{(0)} \in \mathbb{R}^d$, $\beta_0 \in \mathsf{B}$ and $\phi_0 \in \Phi$.
for $k = 0$ **to** $K-1$ **do**
  for $i \in \mathsf{A}_{k+1}$ // On active clients A_{k+1} **do**
    // Warm-start of the SA scheme if possible
    if $k \geq 1$ **then**
      Set $Z_k^{(i,0)} = Z_{k-1}^{(i,M)}$.
    end if
    // Computation of key quantities using MCMC
    for $m = 0$ **to** $M-1$ **do**
      Draw $Z_k^{(i,m+1)} \sim Q_{\gamma,\theta_k}^{(i)}\left(Z_k^{(i,m)}, \cdot\right)$.
      // For Langevin dynamics
      // Draw $\xi_k^{(i,m+1)} \sim \mathrm{N}(0_d, \mathrm{I}_d)$.
      // Set $Z_k^{(i,m+1)} = Z_k^{(i,m)} + \gamma \nabla_z \log p(Z_k^{(i,m)} \mid \mathrm{D}_i, \phi_k, \beta_k) + \sqrt{2\gamma}\xi_k^{(i,m+1)}$.
    end for
    // Communication with the server
    Set $I_k^{(i)} = \frac{1}{M}\sum_{m=1}^M \nabla_\beta \log p\left(Z_k^{(i,m)} \mid \beta_k\right)$.
    Set $J_k^{(i)} = \frac{1}{M}\sum_{m=1}^M \nabla_\phi \log p\left(\mathrm{D}_i \mid Z_k^{(i,m)}, \phi_k\right)$.
    Send $I_k^{(i)}$ and $\mathscr{C}_{k+1}\left(J_k^{(i)}\right)$ to the central server.
  end for
  Set $\beta_{k+1} = \Pi_{\mathsf{B}}\left(\beta_k + \eta_{k+1}^{(1)}\left[\nabla_\beta \log p(\beta_k) + \frac{b}{|\mathsf{A}_{k+1}|}\sum_{i\in\mathsf{A}_{k+1}} I_k^{(i)}\right]\right)$.
  Set $\phi_{k+1} = \Pi_{\Phi}\left(\phi_k + \eta_{k+1}^{(2)}\left[\nabla_\phi \log p(\phi_k) + \frac{b}{|\mathsf{A}_{k+1}|}\sum_{i\in\mathsf{A}_{k+1}} \mathscr{C}_{k+1}\left(J_k^{(i)}\right)\right]\right)$.
  Send $\{\beta_{k+1}, \phi_{k+1}\}$ to clients belonging to A_{k+1}.
end for
Output: $\{\phi_K, \beta_K\}$ and samples $\{Z_{K-1}^{(1:b,m)}\}_{m=1}^M$.

Metropolis acceptance step, overdamped Langevin dynamics (Dalalyan, 2017; Durmus and Moulines, 2017). The bounds we derive allow to showcase explicitly the impact of FL constraints, namely partial participation and compression. Results for general unadjusted Markov kernels are postponed to the supplement.

To show our theoretical results and resort to standard assumptions made in the stochastic approximation literature, we consider a minimisation problem and rewrite the opposite of the objective function (3) for any $\theta \in \Theta$ as

$$f(\theta) = b^{-1}\sum_{i=1}^b f_i(\theta)\,, \quad \text{where } f_i(\theta) = -\log p(\phi) - \log p(\beta) - b\log p(\mathrm{D}_i \mid \phi, \beta)\,. \tag{6}$$

Non-Asymptotic Convergence Bounds. For the sake of better readability, we only detail in the main paper assumptions regarding the objective function, compression operators and the partial participation scenario. Technical assumptions related to the Markov kernels $\{Q_{\gamma,\theta}^{(i)}\}$ are postponed to the supplement. In spirit, we require, for any $i \in [b]$, $\theta \in \Theta$ and γ, that $Q_{\gamma,\theta}^{(i)}$ satisfies some ergodic condition and can provide samples sufficiently close to the local posterior distribution $p(z^{(i)} \mid \mathrm{D}_i, \theta)$. For the sake of simplicity, we also assume that for any $k \in \mathbb{N}^*$, $\eta_k^{(1)} = \eta_k^{(2)} = \eta_k$, see Algorithm 1.

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

H1. *Θ is convex, closed subset of $\mathbb{R}^{d_\Theta}$ and $\Theta \subset \mathrm{B}(0, R_\Theta)$ for $R_\Theta > 0$.*

H2. *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* $\theta_1, \theta_2 \in \Theta$,

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

The assumption below requires compression operators $\{\mathscr{C}_k\}_{k\in\mathbb{N}^*}$ to be unbiased and to have a bounded variance. Such an assumption is for instance verified by stochastic quantisation operators, see Alistarh et al. (2017).

H3. *The compression operators* $\{\mathscr{C}_k\}_{k\in\mathbb{N}^*}$ *are independent and satisfy the following conditions.*

(i) For any $k \in \mathbb{N}^*$, $v \in \mathbb{R}^d$, $\mathbb{E}[\mathscr{C}_k(v)] = v$.
(ii) There exists $\omega \geq 1$, *such that for any* $k \in \mathbb{N}^*$, $v \in \mathbb{R}^d$, $\mathbb{E}[\|\mathscr{C}_k(v) - v\|^2] \leq \omega \|v\|^2$.

We finally assume that each client has probability $p \in (0, 1]$ to be active at each communication round. We would like to point out that this partial participation assumption can be associated to a specific compression operator satisfying **H**3.

H4. *For any* $k \in \mathbb{N}^*$, $\mathsf{A}_k = \{i \in [b] : B_{i,k} = 1\}$ *where for any* $i \in [b]$, $\{B_{i,k} : k \in \mathbb{N}^*\}$ *is a family of i.i.d. Bernouilli random variables with success probability* $p \in (0, 1]$.

Under these assumptions, the next result establishes that $(\bar{\theta}_k)_{k\in\mathbb{N}}$ defined by $\bar{\theta}_k = \sum_{j=1}^k \eta_j \theta_j / (\sum_{j=1}^k \eta_j)$ converges towards an element of $\arg\min_\Theta f$.

Theorem 1. *Assume* ***A****1-****H****4 along with* ***A****8 detailed in the supplement and let for any* $k \in [K]$, $\eta_k \in (0, 1/L_f]$. *Then, for any* $K \in \mathbb{N}^*$, *we have*

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

where E_K *depends linearly on* $(\omega/p)\sum_{k=1}^K \eta_k^2$; *and* $A(\gamma) = C\gamma^\alpha$ *with* $\alpha > 0$ *and* C *is independent of* ω, p *and* (η_k). *Closed-form formulas for these constants are provided in the supplement.*

An interesting feature of Algorithm 1 is that convergence towards a minimum of f is possible and the impact of partial participation and compression vanishes when $\lim_{k\to\infty} \eta_k = 0$. More precisely, $\limsup_{k\to\infty} E_K / (\sum_{k=1}^K \eta_k) = 0$ and $\lim_{\gamma\to 0^+} A(\gamma) = 0$ which shows that we can tend towards a minimum of f with arbitrary precision $\epsilon > 0$ by setting the step-size γ to a small enough value.

4 Related Works

As pointed out in Section 1, many different approaches have been proposed to address personalisation and uncertainty quantification under the federated learning paradigm. This section reviews the main related existing lines of research and shows that the proposed methodology provides many benefits; see Table 1. Interestingly, we also show that `FedPop` encompasses some of the existing FL models.

Bayesian FL. One of our main motivations is the possibility to perform grounded uncertainty quantification in FL by resorting to the Bayesian paradigm. In the recent years, many works have suggested to adapt serial workhorses stochastic simulation approaches such as MCMC or variational inference to the FL setting (Bui et al., 2018; Chen and Chao, 2020; Corinzia et al., 2019; Deng et al., 2021a; El Mekkaoui et al., 2021; Liu and Simeone, 2021a,b; Plassier et al., 2021; Vono et al., 2022). Although some of these approaches address important FL challenges such as the communication bottleneck, partial participation or limited computational device resources, they are not suitable for uncertainty quantification in the cross-device FL scenario. Indeed, all these approaches assume that the posterior distribution targeted by each client is parametrised by a single potentially high-dimensional parameter of size $d_\Phi + d$, see (1). This prevents a sufficient number of samples from being stored locally to perform uncertainty quantification and Bayesian model averaging, especially when the model is a large neural network. In contrast, our approach decouples this unique high-dimensional parameter into a fixed part ϕ and a low-dimensional random part $z^{(i)}$, significantly reducing the memory footprint of local sample storage.

Personalised FL. Beside uncertainty quantification, we also aim at providing each client with a dedicated personalised model. Among the numerous existing personalised FL approaches, those

Table 1: Overview of the main existing personalised FL (top rows) and Bayesian FL (bottom rows) approaches related to the proposed framework. Column "PP" refers to partial participation, "perso." to personalised approaches, "bounds" to available convergence guarantees, "UQ" to available uncertainty quantification, "com." to the scheme (multiple local steps and/or compression) used to address the communication bottleneck and "memory" to the client memory footprint where M stands for the number of samples.

METHOD	PP	PERSO.	BOUNDS	UQ	COM.	MEMORY	FEDPOP INSTANCE
PER-FEDAVG	✓	✓	✓	✗	LOCAL STEPS	$d + d_\Phi$	✗
PFEDME	✗	✓	✓	✗	LOCAL STEPS	$d + d_\Phi$	✗
FEDREP	✓	✓	✓	✗	LOCAL STEPS	$d + d_\Phi$	✓
DITTO	✓	✓	✓	✗	LOCAL STEPS	$d + d_\Phi$	✗
LG-FEDAVG	✓	✓	✓	✗	LOCAL STEPS	$d + d_\Phi$	✗
QLSD	✓	✗	✓	✓	COMPRESSION	$M(d + d_\Phi)$	✗
FSGLD	✗	✗	✓	✓	LOCAL STEPS	$M(d + d_\Phi)$	✗
FEDBE	✓	✗	✗	✓	LOCAL STEPS	$M(d + d_\Phi)$	✗
DG-LMC	✗	✗	✓	✓	LOCAL STEPS	$M(d + d_\Phi)$	✓
FEDPOP	✓	✓	✓	✓	BOTH	$Md + d_\Phi$	–

related to FedPop can be broadly classified into two groups: *meta-learning* and *partially local methods*. Meta-learning based FL methods aim at training a global model conducive to fast training of personalised models. Such a goal can be achieved, for example, by local fine-tuning (Fallah et al., 2020), regularisation of local models towards their average (Hanzely and Richtárik, 2020; Hanzely et al., 2021) – or the opposite (Li et al., 2021), and model interpolation (Liang et al., 2019). On the other hand, FL methods based on partial decoupling take an approach similar to ours by splitting the initial model into a backbone component and a local one aimed at personalisation (Arivazhagan et al., 2019; Collins et al., 2021; Pillutla et al., 2022). This partial decoupling could also enhance privacy as discussed in Singhal et al. (2021). The main difference with FedPop is that such approaches based on empirical risk minimisation cannot provide credibility information.

FedPop: A Compromise between Standard and Personalised FL. Interestingly, we show here that the FedPop framework allows existing FL approaches to be retrieved in certain regimes. To this end, we assume that the prior $p(z^{(i)} \mid \beta)$ is Gaussian with mean μ and covariance matrix $\sigma^2 \mathrm{I}_d$ so that $\beta = \{\mu, \sigma\}$. If $\sigma \to 0^+$, then this Gaussian prior tends towards the Dirac distribution centered at μ and the local likelihood becomes $p(\mathrm{D}_i \mid \phi, \mu)$, which corresponds to the local objective of standard FL approaches such as FedAvg (McMahan et al., 2017). On the other hand, when $\sigma \to \infty$, no common information β is used to locally regress $z^{(i)}$ and we end up with the FedRep algorithm (Collins et al., 2021). This shows that FedPop stands for a subtle compromise between standard and personalised FL which should benefit clients with small data sets by pooling information via a common prior. Finally, in the extreme scenario where ϕ is the null vector, our approach amounts to the Bayesian FL approach DG-LMC proposed in Plassier et al. (2021).

5 Numerical Experiments

In this section, we illustrate the benefits of our methodology on several FL benchmarks associated to both synthetic and real data. Since existing Bayesian FL approaches are not suited for personalisation (see Table 1), we only compare the performances of Algorithm 1 with personalised FL methods. In all our experiments, we use overdamped Langevin dynamics to sample locally and call this specific instance of Algorithm 1, FedSOUL. In addition, we set $p(z^{(i)} \mid \beta) = \mathrm{N}(\mu, \sigma^2 \mathrm{I}_d)$ with $\beta = \{\mu, \sigma\}$ for simplicity. To be comparable with existing personalised FL approaches that only consider periodic communication via multiple local steps, we do not resort to the proposed compression mechanism although the latter could be of interest for real-world applications. Additional details about experimental design are provided in the supplement.

Synthetic Data. We start by showcasing the benefits of FedSOUL for clients having small and highly heterogeneous data sets as pointed out in Section 1 and Section 2. To this end, we consider a similar experimental setting as in Collins et al. (2021) where synthetic observations $\{y_j^{(i)}\}_{j \in [N_i]} \in \mathrm{D}_i$ are

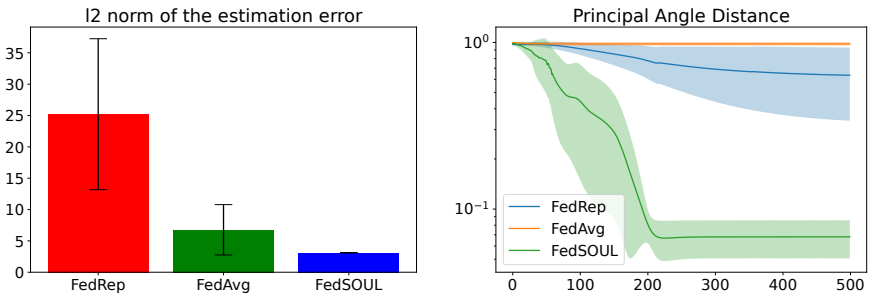

Figure 2: Small data sets - synthetic data.

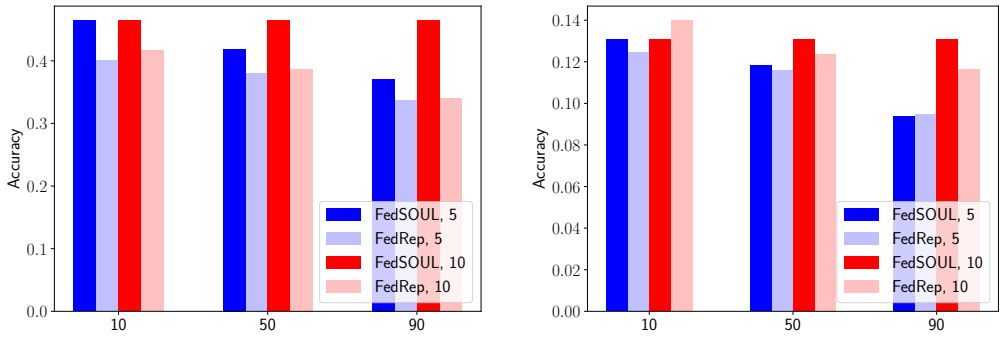

Figure 3: (right) CIFAR-10 with $S = 5$ and (left) CIFAR-100 with $S = 20$. The x-axis refers to the percentage of clients having $N_i \in \{5, 10\}$ images.

generated via the following procedure: $x_j^{(i)} \sim \mathrm{N}(0_k, \mathrm{I}_k)$ and $y_j^{(i)} \sim \mathrm{N}(z_{\text{true}}^{(i)} \phi_{\text{true}}^\top x_j^{(i)}, 0.1)$. The ground-truth parameters $z_{\text{true}}^{(i)} \in \mathbb{R}^d$ and $\phi_{\text{true}} \in \mathbb{R}^{k \times d}$ have been randomly generated beforehand with $(d, k) = (2, 20)$. Compared to Collins et al. (2021), we use heterogeneous data partitions across clients so that 90% of the $b = 100$ clients have small data sets of size 5 and the remaining 10% have data sets of size 10. We compare our results with `FedRep` (Collins et al., 2021) and `FedAvg` (McMahan et al., 2017) since they stand for two limiting instances of the proposed methodology, see Section 4 and Gelman and Hill (2007, Section 12). Figure 2 compares the different approaches by computing the principle angle distance* (respectively the ℓ_2 norm) between ϕ_{true} (respectively $z_{\text{true}}^{(i)}$) and its estimated value; the lesser the better. In contrast to its main competitors and based on both metrics, `FedSOUL` provides an impressive improvement. This illustrates the benefits of the introduction of a common prior $p(z^{(i)} \mid \beta)$ which allows to prevent from overfitting on clients with small data sets while performing personalisation. Additional results with other choices for (b, d, k) and data partitioning strategies are available in the supplement.

Real Data. We consider now real image data sets, namely CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). For our likelihood model defined by $p(\mathrm{D}_i \mid \phi, z^{(i)})$, we use 5-layer convolutional neural networks and perform personalisation for the last layer. We set $b = 100$ for convenience and control data heterogeneity by assigning to each client N_i images belonging to only S different classes.

Small data sets. Under this setting, we first consider (10%, 50%, 90%) of clients having small data sets of size either $N_i = 5$ or $N_i = 10$; while remaining clients have larger data sets of size $N_i = 25$. We compare our approach with `FedRep` since it stands for the state-of-the-art personalised FL approach. The algorithms are trained fulfilling the same computational budget. Figure 3 shows the average accuracy across clients for the two approaches on both CIFAR-10 and CIFAR-100. We can see that `FedSOUL` is consistently better than `FedRep` over different configurations.

*defined in (Collins et al., 2021, Definition 1)

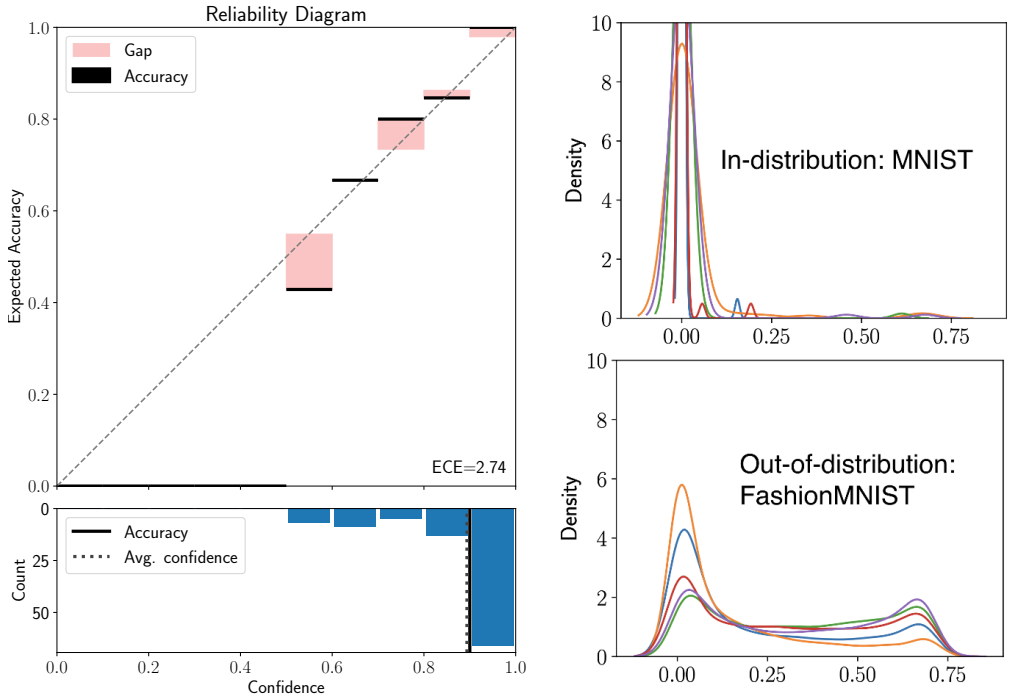

Figure 4: (right) Calibration on CIFAR-10 for a specific client and (left) OOD analysis with MNIST training & FashionMNIST inference – one curve corresponds to one client.

Full data sets. In addition to show that the proposed approach achieves state-of-the-art performances on small data sets (which is common in the cross-device scenario), we now illustrate that `FedSOUL` is also competitive on larger data sets. To this end, we use all training images in CIFAR-10 and CIFAR-100 image data sets and consider the same data partitioning as in Collins et al. (2021). More precisely, in this case the number of observations and the number of classes per client are uniformly shared over the clients. Table 2 shows our results in comparison with state-of-the-art personalised FL approaches. We can see that that our model outperforms other methods on both CIFAR-10 and CIFAR-100 by a large margin. Additional results with other personalised FL algorithms are postponed to the supplement.

Uncertainty Quantification on Real Data. As highlighted in Table 1, one advantage of the proposed approach compared to existing personalised FL methods is the ability to perform uncertainty quantification by sampling locally from the posterior $p(z^{(i)} \mid \mathrm{D}_i, \phi_K, \beta_K)$, see Algorithm 1. We illustrate this feature by computing on CIFAR-10 calibration curves and scores (*e.g.* expected calibration error aka ECE) on a specific client; and by performing an out-of-distribution analysis on MNIST/FashionMNIST data sets. Figure 4 shows that the proposed approach provides relevant uncertainty diagnosis. Additional results on uncertainty quantification can be found in the supplement.

6 Conclusion

In this paper, we proposed a general Bayesian methodology based on a natural mixed-effects modeling approach to model personalisation in federated learning. Our FL method is the first that allows for both personalisation and cheap uncertainty quantification for (cross-device) federated learning. By introducing a common prior on the local parameters, we tackle the local overfitting problem in the scenario where clients have highly heterogeneous and small data sets. In addition, we have shown that the proposed approach has favorable convergence properties. Some limitations of `FedPop` pave the way for more advanced personalised FL approaches. As an example, our model does not allow for training heterogeneous architectures across clients because of the introduced common prior, and

Table 2: Real data - Full data sets. Accuracy (in %) on test samples. FedAvg and SCAFFOLD are not personalised FL approaches but stand for well-known FL benchmarks.

	CIFAR-10		CIFAR-100	
(# clients b, # classes per client S)	(100, 2)	(100, 5)	(100, 5)	(100, 20)
Local learning only	89.79	70.68	75.29	41.29
FedAvg (McMahan et al., 2017)	42.65	51.78	23.94	31.97
SCAFFOLD (Karimireddy et al., 2020)	37.72	47.33	20.32	22.52
LG-FedAvg (Liang et al., 2019)	84.14	63.02	72.44	38.76
Per-FedAvg (Fallah et al., 2020)	82.27	67.20	72.05	52.49
L2GD (Hanzely and Richtárik, 2020)	81.04	59.98	72.13	42.84
APFL (Deng et al., 2021b)	83.77	72.29	78.20	55.44
DITTO (Li et al., 2021)	85.39	70.34	78.91	56.34
FedRep (Collins et al., 2021)	87.70	75.68	79.15	56.10
FedSOUL (this paper)	91.12	79.48	79.56	59.73

only satisfy first-order privacy guarantees. Regarding the latter, further works include for instance deriving differentially private versions of our framework.

Acknowledgments and Disclosure of Funding

The authors acknowledge support from the Lagrange Mathematics and Computing Research Center.

References

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 2017.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated Learning with Personalization Layers. *arXiv preprint arXiv:1912.00818*, 2019.

Yves F. Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(10):1–33, 2017.

Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020.

Thang D. Bui, Cuong V. Nguyen, Siddharth Swaroop, and Richard E. Turner. Partitioned Variational Inference: A unified framework encompassing federated and continual learning. *arXiv preprint arXiv:1811.11206*, 2018.

Hong-You Chen and Wei-Lun Chao. Fedbe: Making bayesian model ensemble applicable to federated learning. *arXiv preprint arXiv:2009.01974*, 2020.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting Shared Representations for Personalized Federated Learning. In *International Conference on Machine Learning*, pages 2089–2099, 2021.

Luca Corinzia, Ami Beuret, and Joachim M. Buhmann. Variational Federated Multi-Task Learning. *arXiv preprint arXiv:1906.06268*, 2019.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society, Series B*, 79(3):651–676, 2017.

Valentin De Bortoli, Alain Durmus, Marcelo Pereyra, and Ana F. Vidal. Efficient stochastic optimisation by unadjusted Langevin Monte Carlo: Application to maximum marginal likelihood and empirical Bayesian estimation. *Statistics and Computing*, 31(3), 2021.

Wei Deng, Yi-An Ma, Zhao Song, Qian Zhang, and Guang Lin. On Convergence of Federated Averaging Langevin Dynamics. *arXiv preprint arXiv:2112.05120*, 2021a.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive Personalized Federated Learning, 2021b.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 06 2017. doi: 10.1214/16-AAP1238.

Khaoula El Mekkaoui, Diego Mesquita, Paul Blomstedt, and Samuel Kaski. Distributed stochastic gradient MCMC for federated learning. In *Conference on Uncertainty in Artificial Intelligence*, 2021.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach. In *Advances in Neural Information Processing Systems*, 2020.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, 2007.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting Gradient-Based Meta-Learning as Hierarchical Bayes. In *International Conference on Learning Representations*, 2018.

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

Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning. *arXiv preprint arXiv:2010.02372*, 2020.

Filip Hanzely, Boxin Zhao, and Mladen Kolar. Personalized federated learning: A unified framework and universal optimization techniques. *arXiv: 2102.09743*, February 2021.

Joey Hong, Branislav Kveton, Manzil Zaheer, and Mohammad Ghavamzadeh. Hierarchical Bayesian Bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic Distributed Learning with Gradient Quantization and Variance Reduction . *arXiv preprint arXiv:1904.05115*, 2019.

Yihan Jiang, Jakub Konevcnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2): 1–210, 2021. ISSN 1935-8237. doi: 10.1561/2200000083.

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*, pages 5132–5143, 2020.

Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32:5917–5928, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Available at http://www.cs.toronto.edu/~kriz/cifar.html, 2009.

Marc Lavielle. *Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools*. Chapman and Hall/CRC, 2014.

Yann LeCun, Sumit Chopra, Raia Hadsell, Fu Jie Huang, and et al. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2006.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In I. Dhillon, D. Papailiopoulos, and V. Sze, editors, *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020. URL https://proceedings.mlsys.org/paper/2020/file/38af86134b65d0f10fe33d30dd76442e-Paper.pdf.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *ICML*, pages 6357–6368, 2021. URL http://proceedings.mlr.press/v139/li21h.html.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. In *NeurIPS 2019 Workshop on Federated Learning*, 2019.

Dongzhu Liu and Osvaldo Simeone. Channel-Driven Monte Carlo Sampling for Bayesian Distributed Learning in Wireless Data Centers. *IEEE Journal on Selected Areas in Communications*, 2021a.

Dongzhu Liu and Osvaldo Simeone. Wireless Federated Langevin Monte Carlo: Repurposing Channel Noise for Bayesian Sampling and Privacy. *arXiv preprint arXiv:2108.07644*, 2021b.

Jeffrey Long. *Longitudinal Data Analysis for the Behavioral Sciences Using R*. Sage Publications, Inc, 2011.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

Constantin Philippenko and Aymeric Dieuleveut. Bidirectional compression in heterogeneous settings for distributed or federated learning with partial participation: tight convergence guarantees . *arXiv preprint arXiv:2006.14591*, 2020.

Krishna Pillutla, Kshitiz Malik, Abdelrahman Mohamed, Michael Rabbat, Maziar Sanjabi, and Lin Xiao. Federated Learning with Partial Model Personalization, 2022. URL https://openreview.net/forum?id=iFf26yMjRdN.

Vincent Plassier, Maxime Vono, Alain Durmus, and Eric Moulines. DG-LMC: a turn-key and scalable synchronous distributed MCMC algorithm via Langevin Monte Carlo within Gibbs. In *International Conference on Machine Learning (ICML)*, 2021.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 09 1951. doi: 10.1214/aoms/1177729586.

C. P. Robert. *The Bayesian Choice: from decision-theoretic foundations to computational implementation*. Springer, New York, 2 edition, 2001.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Karan Singhal, Hakim Sidahmed, Zachary Garrett, Shanshan Wu, John Rush, and Sushant Prakash. Federated reconstruction: Partially local federated learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11220–11232. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/5d44a2b0d85aa1a4dd3f218be6422c66-Paper.pdf.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated multi-task learning. *arXiv preprint arXiv:1705.10467*, 2017.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–17, 2022. doi: 10.1109/TNNLS.2022.3160699.

Maxime Vono, Vincent Plassier, Alain Durmus, Aymeric Dieuleveut, and Eric Moulines. QLSD: Quantised Langevin stochastic dynamics for Bayesian federated learning. In *AISTATS*, 2022.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H. Brendan McMahan, Blaise Aguera y Arcas, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, Suhas Diggavi, Hubert Eichner, Advait Gadhikar, Zachary Garrett, Antonious M. Girgis, Filip Hanzely, Andrew Hard, Chaoyang He, Samuel Horvath, Zhouyuan Huo, Alex Ingerman, Martin Jaggi, Tara Javidi, Peter Kairouz, Satyen Kale, Sai Praneeth Karimireddy, Jakub Konecny, Sanmi Koyejo, Tian Li, Luyang Liu, Mehryar Mohri, Hang Qi, Sashank J. Reddi, Peter Richtarik, Karan Singhal, Virginia Smith, Mahdi Soltanolkotabi, Weikang Song, Ananda Theertha Suresh, Sebastian U. Stich, Ameet Talwalkar, Hongyi Wang, Blake Woodworth, Shanshan Wu, Felix X. Yu, Honglin Yuan, Manzil Zaheer, Mi Zhang, Tong Zhang, Chunxiang Zheng, Chen Zhu, and Wennan Zhu. A Field Guide to Federated Optimization. *arXiv preprint arXiv:2107.06917*, 2021.

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

Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/e1021d43911ca2c1845910d84f40aeae-Paper.pdf.

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. (2021). 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*

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

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. (2017, Lemma 7) we have

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

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

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

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

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

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

Finally, **A**2-(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. (2017, 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 **A**1 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

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

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

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

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

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. (2021, 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. (2021, 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 $C_{\hat{H}} = 8 A_2 \log^{-1}(1/\rho) \rho^{-\bar{\gamma}/4}$.

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. (2021, Lemma 14). □

Lemma S5. *Assume **A**1, **A**2, **A**6 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. (2021, Lemma 15). □

Lemma S6. *Assume **A**1, **A**2, **A**5, **A**6 and **A**7. 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. (2021, Lemma 16). □

Lemma S7. *Assume **A**1, **A**2, **A**6 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. (2021, 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* ***A***8*-(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* ***A***8*. Then,* ***A***6 *and* ***A***7 *are satisfied with* V *defined in* (S24) *and*

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

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

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

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. (2021, Theorem 5). □

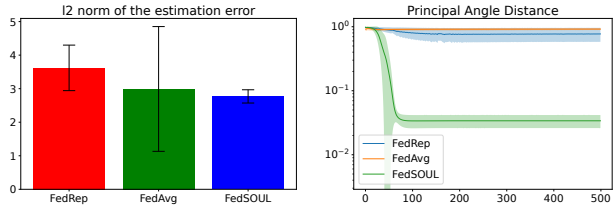

Figure S1: Small data sets - synthetic data. $b = 50$ clients.

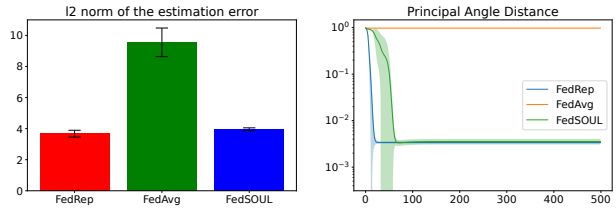

Figure S2: Small data sets - synthetic data. $b = 200$ clients.

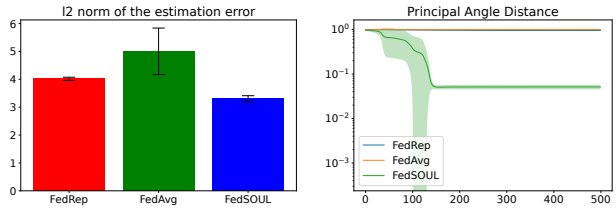

Figure S3: Small data sets - synthetic data. Raw data dimensionality is $k = 50$.

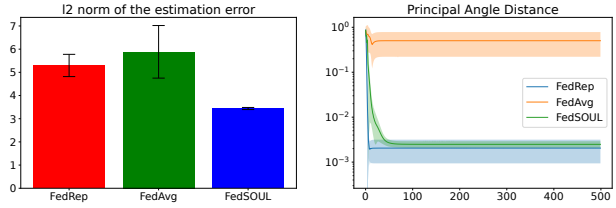

Figure S4: Small data sets - synthetic data. Raw data dimensionality is $k = 5$.

S3 Additional Experiments

In this section, we provide additional experiments. All the experimental details can be found in the "code" folder in the supplement.

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 Collins et al. (2021); Singhal et al. (2021)), 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 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 on the result. Figure S3 and Figure S4 show our results with $k \in \{5, 50\}$. All others parameters are the same as before.

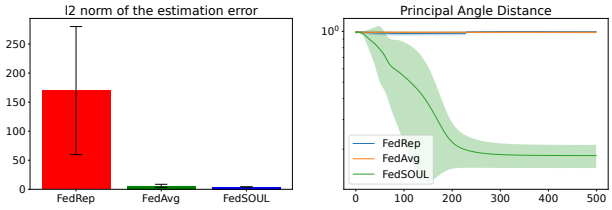

Figure S5: Small data sets - synthetic data. Latent space dimensionality is $d = 5$.

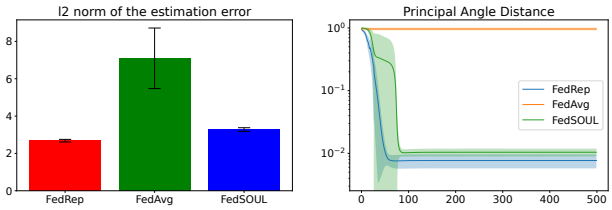

Figure S6: Small data sets - synthetic data. Latent space dimensionality is $d = 2$.

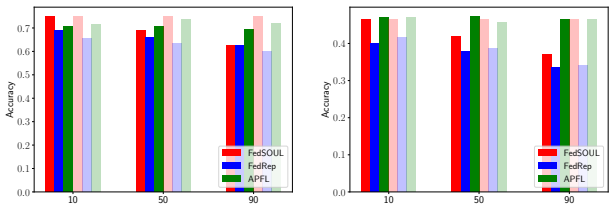

Figure S7: Small image datasets. Minimal local dataset size is 2 (top) or 5 (bottom).

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 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` (Deng et al., 2021b) which is another personalized federated learning approach. We consider CIFAR-10 dataset with 100 clients. Among these clients, there are 10, 50 or 90 which have 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 surprizing, that `APFL` is a very good baseline in these 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 (validational split, but same domain as training data). In the right part we see the distribution for out-of-distribution (SVHN in our case). Contraty to MNIST vs Fashion-MNIST example, here it is not that clear that `FedSOUL` captures uncertainty well.

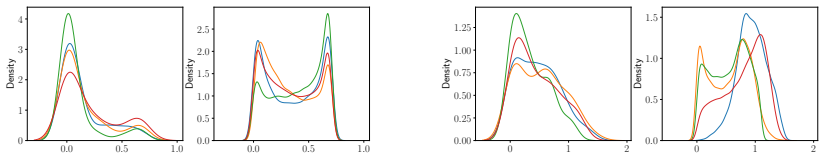

Figure S8: Out-of-distribution detection. CIFAR 10 vs SVHN. 2 classes for model (top) and 5 (bottom).

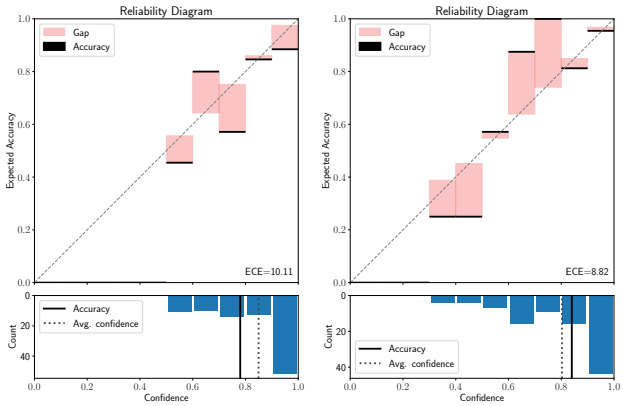

Figure S9: Reliability diagram for CIFAR10. 2 classes for model (top) and 5 (bottom).

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